

Modeling Truck Safety Critical Events

Efficient Bayesian Hierarchical Statistical and Reliability Models

Miao Cai

Department of Epidemiology and Biostatistics
An Oral Comprehensive Exam | October 7, 2019



Dissertation committee

- Chair:
 - Steven E. Rigdon, PhD
- Committee Members:
 - Fadel Megahed, PhD (Farmer School of Business, Miami University)
 - Hong Xian, PhD
- At-large Committee:
 - Kenton Johnson, PhD
 - Juliet Iwelunmor, PhD

1 The problem

Transportation and trucks

Transportation safety deserves attention:

- The 8-th leading cause of death globally in 2016,¹
- 1.4 million people were killed, mostly aged 4 to 44 years old,¹
- a loss of 518 billion dollars.²

Trucks are the backbone of the economy:

- 70% of freights were delivered by trucks,
- 71.3% of domestic goods and 73.1% of value,^{3,4}

Challenges for trucking industry

Drivers:

1. drive alone for long hours,
2. work under time demands, challenging weather and traffic conditions,
3. sleep deprivation and disorders

Trucks:

1. huge weights,
2. large physical dimensions,
3. potentially carry hazardous cargoes.

Truck crash studies

Traditional studies almost exclusively use data that ultimately trace back to **post hoc vehicle inspection, interviews** with survived drivers and witnesses, and **police reports**.^{5,6}

1. rare events → difficulty in estimation,⁷
2. retrospective studies → recall bias,⁸
3. crashes are under-reported → selection bias.^{9,10}

Naturalistic driving studies (NDS)

NDS uses

unobtrusive devices, sensors, and cameras installed on vehicles to proactively collect frequent naturalistic driving behavior and performance data under real-world driving conditions^{5,11}

1. driver-based data, not road segment-based,
2. high-resolution driver behavior and performance data,
3. less costly and difficult per observation.

Safety Critical Events (SCEs)

SCEs are

a chain of adverse events following an initial off-nominal event, which can result in an accident if compounded with additional adverse conditions.¹²

Examples of SCEs are:

1. hard brakes,
2. headways,
3. rolling stability,
4. collision mitigation

The problem

NDSs are relatively *new* and *less studied*. Here are **several problems** in NDS.

1. Are SCEs indicative of **real crashes** among truck drivers?
2. Can we **predict** SCEs?
3. How can we **innovate existing models** to account for features of NDS?

2 Literature review

Association between crashes and SCEs

Examples of studies supporting SCEs:

- hard braking events were significantly associated with collisions and near-crashes,¹³
- a significant positive association between crashes, near crashes, and crash-relevant incidents,¹⁴
- ...

Examples of studies that are against SCEs:

- overspeed negatively associated with injury crashes,¹⁵
- no harm, no validity,¹⁶
- no demonstration on causal link between SCEs and injury crashes.¹⁷

Gaps:

- Limited number of drivers → less convincing (< 100),
- No studies specifically on truck drivers.

Fatigue

The **most important factor** in transportation safety studies. Fatigue is

a multidimensional process that leads to diminished worker performance, which may be a result of prolonged work, psychological, socioeconomic, and environment factors

- 16.5% of fatal traffic accidents,¹⁸
- 12.5% of injuries-related collisions,¹⁸,
- 60% of fatal truck crashes.¹⁹

However, fatigue is hard to measure in transportation safety studies.

- ocular and physiological metrics,
- sleep patterns,
- **cumulative driving time.**

Other risk factors

Four aspects of risk factors are included in previous studies:

- Driver characteristics,
- Weather
- Traffic
- Road features
- ...

Gaps in literature:

1. Lack of **high-resolution** weather and traffic data,
2. No fusion of NDS and **API data**.

Statistical models

- Logistic regression,
- Poisson regression,
- machine learning models,
- ...

Gaps in literature:

1. Road-centric models, not driver-centric models,
2. Maximum likelihood estimation (MLE) limited in rare-event models,
3. Lack of recurrent events models.

Bayesian models

In the Bayesian perspective, parameters are viewed as **random variables** that have probability distributions.²⁰

$$\begin{aligned} p(\theta|\mathbf{X}) &= \frac{p(\theta)p(\mathbf{X}|\theta)}{p(\mathbf{X})} \\ &= \frac{p(\theta)p(\mathbf{X}|\theta)}{\int p(\theta)p(\mathbf{X}|\theta)d\theta} \end{aligned} \tag{1}$$

- $p(\theta)$: subjective priors,
- $p(\mathbf{X}|\theta)$: the likelihood function,
- $p(\mathbf{X}) = \int p(\theta)p(\mathbf{X}|\theta)d\theta$: the normalizing constant, **trickiest** part,
- $p(\theta|\mathbf{X})$: the posterior distribution.

The posterior distribution is a balance between the **prior beliefs** and the **likelihood function**.

Challenges for Bayesian models in a big data setting

Modern Bayesian inferences relies on **Markov chain Monte Carlo (MCMC)** to overcome the intractable denominator issue. However, MCMC is not scalable in the big data setting:

- **Tall data** (a lot of observations),
- **Wide data** (a lot of variables),
- **Correlation between variables** (hierarchical models).²¹

Potential solutions:

- [Hamiltonian Monte Carlo](#),²²
- [Subsampling MCMC](#) such as Energy Conserving Subsampling Hamiltonian Monte Carlo (ECS-HMC).²³

Conceptual framework

1. *Truck Driver Fatigue Model*,²⁴
2. *5× ST-level hierarchy theory in traffic safety*,²⁵
3. *Commercial motor vehicle driver fatigue framework*.⁶

