Modeling safety-critical events using trucking naturalistic driving data: A generalizable framework for data aggregation, fusion, and statistical modeling

Miao Cai[a], Mohammad Ali Alamdar Yazdi[b], Amir Mehdizadeh[c], Qiong Hu[c], Alexander Vinel[c], Karen Davis[d], Fadel Megahed[e], Hong Xian[a], Steven E. Rigdon[a,*]

[a] *Department of Epidemiology and Biostatistics, Saint Louis University, Saint Louis, MO, 63108, United States*
[b] *Carey Business School, Johns Hopkins University, Baltimore, MD, 21218, United States*
[c] *Department of Industrial and Systems Engineering, Auburn University, Auburn, AL, 36849, United States*
[d] *Department of Computer Science and Software Engineering, Miami University, Oxford, OH, 45056, United States*
[e] *Department of Information Systems and Analytics, Miami University, Oxford, OH, 45056, United States*

Abstract

This is an awesome abstract. This is an awesome abstract.

Keywords: Trucking, Naturalistic driving studies, Safety-critical events

1. Introduction

The World Health Organization (WHO, 2018) estimated that road injury claimed around 1.4 million lives globally in 2016, which was the eighth leading cause of death. Among all types of vehicles on road, large trucks are a concern since they are more frequently involved in catastrophic crashes. In the United States, National Highway Traffic Safety Administration (2017) reported that 4.3% of registered vehicles were large trucks or buses, but they account for 12.4% of fatalities associated with vehicles (Hickman et al., 2018). Truck drivers are often on the road for long routes under on-time demands, complex traffic and weather conditions, with little to no supervision and contact with fellow workers. Therefore, a number of studies have been published to predict and reduce crash risk associated with trucks (Cantor et al., 2010; Chen et al., 2015; Dong et al., 2017).

Traditional crash prediction studies collect retrospective reports of crashes in a given road section for a specified time period, match these crash cases with non-crash controls (typically 1 to 4 matching), and then build statistical models, such as logistic regression and neural networks, to study risk factors associated with higher risk of crashes (Blower et al., 2010; Meuleners et al., 2017; Sharwood et al., 2013). This case-control study design is efficient and less time-consuming in trucking safety field since crashes are very rare. However, case-control studies, by nature, are limited in study design. Firstly, it is impossible to estimate and compare the rate of crashes since the number of

*Corresponding Author
Email addresses: miao.cai@slu.edu (Miao Cai), yazdi@jhu.edu (Mohammad Ali Alamdar Yazdi), azm0127@auburn.edu (Amir Mehdizadeh), qzh0011@auburn.edu (Qiong Hu), alexander.vinel@auburn.edu (Alexander Vinel), davisk4@miamioh.edu (Karen Davis), fmegahed@miamioh.edu (Fadel Megahed), hong.xian@slu.edu (Hong Xian), steve.rigdon@slu.edu (Steven E. Rigdon)

non-crashes is unknown. Besides, retrospective reports are often subject to recall and report bias: the drivers may not accurately recall the exact conditions at the time of the event; they may intentionally conceal some critical facts to escape from legal punishment (Dingus et al., 2011; Stern et al., 2019).

Naturalistic driving studies (NDSs) have been emerging in the past decade thanks to the advancement of technology. An NDS continuously collects driving data (including latitude, longitude, and speed) under real-world conditions using on-board unobtrusive equipment (Guo, 2019). In contrast to retrospective reports, an NDS resembles a cohort study: a pre-determined set of drivers are prospectively followed for a certain amount of time. Therefore, NDS comparatively has several advantages. First, NDS collects both crashes and non-crashes, so it is more useful in comparing the rates of events. Second, since vehicle crashes are extremely rare, it may take a huge amount of driving time to have sufficient sample of crashes. Instead, NDS focus safety-critical events (SCEs), which is defined as events that avoid crashes by last-second evasive maneuver (Dingus et al., 2011). SCEs can be 1000 times as high as real crashes and are argued to be good surrogates of crashes (Dingus et al., 2011; Guo et al., 2010). Third, NDS data are collected using programmed instruments or sensors, therefore they are less likely to be subject to human error or manipulation. Lastly, NDS collects data every a few seconds to minutes, and this large-scale high-resolution data provide a promising opportunity to quantifying driving risk (Guo, 2019).

