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

Miao Cai, M.S.

Draft on August 13, 2019

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.

2019

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 and Guizhen Xu, who believe in the power of higher education, hard work, and always support me.

Acknowledgement

I want to thank my PhD 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 FIGURES

List of Tables

Chapter 1

INTRODUCTION

1.1 Transportation safety

Traffic safety is a pressing public health issue that involves huge lives losses and financial burden across the world and in the United States (US). As reported by the World Health Organization (WHO 2018b), 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 (WHO 2018a). 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 (Litman 2013; Evans 2014). Without traffic accidents, these victims could have much longer lives with normal health.

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 (Staton et al. 2016). In high-income countries, the majority 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 (Ameratunga, Hijar, and Norton 2006). 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 the 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 (Peden et al. 2004; Dalal et al. 2013).

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 United States, Bureau of Labor Statistics 2017). The National Safety Council reported that the number of deaths attributable to car crashes will be at least 40,000 in 2018, which is the third straight year that this number is over 40,000 (The National Safety Council 2018). A comparison study of 26 developed countries revealed that 20 to 60 traffic deaths per billion kilometers were reduced from 2011 (Evans 2014). Even though fatality rates attributable to road traffic in the US were reduced by 40% during that period, the rates declined more rapidly in all other 25 countries. Given the 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 (Litman 2013). If the change of traffic fatality rates match those in other unremarkable countries, 20,000 traffic deaths could have been prevented each year (Evans 2014).

The impact of road injuries is even more impactful in developing countries than in developed countries (Goonewardene et al. 2010). Although low- and middle-income countries own only 48% of the registered vehicles in the world, 90% of road traffic fatalities and injuries were estimated to occur in these countries, which continue to be escalating due to rapid urbanization and motorization (Dalal et al. 2013; Staton et al. 2016). Ten developing countries, including Brazil, Cambodia, China, Egypt, India, Kenya, Mexico, Russia, Turkey, and Vietnam, account for almost half of all the road traffic in the world (Hyder et al. 2012). For example, China has around 100,000 traffic-related fatalities each year (Zhang et al. 2010), accounting for around 80% of all accidental deaths, with 87% of them were caused by motor vehicles in 2015 (Jiang et al. 2017). In comparison, 148,707 lives were claimed by road collisions in India in 2015, with the road fatality rate similar to the global average level of 17.4

deaths per 100,000 people (National Crime Records Bureau, Government of India 2015).

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 (Olson et al. 2016; Anderson et al. 2017). However, among all vehicles, large trucks are associated with more catastrophic accidents and therefore 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 (FMCSA 2018a). Among all 4,079 crashes involving large trucks or buses in 2016, 4,564 people (1.12 people per crash) were killed in the accidents (FMCSA 2016). 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 (Neeley and Richardson Jr 2009). 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 US dollars per crash (Zaloshnja, Miller, and others 2008).

The high risk of large trucks is attributed to two aspects of reasons (Huang et al. 2013). First, large truck drivers generally need to drive alone for long routes, under on-time demands, challenging weather and traffic conditions. Professional truck drivers usually need to work in shifts, and sometimes unavoidable late-night or early-morning shifts (Pylkkönen et al. 2015). These late-night or early-morning working shifts have been reported to be associated with sleep deprivation and disorders (Åkerstedt 1988; Mitler et al. 1997; Solomon et al. 2004; Sallinen et al. 2005). Besides, commercial truck drivers are exposed to long route, constant concentration, and overtime work, which intertwines 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 (National

Transportation Safety Board 1990; Mitler et al. 1997).

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 (Lemp, Kockelman, and Unnikrishnan 2011). Large trucks can weight up to 80,000 pounds by federal law, which are twenty times as much as a passenger vehicle (Department of Transportation, Utah 2019). If these trucks travel at the speed of 65 miles per hour on the highway, it will take them 525 feet to stop, which is about two times the length of a football field (Department of Transportation, Utah 2019). 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 the trucks in around 80% of the cases (Neeley and Richardson Jr 2009). In even worse case, commercial trucks crashes can cause massive casualties and regional public health emergency when the carried hazardous materials (such as gasoline and sulfuric acid) are leaked.

The importance of truck industry and the potential catastrophic consequences of truck crashes 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 screen 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) (Theofilatos et al. 2016, 2018). To tackle this rare-event issue, the most common study design is a case-control study that 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 (Braver et al.

1997; Chen Chen and Xie 2014; Meuleners et al. 2015; Née et al. 2019). Unfortunately, a case-control study is limited in estimating the incidence data or overall average treatment effect. It may be contentious in selecting the ratio of controls to cases and how to select these controls (Grimes and Schulz 2005; Sedgwick 2014). 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 (Girotto et al. 2016). Third, crashes are underreported, especially those without injuries or economic losses, as well as those crashes with minor severity (Ye and Lord 2011). The National Highway Traffic Safety Adminitration estimated that 25% of minor-injury crashes and 50% of no-injury crashes were not reported, compared to 100% reporting rate for fatal crashes (Savolainen et al. 2011).

Past truck safety literature almost exclusively focused on crashes, while ignoring the 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 truck crash (Saleh et al. 2013; Janakiraman, Matthews, and Oza 2016). Truck critical events deserve more attention since they occur more frequently than crashes, potentially suggest fatigue and a lapse in performance, and can lead to catastrophic crashes (Dingus, Neale, et al. 2006).

1.4 Proposal

With the rapid development of modern technology, more real-time naturalistic driving and critical events data are collection by commercial truck companies (Janakiraman, Matthews, and Oza 2016). With advanced unobstrusive instrumentation, these data provide a unique opportunity to continuously study real-world driving performance and potential consequences (Dingus et al. 2016). Naturalistic driving data can be further powered to examine the risk factors associated with truck crash, by continuously collecting high resolution data on the risk factors, such as driver behavior, vehicle condition, traffic, weather, and road geometry (Guo 2019).

This prospectus proposal focuses on Bayesian statistical models in a large truck naturalistic driving data. The research goals are:

1) quantify the association between truck crashes and SCEs;
2) construct scalable Bayesian hierarchical models for truck critical events;
3) create an innovative reliability model to account for short rests between different trips.

I believe that this work will contribute to statistical theories in constructing scalable Bayesian hierarchical statistical models using modern Markov chain Monte Carlo simulations (MCMC). Realistically, these statistical models will provide insights into the functional relationship between driver characteristics, traffic, weather, and other real-time driving environment. These statistical models can be further used to provide data-driven justification to optimize trucking routes and minimize unsafe driving behaviors.

Chapter 2

LITERATURE REVIEW

2.1 Naturalistic driving Study (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 (Hickman, Hanowski, and Bocanegra 2018; Stern et al. 2019). 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 FMCSA (2018b), large truck and bus fatalities in 2017 were 0.156 per million travelled vehicle miles, which was a 6.8 percent increase from 2016. This rareness poses a challenge to infer unbiased estimates using traditional statistical models (Guo et al. 2010; Theofilatos et al. 2018). Secondly, most truck crash data reply on post hoc police reports. Although these data are generally accurate and detailed by police officers, they are limited in determining the information of the driver in the meaningful time period leading up to the crash (Dingus, Hanowski, and Klauer 2011). Some of the critical factors, such as distraction, are not reported or cannot be determined due to a varitety of reasons (Dingus, Hanowski, and Klauer 2011), 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 (Ye and Lord 2011; Stern et al. 2019). It is estimated that 25% of minor-injury and 50% of non-injury crashes were not reported, while 100% of fatal crashes were reported (Savolainen et al. 2011).

In light of the aforementioned limitations, a growing number of naturalistic driving studies have been initiated worldwide to identify crash causation and develop effective action to improve traffic safety (Klauer et al. 2009; Hickman, Hanowski, and Bocanegra 2018; Guo 2019), such as the 100-Car NDS by the Virginia Tech Transportation Institute (Dingus, Klauer, et al. 2006), the second Strategic Highway Research Program including more than 3,400 drivers (Ghasemzadeh and Ahmed 2017), and the UDRIVE NDS in Europe (Eenink et al. 2014; Barnard et al. 2016). There are also a few other NDS that target at specific sub-populations, such as the 40-Teen NDS (Alden et al. 2016), the Older Driver Fitness-to-Drive NDS (Guo, Fang, and Antin 2015), and the Commercial Truck Driver NDS (Sparrow et al. 2016).

NDS use unobstrusive devices, sensors, and cameras installed on vehicles to proactively collect frequent naturalistic driving behavior and and performance data under real-world driving conditions (Hickman, Hanowski, and Bocanegra 2018; Guo 2019). Compared with tradtional post-hoc crash data that are road segment-based, NDS collect driver-based data which are more useful in comparing the rate of unsafe events 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 occurance of crashes or SCE without information bias or selection bias (Guo et al. 2010). Third, collecting naturalistic data is considerably less costly and difficult per observation compared to tradtional 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. Guo (2019) 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 in North Virginia, the United States (Neale et al. 2005). The research team continuously followed the 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, Dingus, Klauer, et al. (2006) 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.

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 (Dingus, Hanowski, and Klauer 2011). Although near-crashes or SCEs were not real crashes, these studies suggested that they are highly correlated with crashes (Dingus, Klauer, et al. 2006; Guo et al. 2010; Dingus, Hanowski, and Klauer 2011). 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 (Jansen and Simone Wesseling 2018; Mollicone et al. 2019).

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 Saleh et al. (2013). 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 (Saleh et al. 2013). 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 studies in certain safety science in which the accidents are extremely rare, such as nuclear industry (Smith and Borgonovo 2007), chemistry (Phimister et al. 2003), and aerospace industry (Kirwan, Gibson, and Hickling 2008).

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 (Guo 2019). 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 (Guo et al. 2010). 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.

Guo et al. (2010) 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

Gordon et al. (2011) 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. The found that time to edge crossing and lane-departure warning were two useful surrogates for crashes on rural nonfreeway roads, while lane deviation was a poor surrogate for lane-departure crashes.

Simons-Morton et al. (2012) examined whether elevated gravitational-force predicts crashes and near crashes among 42 newly licensed teenage drivers in Virginia. The study used the Naturalistic Teenage Driving Study that followed the recruited drivers for 18 months. A logistic regression estimated with generalized estimating equations to account for the within-subject correlation among different months. It was found that the rate of elevated gravitational-force events was positively associated with the rate of crashes and near-crashes (odds ratio = 1.07, 95% confidence interval: 1.02, 1.12), with the area under curve (AUC) value of 0.76.

Wu and Jovanis (2012) proposed a conceptual framework to estimate the crash-to-surrogate ratio π and used the 100-Car study to test the framework. 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 Wu and Jovanis (2013) developed diagnostic procedures to screen crashes and near misses under naturalistic study settings. The study applied the 100-Car Natural Driving Study on the proposed framework 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.

Guo and Fang (2013) attempted to identify risk factors of driving at drivers level and predict high-risk drivers based on the 100-Car study. The study used a negative binomial regression to examine the potential four risk factors of crashes and near-crashes. Besides, they used a K-mean clustering to classify the drivers into high-, moderate-, and low-risk groups based on crash and near-crash rates, and applied two logistic regressions to predict high- or moderate-risk drivers. The results confirmed that critical-incident event rates were significantly associated with individual driving risk. The two logistic regressions high achieved AUC values of 0.938 and 0.93. They also highlighted that it was a first-step study and more studies with larger and representative data were needed to confirm the association. A similar study by Wu, Aguero-Valverde, and Jovanis (2014) also used the 100-Car study. This study used a Bayesian multivariate Poisson log-normal model to simultaneously account for crash frequency and severity. They also found a significant positive association between crashes, near crashes, and crash-relevant incidents.

Pande et al. (2017) used linear referencing to link Global Positioning System (GPS) data with roadway features on 39 segments of Highway 101 in California. Negative binomial models and random-effects negative models 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.

Gitelman et al. (2018) used in-vehicle data recorders (IVDR) data collected on 3500 segments of interurban 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 a number of 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. In contrast, 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 were not a good surrogate for crashes, possibly due to its rough definition.

However, NDS is not bought without a doubt. Knipling (2015) 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 and develop interventions. Crash harm is defined as property damage, injury, income lost, and all other consequences of different severities (Zaloshnja and Miller 2007). NDS often use SCEs as surrogates of crashes, but very few or no crashes, let alone human harm. Therefore, Knipling (2015) argues that SCEs are not an appropriate part of the Heinrich's Triangle and researchers generally cannot derive valid quantitative conclusions on causations of harm based on NDS datasets.

Another study by Knipling (2017) specifically targeted Hour-of-Service rule research, such as Blanco et al. (2011) and Hanowski et al. (2008), 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 as 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 by National Sleep Foundation that approximately 32% of drivers drive with fatigue over twice a month (National Sleep Foundation 2008). 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 (American Automobile Association Foundation for Traffic Safety 2010). The National Highway Traffic Safety Administration (NHTSA) estimated that 60% of fatal truck crashes were attributable to the driver falling asleep while driving (Craye et al. 2016; Cavuoto and Megahed 2017). The FMCSA estimated that the causal role of fatigue is around five times higher in fatal than in property damage truck crashes (Knipling 2017).

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 (Yung 2016; Cavuoto and Megahed 2017). 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 succint definition on fatigue since it involves interactions between biological, behavior, and psychological process. A comprehensive review on fatigue definition and measurement is provided by Yung (2016).

The mechanism of fatigue leading to traffic safety events is that the driver's capability to stay alert to ambient traffic and pedestrians will be largely impaired. The reaction time is subsequently prolonged in that situation (Zhang et al. 2014). 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% (MacLean, Davies, and Thiele 2003). 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 (Knipling and Wang 1994). This is especially concerning because fatigue driving safety critical events 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, little has been done about drowsy driving since there is no simple way to objectively measure fatigue driving (Dement 1997). In view of the difficulty of measuring fatigue, researchers have 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.