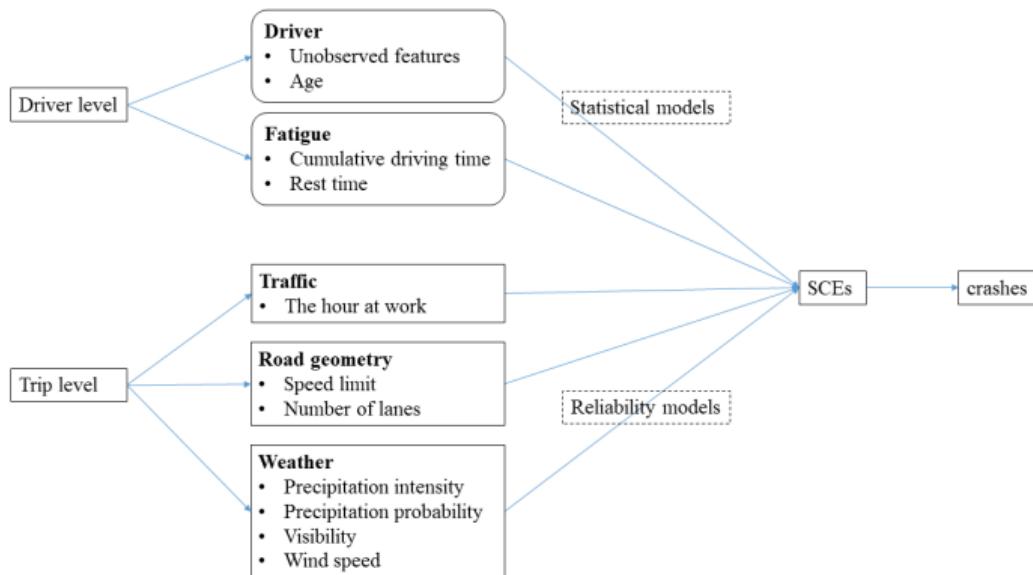


Figure 1: Conceptual model. SCEs represent safety critical events.

3 Research aims

Overall aim

Gaps in previous literature:

1. The association between **crashes and SCEs** has not been confirmed among truck drivers,
2. Difficulty in **fusing high-resolution NDS and API data**,
3. Bayesian inference is not **scalable** in tall and wide NDS data setting,
4. **Recurrent events models** were not widely applied in NDS data.

The overarching goal of this proposed dissertation is to construct **scalable Bayesian hierarchical models** for NDS data and understand how **cumulative driving time** and other environmental factors will impact the performance of truck drivers.

Aim 1

To examine the association between truck crashes and SCEs using a Bayesian Gamma-Poisson regression.

I hypothesize that the rate of crashes is positively associated with the rate of SCEs among the truck drivers controlling for the miles driven and other covariates.

Aim 2

To construct three scalable Bayesian hierarchical models to identify potential risk factors for SCEs.

I hypothesize that the patterns of SCEs vary significantly from drivers to drivers and can be predicted using cumulative driving time, weather, road geometry, driver's age, and other factors}.

- 2a) Bayesian hierarchical **logistic** regression,
- 2b) Bayesian hierarchical **Poisson** regression,
- 2c) Bayesian hierarchical **non-homogeneous Poisson process (NHPP)** with the power law process (PLP) intensity function.

Aim 3

To propose an innovative reliability model that accounts for both within shift cumulative driving time and between-trip rest time.

I hypothesize that **between-trip rest time** can **recover** the intensity function by **some proportion κ** , and intensity function varies significantly from drivers to drivers.

4 Data

Data sources

1. **Real-time ping:** vehicle number, date and time, latitude, longitude, driver identification number (ID), and speed at that second (every 2-10 minutes), ~1.4 billion pings (150 GB .csv file),
2. **Truck crashes and SCEs:** hard brakes, headways, and rolling stability were collected if kinematic thresholds were met,
3. **Driver demographics:** age,
4. **Weather from the DarkSky API** (500 drivers): precipitation intensity, precipitation probability, wind speed, and visibility,
5. **Road geometry from the OpenStreetMap** (500 drivers): speed limits and the number of lanes.

Demonstration of data I

Table 1: A demonstration of ping data

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

Demonstration of data II

Table 2: A demonstration of 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