However, many issues arise given the characteristics of NDSs. First, the sheer volume of NDS data creates a challenge to data management and aggregation (Mannering and Bhat, 2014). For example, a NDS data set can have billions rows of real-time speeds and locations, and it is important to have scalable and high-performance tools to aggregate these data into units that fit into the framework of statistical modeling. Second, routinely collected NDS data only have vehicle driving data. Crucial environmental variables such as weather and traffic need to be accessed from other sources and merged back to the driving data. Third, even with these data sources, management, and aggregation issues solved, scalable statistical models that account for the characteristics of NDS are needed to analyze the aggregated data.

A brief review of previous NDS analytic studies.

With increasing vehicle and insurance companies collecting NDS data on a regular basis, a scalable and generalizable analyzing framework serves as a pattern for follow-up researchers to better understand NDS data and gain insights into transportation safety. In this paper, we proposed a framework for data collection, aggregation, fusing, and statistical modeling, which is demonstrated in a case study. Although the NDS data used in this study were from large commercial truck drivers, the framework is generalizable to other drivers since the data collected among different drivers are similar.

2. Data

The data were collected by a leading freight shipping trucking company (we will name it as Company A for confidentiality reasons) in the United States. From April 2015 to March 2016, Company A installed in-vehicle data acquisition systems (DAGs) to all their trucks, which collect real-time *ping* and *SCEs* data. For demonstration purposes, we selected 496 regional drivers who move freights in a region that can include surrounding states. Apart from these vehicle driving data, demographic variables including age, gender, and race were also provided by Company A. The names of the drivers were not provided to the research team to ensure confidentiality, while a unique identification number was provided for each driver to link the three data sources. The study protocol was reviewed and approved by the Institutional Review Board of Saint Louis University.

2.1. ping and SCEs data

Every a couple of seconds to minutes, the DAG collects the date and time (year, month, day, hour, minute, and second), latitude and longitude (specific to five decimal places), driver identity number, and speed at that second. In total, 13,187,289 pings were provided to the research team.

SCEs. Besides, the company also regularly collected real-time GPS location and time-stamped critical events data for all their trucks. There were 12,458 critical events occurred to these 498 truck drivers during the study period. Four types of critical events were recorded in this critical events data. The number of SCEs.

2.2. Weather

Apart from driver's characteristics and driving condition, weather also poses a threat on truck crashes and injuries (Naik et al., 2016; Uddin and Huynh, 2017; Zhu and Srinivasan, 2011). We obtained historic weather data from the DarkSky Application Programming Interface (API), which allows us to query real-time and hour-by-hour nationwide historic weather conditions according to latitude, longitude, date, and time (The Dark Sky Company, LLC, 2019). The variables included visibility, precipitation probability[1] and intensity, temperature, wind and others.

2.3. Other available sources

Traffic and road geometry can be collected from Google map API and OpenStreet API.

[1] Ideally, historic precipitation at a specific location and time should be yes or not. However, in reality, since the weather stations are distributed not densely enough to record the exact weather conditions in every latitude and longitude in the US, the DarkSky API uses their algorithms to infer the probability of precipitation in each location.