Cumulative driving time, has also been a measure of driver's fatigue level, especially among NDSs. For example, Nakayama (2002) found that there was a significant increase in the fatigue of drivers after 12 hours of continuous driving. Jovanis, Wu, and Chen (2011) 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 (224 crash-involved drivers and 464 randomly sampled non-crash drivers from the same terminal in the same month), with the highest odds in the 11-th hour. From the fifth hour to the 11th hour, the odds of crashes were consistently increasing (Jovanis, Wu, and Chen 2012). In contrast, Soccolich et al. (2013) found no significant difference in safety outcomes between the 11th driving hour and driving hours 8, 9, or 10 using from the Naturalistic Truck Driving Study data. Desipte the fact that working hours were not significant, they suggested an interaction between driving hours and working hours: a work day that starts with several hours of non-driving work and then followed by 14 hours of driving was significantly associated with risk of safety-critical events. Another study by Mollicone et al. (2019) used a Poisson regression to quantify the association between driver performance and predicted fatigue level based on natualistic 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 McCauley et al. (2013). 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% confidence interval 1.013-1.146).

Numerous studies used ocular metrics, such as eye closure and blinking patterns, to detect driver's fatigue (Golias and Mishra 2013). For example, Jackson et al. (2016) measured the eyelid closure of 22 healthy participants with 24 hours of sleep deprivation. They found a significant increase in proportions of time with closed eyeclids after the sleep deprivation, which were significantly associated with crashes, Psychomotor Vigilance Task lapses, and subjective sleepiness scores (95% confidence interval of Spearman's correlation coefficient was 0.46 to 0.69). Other studies used physiological measures, such as electroencephalogram (EEG) and heart rate, as indicators of fatigue (Golias and Mishra 2013). It has been demonstrated that awake but drowsy truck drivers during working have increased alpha and theta power densities in EEG (Kecklund and Åkerstedt 1993; Cajochen et al. 1995; Mitler et al. 1997). The variability of heart rates has been identified as passive proxy of fatigue among drivers (Mulder 1992; Patel et al. 2011).

Lack of sleep or specific sleep patterns have also been used as proxies of fatigue. For example, G. X. Chen et al. (2016) 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 a.m. and 5 a.m. were associated with increased safety-critical event rates. Sparrow et al. (2016) reported that truck drivers with a retart break of only one nighttime period (defined as 1 a.m. to 5 a.m.) experienced more lapses of attention, elevated lane deviation at night, and higher sleepiness measured by subjective questionnaires. A naturalistic driving study by Wu, Aguero-Valverde, and Jovanis (2014) found that more sleeping hours was found to be beneficial to reduce near crashes.

A significant amount of research emphasizes the association between time of the day (such as night driving) and the development of fatigue (Cavuoto and Megahed 2017). Night drving is often accompanied by changes in shift scheduling, inadequate sleep, sleep apnea and disorder. For example, Pack et al. (1995) reported that the crashes in which the drivers fell asleep occurred primarily from mid-night to 7 a.m. and from 2 p.m. to 4 p.m.. Mitler et al. (1997) monitored 80 truck drivers in North America using 24-hour electrophysiologic measures. They found that drivers an average had 5.18 hours of sleep in bed per day 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. Pahukula, Hernandez, and Unnikrishnan (2015) 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 a.m. to 4 a.m.), morning (5 a.m. to 9 a.m.), mid-day (10 a.m. to 3 p.m.), afternoon (4 p.m. to 8 p.m.), and evening (9 p.m. to 11 p.m.). The results revealed major differences in traffic flow, light conditions, surface conditions, time of year, and the percentage of trucks on road among the five different models.

2.4.2 Driver characteristics

Young and older drivers have been reported to have higher risk of crashes or SCEs. The reasons for these 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 the intense schedule required by the employer company, which may increase the likelihood to be sleepy or fatigued. Duke, Guest, and Boggess (2010) 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 more vulnerable regarding the experience of driving. Clarke et al. (2006) 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 with male drivers aged 20 -25 years. Campbell (1991) 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. Pack et al. (1995) 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. Otmani, Rogé, and Muzet (2005) 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. Wu, Aguero-Valverde, and Jovanis (2014) used the 100-Car Naturalistic Driving Study dataset and reported that drivers under the age of 25 were more likely to have crashes and safety-related events.

To meet the huge demand services and supply chain management, it is common to extend the retirement age or reemploy retired workers, especially in developing countries (Popkin et al. 2008). 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 (Di Milia et al. 2011). Aged drivers are associated with eyesight diseases or functionality impairment, such as cataracts, narrowed peripheral vision and decreasing visual acuity (Di Milia et al. 2011). In addition, working for truck companies often means irregular shifts and taking the night schedules, which disrupt the circadian time-keeping systems, especially for the aged workers (Moneta et al. 1996). It is indicated by research 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 obvious (Di Milia et al. 2011).

Another risk factor for truck crashes at driver's level is gender. Gender has been suggested to be related with outcomes in medical treatment, education, sports and other fields, and there is no exception for truck drivers' safety. In the first place, women are more likely to suffer from fatigue compared with men. Fjell et al. (2008) found that women in general have 1.4 times higher chance of complaining of fatigue than men. However, females are found to have longer sleeping hours than their male counterparts of the same race (Lauderdale et al. 2006). This study also found that the mean sleep hours for white females was 6.7 hours compared with 6.1 hours for white males, and 5.9 hours for black female compared with 5.1 hours for black males even after adjusting for socioeconomic status, lifestyle and sleep apnea (Lauderdale et al. 2006). Gender differences are huge in terms of working conditions. Females had significantly fewer working hours per week, with 47 hours versus 52 hours per week (Rotenberg et al. 2008). In general, women tend to work fewer hours within a week but are more prone to feel fatigue and have a higher risk of traffic incidences.

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 (Wang, Quddus, and Ison 2013; Theofilatos and Yannis 2014; Roshandel, Zheng, and Washington 2015).

Traffic variables include flow (traffic volume), occupancy/density, and speed (Wang, Quddus, and Ison 2013; Theofilatos and Yannis 2014). 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, Dong et al. (2017) 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. Theofilatos et al. (2018) used hourly aggregated traffic variables, including flow, occupancy, mean time speed, and percentage of trucks to predict crash occurence 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, Kamla, Parry, and Dawson (2019) focused on 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 have reported to have both direct and indirect effects on traffic safety events. Theofilatos and Yannis (2014) 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 (Baker and Reynolds 1992; Chang and Chen 2005; Al-Ghamdi 2007). It is to noted that the cumulative time of driving in such extreme weather conditions could potentially increase the chances of safety critical events. The positive linear relationship between precipitation and traffic accidents can be observed in both driver accidents and pedestrian accidents (Al-Ghamdi 2007; Graham and Glaister 2003). Naik et al. (2016) 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. Abdel-Aty et al. (2012) 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 safety critical events increase in hot weather (Kjellstrom et al. 2009; Basagaña et al. 2015). Leard and Roth 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 (Leard, Roth, and others 2015). A literature review by Kampe, Kovats, and Hajat (2016) found that 11 out of 13 studies indicated an increase in unintentional injuries associated with high temperatures. In constrast, 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 (Lemp, Kockelman, and Unnikrishnan 2011). For example, Ahmed et al. (2018) 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, it is expected that road characteristics are potential risk factors of road safety (Wang, Quddus, and Ison 2013). 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, Zhu and Srinivasan (2011) 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 (Noland and Oh 2004; Kononov, Bailey, and Allery 2008; Zhu and Srinivasan 2011; Islam, Jones, and Dye 2014). This reversed relationship could possibly be explained by an increased chance of lane-change-related conflict opportunities (Wang, Quddus, and Ison 2013).

It is generally believed that lower speed limits can reduce the chances of traffic crashes, as well as the severity of crashes. For example, Neeley and Richardson Jr (2009) 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 Davis et al. (2015) 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. Dong et al. (2015) used zero-inflated negative binomial models to examine the effects of road geometry fratures 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. Islam, Jones, and Dye (2014) 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 (Shmueli and others 2010; Breiman and others 2001). The pro-explanation culture has long been adopted by most desciplines, such as epidemiology, economics, and psychology. In these desciplines, 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 blackbox 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 conceptual framework driven. 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 for are logistic regression and Poisson regression. Logistic regression is commonly used to predict crash likelihood (probability) using real-time data, for example traffic and weather at 5-minute intervals (Wang, Abdel-Aty, and Lee 2017). In contrast, Poisson regression is used to predict the crash frequency (the number of crashes) within a time period using aggregated data such as average data traffic and precipitation. I 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 in 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 (Savolainen et al. 2011). For example, Cong Chen et al. (2016) 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-level 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. Considering the rare-event natural of crashes, Theofilatos et al. (2016) 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 Moudon et al. (2011), Meuleners et al. (2017), Ahmed et al. (2018). There are two excellent systematic reviews on traffic crash likelihood predictions by Roshandel, Zheng, and Washington (2015) and Xu et al. (2015).

Other variants of a binary logistic regression are binary probit models (Lee and Abdel-Aty 2008; Yu and Abdel-Aty 2014), ordered logistic or probit models (Xie, Zhang, and Liang 2009; Zhu and Srinivasan 2011), multinomial logit models (Ye and Lord 2011). There are only minor difference 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 $\mathrm{erf}(x)$ is an integral without an analytical solution: $\mathrm{erf}(x) = \frac{2}{\pi}\int_0^x e^{-t^2}dt$.

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

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

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 (Zhu and Srinivasan 2011). These ordered models account for the ranked nature of different severity levels but make the proportional odds assumption (Rifaat, Tay, and De Barros 2012). 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.

On the other hand, the parameterization of a Poisson regression is shown in model (2.4).

$$\begin{aligned} Y_i^\star &\sim \mathrm{Poisson}(\mu_i) \\ \log \mu_i &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \end{aligned} \tag{2.4}$$

Where $Y_i^\star$ is the number of events in the i-th observation, which must be a non-negative integer. μ_i is the mean and variance 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 the Poisson distribution, which is often violated in real-life data. When the variance of the data is greater than expected, it is called overdispersion. Otherwise, it is called underdispersion. Overdispersion is much more common than underdispersion in statistical practice.

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 well handles count data that are right-skewed, long tailed, and only have non-negative integer values. For example, Cantor et al. (2010) 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 includes negative binomial models, quasi-Poisson models, and zero-inflated Poisson or negative binomial models (Lord 2006; Mohammadi, Samaranayake, and Bham 2014). Negative binomial or quasi-Poisson models are developed to account for the overdispersion and underdispersion in count data, for which a Poisson model fails to account. Zero-inflated Poisson or negative binomial models are developed to account for the feature of rare events in traffic crash data (Lord, Washington, and Ivan 2005, 2007; Washington, Karlaftis, and Mannering 2010; Dong et al. 2014). There is an excellent review paper on statistical models for crash frequency data by Lord and Mannering (2010).

Recently, recurrent event models have also been applied to model the change in intensity of SCEs in the traffic safety field. For example, Liu, Guo, and Hanowski (2019) proposed to use a mixed-effects Poisson process (a recurrent-event model) to model unintentional lane deviation events, with the baseline intensity and time-varying coefficients modeled by penalized B-splines. The 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

In contrast to traditional frequentist models that view parameters as unknown but fixed values, Bayesian models view parameters as random variables that have probability distributions (Gelman et al. 2013). Researchers 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 an unconditional distribution that is a compromise between the prior beliefs and the data. This compromise is given analytically by the Bayes Theorem (Equation (2.5)).

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

Where $p(\mathbf{X}|\theta)$ is the likelihood function, which reflects the data generating process that gives rise to the data observed. The denominator $\int p(\theta)p(\mathbf{X}|\theta)d\theta$ is a normalizing constant that makes the posterior distribution integrates to one. The prior and likelihood function are straightforward since they both have analytical forms. The trickiest part of Bayesian inference is the normalizing constant in the denominator (Gelman et al. 2013; Kruschke 2014).

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

There are several strengths of Bayesian models over traditional Frequentist models. First, the probabilistic distribution of parameters, posterior credibles intervals, and posterior predictive distributions account for the uncertainty in parameters and the data generating process. 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 (regularizing priors) is particularly useful for estimation in high-dimensional, sparse data settings, and complex statistical models such as hierarchical models (Betancourt and Girolami 2015; McElreath 2018). Lastly, Bayesian models are scalable to complex data generating process. This is because modern Bayesian estimation is powered by numerical methods and simulation, which in essence only requires researchers to specify the priors and likelihood function. The difficulty of written the likelihood function is minimal compared to traditional Frequentist approaches such as the maximum likelihood estimation, which scales with the complexity of models (Lambert 2018).

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 study 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. 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 crash history in the past year (Cantor et al. 2010). 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 associated with vehicle drivers (Huang and Abdel-Aty 2010).

In the Bayesian perspective, a hierarchical model is a statistical model with the probability distribution of one parameter depends on another parameter (Kruschke and Vanpaemel 2015). 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) (Kruschke and Vanpaemel 2015).

Model (2.6) demonstrates a random-intercept hierarchical logitistic regression that predicts the likelihood of safety events. The outcome $Y_{i,d(i)}$ is a binary variable that indicates whether a safety event occurred or not, and it has a Bernoulli distribution with the mean parameter $p_{i,d(i)}$. The logit transformation of $p_{i,d(i)}$ can then be predicted by k variables $x_1, x_2, \cdots, x_k$. The random intercept $\beta_{0,d(i)}$ determines that this is hierarchical model since they vary across different drivers $d(i)$. This model assumes that these random intercepts are sampled from a population of drivers with the mean of μ_0 and standard deviation of σ_0, which are known as hyperparameters.

$$\begin{aligned} Y_{i,d(i)} &\sim \text{Bernoulli}(p_{i,d(i)}) \\ \log \frac{p_{i,d(i)}}{1-p_{i,d(i)}} &= \beta_{0,d(i)} + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i} \\ \beta_{0,d(i)} &\sim N(\mu_0, \sigma_0^2), \quad k = 1, 2, \cdots, D \end{aligned} \tag{2.6}$$

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 (McElreath 2018). 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 (Gelman and Hill 2006; Lambert 2018).

Hierarchical models also come with costs. They are particularly known for its complexity to estimate to coefficients in both Frequetist maximum likelihood 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) sampling to overcome the intractable denominator in the Bayes Theroem (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 (Kruschke 2014). 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 (2.7).

$$P\left(X_i = x_i | X_1 = x_1, X_2 = x_2, \ldots, X_{i-1} = x_{i-1}\right) = P\left(X_i = x_i | X_{i-1} = x_{i-1}\right) \tag{2.7}$$

Integrating Markov chains and Monte Carlo simulations, the MCMC method can characterize an unknown unconditional distribution without knowing its all mathematical properties by sampling from the distribution (Van Ravenzwaaij, Cassey, and Brown 2018). It has been widely applied in multiple fields such as statistics, physics, chemistry, and computer science (Craiu and Rosenthal 2014). 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 mean, standard deviation, and intervals).