Demonstration of data III

Table 3: A demonstration of crashes table

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

Demonstration of data IV

Table 4: A demonstration of drivers table

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

Table 5: A demonstration of weather data from the DarkSky API

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

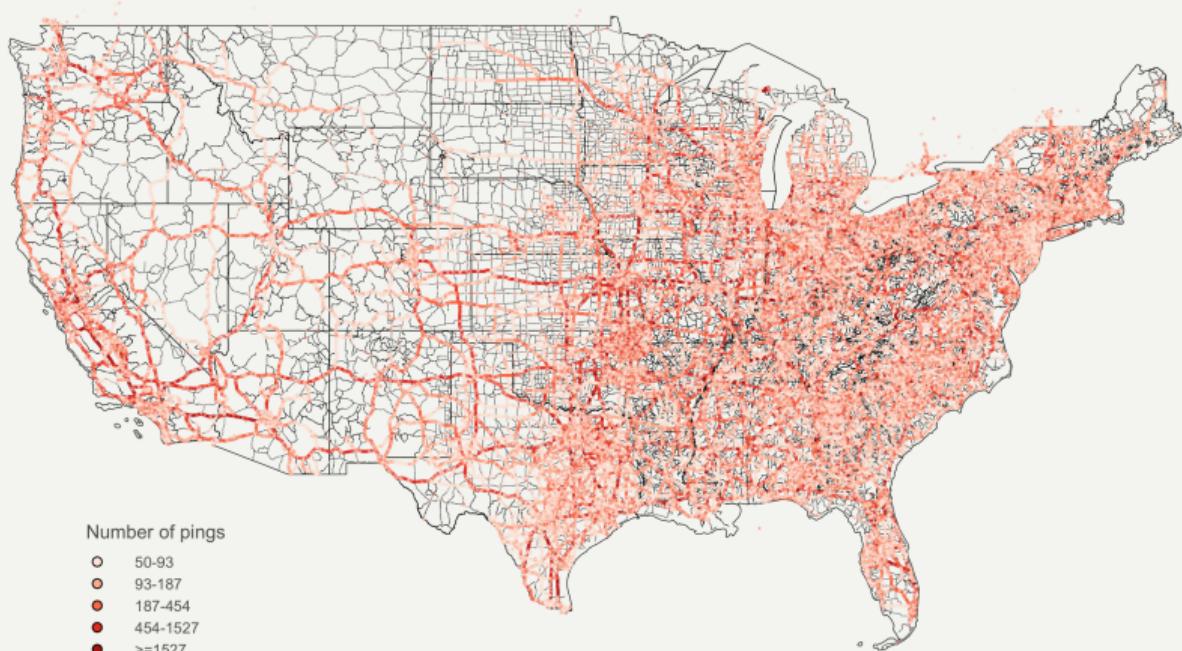
Demonstration of data V

Table 6: A demonstration of road geometry data from the OpenStreetMap API

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

Geographical distribution of active moving pings

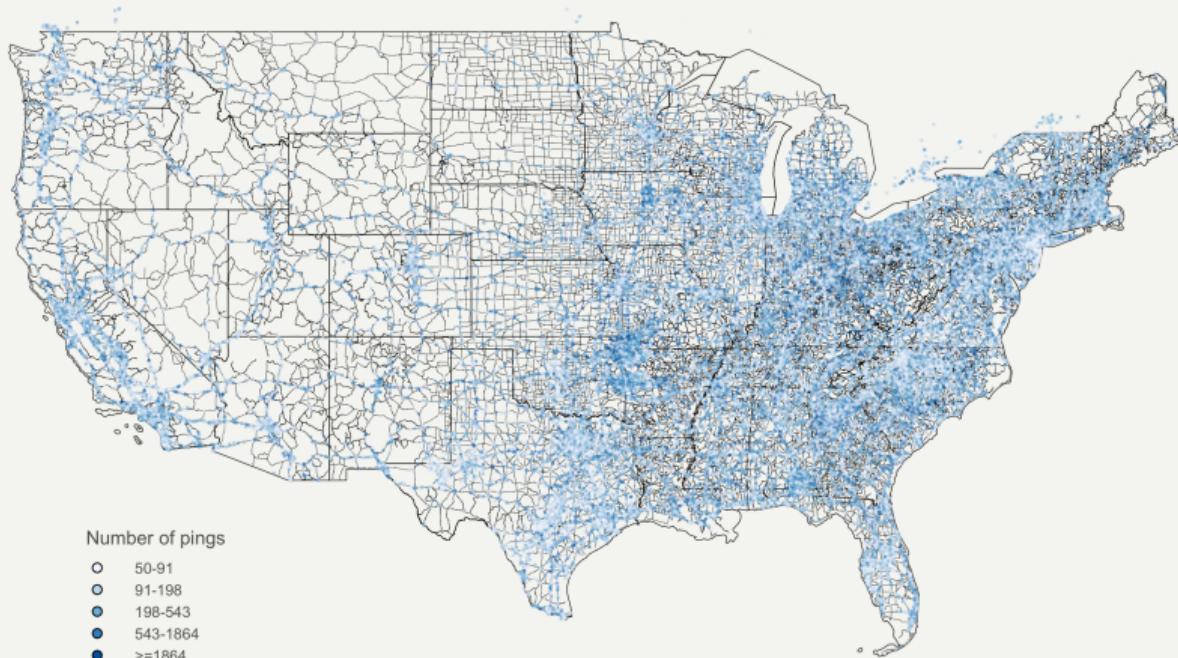
A large commercial truck NDS data set in USA, 2015-2016



The grey line are major highways in the USA. Only locations with at least 50 pings were shown.

Geographical distribution of stopped pings

A large commercial truck NDS data set in USA, 2015-2016



The grey lines are major highways in the USA. Only locations with at least 50 pings were shown.

Data aggregation

1. **Trip:** 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 (~200,000 rows),
2. **30-minute intervals:** as the length of a trip can vary significantly from 5 minutes to more than 8 hours, I will transform the trips data into **standardized 30-minute fixed intervals** according to the starting and ending time of trips (~1 million rows),
3. **Shift:** the trips data will be further divided into different shifts if the specific driver was **not moving for eight hours**.