3. Data preparation

3.1. Data aggregation

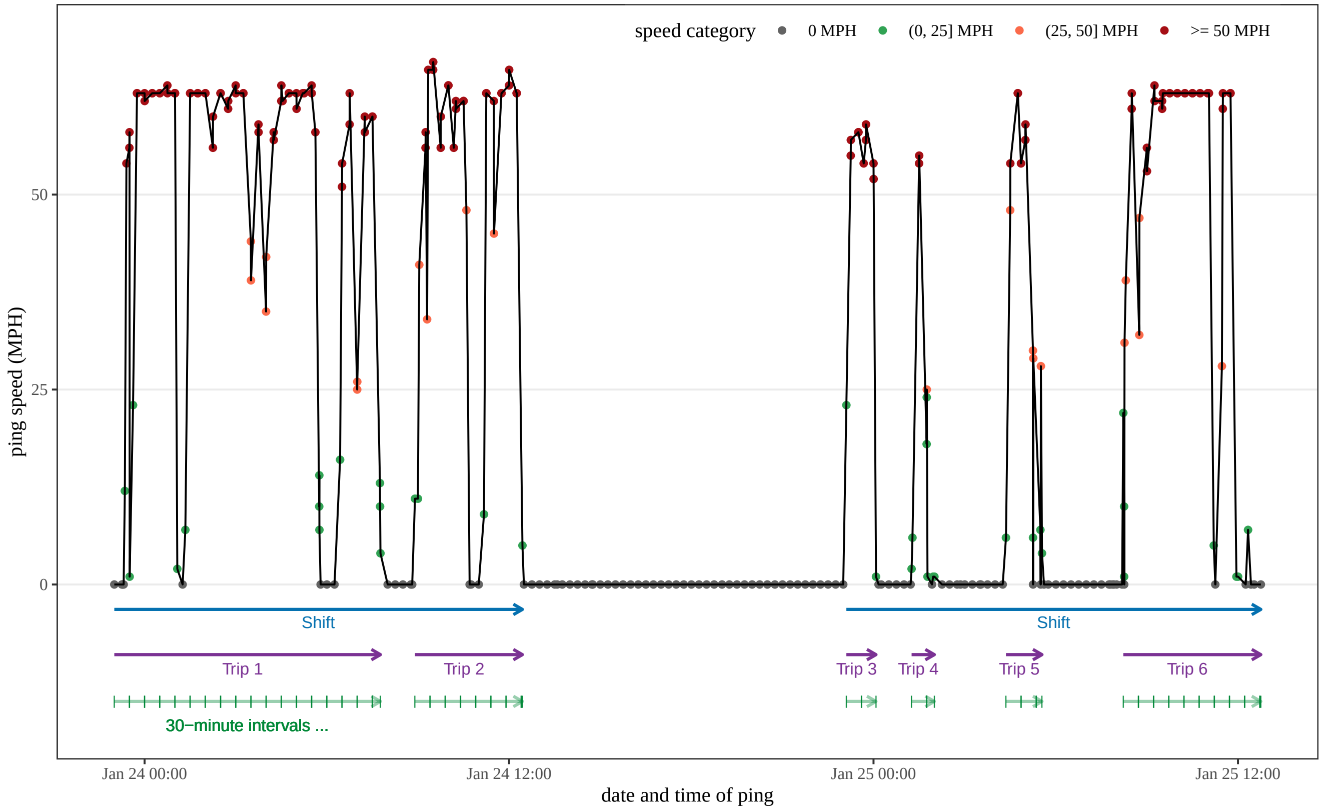

Figure 1: Data aggregation process from pings to shifts, trips, and 30-minute intervals.

To shrink the large size of over 10 million ping data, we rounded the GPS coordinates to the second decimal places, which are worth up to 1.1 kilometers, and we also round the time to the nearest hour. We then queried weather variables from the DarkSky API using the approximated latitudes, longitudes, date and hour. The weather variables used in this study include precipitation probability, precipitation intensity, and visibility.

For each of the truck drivers, if the ping data showed that the truck was not moving for more than 20 minutes, the ping data were separated into two different trips. These ping data were then aggregated into different trips. A **trip** is therefore defined as a continuous period of driving without stop. As Table demonstrates, each row is a trip. The average length of a trip in this study is 2.31 hours with the standard deviation of 1.8 hours.

After the ping data were aggregated into trips, these trips data were then further divided into different shifts according to an eight-hour rest time for each driver. A **shift** is defined as a long period of driving with potentially less than 8 hours' stops. The Shift_ID column in shows different shifts, separated by an eight-hour threshold. The average length of a shift in this study is 8.42 hours with the standard deviation of 2.45 hours.

3.2. Cumulative driving time as a measure of fatigue

Fatigue has been reported to be the most important predictor to truck crashes, considering that truck drivers are exposed to long routes and lone working environment Stern et al. (2019).

Driver's fatigue is difficult to measure in real life. In this study, we attempt to use three proxies to measure the fatigue of the truck drivers: cumulative driving time in a shift, the rest time before a shift, and the rest time before a trip.