The first proposal of using MCMC dates to the paper by Metropolis et al. (1953), 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 (2.8).

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

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.

After decades of successful empirical trials in physics, Hastings (1970) 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 (2.9). This

Metropolis-Hasting (MH) algorithm is the most classic and widely-known MCMC algorithm used in multiple 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:

$$\alpha^{\star} = \min\left(1, \frac{p\left(\theta^{(\text{prop})}|\boldsymbol{x}\right)}{p\left(\theta^{(t)}|\boldsymbol{x}\right)} \frac{q\left(\theta^{(t)}|\theta^{(\text{prop})}\right)}{q\left(\theta^{(\text{prop})}|\theta^{(t)}\right)}\right) \tag{2.9}$$

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 choice of 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 very 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 (Geman and Geman 1987; Gelfand and Smith 1990; Lambert 2018). The development of the software *Bayesian inference Using Gibbs Sampler (BUGS)* (Lunn et al. 2000, 2009) was critical in increasing the popularity of applied Bayesian analyses considering its support for a wide variety of statistical distributions, automatic application of the Gibbs Sampler, and numerous textbooks, tutorials and discussion.

Suppose $\theta = [\theta_1, \theta_2, \cdots, \theta_k]$ is a k-dimensional parameter. Let $\mathbf{X}$ denote the data. The *Gibss 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 for a sufficient number of iterations.

The generality of the M-H algorithm and Gibbs sampler and the simplicity in software packages in `R` or `BUGS` help them gain popularity among applied researchers in the recent 30 years. However, as more and more data are available in applied field, the performance of the two most popular MCMC methods has been widely criticized (Betancourt 2019). The performance of the M-H algorithm crucially depends on the proposal distribution. An efficient proposal distribution in M-H algorithm should generate random draws with less auto-correlation, which enables more effective exploration of the parameter space (Quiroz 2015). On the other hand, the performance of the Gibss Sampler crucially depends on the parameter structure. If there is a significant correlation between parameter estimates, the Gibbs Sampler will become very inefficient as the geometry of the distribution is not aligned with the stepping directions of each sampler (Lambert 2018).

2.6 Scalable Bayesian models

Recent ten years witnessed an explosive growth of data size and dimensionality. This poses a major challenge to Bayesian methods 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 (Bardenet, Doucet, and Holmes 2017). In applied analysis, researchers often need to set thousands of iterations to reach stable posterior distribution, which takes hours or days to implement a single model. Besides, when the researchers have high dimensional data where high-probability regions are concentrated on a extremely limited region of sample space, it would very hard for random-walk MCMC to generate samples from these small regions (Barp et al. 2018). 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 the case of many parameters, neither the M-H algorithm or Gibbs sampler can efficiently generate samples from the posterior distribtion. All the aforementioned 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)

The M-H algorithm and Gibbs sampler can be very inefficient in big data settings because of sparse high-density parameter space, high costs of evaluating the entire data at each step, or a high correlation between parameters. Originally proposed by Duane et al. (1987) 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, thus more efficiently explores the high-density regions in big data settings compared to the traditional M-H algorithm or the Gibbs sampler (Betancourt 2017; Wang, Broccardo, and Song 2019). Although HMC was originally proposed in 1987 (Duane et al. 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) (Hoffman and Gelman 2014) and the statistical programming language `Stan` (Carpenter et al. 2017).

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 (Neal and others 2011; Betancourt 2017):

$$\pi(\boldsymbol{p}, \boldsymbol{q}) = \pi(\boldsymbol{p}|\boldsymbol{q})\pi(\boldsymbol{q})$$

This joint distribution can also be defined in terms of the *Hamiltonian*:

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

After a little bit of transformation, we have:

$$\begin{aligned} H(\boldsymbol{p}, \boldsymbol{q}) &= -\log \pi(\boldsymbol{p}, \boldsymbol{q}) \\ &= -\log \pi(\boldsymbol{p}|\boldsymbol{q}) - \log \pi(\boldsymbol{q}) \\ &= K(\boldsymbol{p}, \boldsymbol{q}) + V(\boldsymbol{q}) \end{aligned} \tag{2.10}$$

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*:

$$\begin{aligned} \frac{d\boldsymbol{q}}{dt} &= \frac{\partial H}{\boldsymbol{p}} = \frac{\partial K}{\boldsymbol{p}} \\ \frac{d\boldsymbol{p}}{dt} &= -\frac{\partial H}{\boldsymbol{q}} = -\frac{\partial K}{\boldsymbol{q}} - \frac{\partial V}{\boldsymbol{q}} \end{aligned} \tag{2.11}$$

It turns out that we can randomly generate high density proposals in the parameters space by taking advantage of the Hamiltonian system. Here is the general idea if the *HMC algorithm* (Lambert 2018):

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:

$$\alpha^H = \min\left(1, \frac{p\left(\theta^{(\text{prop})}|\boldsymbol{x}\right) p(\theta^{(\text{prop})})}{p\left(\theta^{(t)}|\boldsymbol{x}\right) p(\theta^{(t)})} \frac{q(m')}{q(m)}\right) \tag{2.12}$$

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 (2.12)).

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 (Neal and others 2011; Monnahan, Thorson, and Branch 2017). 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.

The No-U-Turn Sampler (NUTS) proposed by Hoffman and Gelman (2014) solves the difficulty of hand tuning ϵ and T in static HMC. NUTS calculates the optimal step size ϵ

and number of steps L through a tree building algorithm (Monnahan, Thorson, and Branch 2017). 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 automatically 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 algorithms, as most of them 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 a subset of data. 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 (Quiroz, Villani, et al. 2018).

The first well-known subsampling MCMC algorithm is the firefly MCMC by Maclaurin and Adams (2015), 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. Korattikara, Chen, and Welling (2014) 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 Bardenet, Doucet, and Holmes (2014) and Bardenet, Doucet, and Holmes (2017). 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 (Quiroz, Villani, et al. 2018; Quiroz et al. 2019). 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 (Quiroz et al. 2016), delayed acceptance (Quiroz, Tran, et al. 2018), noisy MCMC (Alquier et al. 2016), and zig-zag process MCMC (Bierkens et al. 2019).

Apart from the subsampling MCMC algorithms being mentioned, subsampling MCMC using the Hamiltonian mechanism deserves special attention as it efficiently explores the posterior in high-dimensional parameter space. However, HMC is especially inefficiently in the case of tall data as the gradient will be very expensive. T. Chen, Fox, and Guestrin (2014) proposed a stochastic gradient HMC, which introduces a friction term that counteracts the effects of noisy gradient. In contrast, Betancourt (2015) argued that the stochastic gradient HMC proposed by T. Chen, Fox, and Guestrin (2014) 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 recent paper by Dang et al. (2019) extended the PMCMC algorithm by Quiroz et al. (2019) to HMC via introducing a fictitious momentum vector $\vec{p}$, which has the same dimension as the parameter vector θ.

2.7 Conceptual framework

Cantor et al. (2010) suggested three factors that cause truck crashes: driver factors, vehicle factors (type and condition), and environmental factors.

Roshandel, Zheng, and Washington (2015) proposed five factors that affect traffic safety: (a) behavioral characteristics of the driver, e.g., impairment, fatigue, distractions; (b) vehicle

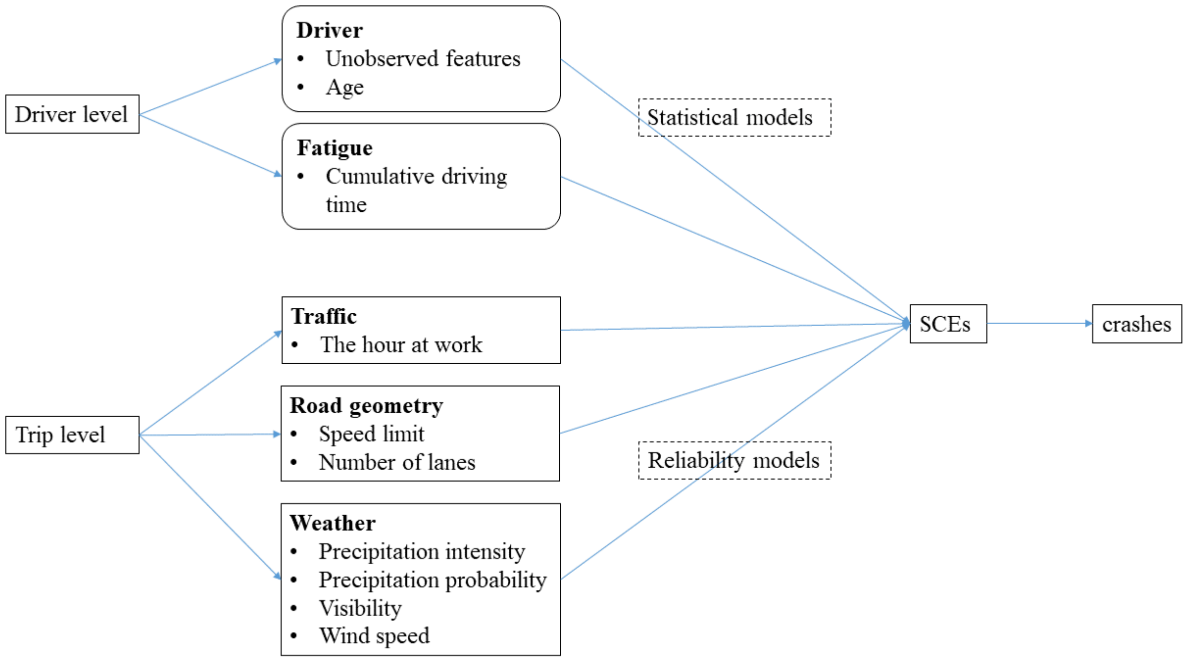

Figure 2.1: *Conceptual model. SCEs represent safety critical events.*

— the condition of the vehicle; (c) traffic — the traffic conditions; (d) geometry — geometric characteristics of the road, e.g. curve, hill, ramps, etc.; and (e) environmental — characteristics of the surrounding environment, such as weather conditions (rain, snow, night-time driving, etc.). Traffic conditions are the most studied of these and we focus on discussing them in this subsection.

As shown in Figure 2.1, the conceptual model in this study is based on the framework proposed by Stern et al. (2019).

2.8 Research aims

Describes the research problems to be addressed and why these are important issues.

The overarching goal of this proposed study is to

1. **Aim1**: to examine the association between truck crashes and critical events using a Gamma-Poisson regression.

2. **Aim2**: to construct three scalable Bayesian hierarchical models for SCEs: logistic regression, Poisson regression, and non-homogeneous Poisson process.

3. **Aim3**: to propose an innovative reliability model that accounts for both within shift cumulative driving time and within-shift between-trip rest time.

Gaps in literature

- A focus on crashes instead of precursors of crashes
- A focus on road segments rather than drivers
- A focus on case-control comparison given the rareness of truck crashes rather than rates
- A focus on small-scale data
- A focus on traditional statistical models instead of recurrent events models

Chapter 3

METHODS

3.1 Data sources

3.1.1 Real-time ping

The J.B. Hunt Transport Services, a commercial trucking and transportation company in the United States, will provid 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.

3.1.2 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 thresholds with regard to 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.

3.1.3 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.

3.1.4 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 (The Dark Sky Company, LLC 2019). 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 (The Dark Sky API 2019). Among several different weather data providers I tested, the `Dark Sky API` provides the most accurate and complete weather variables.

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 scaled 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 give million unique combinations (Rudis 2018). The weather data for these combinations will then be merged back to the original ping data. A minimal example of R code to retrieve weather data from the DarkSky API can be found in Appendix 8.

3.1.5 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 (Wikipedia contributors 2019). 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 (Eugster and Schlesinger 2013). 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. The R code to retrieve road geometry data can be found in Appendix 8.

3.2 Data aggregation and merging

3.3 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. The cohort will be all drivers with at more than 100 pings. Drivers with less than 100 real-time pings will be recognized as potential outliers and excluded

Table 1: ping data

ping_time	speed	latitude	longitude	driver
2015-10-23 08:00:00	0	33.94360	-118.1681	canj1
2015-10-23 08:08:10	0	33.94358	-118.1681	canj1
2015-10-23 08:09:26	5	33.94288	-118.1681	canj1
2015-10-23 08:22:58	4	33.97146	-118.1677	canj1
2015-10-23 08:23:12	8	33.97178	-118.1677	canj1

Table 2: 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

Table 3: 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

Table 4: Transformed shifts data

driver	shift_ID	shift_start	shift_end	shift_length
canj1	1	2015-10-23 08:09:26	2015-10-23 18:37:56	628
canj1	2	2015-10-26 07:49:04	2015-10-26 15:06:58	437
canj1	3	2015-10-27 01:59:48	2015-10-27 07:58:56	359
canj1	4	2015-10-28 08:05:08	2015-10-28 20:20:32	735
canj1	6	2015-10-30 09:27:12	2015-10-30 21:18:22	711

Table 5: 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 6: drivers

driver	age
canj1	46
farj7	54
gres0	55
hunt	48
kell0	51
lewr10	27
rice30	34
smiv	49
sunc	37
woow59	24

Table 7: Road geometry 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

Table 8: weather 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

Figure 3.1: *Flow chart of data aggregation and merging*

from analysis.

3.3.1 Data reduction

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*:

- For each of the truck drivers, if the real-time ping data showed that the truck was not moving for more than 20 minutes, the ping data will be separated into two different *trips*.
- These trips data will be further divided into different *shifts* if the specific driver was not moving for eight hours.

Therefore, a *trip* is defined as a continuous period of driving without stopping for more than 20 minutes. a *shift* is defined as a long period of driving without stopping for more than 8 hours.

3.3.2 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.

3.3.3 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 (Lord and Mannering 2010). 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 (Center 1987). The Bayesian statistical models will be conducted using the `rstan` package in R 3.5.1 (Stan Development Team 2018; R Core Team 2018).