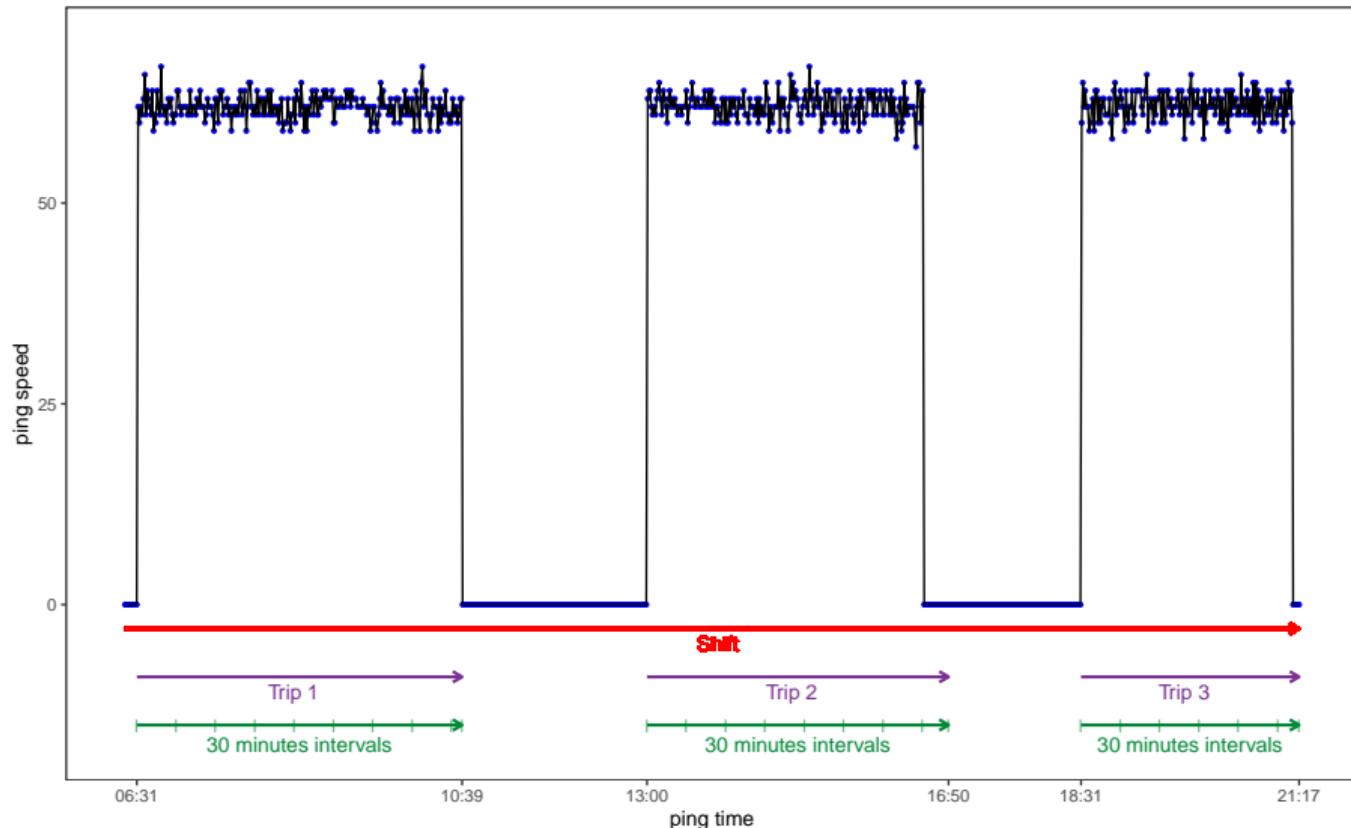


Table 1: ping data

trip_id	ping_time	speed	latitude	longitude	driver
100160724	2015-10-23 08:09:26	5	33.94288	-118.1681	canj1
100160724	2015-10-23 08:22:58	4	33.97146	-118.1677	canj1
100160724	2015-10-23 08:23:12	8	33.97178	-118.1677	canj1
100160724	2015-10-23 08:23:30	4	33.97233	-118.1678	canj1
100160725	2015-10-23 08:09:00	60	34.00753	-118.1844	canj1
100160725	2015-10-23 08:04:34	16	34.00753	-118.1844	canj1
100160725	2015-10-23 09:08:00	16	34.00984	-118.1939	canj1
100160725	2015-10-23 09:08:00	16	34.00982	-118.1938	canj1
100160725	2015-10-23 09:08:08	23	34.01016	-118.1940	canj1
100160725	2015-10-23 09:08:10	10	34.01016	-118.1940	canj1
100160725	2015-10-23 09:08:10	20	34.01024	-118.1951	canj1
100160725	2015-10-23 09:08:24	21	34.01008	-118.1964	canj1
100160725	2015-10-23 09:22:56	4	33.99553	-118.1747	canj1
100160725	2015-10-23 09:23:12	2	33.99721	-118.1748	canj1
100160725	2015-10-23 09:23:38	2	33.99721	-118.1748	canj1
100160725	2015-10-23 09:29:20	0	33.99440	-118.1758	canj1
100160725	2015-10-23 09:38:10	0	33.99441	-118.1758	canj1
100160725	2015-10-23 09:43:02	9	33.99571	-118.1757	canj1
100160725	2015-10-23 09:52:22	6	33.99815	-118.1819	canj1
100160725	2015-10-23 09:53:12	8	33.99812	-118.1816	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
kelli0	51
lewri10	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

Data demonstration I

Table 7: 30 minutes intervals data for hierarchical logistic and Poisson regression

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

Table 8: shifts data for hierarchical non-homogeneous Poisson process

driver	start_time	end_time	shift_length	n_SCE	SCE_time	SCE_type
canj1	2015-10-23T08:09:26Z	2015-10-23T18:37:56Z	628	1	2015-10-23 14:46:08	HB
canj1	2015-10-26T07:49:04Z	2015-10-26T15:06:58Z	437	1	2015-10-26 15:06:03	HB
canj1	2015-10-27T01:59:48Z	2015-10-27T07:58:56Z	359	0	NA	NA
canj1	2015-10-28T08:05:08Z	2015-10-28T20:20:32Z	735	2	2015-10-28 11:58:24;2015-10-28 17:42:36	HB;HB
canj1	2015-10-30T09:27:12Z	2015-10-30T21:18:22Z	711	0	NA	NA

Data demonstration II

Table 9: SCEs data for hierarchical non-homogeneous Poisson process

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

5 Methods

Aim 1 - Data and variables

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

- **Data:**
 - over 50,000 commercial truck drivers,
 - 1,494,678,173 pings,
 - 35,008 crashes, 480,331 SCEs
- **outcome variable:** the number of crashes for each driver.
- **The primary independent variable:** 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:** the total miles driven, the percent of night driving, and the age of the drivers.

Aim 1 - Gamma-Poisson model

Here is how the proposed Gamma-Poisson model will be implemented. Let us assume that:

$$\begin{aligned}\lambda &\sim \text{Gamma}(\alpha, \beta) \\ X|\lambda &\sim \text{Poisson}(\lambda)\end{aligned}$$

Then we have:

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

The Gamma-Poisson distribution is a α -parameter distribution. The log-linear Gamma-Poisson model will be specified as:

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

- \mathbf{X} is the predictor variables matrix, including age, gender, mean speed, business unit, and driver types,
- γ is the associated $2 * 1$ parameter vector,
- m : is the total miles driven as an offset term,
- α : a fixed overdispersion parameter.

Aim 1 - potential problems and alternative plans

The sheer size of the original ping data may be a problem in aim 1: the ping data has 1,494,678,173 rows and 9 columns, (>140 gigabytes (GB) . csv).

Although I will use the OSC server that has Random-Access Memory (RAM) of more than 500 GB, it may still be hard to read and process this giant file.