4. Methodology

4.1. Statistical models

Traditional statistical models assume that observations are independent from each other given their predictor variables. However, natural data are almost never independent given the predictor variables. In the example of truck driver's safety events, if we assume the external traffic, weather and driver's socioeconomic status are fixed, truck drivers may exhibit similar driving patterns in multiple trips, and then drivers hired by the same company may share similar culture and safety atmospheres. Therefore, traffic accidents are naturally nested within drivers and drivers are nested within companies. Traditional statistical models that assume independence between observations are not appropriate in this case since objects tend to be similar within a group. Hierarchical models, also known as multilevel model, random-effects model or mixed model, have been developed to allow for the nested nature of data. Instead of assuming independence given predictor variables, hierarchical models assume conditional independence. Hierarchical models are advocated to be the default method since they can produce more precise prediction and more robust results than traditional models.

Random-effects models (Han et al., 2018; Pantangi et al., 2019).

Here we model the probability of a critical event occurred using a Bayesian hierarchical Bernoulli regression. We categorized the number of safety events during a trip into a binary variable Y with the value of either 0 or 1, where 0 indicated that no critical event occurred during that trip while 1 indicated that at least 1 critical event occurred during the trip. Since each trip i has a different travel time t_i, we derived the Bernoulli distribution parameter p_i using the probability density function of the Poisson distribution, with the parameter λ_i equaled a linear combination of β_i and x_i.

$$
\begin{aligned}
P_i &= P(\text{at least one event in trip i}) \\
&= 1 - P(\text{no event in trip i}) \\
&= 1 - \frac{e^{-t_i\lambda_i}(t_i\lambda_i)^0)}{0!} \\
&= 1 - \exp(-t_i\lambda_i) \\
&= 1 - \exp(-t_i e^{\beta_0+\beta_i x_i})
\end{aligned}
\tag{1}
$$

Transform that into a linear function of β_i, x_i and t_i

$$\begin{aligned} 1 - P_i &= \text{EXP}(-t_i e^{\beta_0 + \beta_i x_i}) \\ \log(1 - P_i) &= -t_i e^{\beta_0 + \beta_i x_i} \\ \log \frac{1}{1 - P_i} &= e^{\beta_0 + \beta_i x_i + \log(t_i)} \\ \log\left(\log \frac{1}{1 - P_i}\right) &= \beta_0 + \beta_i x_i + \log(t_i) \end{aligned} \tag{2}$$

Then, the random effects logistic model is

$$\begin{aligned} Y_i &\sim \text{Bern}(P_i) \\ \log\left(\log \frac{1}{1 - P_i}\right) &= \beta_{0,d(i)} + \beta_{1,d(i)} \cdot \text{CT}_i + \xi \cdot \mathbf{W} + \nu \cdot \mathbf{D_i} + \log(t_i) \end{aligned} \tag{3}$$

Here the trip is indexed by i, Y_i is the binary outcome variable of whether at least one critical event occurred in trip i; $d(i)$ is the driver for trip i, $\beta_{0,d(i)}$ is the random intercept for driver $d(i)$; $\beta_{1,d(i)}$ is the random slope for the cumulative time (CTi) of driving in the shift (the sum of driving time for all previous trips) for driver $d(i)$; $\mathbf{W}$ is a vector of external environment fixed effects, including precipitation intensity and probability, visibility, and whether it was sunrise or sunset time; $\mathbf{D}_i$ are driver level fixed effects, including age group and business unit; t_i is the travel time for the trip i.

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.

$$\begin{aligned} \mu_0 &\sim N(0, 10^2) \\ \mu_1 &\sim N(0, 10^2) \\ \sigma_0 &\sim \text{GAMMA}(1, 1) \\ \sigma_1 &\sim \text{GAMMA}(1, 1) \end{aligned} \tag{4}$$

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.