3.4 Analytical Plan for Aim 2

The purpose of aim 2 is to develop three hierarchical Bayesian statistical and reliability models for the SCEs of truck drivers. Vehicle drivers will be viewed as the sampling unit. The workflow is to sample a certain number of drivers from a population of drivers, observe their driving trips or shifts for a specific period, then compare the safety events with non-events, and make conclusions on risk factors associated with these safety events. Bayesian hierarchical logistic regression, Poisson regression, and NHPP that accounts for both driver-level variation and trip-level variation will be used.

3.4.1 Data aggregation

3.4.2 Logistic regression

Here the probability of a critical event occurred will be modeled using a Bayesian hierarchical logistic regression. I will categorize the number of safety events during a trip into a binary variable $Y_{i,d}$ of either 0 or 1, where 0 indicates that no critical event occurred during that trip while 1 indicates that at least 1 critical event occurred during the trip.

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

We assume that the drivers are random effects, and we assume exchangeable priors of the form

$$\beta_{0,d(1)}, \beta_{0,d(2)}, \ldots, \beta_{0,d(n)} \sim \text{i.i.d.} N(\mu_0, \sigma_0^2)$$

and

$$\beta_{1,d(1)}, \beta_{1,d(2)}, \ldots, \beta_{1,d(n)} \sim \text{i.i.d.} N(\mu_1, \sigma_1^2)$$

The parameters μ_0, σ_0, μ_1, and σ_1 are hyperparameters with priors. Since we do not have much prior knowledge on the hyperparameters, we assigned diffuse priors for these hyperparameters.

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

Since μ_0 and μ_1 can be any real number, so we assigned two normal distributions with mean of 0 and standard deviation of 10 as the priors for these two hyperparameters. In comparison, σ_0 and σ_1 must be strictly positive, so we assigned $\text{GAMMA}(1, 1)$ with wide distribution on positive real numbers as their priors.

3.4.3 Poisson regression

Since logistic regression ignores the intensity of the critical events with any number greater than 0 categorized into 1, I adopt a Bayesian hierarchical Poisson regression to model the effect of cumulative driving time on the occurrence of critical events. Each driver has a random intercept and a random slope on cumulative driving time.

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

Where N is the number of critical events for driver $d(i)$ in time interval j, and it has a Poisson distribution with parameter λ. The other variables are identical as those described in Equation .

3.4.4 Non-homogeneous Poisson process (NHPP)

Mean function of a point process:

$$\Lambda(t) = E(N(t))$$

$\Lambda(t)$ is the expected number of failures through time t.

Rate of Occurence of Failures (ROCOF): When Λ is differentiable, the ROCOF is:

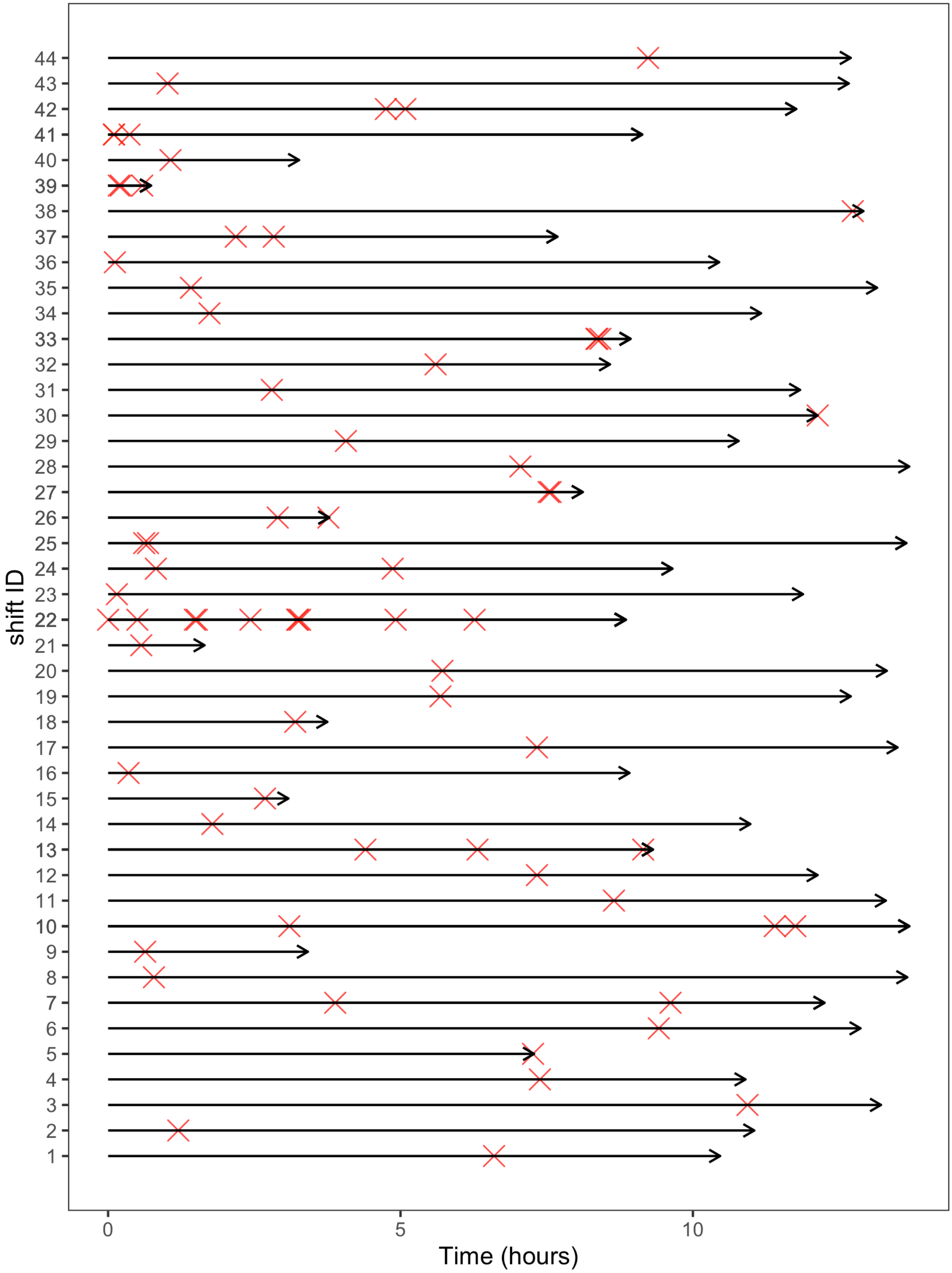

Figure 3.2: *An arrow plot of time to SCEs in each shift*

$$\mu(t) = \frac{d}{dt}\Lambda(t)$$

The ROCOF can be interpreted as the instantaneous rate of change in the expected number of failures.

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 intensity function.

Nonhomogeneous Poisson Process (NHPP): The NHPP 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 of a NHPP is:

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

, Where $\beta > 0$ and $\theta > 0$, the process is called the power law process (PLP).

Therefore, 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, the data are said to be failure truncated.
2. **Time truncation**: Data are said to be time truncated when testing stops at a predetermined time t.

In a time truncated case, the joint likelihood function for $f(n, t_1, t_2, \cdots, t_n)$ is (the prove can be found in the Appendix):

$$
\begin{aligned}
f(n, t_1, t_2, \cdots, t_n) &= f(n) f(t_1, t_2, \cdots, t_n | n) \\
&= \frac{e^{-\int_0^\tau \lambda(u)du} [\int_0^\tau \lambda(u)du]^n}{n!} n! \frac{\prod_{i=1}^n \lambda(t_i)}{[\Lambda(\tau)]^n} \\
&= \Big(\prod_{i=1}^n \lambda(t_i) \Big) e^{-\int_0^\tau \lambda(u)du} \\
&= \Big(\prod_{i=1}^n \frac{\beta}{\theta} (\frac{t_i}{\theta})^{\beta-1} \Big) e^{-(\tau/\theta)^\beta}, \\
n = 0, 1, 2, \cdots, &\quad 0 < t_1 < t_2 < \cdots < t_n
\end{aligned}
\tag{3.1}
$$

The log likelihood function l is then:

$$
\begin{aligned}
l &= \log \Bigg(\Big(\prod_{i=1}^n \frac{\beta}{\theta} (\frac{t_i}{\theta})^{\beta-1} \Big) e^{-(\tau/\theta)^\beta} \Bigg) \\
&= \sum_{i=1}^n \log \Big(\frac{\beta}{\theta} (\frac{t_i}{\theta})^{\beta-1} \Big) - (\frac{\tau}{\theta})^\beta \\
&= n \log \beta - n\beta \log \theta + (\beta - 1) \Bigg(\sum_{i=1}^n \log t_i \Bigg) - \Big(\frac{\tau}{\theta} \Big)^\beta
\end{aligned}
\tag{3.2}
$$

Despite Poisson regression consider the frequency of SCEs in a given interval, it assumes that the intensity of events is a constant, which may not be true in real-life transportation practice. Here we presented a reliability model, a non-homogeneous Poisson process (NHPP) with a power law process (PLP) based on the merged shifts data set. we aim to answer if SCEs occurred more frequently at early stages of shifts, towards the end of shifts, or does not show significant patterns.

Let $T_{d,s,i}$ denote 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:

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

- $d = 1, 2, \cdots, D$.

We assumes 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~3.3.

$$
\begin{aligned}
T_{d,s,1}, T_{d,s,2}, \cdots, T_{d,s,n_{d,s}} &\sim \text{PLP}(\beta, \theta_{d,s}) \\
\beta &\sim \text{Gamma}(1, 1) \\
\log \theta_{d,s} &= \gamma_{0d} + \gamma_1 x_{d,s,1} + \gamma_2 x_{d,s,2} + \cdots + \gamma_k x_{d,s,k} \\
\gamma_{01}, \gamma_{02}, \cdots, \gamma_{0D} &\sim \text{i.i.d. } N(\mu_0, \sigma_0^2) \\
\gamma_1, \gamma_2, \cdots, \gamma_k &\sim \text{i.i.d. } N(0, 10^2) \\
\mu_0 &\sim N(0, 10^2) \\
\sigma_0 &\sim \text{Gamma}(1, 1)
\end{aligned}
\tag{3.3}
$$

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.

3.5 Analytical Plan for Aim 3

Aim 3 seeks to innovate the NHPP proposed in Aim 2 by accounting for the rest time within a shift. One more parameter, the percent of reliability recovery during a break within a shift, will be estimated for in this model.

1.

Chapter 4

THE PROBABLE CONTENT

The contents will follow the standard format for a traditional dissertation, as per guidelines set by the College for Public Health and Social Justice and Saint Louis University's Office of Graduate Education.

Chapter 1: Introduction: The problem

1. Transportation safety
2. Truck safety
3. Modern Truck Safety Studies

Chapter 2: Literature review

1. Naturalistic driving Study (NDS)
2. Safety-critical events (SCEs)
3. Crashes and SCEs
4. Risk factors for traffic safety
5. Predictive models a. Overview b. Bayesian models c. Hierarchical models d. Markov chain Monte Carlo (MCMC)
6. Scalable Bayesian models a. Hamiltonian Monte Carlo (HMC) b. Subsampling MCMC

Chapter 3: Aim 1 - truck crashes and critical events

1. Introduction
2. Data sources
3. Methods
4. Results a. A simulation study on Gamma-Poisson models b. Real-world application on all SCEs c. Real-world application on different types of SCEs
5. Discussion

Chapter 4: Aim 2 - Bayesian hierarchical models for SCEs

1. Introduction
2. Data sources
3. Methods a. Logistic regression b. Poisson regression c. Non-homogeneous Poisson process (NHPP)
4. Results a. Logistic regression b. Poisson regression c. NHPP
   i. A simulation study on NHPP
   ii. Real-world application on NHPP
5. Discussion

Chapter 5: Aim 3 - an innovation on NHPP to account for within shift rest time

1. Introduction
2. Data sources
3. Methods
4. Results a. A simulation study on this new method b. Real-world application on this new method c. Real-world application on this new method stratified by SCE types
5. Discussion

Chapter 6: Discussion

1. Conclusion

2. Strengths and limitation

3. Future research

Chapter 5

TRUCK CRASHES AND CRITICAL EVENTS

Chapter 6

THREE STATISTICAL MODELS PREDICTING TRUCK CRITICAL EVENTS

6.1 Hierarchical logistic model

6.1.1 Model set up

6.1.2 Bayesian estimation based on simulated data

6.2 Hierarchical Poisson model

6.3 Hierarchical power law process

*Mean function of a point process**:

$$\Lambda(t) = E(N(t))$$

$\Lambda(t)$ is the expected number of failures through time t.

Rate of Occurence of Failures (ROCOF): When Λ 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.

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 intensity function.

Nonhomogeneous Poisson Process (NHPP): The NHPP is a Poisson process whose intensity function is non-constant.

Power law process (PLP): When the intensity function of a NHPP is:

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

Where $\beta > 0$ and $\theta > 0$, the process is called the power law process (PLP).