1. If the OSC server cannot handle the data correctly, I will separate the single giant csv file into **several small csv files** according to driver ID, then aggregate the pings to trips for each small csv file.
2. After the ping data are aggregated to trips, it is unlikely that the log-linear Gamma-Poisson model fail. In that unlikely event, I can turn to **negative binomial models** or use **traditional MLE estimates** instead of Bayesian estimation.

Aim 2 Overview

*The purpose of aim 2 is to develop three **scalable hierarchical Bayesian statistical and reliability models** for the SCEs of truck drivers and identify potential risk factors.*

- **Data:** 30-minute intervals for 496 drivers,
- **Outcome:**
 - whether SCEs occurred or not (binary variable),
 - the number of SCEs (count variable),
 - the time to each SCE (in minutes),
- **Predictors:**
 - driver-level random-effects,
 - age,
 - cumulative driving time,
 - weather,
 - road geometry,
 - mean speed,
 - speed variation,

Aim 2a) Bayesian hierarchical logistic regression

Two-level model: 1) 30-minute interval level i , 2) driver level $d(i)$.

$$Y_i \sim \text{Bernoulli}(p_i)$$

$$\log \frac{p_i}{1 - p_i} = \beta_{0,d(i)} + \beta_{1,d(i)} \cdot CT_i + \sum_{j=1}^J x_{ij} \beta_j \quad (2)$$

$$\beta_{0,d} \sim \text{i.i.d. } N(\mu_0, \sigma_0^2), \quad d = 1, 2, \dots, D$$

$$\beta_{1,d} \sim \text{i.i.d. } N(\mu_1, \sigma_1^2), \quad d = 1, 2, \dots, D$$

- Y_i : whether SCEs occurred in the 30-minute interval or not (binary),
- $\beta_{0,d(i)}$: random intercepts for each driver, $\beta_{1,d(i)}$ is random slopes for cumulative driving time CT_i ,
- $\beta_2, \beta_3, \dots, \beta_J$: fixed parameters for covariates x_{ij} .
- μ_0, σ_0 : hyper-parameters for random intercepts $\beta_{0,d}$,
- μ_1, σ_1 : hyper-parameters for random slopes $\beta_{1,d}$.

Aim 2a) Priors

Since we do not have much prior knowledge on the parameters, I will assign weakly informative priors²⁶ for these parameters:

$$\begin{aligned}\mu_0 &\sim N(0, 5^2) \\ \mu_1 &\sim N(0, 5^2) \\ \sigma_0 &\sim \text{Gamma}(1, 1) \\ \sigma_1 &\sim \text{Gamma}(1, 1) \\ \beta_2, \beta_3, \dots, \beta_J &\sim N(0, 10^2)\end{aligned}\tag{3}$$

The priors for the hyperpriors need to be relatively more restrictive than priors for fixed-effects parameters $\beta_2, \beta_3, \dots, \beta_J^{20}$.

Aim 2b) Model 2: Bayesian hierarchical Poisson regression

Two-level model: 1) 30-minute interval level i , 2) driver level $d(i)$.

$$N_i \sim \text{Poisson}(T_i \cdot \lambda_i)$$

$$\log \lambda_i = \beta_{0,d(i)} + \beta_{1,d(i)} \cdot \text{CT}_i + \sum_{j=1}^J x_{ij} \beta_j \quad (4)$$

$$\beta_{0,d} \sim \text{i.i.d. } N(\mu_0, \sigma_0^2), \quad d = 1, 2, \dots, D$$

$$\beta_{1,d} \sim \text{i.i.d. } N(\mu_1, \sigma_1^2), \quad d = 1, 2, \dots, D$$

- Y_i : the number of SCEs occurred in the 30-minute interval,
- T_i : length of the 30-minute interval,
- The other components are the same as those in hierarchical logistic regression.

The scalable Bayesian statistical and reliability models will be conducted using the **HMC-ECS algorithm** (self-defined functions in Python 3.6.0) or **HMC** (the `rstan` package in statistical computing environment R 3.6.0)^{23,27,28}.

Aim 2c) motivation for recurrent event models

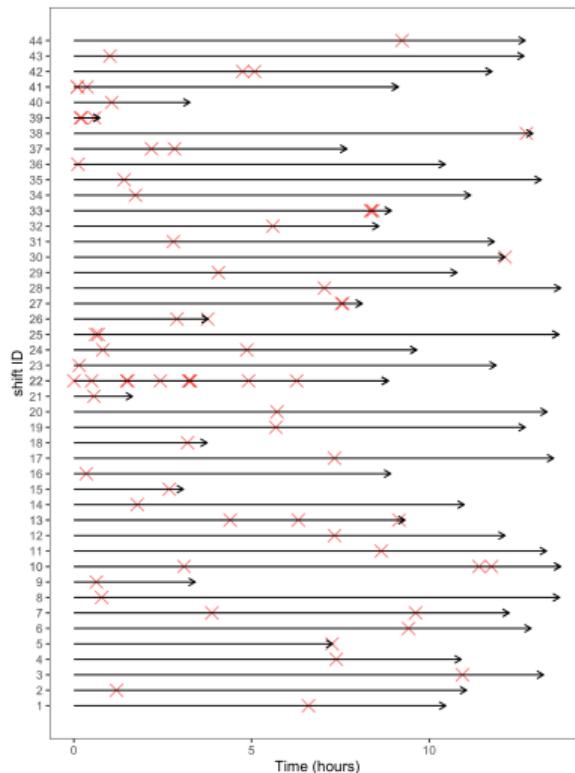


Figure 2: An arrow plot of time to SCEs in each shift

Aim 2c) Theories on NHPP and PLP

Nonhomogeneous Poisson Process (NHPP):

a Poisson process whose intensity function is non-constant.

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

$$\lambda(t) = \frac{\beta}{\theta} \left(\frac{t}{\theta} \right)^{\beta-1}, \quad \beta > 0, \theta > 0 \quad (5)$$

- $\beta > 1$: intensity increasing → reliability deteriorating,
- $\beta = 1$: constant intensity → reliability not changing,
- $\beta < 1$: intensity decreasing → reliability improving,
- θ : scale parameter.