5. Results

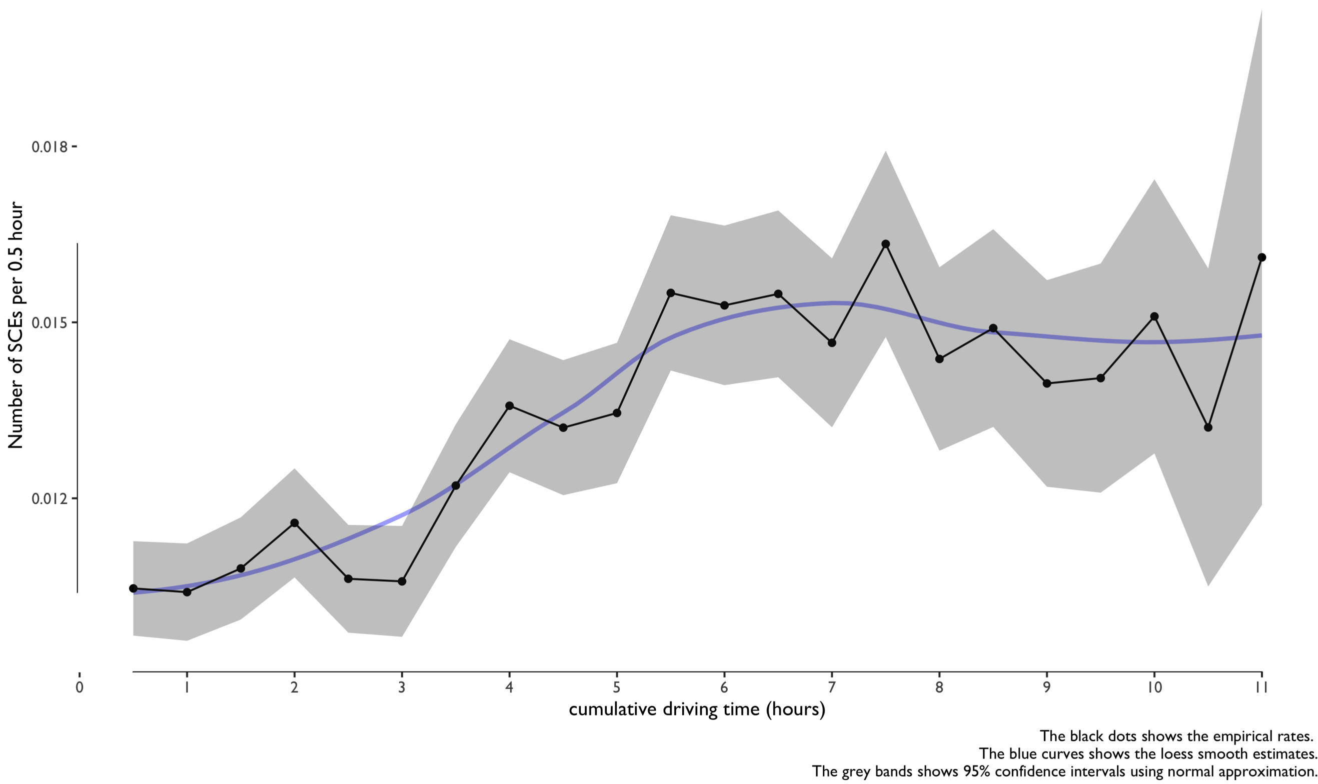

Figure 2: The rate of safety critical events and cumulative driving time

6. Discussion

7. Conclusions

Subsampling MCMC (Dang et al., 2019; Quiroz et al., 2019, 2018, 2016).

Acknowledgement

This work was supported in part by the National Science Foundation (CMMI-1635927 and CMMI-1634992), the Ohio Supercomputer Center (PMIU0138 and PMIU0162), the American Society of Safety Professionals (ASSP) Foundation, the University of Cincinnati Education and Research Center Pilot Research Project Training Program, and the Transportation Informatics Tier I University Transportation Center (TransInfo). We also thank the DarkSky company for providing us five million free calls to their historic weather API.

References

Blower, D., Green, P.E., Matteson, A., 2010. Condition of trucks and truck crash involvement: Evidence from the large truck crash causation study. Transportation Research Record 2194, 21–28.

Cantor, D.E., Corsi, T.M., Grimm, C.M., Özpolat, K., 2010. A driver focused truck crash prediction model. Transportation Research Part E: Logistics and Transportation Review 46, 683–692.