Therefore, the mean function $\Lambda(t)$ is the integral of the intensity function:

$$\Lambda(t) = \int_0^t \lambda(t)dt = \int_0^t \frac{\beta}{\theta}\left(\frac{t}{\theta}\right)^{\beta-1} = \left(\frac{t}{\theta}\right)^{\beta}$$

6.3.1

6.3.2 Bayesian estimation based on simulated data

Chapter 7

HIERARCHICAL BAYESIAN MODELS USING SUBSAMPLING MARKOVE CHAIN MONTE CARLO METHODS

- Comparison with results given by frequentist models using REML (`lmer4`)
- Comparison with results given by Bayesian models using HMC (`rstan`)
- Comparison with results given Bayesian models using INLA (`INLA`)

Chapter 8

DISCUSSION

APPENDIX

Query weather data

```
gps_sample =
  structure(list(
    from_lat = c(41.3473127, 41.8189037, 32.8258477, 40.6776808,
                 40.2366043, 41.3945561, 32.6320605, 40.5413856,
                 33.6287422, 40.0692742, 41.347986, 37.7781459,
                 43.0843081, 41.48026, 43.495149, 41.5228684,
                 41.5763081, 47.6728665, 41.0918361, 41.1537819),
    from_lon = c(-74.2850908, -73.0835104, -97.0306677, -75.1450753,
                 -76.9367494, -72.8589916, -96.8538145, -74.8547061,
                 -113.7671634, -76.762612, -74.284785, -77.4615586,
                 -76.0977384, -73.2107541, -73.7727896, -74.0739204,
                 -88.1529175, -117.3224667, -74.1554972, -74.1887031),
    beg_time = structure(
      c(1453101738, 1437508088, 1436195038, 1435243088, 1454270680,
        1432210106, 1438937772, 1446486480, 1450191622, 1449848630,
        1457597084, 1432870446, 1457968284, 1451298724, 1431503502,
        1443416864, 1438306368, 1445540454, 1452619392, 1436091072),
      class = c("POSIXct", "POSIXt"), tzone = "UTC")),
    .Names = c("from_lat", "from_lon", "beg_time"),
    row.names = c(NA, 20L),
    class = c("tbl_df", "tbl", "data.frame"))
gps_sample
```

```
library(darksky)
add_var = function(dat){
  dat[,c("time", "summary", "icon", "precipIntensity",
         "precipProbability", "temperature", "apparentTemperature",
         "dewPoint", "humidity", "pressure", "windSpeed", "windGust",
         "windBearing", "cloudCover", "visibility")] = NA
  return(dat)
}

for(i in 1:nrow(gps_sample)){
```

```
  t = get_forecast_for(gps_sample$from_lat[i], gps_sample$from_lon[i],
                       gps_sample$beg_time[i])
  gps_sample$summary[i] = ifelse(is.null(t[[3]]$summary), NA,
                              t[[3]]$summary)
  gps_sample$icon[i] = ifelse(is.null(t[[3]]$icon), NA, t[[3]]$icon)
  gps_sample$precipIntensity[i] = ifelse(is.null(t[[3]]$precipIntensity),
                                      NA, t[[3]]$precipIntensity)
  gps_sample$precipProbability[i] = ifelse(is.null(t[[3]]$precipProbability),
                                        NA, t[[3]]$precipProbability)
  gps_sample$temperature[i] = ifelse(is.null(t[[3]]$temperature), NA,
                                  t[[3]]$temperature)
  gps_sample$apparentTemperature[i] = ifelse(is.null(
    t[[3]]$apparentTemperature), NA, t[[3]]$apparentTemperature)
  gps_sample$dewPoint[i] = ifelse(is.null(t[[3]]$dewPoint), NA,
                               t[[3]]$dewPoint)
  gps_sample$humidity[i] = ifelse(is.null(t[[3]]$humidity), NA,
                               t[[3]]$humidity)
  gps_sample$pressure[i] = ifelse(is.null(t[[3]]$pressure), NA,
                               t[[3]]$pressure)
  gps_sample$windSpeed[i] = ifelse(is.null(t[[3]]$windSpeed), NA,
                                t[[3]]$windSpeed)
  gps_sample$windGust[i] = ifelse(is.null(t[[3]]$windGust), NA,
                               t[[3]]$windGust)
  gps_sample$windBearing[i] = ifelse(is.null(t[[3]]$windBearing), NA,
                                  t[[3]]$windBearing)
  gps_sample$cloudCover[i] = ifelse(is.null(t[[3]]$cloudCover), NA,
                                 t[[3]]$cloudCover)
  gps_sample$visibility[i] = ifelse(is.null(t[[3]]$visibility), NA,
                                 t[[3]]$visibility)
}
```

Query road geometry data

```
pacman::p_load(osmar, stringr)
src <- osmsource_api(url = "https://api.openstreetmap.org/api/0.6/")
road_data = function(i = 5, width = 100, data = df3){
  bb <- center_bbox(data$lon_short[i], data$lat_short[i],
                    width, width)
  ua = get_osm(bb, source = src)
  ua
  road_inf <- data.frame(ua$ways$tags)
  colnames(road_inf) <- c("ID", "Key", "Value")
  road_inf$Key <- as.character(road_inf$Key)
  road_inf$Value <- as.character(road_inf$Value)
  row_speed <-  which(road_inf$Key == "maxspeed", arr.ind=TRUE)
  row_lane <- which(road_inf$Key == "lanes", arr.ind=TRUE)

  max_speed <- as.numeric(str_extract(road_inf[row_speed,"Value"],
                                      "[[:digit:]]+"))
  num_lanes <- as.numeric(str_extract(road_inf[row_lane,"Value"],
                                      "[[:digit:]]+"))
  return(c(mean(max_speed), mean(num_lanes)))
}

loop_data = function(start_index = 1, loop_length = 100000){
  end_index = start_index + loop_length
  out_data = data.frame(matrix(0, ncol = 2, nrow = loop_length))
  df_index_diff = start_index-1

  for (i in start_index:end_index) {
    out_data[i-df_index_diff,] = road_data(i, data = df)
    print(paste0(end_index - i, " remained (",
                 round((end_index - i)*100/
                         (end_index-df_index_diff), 3), "%)"))
  }

  return(out_data)
}
```

```
df = data.table::fread("data/20190605_ping_compressed_3digits.csv")
dfcontainer12 = loop_data(start_index = 1)
```

8.1 Likelihood function of a NHPP

The first event: The cumulative density function (cdf) of time to the first event is $F(t_1)$:

$$F_1(t_1) = P(T_1 \leq t_1) = 1 - S(t_1)$$

.

The survival function for the first event $S_1(t_1)$ is:

$$\begin{aligned}
S_1(t_1) &= P(T_1 > t_1) \\
&= P(N(0, t_1) = 0) \quad N\text{is the number of events} \\
&= e^{-\int_0^{t_1} \lambda_u du} (e^{-\int_0^{t_1} \lambda_u du})^0 / 0! \\
&= e^{-\int_0^{t_1} \lambda_u du}
\end{aligned}$$

The probability density function (pdf) of time to the first event can be calculated by taking the first order derivative of the cdf $F_1(t_1)$:

$$\begin{aligned}
f_1(t_1) &= \frac{d}{dt_1} F_1(t_1) \\
&= \frac{d}{dt_1} [1 - S_1(t_1)] \\
&= -\frac{d}{dt_1} S_1(t_1) \\
&= -\frac{d}{dt_1} e^{-\int_0^{t_1} \lambda(u) du} \\
&= -(-\lambda_{t_1}) e^{-\int_0^{t_1} \lambda(u) du} \\
&= \lambda(t_1) e^{-\int_0^{t_1} \lambda(u) du}
\end{aligned}$$

If this NHPP is a PLL, we plug in the intensity function $\lambda(t) = (\beta/\theta)(t/\theta)^{\beta-1}$, then we have:

$$f_1(t_1) = \frac{\beta}{\theta}(\frac{t_1}{\theta})^{\beta-1}e^{-(\frac{t_1}{\theta})^\beta}, \quad t_1 > 0$$

This pdf is identical with the pdf of Weibull distribution, so we have:

$$T_1 \sim \text{Weibull}(\beta, \theta)$$

The second event: the Survival function of the second event given the first event occurred at t_2 is:

$$\begin{aligned}
S_2(t_2|t_1) &= P(T_2 > t_2|T_1 = t) \\
&= P(N(t_1, t_2) = 0|T_1 = t_1) \\
&= e^{-\int_{t_1}^{t_2}\lambda_u du}[\int_{t_1}^{t_2}\lambda_u du]^0/0! \\
&= e^{-\int_{t_1}^{t_2}\lambda_u du}
\end{aligned}$$

The we can derive the pdf of t_2 conditioned on t_1

$$\begin{aligned}
f(t_2|t_1) &= -\frac{d}{dt_2}S_2(t_2) \\
&= -\frac{d}{dt_2}e^{-\int_{t_1}^{t_2}\lambda(u)du} \\
&= \lambda(t_2)e^{-\int_{t_1}^{t_2}\lambda(u)du} \\
&= \frac{\beta}{\theta}(\frac{t_2}{\theta})^{\beta-1}e^{-[(\frac{t_2}{\theta})^\beta-(\frac{t_1}{\theta})^\beta]} \\
&= \frac{\frac{\beta}{\theta}(\frac{t_2}{\theta})^{\beta-1}e^{-(t_2/\theta)^\beta}}{e^{-(t_1/\theta)^\beta}}, \quad t_2 > t_1
\end{aligned} \tag{8.1}$$

Failure truncated case: in this case, we know that the total number of events n before the experiment starts. Therefore, we can get the joint likelihood function for $t_1 < t_2 < \cdots < t_n$

in the failure truncated case based on Equation 8.1.

$$
\begin{aligned}
f(t_1, t_2, \cdots, t_n) &= f(t_1)f(t_2|t_1)f(t_3|t_1,t_2)\cdots f(t_n|t_1,t_2,\ldots,t_{n-1}) \\
&= \lambda(t_1)e^{-\int_0^{t_1}\dot{\lambda}(u)du}\lambda(t_2)e^{-\int_{t_1}^{t_2}\dot{\lambda}(u)du}\cdots\lambda(t_n)e^{-\int_{t_{n-1}}^{t_n}\lambda(u)du} \\
&= \Big(\prod_{i=1}^{n}\lambda(t_i)\Big)e^{-\int_0^{t}\lambda(u)du} \\
&= \Big(\prod_{i=1}^{n}\frac{\beta}{\theta}(\frac{t_i}{\theta})^{\beta-1}\Big)e^{-(t_n/\theta)^{\beta}}, \quad t_1 < t_2 < \cdots < t_n
\end{aligned}
\tag{8.2}
$$

The log-likelihood function in the failure truncated case is therefore:

$$
\log \ell = n\log\beta - n\beta\log\theta + (\beta-1)\Bigg(\sum_{i=1}^{n}\log t_i\Bigg) - \Big(\frac{t_n}{\theta}\Big)^{\beta}
$$

Time truncated case: in this case, we assume that the truncated time is τ. The derivation of $f(t_1, t_2, \cdots, t_n|n)$ is messy in math, we directly give the conclusion here:

$$
f(t_1, t_2, \cdots, t_n|n) = n!\prod_{i=1}^{n}\frac{\lambda(t_i)}{\Lambda(\tau)}
$$

Therefore, the joint likelihood function for $f(n, t_1, t_2, \cdots, t_n)$ is:

$$
\begin{aligned}
f(n, t_1, t_2, \cdots, t_n) &= f(n)f(t_1, t_2, \cdots, t_n|n) \\
&= \frac{e^{-\int_0^{\tau}\lambda(u)du}[\int_0^{\tau}\lambda(u)du]^n}{n!}n!\frac{\prod_{i=1}^{n}\lambda(t_i)}{[\Lambda(\tau)]^n} \\
&= \Big(\prod_{i=1}^{n}\lambda(t_i)\Big)e^{-\int_0^{\tau}\lambda(u)du} \\
&= \Big(\prod_{i=1}^{n}\frac{\beta}{\theta}(\frac{t_i}{\theta})^{\beta-1}\Big)e^{-(\tau/\theta)^{\beta}}, \\
& n = 0, 1, 2, \cdots, \quad 0 < t_1 < t_2 < \cdots < t_n
\end{aligned}
\tag{8.3}
$$

The log likelihood function l is then:

$$
\begin{aligned}
l &= \log\left(\Big(\prod_{i=1}^{n}\frac{\beta}{\theta}(\frac{t_i}{\theta})^{\beta-1}\Big)e^{-(\tau/\theta)^{\beta}}\right) \\
&= \sum_{i=1}^{n}\log\Big(\frac{\beta}{\theta}(\frac{t_i}{\theta})^{\beta-1}\Big) - (\frac{\tau}{\theta})^{\beta} \\
&= n\log\beta - n\beta\log\theta + (\beta-1)\bigg(\sum_{i=1}^{n}\log t_i\bigg) - \Big(\frac{\tau}{\theta}\Big)^{\beta}
\end{aligned}
\tag{8.4}
$$

Bibliography

Abdel-Aty, Mohamed A, Hany M Hassan, Mohamed Ahmed, and Ali S Al-Ghamdi. 2012. "Real-Time Prediction of Visibility Related Crashes." *Transportation Research Part C: Emerging Technologies* 24: 288–98.

Ahmed, Mohamed M, Rebecca Franke, Khaled Ksaibati, and Debbie S Shinstine. 2018. "Effects of Truck Traffic on Crash Injury Severity on Rural Highways in Wyoming Using Bayesian Binary Logit Models." *Accident Analysis & Prevention* 117: 106–13.

Alden, Andrew Scott, Brian Mayer, Patrick Mcgowen, Rini Sherony, and Hiroyuki Takahashi. 2016. "Animal-Vehicle Encounter Naturalistic Driving Data Collection and Photogrammetric Analysis." SAE Technical Paper.

Al-Ghamdi, Ali S. 2007. "Experimental Evaluation of Fog Warning System." *Accident Analysis & Prevention* 39 (6): 1065–72.

Alquier, Pierre, Nial Friel, Richard Everitt, and Aidan Boland. 2016. "Noisy Monte Carlo: Convergence of Markov Chains with Approximate Transition Kernels." *Statistics and Computing* 26 (1-2): 29–47.

Ameratunga, Shanthi, Martha Hijar, and Robyn Norton. 2006. "Road-Traffic Injuries: Confronting Disparities to Address a Global-Health Problem." *The Lancet* 367 (9521): 1533–40.

American Automobile Association Foundation for Traffic Safety. 2010. "Asleep at the Wheel: The Prevalence and Impact of Drowsy Driving." https://www.aaafoundation.org/sites/default/files/2010DrowsyDrivingReport_1.pdf.

Anderson, Jason R, Jeffrey D Ogden, William A Cunningham, and Christine Schubert-

Kabban. 2017. "An Exploratory Study of Hours of Service and Its Safety Impact on Motorists." *Transport Policy* 53: 161–74.

Åkerstedt, Torbjörn. 1988. "Sleepiness as a Consequence of Shift Work." *Sleep* 11 (1): 17–34.

Baker, CJ, and Sheila Reynolds. 1992. "Wind-Induced Accidents of Road Vehicles." *Accident Analysis & Prevention* 24 (6): 559–75.

Bardenet, Rémi, Arnaud Doucet, and Chris Holmes. 2014. "Towards Scaling up Markov Chain Monte Carlo: An Adaptive Subsampling Approach." In.

———. 2017. "On Markov Chain Monte Carlo Methods for Tall Data." *The Journal of Machine Learning Research* 18 (1): 1515–57.