There are two forms of truncation in a NHPP:

1. *Failure truncation:* when testing stops after a predetermined number of failures,
2. *Time truncation:* when testing stops at a predetermined time t .

Aim 2c) Intensity function of NHPP

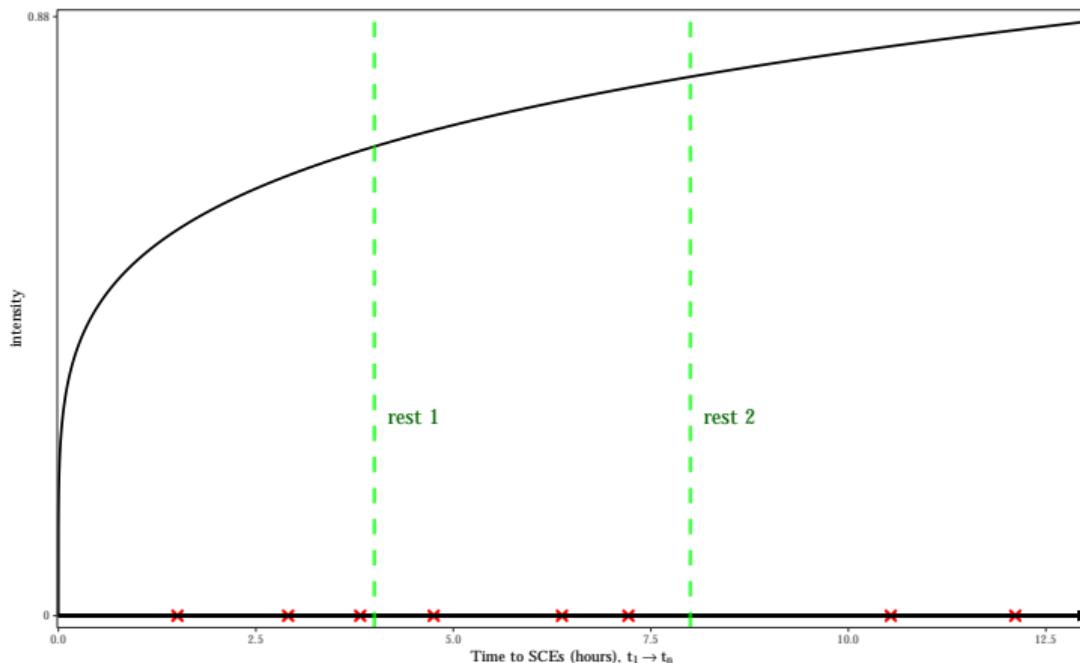


Figure 3: Intensity function, time to SCEs, and rest time within a shift generated from a NHPP with a PLP intensity function, $\beta = 1.2, \theta = 2$

Aim 2c) Notations

Let $T_{d,s,i}$ denotes the time to the d -th driver's s -th shift's i -th critical event. The total number critical events of d -th driver's s -th shift is $n_{d,s}$. The ranges of these notations are:

- $i = 1, 2, \dots, n_{d,S_d}$: SCE ID,
- $s = 1, 2, \dots, S_d$: shift ID,
- $d = 1, 2, \dots, D$: driver ID.

Aim 2c) Bayesian hierarchical NHPP with PLP intensity function

Assume the time to SCEs 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 .

$$\begin{aligned} T_{d,s,1}, T_{d,s,2}, \dots, T_{d,s,n_{d,s}} &\sim \text{PLP}(\beta, \theta_{d,s}) \\ \beta &\sim \text{Gamma}(1, 1) \\ \log \theta_{d,s} &= \gamma_{0d} + \gamma_1 x_{d,s,1} + \gamma_2 x_{d,s,2} + \dots + \gamma_k x_{d,s,k} \\ \gamma_{01}, \gamma_{02}, \dots, \gamma_{0D} &\sim \text{i.i.d. } N(\mu_0, \sigma_0^2) \\ \gamma_1, \gamma_2, \dots, \gamma_k &\sim \text{i.i.d. } N(0, 10^2) \\ \mu_0 &\sim N(0, 10^2) \\ \sigma_0 &\sim \text{Gamma}(1, 1) \end{aligned} \tag{6}$$

The shape parameter β shows the reliability changes of drivers:

- $\beta > 1$: intensity increasing → reliability deteriorating,
- $\beta = 1$: constant intensity → reliability not changing,
- $\beta < 1$: intensity decreasing → reliability improving,

Aim 2c) Potential problems and alternative plans

The sheer size of the 30-minute interval table and merged shifts table may be a problem in this aim.

- The 30-minute interval table: ~one million rows and 10 variables,
- Merged shift table: ~200,000 rows and 10 variables.
- 496 random intercepts and slopes

Although I propose to use the HMC-ECS to estimate the random effect, there are still chances that the model does not work. In that case, I will **sample 50 to 200 typical drivers**, then conduct the analysis based on this smaller sample data. In the unlikely event that the models still fails based on this smaller data, I can restrict the hierarchical models to **random intercepts only model** or use **traditional MLE** instead of Bayesian estimation.

Aim 3

Aim 3 seeks to innovate the NHPP with a PLP intensity function proposed in Aim 2 by adding one more parameter κ .

I propose to account for the rest time within a shift by adding one **more parameter κ , the percent of reliability recovery** for each a break within a shift.

This new reliability model (**jump-point PLP (JPLP)**) will be between a *NHPP* where the intensity function is not influenced by between-trip rests ("as bad as old"), and a *renewal process* where the intensity function is fully recovered by between-trip rests ("as good as new").