Chen, C., Zhang, G., Tian, Z., Bogus, S.M., Yang, Y., 2015. Hierarchical bayesian random intercept model-based cross-level interaction decomposition for truck driver injury severity investigations. Accident Analysis & Prevention 85, 186–198.

Dang, K.-D., Quiroz, M., Kohn, R., Tran, M.-N., Villani, M., 2019. Hamiltonian Monte Carlo with energy conserving subsampling. Journal of Machine Learning Research 20, 1–31.

Dingus, T.A., Hanowski, R.J., Klauer, S.G., 2011. Estimating crash risk. Ergonomics in Design 19, 8–12.

Dong, C., Dong, Q., Huang, B., Hu, W., Nambisan, S.S., 2017. Estimating factors contributing to frequency and severity of large truck–involved crashes. Journal of Transportation Engineering, Part A: Systems 143, 04017032.

Guo, F., 2019. Statistical methods for naturalistic driving studies. Annual Review of Statistics and Its Application 6, 309–328.

Guo, F., Klauer, S.G., Hankey, J.M., Dingus, T.A., 2010. Near crashes as crash surrogate for naturalistic driving studies. Transportation Research Record 2147, 66–74.

Han, C., Huang, H., Lee, J., Wang, J., 2018. Investigating varying effect of road-level factors on crash frequency across regions: A bayesian hierarchical random parameter modeling approach. Analytic methods in accident research 20, 81–91.

Hickman, J.S., Hanowski, R.J., Bocanegra, J., 2018. A synthetic approach to compare the large truck crash causation study and naturalistic driving data. Accident Analysis & Prevention 112, 11–14.

Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: Methodological frontier and future directions. Analytic methods in accident research 1, 1–22.

Meuleners, L., Fraser, M.L., Govorko, M.H., Stevenson, M.R., 2017. Determinants of the occupational environment and heavy vehicle crashes in western australia: A case–control study. Accident Analysis & Prevention 99, 452–458.

Naik, B., Tung, L.-W., Zhao, S., Khattak, A.J., 2016. Weather impacts on single-vehicle truck crash injury severity. Journal of Safety Research 58, 57–65.

National Highway Traffic Safety Administration, 2017. A Compilation of Motor Vehicle Crash Data from the Fatality Analysis Reporting System and the General Estimates System.

Pantangi, S.S., Fountas, G., Sarwar, M.T., Anastasopoulos, P.C., Blatt, A., Majka, K., Pierowicz, J., Mohan, S.B., 2019. A preliminary investigation of the effectiveness of high visibility enforcement programs using naturalistic driving study data: A grouped random parameters approach. Analytic Methods in Accident Research 21, 1–12.

Quiroz, M., Kohn, R., Villani, M., Tran, M.-N., 2019. Speeding up MCMC by efficient data subsampling. Journal of the American Statistical Association 114, 831–843.

Quiroz, M., Tran, M.-N., Villani, M., Kohn, R., 2018. Speeding up MCMC by delayed acceptance and data subsampling. Journal of Computational and Graphical Statistics 27, 12–22.

Quiroz, M., Tran, M.-N., Villani, M., Kohn, R., Dang, K.-D., 2016. The block-Poisson estimator for optimally tuned exact subsampling MCMC. arXiv preprint arXiv:1603.08232.

Sharwood, L.N., Elkington, J., Meuleners, L., Ivers, R., Boufous, S., Stevenson, M., 2013. Use of caffeinated substances and risk of crashes in long distance drivers of commercial vehicles: Case-control study. BMJ 346, f1140.

Stern, H.S., Blower, D., Cohen, M.L., Czeisler, C.A., Dinges, D.F., Greenhouse, J.B., Guo, F., Hanowski, R.J., Hartenbaum, N.P., Krueger, G.P., others, 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 Company, LLC, 2019. Dark Sky API — Overview.

Uddin, M., Huynh, N., 2017. Truck-involved crashes injury severity analysis for different lighting conditions on rural and urban roadways. Accident Analysis & Prevention 108, 44–55.

WHO, 2018. The top 10 causes of death.

Zhu, X., Srinivasan, S., 2011. A comprehensive analysis of factors influencing the injury severity of large-truck crashes. Accident Analysis & Prevention 43, 49–57.