Barnard, Yvonne, Fabian Utesch, Nicole van Nes, Rob Eenink, and Martin Baumann. 2016. "The Study Design of Udrive: The Naturalistic Driving Study Across Europe for Cars, Trucks and Scooters." *European Transport Research Review* 8 (2): 14.

Barp, Alessandro, François-Xavier Briol, Anthony D Kennedy, and Mark Girolami. 2018. "Geometry and Dynamics for Markov Chain Monte Carlo." *Annual Review of Statistics and Its Application* 5: 451–71.

Basagaña, Xavier, Juan Pablo Escalera-Antezana, Payam Dadvand, Òscar Llatje, Jose Barrera-Gómez, Jordi Cunillera, Mercedes Medina-Ramón, and Katherine Pérez. 2015. "High Ambient Temperatures and Risk of Motor Vehicle Crashes in Catalonia, Spain (2000–2011): A Time-Series Analysis." *Environmental Health Perspectives* 123 (12): 1309–16.

Betancourt, Michael. 2015. "The Fundamental Incompatibility of Scalable Hamiltonian Monte Carlo and Naive Data Subsampling." In *International Conference on Machine Learning*, 533–40.

———. 2017. "A Conceptual Introduction to Hamiltonian Monte Carlo." *arXiv Preprint arXiv:1701.02434*.

———. 2019. "The Convergence of Markov Chain Monte Carlo Methods: From the Metropolis Method to Hamiltonian Monte Carlo." *Annalen Der Physik* 531 (3): 1700214.

Betancourt, Michael, and Mark Girolami. 2015. "Hamiltonian Monte Carlo for Hierarchical Models." *Current Trends in Bayesian Methodology with Applications* 79: 30.

Bierkens, Joris, Paul Fearnhead, Gareth Roberts, and others. 2019. "The Zig-Zag Process and Super-Efficient Sampling for Bayesian Analysis of Big Data." *The Annals of Statistics* 47 (3): 1288–1320.

Blanco, Myra, Richard J Hanowski, Rebecca Lynn Olson, Justin F Morgan, Susan A Soccolich, Shih-Ching Wu, and others. 2011. "The Impact of Driving, Non-Driving Work, and Rest Breaks on Driving Performance in Commercial Vehicle Operations."

Braver, Elisa R, Paul L Zador, Denise Thum, Eric L Mitter, Herbert M Baum, and Frank J Vilardo. 1997. "Tractor-Trailer Crashes in Indiana: A Case-Control Study of the Role of Truck Configuration." *Accident Analysis & Prevention* 29 (1): 79–96.

Breiman, Leo, and others. 2001. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." *Statistical Science* 16 (3): 199–231.

Cajochen, Christian, Daniel P Brunner, Kurt Krauchi, Peter Graw, and Anna Wirz-Justice. 1995. "Power Density in Theta/Alpha Frequencies of the Waking Eeg Progressively Increases During Sustained Wakefulness." *Sleep* 18 (10): 890–94.

Campbell, Kenneth L. 1991. "Fatal Accident Involvement Rates by Driver Age for Large Trucks." *Accident Analysis & Prevention* 23 (4): 287–95.

Cantor, David E, Thomas M Corsi, Curtis M Grimm, and Koray Özpolat. 2010. "A Driver Focused Truck Crash Prediction Model." *Transportation Research Part E: Logistics and Transportation Review* 46 (5): 683–92.

Carpenter, Bob, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76 (1).

Cavuoto, Lora, and Fadel Megahed. 2017. "Understanding Fatigue: Implications for Worker Safety." *Professional Safety* 62 (12): 16–19.

Center, Ohio Supercomputer. 1987. "Ohio Supercomputer Center." http://osc.edu/

ark:/19495/f5s1ph73.

Chang, Li-Yen, and Wen-Chieh Chen. 2005. "Data Mining of Tree-Based Models to Analyze Freeway Accident Frequency." *Journal of Safety Research* 36 (4): 365–75.

Chen, Chen, and Yuanchang Xie. 2014. "Modeling the Safety Impacts of Driving Hours and Rest Breaks on Truck Drivers Considering Time-Dependent Covariates." *Journal of Safety Research* 51: 57–63.

Chen, Cong, Guohui Zhang, Xiaoyue Cathy Liu, Yusheng Ci, Helai Huang, Jianming Ma, Yanyan Chen, and Hongzhi Guan. 2016. "Driver Injury Severity Outcome Analysis in Rural Interstate Highway Crashes: A Two-Level Bayesian Logistic Regression Interpretation." *Accident Analysis & Prevention* 97: 69–78.

Chen, Guang Xiang, Youjia Fang, Feng Guo, and Richard J Hanowski. 2016. "The Influence of Daily Sleep Patterns of Commercial Truck Drivers on Driving Performance." *Accident Analysis & Prevention* 91: 55–63.

Chen, Tianqi, Emily Fox, and Carlos Guestrin. 2014. "Stochastic Gradient Hamiltonian Monte Carlo." In *International Conference on Machine Learning*, 1683–91.

Clarke, David D, Patrick Ward, Craig Bartle, and Wendy Truman. 2006. "Young Driver Accidents in the Uk: The Influence of Age, Experience, and Time of Day." *Accident Analysis & Prevention* 38 (5): 871–78.

Craiu, Radu V, and Jeffrey S Rosenthal. 2014. "Bayesian Computation via Markov Chain Monte Carlo." *Annual Review of Statistics and Its Application* 1: 179–201.

Craye, Céline, Abdullah Rashwan, Mohamed S Kamel, and Fakhri Karray. 2016. "A Multi-Modal Driver Fatigue and Distraction Assessment System." *International Journal of Intelligent Transportation Systems Research* 14 (3): 173–94.

Dalal, Koustuv, Zhiquin Lin, Mervyn Gifford, and Leif Svanström. 2013. "Economics of Global Burden of Road Traffic Injuries and Their Relationship with Health System Variables." *International Journal of Preventive Medicine* 4 (12): 1442.

Dang, Khue-Dung, Matias Quiroz, Robert Kohn, Minh-Ngoc Tran, and Mattias Villani.

2019. "Hamiltonian Monte Carlo with Energy Conserving Subsampling." *Journal of Machine Learning Research* 20 (100): 1–31.

Davis, Amelia, Elizabeth Hacker, Peter T Savolainen, and Timothy J Gates. 2015. "Longitudinal Analysis of Rural Interstate Fatalities in Relation to Speed Limit Policies." *Transportation Research Record* 2514 (1): 21–31.

Dement, William C. 1997. "The Perils of Drowsy Driving." *The New England Journal of Medicine* 337 (11): 783–84.

Department of Transportation, Utah. 2019. "TRUCKS Need More Time to Stop." https://www.udot.utah.gov/trucksmart/motorist-home/stopping-distances/.

Di Milia, Lee, Michael H Smolensky, Giovanni Costa, Heidi D Howarth, Maurice M Ohayon, and Pierre Philip. 2011. "Demographic Factors, Fatigue, and Driving Accidents: An Examination of the Published Literature." *Accident Analysis & Prevention* 43 (2): 516–32.

Dingus, Thomas A, Feng Guo, Suzie Lee, Jonathan F Antin, Miguel Perez, Mindy Buchanan-King, and Jonathan Hankey. 2016. "Driver Crash Risk Factors and Prevalence Evaluation Using Naturalistic Driving Data." *Proceedings of the National Academy of Sciences* 113 (10): 2636–41.

Dingus, Thomas A, Richard J Hanowski, and Sheila G Klauer. 2011. "Estimating Crash Risk." *Ergonomics in Design* 19 (4): 8–12.

Dingus, Thomas A, Sheila G Klauer, Vicki Lewis Neale, Andy Petersen, Suzanne E Lee, Jeremy Sudweeks, Miguel A Perez, et al. 2006. "The 100-Car Naturalistic Driving Study. Phase 2: Results of the 100-Car Field Experiment." United States. Department of Transportation. National Highway Traffic Safety ....

Dingus, Thomas A, Vicki L Neale, Sheila G Klauer, Andrew D Petersen, and Robert J Carroll. 2006. "The Development of a Naturalistic Data Collection System to Perform Critical Incident Analysis: An Investigation of Safety and Fatigue Issues in Long-Haul Trucking." *Accident Analysis & Prevention* 38 (6): 1127–36.

Dong, Chunjiao, David B Clarke, Xuedong Yan, Asad Khattak, and Baoshan Huang. 2014. "Multivariate Random-Parameters Zero-Inflated Negative Binomial Regression Model: An Application to Estimate Crash Frequencies at Intersections." *Accident Analysis & Prevention* 70: 320–29.

Dong, Chunjiao, Qiao Dong, Baoshan Huang, Wei Hu, and Shashi S Nambisan. 2017. "Estimating Factors Contributing to Frequency and Severity of Large Truck–Involved Crashes." *Journal of Transportation Engineering, Part A: Systems* 143 (8): 04017032.

Dong, Chunjiao, Shashi S Nambisan, Stephen H Richards, and Zhuanglin Ma. 2015. "Assessment of the Effects of Highway Geometric Design Features on the Frequency of Truck Involved Crashes Using Bivariate Regression." *Transportation Research Part A: Policy and Practice* 75: 30–41.

Duane, Simon, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. 1987. "Hybrid Monte Carlo." *Physics Letters B* 195 (2): 216–22.

Duke, Janine, Maya Guest, and May Boggess. 2010. "Age-Related Safety in Professional Heavy Vehicle Drivers: A Literature Review." *Accident Analysis & Prevention* 42 (2): 364–71.

Eenink, Rob, Yvonne Barnard, Martin Baumann, Xavier Augros, and Fabian Utesch. 2014. "UDRIVE: The European Naturalistic Driving Study." In *Proceedings of Transport Research Arena*. IFSTTAR.

Eugster, Manuel JA, and Thomas Schlesinger. 2013. "Osmar: OpenStreetMap and R." *The R Journal* 5 (1): 53–63.

Evans, Leonard. 2014. "Traffic Fatality Reductions: United States Compared with 25 Other Countries." *American Journal of Public Health* 104 (8): 1501–7.

Fjell, Ylva, Kristina Alexanderson, Mikael Nordenmark, and Carina Bildt. 2008. "Perceived Physical Strain in Paid and Unpaid Work and the Work-Home Interface: The Associations with Musculoskeletal Pain and Fatigue Among Public Employees." *Women & Health* 47 (1): 21–44.

FMCSA. 2016. "Fatal occupational injuries by event, 2016." https://www.fmcsa.dot.gov/sites/fmcsa.dot.gov/files/docs/safety/data-and-statistics/84856/cmvtrafficsafetyfact2017.pdf.

———. 2018a. "Large Truck and Bus Crash Facts 2016." https://www.fmcsa.dot.gov/sites/fmcsa.dot.gov/files/docs/safety/data-and-statistics/398686/ltbcf-2016-final-508c-may-2018.pdf.

———. 2018b. "Large Truck and Bus Crash Facts 2017." https://www.fmcsa.dot.gov/safety/data-and-statistics/large-truck-and-bus-crash-facts-2017.

Gelfand, Alan E, and Adrian FM Smith. 1990. "Sampling-Based Approaches to Calculating Marginal Densities." *Journal of the American Statistical Association* 85 (410): 398–409.

Gelman, Andrew, and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge university press.

Gelman, Andrew, Hal S Stern, John B Carlin, David B Dunson, Aki Vehtari, and Donald B Rubin. 2013. *Bayesian Data Analysis*. Chapman; Hall/CRC.

Geman, Stuart, and Donald Geman. 1987. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images." In *Readings in Computer Vision*, 564–84. Elsevier.

Ghasemzadeh, Ali, and Mohamed M Ahmed. 2017. "A Probit-Decision Tree Approach to Analyze Effects of Adverse Weather Conditions on Work Zone Crash Severity Using Second Strategic Highway Research Program Roadway Information Dataset."

Girotto, Edmarlon, Selma Maffei de Andrade, Alberto Durán González, and Arthur Eumann Mesas. 2016. "Professional Experience and Traffic Accidents/Near-Miss Accidents Among Truck Drivers." *Accident Analysis & Prevention* 95: 299–304.

Gitelman, Victoria, Shlomo Bekhor, Etti Doveh, Fany Pesahov, Roby Carmel, and Smadar Morik. 2018. "Exploring Relationships Between Driving Events Identified by in-Vehicle Data Recorders, Infrastructure Characteristics and Road Crashes." *Transportation Research Part C: Emerging Technologies* 91: 156–75.

Golias, Mihalis, and S Mishra. 2013. "Evaluating the Hours-of-Service Rule via Gps/Gis

Truck Trip Data." *Draft Report. Intermodal Freight Transportation Institute, University of Memphis, Tenn.*

Goonewardene, Sanchia S, Khalid Baloch, Keith Porter, Ian Sargeant, and Gamini Punchihewa. 2010. "Road Traffic Collisions—Case Fatality Rate, Crash Injury Rate, and Number of Motor Vehicles: Time Trends Between a Developed and Developing Country." *The American Surgeon* 76 (9): 977–81.

Gordon, Tim J, Lidia P Kostyniuk, Paul E Green, Michelle A Barnes, Daniel Blower, Adam D Blankespoor, and Scott E Bogard. 2011. "Analysis of Crash Rates and Surrogate Events: Unified Approach." *Transportation Research Record* 2237 (1): 1–9.

Graham, Daniel J, and Stephen Glaister. 2003. "Spatial Variation in Road Pedestrian Casualties: The Role of Urban Scale, Density and Land-Use Mix." *Urban Studies* 40 (8): 1591–1607.

Grimes, David A, and Kenneth F Schulz. 2005. "Compared to What? Finding Controls for Case-Control Studies." *The Lancet* 365 (9468): 1429–33.

Guo, Feng. 2019. "Statistical Methods for Naturalistic Driving Studies." *Annual Review of Statistics and Its Application* 6: 309–28.

Guo, Feng, and Youjia Fang. 2013. "Individual Driver Risk Assessment Using Naturalistic Driving Data." *Accident Analysis & Prevention* 61: 3–9.

Guo, Feng, Youjia Fang, and Jonathan F Antin. 2015. "Older Driver Fitness-to-Drive Evaluation Using Naturalistic Driving Data." *Journal of Safety Research* 54: 49–e29.