Aim 3 - intensity function of NHPP

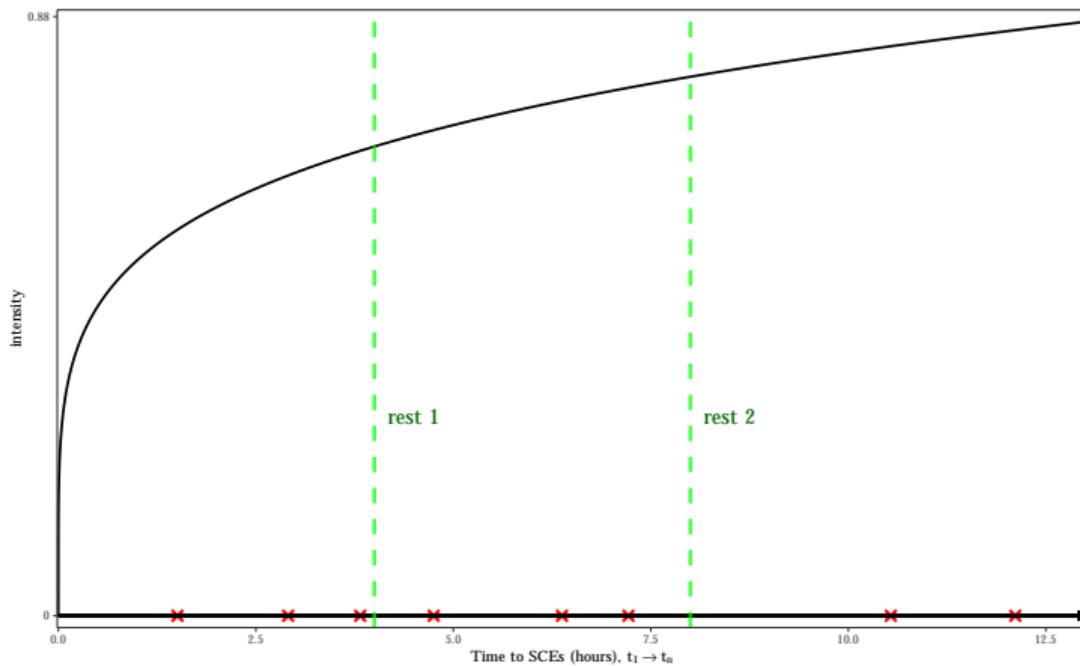


Figure 4: Intensity function, time to SCEs, and rest time within a shift generated from a NHPP with a PLP intensity function, $\beta = 1.2, \theta = 2$

Aim 3 - intensity function of proposed jump-point PLP (JPLP)

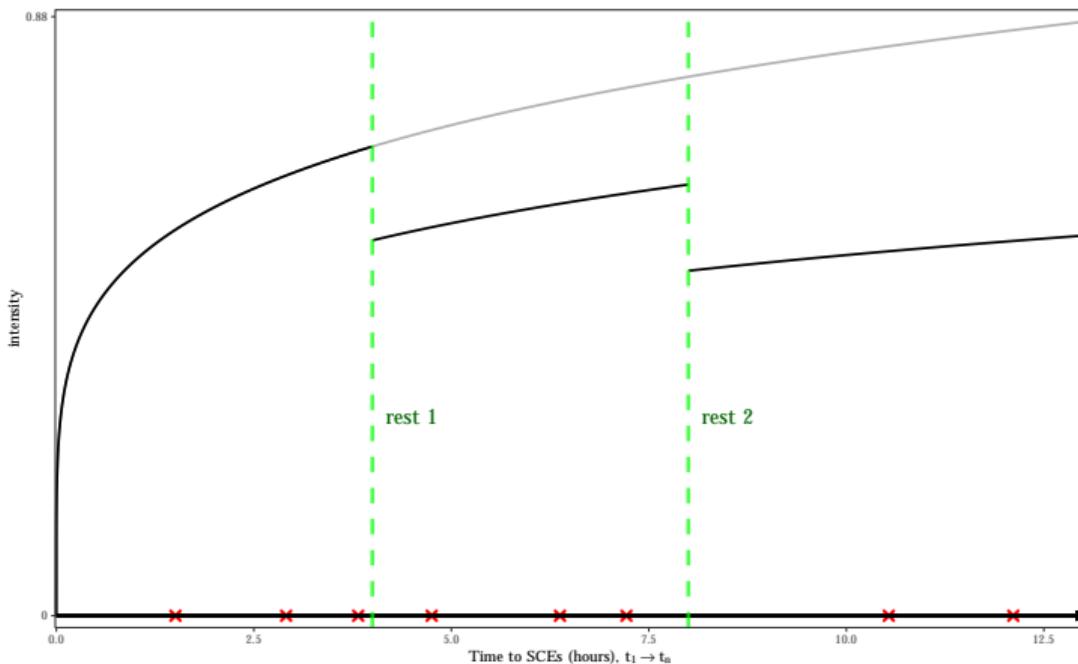


Figure 5: Intensity function, time to SCEs, and rest time within a shift with a jump-point PLP intensity function, $\beta = 1.2$, $\theta = 2$, $\kappa = 0.8$

Aim 3 - JPLP

JPLP: **an added parameter** κ based on Bayesian hierarchical PLP.

$$T_{d,s,1}, T_{d,s,2}, \dots, T_{d,s,n_{d,s}} \sim \text{JPLP}(\beta, \theta_{d,s}, \kappa)$$

$$\beta \sim \text{Gamma}(1, 1)$$

$$\log \theta_{d,s} = \gamma_{0d} + \gamma_1 x_{d,s,1} + \gamma_2 x_{d,s,2} + \dots + \gamma_k x_{d,s,k}$$

$$\kappa \sim \text{Uniform}(0, 1)$$

$$\gamma_{01}, \gamma_{02}, \dots, \gamma_{0D} \sim \text{i.i.d. } N(\mu_0, \sigma_0^2)$$

$$\gamma_1, \gamma_2, \dots, \gamma_k \sim \text{i.i.d. } N(0, 10^2)$$

$$\mu_0 \sim N(0, 5^2)$$

$$\sigma_0 \sim \text{Gamma}(1, 1)$$

(7)

- β : shape parameter that reflects the reliability changes of drivers,
- $\theta_{d,s}$: a scale parameter,
- κ : **the percent of intensity function recovery** once the driver takes a break.