Guo, Feng, Sheila G Klauer, Jonathan M Hankey, and Thomas A Dingus. 2010. "Near Crashes as Crash Surrogate for Naturalistic Driving Studies." *Transportation Research Record* 2147 (1): 66–74.

Hanowski, Richard Joseph, RL Olson, J Bocanegra, and Jeffrey S Hickman. 2008. "Analysis of Risk as a Function of Driving-Hour: Assessment of Driving-Hours 1 Through 11." US Department. of Transportation, Federal Motor Carrier Safety Administration.

Hastings, W Keith. 1970. "Monte Carlo Sampling Methods Using Markov Chains and

Their Applications."

Hickman, Jeffrey S, Richard J Hanowski, and Joseph Bocanegra. 2018. "A Synthetic Approach to Compare the Large Truck Crash Causation Study and Naturalistic Driving Data." *Accident Analysis & Prevention* 112: 11–14.

Hoffman, Matthew D, and Andrew Gelman. 2014. "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." *Journal of Machine Learning Research* 15 (1): 1593–1623.

Huang, Helai, and Mohamed Abdel-Aty. 2010. "Multilevel Data and Bayesian Analysis in Traffic Safety." *Accident Analysis & Prevention* 42 (6): 1556–65.

Huang, Yueng-hsiang, Dov Zohar, Michelle M Robertson, Angela Garabet, Jin Lee, and Lauren A Murphy. 2013. "Development and Validation of Safety Climate Scales for Lone Workers Using Truck Drivers as Exemplar." *Transportation Research Part F: Traffic Psychology and Behaviour* 17: 5–19.

Hyder, Adnan A, Katharine A Allen, Gayle Di Pietro, Claudia A Adriazola, Rochelle Sobel, Kelly Larson, and Margie Peden. 2012. "Addressing the Implementation Gap in Global Road Safety: Exploring Features of an Effective Response and Introducing a 10-Country Program." *American Journal of Public Health* 102 (6): 1061–7.

Islam, Samantha, Steven L Jones, and Daniel Dye. 2014. "Comprehensive Analysis of Single-and Multi-Vehicle Large Truck at-Fault Crashes on Rural and Urban Roadways in Alabama." *Accident Analysis & Prevention* 67: 148–58.

Jackson, Melinda L, Gerard A Kennedy, Catherine Clarke, Melissa Gullo, Philip Swann, Luke A Downey, Amie C Hayley, Rob J Pierce, and Mark E Howard. 2016. "The Utility of Automated Measures of Ocular Metrics for Detecting Driver Drowsiness During Extended Wakefulness." *Accident Analysis & Prevention* 87: 127–33.

Janakiraman, Vijay Manikandan, Bryan Matthews, and Nikunj Oza. 2016. "Discovery of Precursors to Adverse Events Using Time Series Data." In *Proceedings of the 2016 Siam International Conference on Data Mining*, 639–47. SIAM.

Jansen, Reinier J, and SWOV Simone Wesseling. 2018. "Harsh Braking by Truck Drivers: A Comparison of Thresholds and Driving Contexts Using Naturalistic Driving Data." In *Proceedings of the 6th Humanist Conference, Disponibile Al Link Https://Bit. Ly/2C2Bw3Z [Online*.

Jiang, Baoguo, Song Liang, Zhong-Ren Peng, Haozhe Cong, Morgan Levy, Qu Cheng, Tianbing Wang, and Justin V Remais. 2017. "Transport and Public Health in China: The Road to a Healthy Future." *The Lancet* 390 (10104): 1781–91.

Jovanis, Paul P, Kun-Feng Wu, and Chen Chen. 2011. "Hours of Service and Driver Fatigue: Driver Characteristics Research."

———. 2012. "Effects of Hours of Service and Driving Patterns on Motor Carrier Crashes." *Transportation Research Record* 2281 (1): 119–27.

Kamla, Jwan, Tony Parry, and Andrew Dawson. 2019. "Analysing Truck Harsh Braking Incidents to Study Roundabout Accident Risk." *Accident Analysis & Prevention* 122: 365–77.

Kampe, Eveline Otte im, Sari Kovats, and Shakoor Hajat. 2016. "Impact of High Ambient Temperature on Unintentional Injuries in High-Income Countries: A Narrative Systematic Literature Review." *BMJ Open* 6 (2): e010399.

Kecklund, Göran, and Torbjörn Åkerstedt. 1993. "Sleepiness in Long Distance Truck Driving: An Ambulatory Eeg Study of Night Driving." *Ergonomics* 36 (9): 1007–17.

Kirwan, Barry, W Huw Gibson, and Brian Hickling. 2008. "Human Error Data Collection as a Precursor to the Development of a Human Reliability Assessment Capability in Air Traffic Management." *Reliability Engineering & System Safety* 93 (2): 217–33.

Kjellstrom, Tord, R Sari Kovats, Simon J Lloyd, Tom Holt, and Richard SJ Tol. 2009. "The Direct Impact of Climate Change on Regional Labor Productivity." *Archives of Environmental & Occupational Health* 64 (4): 217–27.

Klauer, Sheila G, Thomas A Dingus, Vicki L Neale, Jeremy D Sudweeks, and David J Ramsey. 2009. "Comparing Real-World Behaviors of Drivers with High Versus Low Rates of Crashes and Near Crashes."

Knipling, Ronald R. 2015. "Naturalistic Driving Events: No Harm, No Foul, No Validity."

———. 2017. "Threats to Scientific Validity in Truck Driver Hours-of-Service Studies."

Knipling, Ronald R, and Jing-Shiarn Wang. 1994. *Crashes and Fatalities Related to Driver Drowsiness/Fatigue*. National Highway Traffic Safety Administration Washington, DC.

Kononov, Jake, Barbara Bailey, and Bryan K Allery. 2008. "Relationships Between Safety and Both Congestion and Number of Lanes on Urban Freeways." *Transportation Research Record* 2083 (1): 26–39.

Korattikara, Anoop, Yutian Chen, and Max Welling. 2014. "Austerity in Mcmc Land: Cutting the Metropolis-Hastings Budget." In *International Conference on Machine Learning*, 181–89.

Kruschke, John. 2014. *Doing Bayesian Data Analysis: A Tutorial with R, Jags, and Stan*. Academic Press.

Kruschke, John K, and Wolf Vanpaemel. 2015. "Bayesian Estimation in Hierarchical Models." *The Oxford Handbook of Computational and Mathematical Psychology*, 279–99.

Lambert, Ben. 2018. *A Student's Guide to Bayesian Statistics*. Sage.

Lauderdale, Diane S, Kristen L Knutson, Lijing L Yan, Paul J Rathouz, Stephen B Hulley, Steve Sidney, and Kiang Liu. 2006. "Objectively Measured Sleep Characteristics Among Early-Middle-Aged Adults: The Cardia Study." *American Journal of Epidemiology* 164 (1): 5–16.

Leard, Benjamin, Kevin Roth, and others. 2015. "Weather, Traffic Accidents, and Climate Change." *Resources for the Future Discussion Paper*, 15–19.

Lee, Chris, and Mohamed Abdel-Aty. 2008. "Presence of Passengers: Does It Increase or Reduce Driver's Crash Potential?" *Accident Analysis & Prevention* 40 (5): 1703–12.

Lemp, Jason D, Kara M Kockelman, and Avinash Unnikrishnan. 2011. "Analysis of Large Truck Crash Severity Using Heteroskedastic Ordered Probit Models." *Accident Analysis & Prevention* 43 (1): 370–80.

Litman, Todd. 2013. "Transportation and Public Health." *Annual Review of Public Health* 34: 217–33.

Liu, Yi, Feng Guo, and Richard J Hanowski. 2019. "Assessing the Impact of Sleep Time on Truck Driver Performance Using a Recurrent Event Model." *Statistics in Medicine*.

Lord, Dominique. 2006. "Modeling Motor Vehicle Crashes Using Poisson-Gamma Models: Examining the Effects of Low Sample Mean Values and Small Sample Size on the Estimation of the Fixed Dispersion Parameter." *Accident Analysis & Prevention* 38 (4): 751–66.

Lord, Dominique, and Fred Mannering. 2010. "The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives." *Transportation Research Part A: Policy and Practice* 44 (5): 291–305.

Lord, Dominique, Simon Washington, and John N Ivan. 2007. "Further Notes on the Application of Zero-Inflated Models in Highway Safety." *Accident Analysis & Prevention* 39 (1): 53–57.

Lord, Dominique, Simon P Washington, and John N Ivan. 2005. "Poisson, Poisson-Gamma and Zero-Inflated Regression Models of Motor Vehicle Crashes: Balancing Statistical Fit and Theory." *Accident Analysis & Prevention* 37 (1): 35–46.

Lunn, David J, Andrew Thomas, Nicky Best, and David Spiegelhalter. 2000. "WinBUGS-a Bayesian Modelling Framework: Concepts, Structure, and Extensibility." *Statistics and Computing* 10 (4): 325–37.

Lunn, David, David Spiegelhalter, Andrew Thomas, and Nicky Best. 2009. "The Bugs Project: Evolution, Critique and Future Directions." *Statistics in Medicine* 28 (25): 3049–67.

Maclaurin, Dougal, and Ryan Prescott Adams. 2015. "Firefly Monte Carlo: Exact Mcmc with Subsets of Data." In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

MacLean, Alistair W, David RT Davies, and Kris Thiele. 2003. "The Hazards and Prevention of Driving While Sleepy." *Sleep Medicine Reviews* 7 (6): 507–21.

McCauley, Peter, Leonid V Kalachev, Daniel J Mollicone, Siobhan Banks, David F

Dinges, and Hans PA Van Dongen. 2013. "Dynamic Circadian Modulation in a Biomathematical Model for the Effects of Sleep and Sleep Loss on Waking Neurobehavioral Performance." *Sleep* 36 (12): 1987–97.

McElreath, Richard. 2018. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman; Hall/CRC.

Metropolis, Nicholas, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. 1953. "Equation of State Calculations by Fast Computing Machines." *The Journal of Chemical Physics* 21 (6): 1087–92.

Meuleners, Lynn, Michelle L Fraser, Matthew H Govorko, and Mark R Stevenson. 2015. "Obstructive Sleep Apnea, Health-Related Factors, and Long Distance Heavy Vehicle Crashes in Western Australia: A Case Control Study." *Journal of Clinical Sleep Medicine* 11 (04): 413–18.

———. 2017. "Determinants of the Occupational Environment and Heavy Vehicle Crashes in Western Australia: A Case–Control Study." *Accident Analysis & Prevention* 99: 452–58.

Mitler, Merrill M, James C Miller, Jeffrey J Lipsitz, James K Walsh, and C Dennis Wylie. 1997. "The Sleep of Long-Haul Truck Drivers." *New England Journal of Medicine* 337 (11): 755–62.

Mohammadi, Mojtaba A, VA Samaranayake, and Ghulam H Bham. 2014. "Crash Frequency Modeling Using Negative Binomial Models: An Application of Generalized Estimating Equation to Longitudinal Data." *Analytic Methods in Accident Research* 2: 52–69.

Mollicone, Daniel, Kevin Kan, Chris Mott, Rachel Bartels, Steve Bruneau, Matthew van Wollen, Amy R Sparrow, and Hans PA Van Dongen. 2019. "Predicting Performance and Safety Based on Driver Fatigue." *Accident Analysis & Prevention* 126: 142–45.

Moneta, Giovanni B, Annette Leclerc, Jean-François Chastang, Patrick Dang Tran, and Marcel Goldberg. 1996. "Time-Trend of Sleep Disorder in Relation to Night Work: A Study of Sequential 1-Year Prevalences Within the Gazel Cohort." *Journal of Clinical Epidemiology*

49 (10): 1133–41.

Monnahan, Cole C, James T Thorson, and Trevor A Branch. 2017. "Faster Estimation of Bayesian Models in Ecology Using Hamiltonian Monte Carlo." *Methods in Ecology and Evolution* 8 (3): 339–48.

Moudon, Anne Vernez, Lin Lin, Junfeng Jiao, Philip Hurvitz, and Paula Reeves. 2011. "The Risk of Pedestrian Injury and Fatality in Collisions with Motor Vehicles, a Social Ecological Study of State Routes and City Streets in King County, Washington." *Accident Analysis & Prevention* 43 (1): 11–24.

Mulder, LJM. 1992. "Measurement and Analysis Methods of Heart Rate and Respiration for Use in Applied Environments." *Biological Psychology* 34 (2-3): 205–36.

Naik, Bhaven, Li-Wei Tung, Shanshan Zhao, and Aemal J Khattak. 2016. "Weather Impacts on Single-Vehicle Truck Crash Injury Severity." *Journal of Safety Research* 58: 57–65.

Nakayama, Haruyuki. 2002. "TRIAL Measurments of Driver Fatigue in Extended Driving Condition." In *9th World Congress on Intelligent Transport Systemsits America, Its Japan, Ertico (Intelligent Transport Systems and Services-Europe)*.

National Crime Records Bureau, Government of India. 2015. "NCRB 2016 Report, Chapter 1A: Traffic Accidents." http://ncrb.gov.in/StatPublications/ADSI/ADSI2015/chapter-1A%20traffic%20accidents.pdf.

National Sleep Foundation. 2008. "2008 State of the States Report on Drowsy Driving." http://drowsydriving.org/resources/2008-state-of-the-states-report-on-drowsy-driving/.

National Transportation Safety Board. 1990. "Safety Study: Fatigue, Alcohol, Other Drugs, and Medical Factors in Fatal-to-the-Driver Heavy Truck Crashes." National Transportation Safety Board.

Neal, Radford M, and others. 2011. "MCMC Using Hamiltonian Dynamics." *Handbook of Markov Chain Monte Carlo* 2 (11): 2.

Neale, Vicki L, Thomas A Dingus, Sheila G Klauer, Jeremy Sudweeks, and Michael Goodman. 2005. "An Overview of the 100-Car Naturalistic Study and Findings." *National Highway Traffic Safety Administration, Paper* 5: 0400.

Neeley, Grant W, and Lilliard E Richardson Jr. 2009. "The Effect of State Regulations on Truck-Crash Fatalities." *American Journal of Public Health* 99 (3): 408–15.

Née, Mélanie, Benjamin Contrand, Ludivine Orriols, Cédric Gil-Jardiné, Cedric Galéra, and Emmanuel Lagarde. 2019. "Road Safety and Distraction, Results from a Responsibility Case-Control Study Among a Sample of Road Users Interviewed at the Emergency Room." *Accident Analysis & Prevention* 122: 19–24.

Noland, Robert B, and Lyoong Oh. 2004. "The Effect of Infrastructure and Demographic Change on Traffic-Related Fatalities and Crashes: A Case Study of Illinois County-Level Data." *Accident Analysis & Prevention* 36 (4): 525–32.

Olson, Ryan, Brad Wipfli, Sharon V Thompson, Diane L Elliot, W Kent Anger, Todd Bodner, Leslie B Hammer, and Nancy A Perrin. 2016. "Weight Control Intervention for Truck Drivers: The Shift Randomized Controlled Trial, United States." *American Journal of Public Health* 106 (9): 1698–1706.

Otmani, Sarah, Joceline Rogé, and Alain Muzet. 2005. "Sleepiness in Professional Drivers: Effect of Age and Time of Day." *Accident Analysis & Prevention* 37 (5): 930–37.

Pack, Allan I, Andrew M Pack, Eric Rodgman, Andrew Cucchiara, David F Dinges, and C William Schwab. 1995. "Characteristics of Crashes Attributed to the Driver Having Fallen Asleep." *Accident Analysis & Prevention* 27 (6): 769–75.

Pahukula, Jasmine, Salvador Hernandez, and Avinash Unnikrishnan. 2015. "A Time of Day Analysis of Crashes Involving Large Trucks in Urban Areas." *Accident Analysis & Prevention* 75: 155–63.

Pande, Anurag, Sai Chand, Neeraj Saxena, Vinayak Dixit, James Loy, Brian Wolshon, and Joshua D Kent. 2017. "A Preliminary Investigation of the Relationships Between Historical Crash and Naturalistic Driving." *Accident Analysis & Prevention* 101: 107–16.

Patel, Miteshkumar, Sara KL Lal, Diarmuid Kavanagh, and Peter Rossiter. 2011. "Applying Neural Network Analysis on Heart Rate Variability Data to Assess Driver Fatigue." *Expert Systems with Applications* 38 (6): 7235–42.

Peden, Margie, Richard Scurfield, David Sleet, Dinesh Mohan, Adnan A Hyder, Eva Jarawan, Colin D Mathers, and others. 2004. "World Report on Road Traffic Injury Prevention." World Health Organization Geneva.

Phimister, James R, Ulku Oktem, Paul R Kleindorfer, and Howard Kunreuther. 2003. "Near-Miss Incident Management in the Chemical Process Industry." *Risk Analysis: An International Journal* 23 (3): 445–59.

Popkin, Stephen M, Stephanie L Morrow, Tara E Di Domenico, and Heidi D Howarth. 2008. "Age Is More Than Just a Number: Implications for an Aging Workforce in the Us Transportation Sector." *Applied Ergonomics* 39 (5): 542–49.

Pylkkönen, M, M Sihvola, HK Hyvärinen, S Puttonen, C Hublin, and M Sallinen. 2015. "Sleepiness, Sleep, and Use of Sleepiness Countermeasures in Shift-Working Long-Haul Truck Drivers." *Accident Analysis & Prevention* 80: 201–10.

Quiroz, Matias. 2015. "Bayesian Inference in Large Data Problems." PhD thesis, Department of Statistics, Stockholm University.

Quiroz, Matias, Robert Kohn, Mattias Villani, and Minh-Ngoc Tran. 2019. "Speeding up Mcmc by Efficient Data Subsampling." *Journal of the American Statistical Association* 114 (526): 831–43.

Quiroz, Matias, Minh-Ngoc Tran, Mattias Villani, and Robert Kohn. 2018. "Speeding up Mcmc by Delayed Acceptance and Data Subsampling." *Journal of Computational and Graphical Statistics* 27 (1): 12–22.

Quiroz, Matias, Minh-Ngoc Tran, Mattias Villani, Robert Kohn, and Khue-Dung Dang. 2016. "The Block-Poisson Estimator for Optimally Tuned Exact Subsampling Mcmc." *arXiv Preprint arXiv:1603.08232*.

Quiroz, Matias, Mattias Villani, Robert Kohn, Minh-Ngoc Tran, and Khue-Dung Dang.

2018. “Subsampling Mcmc-an Introduction for the Survey Statistician.” *Sankhya A*, 1–37.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rifaat, Shakil Mohammad, Richard Tay, and Alexandre De Barros. 2012. “Severity of Motorcycle Crashes in Calgary.” *Accident Analysis & Prevention* 49: 44–49.

Roshandel, Saman, Zuduo Zheng, and Simon Washington. 2015. “Impact of Real-Time Traffic Characteristics on Freeway Crash Occurrence: Systematic Review and Meta-Analysis.” *Accident Analysis & Prevention* 79: 198–211.

Rotenberg, Lúcia, Luciana Fernandes Portela, Bahby Banks, Rosane Harter Griep, Frida Marina Fischer, and Paul Landsbergis. 2008. “A Gender Approach to Work Ability and Its Relationship to Professional and Domestic Work Hours Among Nursing Personnel.” *Applied Ergonomics* 39 (5): 646–52.

Rudis, Bob. 2018. “darksky: An R interface to the Dark Sky API.” GitHub. https://github.com/hrbrmstr/darksky.

Saleh, Joseph H, Elizabeth A Saltmarsh, Francesca M Favaro, and Loic Brevault. 2013. “Accident Precursors, Near Misses, and Warning Signs: Critical Review and Formal Definitions Within the Framework of Discrete Event Systems.” *Reliability Engineering & System Safety* 114: 148–54.

Sallinen, Mikael, Mikko HÄRMÄ, Pertti Mutanen, Riikka RANTA, Jussi Virkkala, and Kiti MÜLLER. 2005. “Sleepiness in Various Shift Combinations of Irregular Shift Systems.” *Industrial Health* 43 (1): 114–22.

Savolainen, Peter T, Fred L Mannering, Dominique Lord, and Mohammed A Quddus. 2011. “The Statistical Analysis of Highway Crash-Injury Severities: A Review and Assessment of Methodological Alternatives.” *Accident Analysis & Prevention* 43 (5): 1666–76.

Sedgwick, Philip. 2014. “Case-Control Studies: Advantages and Disadvantages.” *Bmj* 348: f7707.

Shmueli, Galit, and others. 2010. “To Explain or to Predict?” *Statistical Science* 25 (3):

289–310.

Simons-Morton, Bruce G, Zhiwei Zhang, John C Jackson, and Paul S Albert. 2012. "Do Elevated Gravitational-Force Events While Driving Predict Crashes and Near Crashes?" *American Journal of Epidemiology* 175 (10): 1075–9.

Smith, Clyde L, and Emanuele Borgonovo. 2007. "Decision Making During Nuclear Power Plant Incidents—a New Approach to the Evaluation of Precursor Events." *Risk Analysis: An International Journal* 27 (4): 1027–42.

Soccolich, Susan A, Myra Blanco, Richard J Hanowski, Rebecca L Olson, Justin F Morgan, Feng Guo, and Shih-Ching Wu. 2013. "An Analysis of Driving and Working Hour on Commercial Motor Vehicle Driver Safety Using Naturalistic Data Collection." *Accident Analysis & Prevention* 58: 249–58.

Solomon, Andrew J, John T Doucette, Elizabeth Garland, and Thomas McGinn. 2004. "Healthcare and the Long Haul: Long Distance Truck Drivers—a Medically Underserved Population." *American Journal of Industrial Medicine* 46 (5): 463–71.

Sparrow, Amy R, Daniel J Mollicone, Kevin Kan, Rachel Bartels, Brieann C Satterfield, Samantha M Riedy, Aaron Unice, and Hans PA Van Dongen. 2016. "Naturalistic Field Study of the Restart Break in Us Commercial Motor Vehicle Drivers: Truck Driving, Sleep, and Fatigue." *Accident Analysis & Prevention* 93: 55–64.

Stan Development Team. 2018. "RStan: The R Interface to Stan." http://mc-stan.org/.

Staton, Catherine, Joao Vissoci, Enying Gong, Nicole Toomey, Rebeccah Wafula, Jihad Abdelgadir, Yi Zhou, et al. 2016. "Road Traffic Injury Prevention Initiatives: A Systematic Review and Metasummary of Effectiveness in Low and Middle Income Countries." *PLoS One* 11 (1): e0144971.

Stern, Hal S, Daniel Blower, Michael L Cohen, Charles A Czeisler, David F Dinges, Joel B Greenhouse, Feng Guo, et al. 2019. "Data and Methods for Studying Commercial Motor Vehicle Driver Fatigue, Highway Safety and Long-Term Driver Health." *Accident Analysis &*

Prevention 126: 37–42.

The Dark Sky API. 2019. "Data Sources." https://darksky.net/dev/docs/sources.

The Dark Sky Company, LLC. 2019. "Dark Sky API — Overview." https://darksky.net/dev/docs.

The National Safety Council. 2018. "Vehicle Deaths Estimated at 40,000 for Third Straight Year." https://www.nsc.org/road-safety/safety-topics/fatality-estimates.

Theofilatos, Athanasios, and George Yannis. 2014. "A Review of the Effect of Traffic and Weather Characteristics on Road Safety." *Accident Analysis & Prevention* 72: 244–56.

Theofilatos, Athanasios, George Yannis, Pantelis Kopelias, and Fanis Papadimitriou. 2016. "Predicting Road Accidents: A Rare-Events Modeling Approach." *Transportation Research Procedia* 14: 3399–3405.

———. 2018. "Impact of Real-Time Traffic Characteristics on Crash Occurrence: Preliminary Results of the Case of Rare Events." *Accident Analysis & Prevention*.

The United States, Bureau of Labor Statistics. 2017. "Fatal occupational injuries by event, 2017." https://www.bls.gov/charts/census-of-fatal-occupational-injuries/fatal-occupational-injuries-by-event-drilldown.htm.

Van Ravenzwaaij, Don, Pete Cassey, and Scott D Brown. 2018. "A Simple Introduction to Markov Chain Monte–Carlo Sampling." *Psychonomic Bulletin & Review* 25 (1): 143–54.

Wang, Chao, Mohammed A Quddus, and Stephen G Ison. 2013. "The Effect of Traffic and Road Characteristics on Road Safety: A Review and Future Research Direction." *Safety Science* 57: 264–75.

Wang, Ling, Mohamed Abdel-Aty, and Jaeyoung Lee. 2017. "Safety Analytics for Integrating Crash Frequency and Real-Time Risk Modeling for Expressways." *Accident Analysis & Prevention* 104: 58–64.

Wang, Ziqi, Marco Broccardo, and Junho Song. 2019. "Hamiltonian Monte Carlo Methods for Subset Simulation in Reliability Analysis." *Structural Safety* 76: 51–67.

Washington, Simon P, Matthew G Karlaftis, and Fred Mannering. 2010. *Statistical and*

Econometric Methods for Transportation Data Analysis. Chapman; Hall/CRC.

WHO. 2018a. "Road traffic injuries." http://www.who.int/mediacentre/factsheets/fs358/en/.

———. 2018b. "The Top 10 Causes of Death." http://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.

Wikipedia contributors. 2019. "OpenStreetMap — Wikipedia, the Free Encyclopedia." https://en.wikipedia.org/w/index.php?title=OpenStreetMap&oldid=900226891.

Wu, Kun-Feng, Jonathan Aguero-Valverde, and Paul P Jovanis. 2014. "Using Naturalistic Driving Data to Explore the Association Between Traffic Safety-Related Events and Crash Risk at Driver Level." *Accident Analysis & Prevention* 72: 210–18.

Wu, Kun-Feng, and Paul P Jovanis. 2012. "Crashes and Crash-Surrogate Events: Exploratory Modeling with Naturalistic Driving Data." *Accident Analysis & Prevention* 45: 507–16.

———. 2013. "Defining and Screening Crash Surrogate Events Using Naturalistic Driving Data." *Accident Analysis & Prevention* 61: 10–22.

Xie, Yuanchang, Yunlong Zhang, and Faming Liang. 2009. "Crash Injury Severity Analysis Using Bayesian Ordered Probit Models." *Journal of Transportation Engineering* 135 (1): 18–25.

Xu, Chengcheng, Wei Wang, Pan Liu, and Zhibin Li. 2015. "Calibration of Crash Risk Models on Freeways with Limited Real-Time Traffic Data Using Bayesian Meta-Analysis and Bayesian Inference Approach." *Accident Analysis & Prevention* 85: 207–18.

Ye, Fan, and Dominique Lord. 2011. "Investigation of Effects of Underreporting Crash Data on Three Commonly Used Traffic Crash Severity Models: Multinomial Logit, Ordered Probit, and Mixed Logit." *Transportation Research Record* 2241 (1): 51–58.

Yu, Rongjie, and Mohamed Abdel-Aty. 2014. "Using Hierarchical Bayesian Binary Probit Models to Analyze Crash Injury Severity on High Speed Facilities with Real-Time Traffic Data." *Accident Analysis & Prevention* 62: 161–67.

Yung, Marcus. 2016. "Fatigue at the Workplace: Measurement and Temporal Development."

Zaloshnja, Eduard, Ted Miller, and others. 2008. "Unit Costs of Medium and Heavy Truck Crashes." The United States. Federal Motor Carrier Safety Administration.

Zaloshnja, E, and TA Miller. 2007. "Unit Costs of Medium and Heavy Trucks." Report No. FMCSA-RRA-07-034). Washington, DC: Federal Motor Carrier Safety ….

Zhang, Wei, Omer Tsimhoni, Michael Sivak, and Michael J Flannagan. 2010. "Road Safety in China: Analysis of Current Challenges." *Journal of Safety Research* 41 (1): 25–30.

Zhang, Xingjian, Xiaohua Zhao, Hongji Du, and Jian Rong. 2014. "A Study on the Effects of Fatigue Driving and Drunk Driving on Drivers' Physical Characteristics." *Traffic Injury Prevention* 15 (8): 801–8.

Zhu, Xiaoyu, and Sivaramakrishnan Srinivasan. 2011. "A Comprehensive Analysis of Factors Influencing the Injury Severity of Large-Truck Crashes." *Accident Analysis & Prevention* 43 (1): 49–57.

Vita Auctoris

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