Aim 3 - potential problems and alternative plans

In the unlikely event that the JPLP fails to be models, I will use the **modulated PLP** proposed by Black and Rigdon (1996)²⁹. The modulated PLP has well-defined data generating process, intensity function, and likelihood functions. If the JPLP does not work, I will revise the modulated PLP into **a hierarchical modulated PLP**.

The hierarchical JPLP and hierarchical modulated PLP will be estimated using Stan programs by adding self-defined likelihood function, which can be accessed via the `rstan` package in statistical computing environment R 3.6.0 on the OSC^{27,28,30}.

Conclusion and implications

- This work will illustrate the relationship between crashes and SCEs,
- The work will provide estimates of cumulative driving time on the risk of SCEs,
- The proposed JPLP will be an innovative reliability model for NDS datasets,
- The fusion of high-resolutional API and NDS data sets is an exciting opportunity,
- The subsampling MCMC can be a useful solution for wide and long NDS datasets.

References I

- 1 WHO. The top 10 causes of death. 2018.
- 2 Dalal K, Lin Z, Gifford M, Svanström L. Economics of global burden of road traffic injuries and their relationship with health system variables. *International journal of preventive medicine* 2013; **4**: 1442.
- 3 Olson R, Wipfli B, Thompson SV *et al*. Weight control intervention for truck drivers: The shift randomized controlled trial, united states. *American journal of public health* 2016; **106**: 1698–706.
- 4 Anderson JR, Ogden JD, Cunningham WA, Schubert-Kabban C. An exploratory study of hours of service and its safety impact on motorists. *Transport Policy* 2017; **53**: 161–74.
- 5 Hickman JS, Hanowski RJ, Bocanegra J. A synthetic approach to compare the large truck crash causation study and naturalistic driving data. *Accident Analysis & Prevention* 2018; **112**: 11–4.
- 6 Stern HS, Blower D, Cohen ML *et al*. Data and methods for studying commercial motor vehicle driver fatigue, highway safety and long-term driver health. *Accident Analysis & Prevention* 2019; **126**: 37–42.

References II

- 7 Theofilatos A, Yannis G, Koppelias P, Papadimitriou F. Impact of real-time traffic characteristics on crash occurrence: Preliminary results of the case of rare events. *Accident Analysis & Prevention* 2018.
- 8 Girotto E, Andrade SM de, González AD, Mesas AE. Professional experience and traffic accidents/near-miss accidents among truck drivers. *Accident Analysis & Prevention* 2016; **95**: 299–304.
- 9 Ye F, Lord D. 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* 2011; **2241**: 51–8.
- 10 Savolainen PT, Mannering FL, Lord D, Qudus MA. The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. *Accident Analysis & Prevention* 2011; **43**: 1666–76.
- 11 Guo F. Statistical methods for naturalistic driving studies. *Annual review of statistics and its application* 2019; **6**: 309–28.

References III

- 12 Saleh JH, Saltmarsh EA, Favaro FM, Brevault L. Accident precursors, near misses, and warning signs: Critical review and formal definitions within the framework of discrete event systems. *Reliability Engineering & System Safety* 2013; **114**: 148–54.
- 13 Dingus TA, Klauer SG, Neale VL *et al.* The 100-car naturalistic driving study. Phase 2: Results of the 100-car field experiment. United States. Department of Transportation. National Highway Traffic Safety ..., 2006.
- 14 Wu K-F, Aguero-Valverde J, Jovanis PP. Using naturalistic driving data to explore the association between traffic safety-related events and crash risk at driver level. *Accident Analysis & Prevention* 2014; **72**: 210–8.
- 15 Gitelman V, Bekhor S, Doveh E, Pesahov F, Carmel R, Morik S. Exploring relationships between driving events identified by in-vehicle data recorders, infrastructure characteristics and road crashes. *Transportation research part C: emerging technologies* 2018; **91**: 156–75.
- 16 Knipling RR. Naturalistic driving events: No harm, no foul, no validity. 2015.
- 17 Knipling RR. Threats to scientific validity in truck driver hours-of-service studies. 2017.
- 18 American Automobile Association Foundation for Traffic Safety. Asleep at the Wheel: The Prevalence and Impact of Drowsy Driving. 2010.

References IV

- 19 Cavuoto L, Megahed F. Understanding fatigue: Implications for worker safety. *Professional Safety* 2017; **62**: 16–9.
- 20 Gelman A, Stern HS, Carlin JB, Dunson DB, Vehtari A, Rubin DB. Bayesian data analysis. Chapman; Hall/CRC, 2013.
- 21 Kruschke JK, Vanpaemel W. Bayesian estimation in hierarchical models. *The Oxford handbook of computational and mathematical psychology* 2015;: 279–99.
- 22 Betancourt M. A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:170102434* 2017.
- 23 Dang K-D, Quiroz M, Kohn R, Tran M-N, Villani M. Hamiltonian monte carlo with energy conserving subsampling. *Journal of Machine Learning Research* 2019; **20**: 1–31.
- 24 Crum MR, Morrow PC. The influence of carrier scheduling practices on truck driver fatigue. *Transportation Journal* 2002;: 20–41.
- 25 Huang H, Abdel-Aty M. Multilevel data and bayesian analysis in traffic safety. *Accident Analysis & Prevention* 2010; **42**: 1556–65.
- 26 Gelman A, Simpson D, Betancourt M. The prior can often only be understood in the context of the likelihood. *Entropy* 2017; **19**: 555.

References V

- 27 Stan Development Team. RStan: The R interface to Stan. 2018.
<http://mc-stan.org/>.
- 28 R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2018 <https://www.R-project.org/>.
- 29 Black SE, Rigdon SE. Statistical inference for a modulated power law process. *Journal of Quality Technology* 1996; **28**: 81–90.
- 30 Center OS. Ohio supercomputer center. 1987.
<http://osc.edu/ark:/19495/f5s1ph73>.