

# Road Traffic Crash Data: An Overview on Sources, Problems, and Collection Methods

Azad Abdulhafedh

University of Missouri-Columbia, MO, USA

Email: asa8cd@mail.missouri.edu

**How to cite this paper:** Abdulhafedh, A. (2017) Road Traffic Crash Data: An Overview on Sources, Problems, and Collection Methods. *Journal of Transportation Technologies*, 7, 206-219.

<https://doi.org/10.4236/jtts.2017.72015>

**Received:** December 19, 2016

**Accepted:** April 27, 2017

**Published:** April 30, 2017

Copyright © 2017 by author and Scientific Research Publishing Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

---

## Abstract

Road traffic crash data are useful tools to support the development, implementation, and assessment of highway safety programs that tend to reduce road traffic crashes. Collecting road traffic crash data aims at gaining a better understanding of road traffic operational problems, locating hazardous road sections, identifying risk factors, developing accurate diagnosis and remedial measures, and evaluating the effectiveness of road safety programs. Furthermore, they can be used by many agencies and businesses such as: law enforcements to identify persons at fault in road traffic crashes; insurers seeking facts about traffic crash claims; road safety researchers to access traffic crash reliable database; decision makers to develop long-term, statewide strategic plans for traffic and highway safety; and highway safety administrators to help educate the public. Given the practical importance of vehicle crash data, this paper presents an overview of the sources, trends and problems associated with road traffic crash data.

## Keywords

Road Safety, Vehicle Crash Data, Over-Dispersion, Under-Dispersion, Under-Reporting, FARS, NASS, HSIS

---

## 1. Introduction

Throughout the world, cars, buses, trucks, motorcycles, pedestrians, animals, taxis and other categories of travelers, share the roadways, contributing to economic and social development in many countries. Yet each year, many vehicles are involved in crashes that are responsible for millions of deaths and injuries. Globally, every year, about 1.25 million people are killed in motor vehicle crash-

---

\*PhD in Civil Engineering.

es and approximately 50 million more are injured. Vehicular crashes are the world's leading cause of death for individuals between the ages of one and twenty-nine [1]. Following current trends, about two million people could be expected to be killed in motor vehicle crashes each year by 2030 [1]. Currently, road crashes are ranked as the ninth most serious cause of death in the world, and without new initiatives to improve road safety, fatal crashes will likely rise to the third place by the year 2020 [1]. In developed countries, road traffic death rates have decreased since the 1960s because of successful interventions such as seat belt safety laws, enforcement of speed limits, warnings about the dangers of mixing alcohol consumption with driving, and safer design and use of roads and vehicles. For example, road traffic fatalities have declined by about 25.0 percent in the United States from 2005 to 2014 and the number of people injured has decreased 13.0 percent from 2005 to 2014 [2]. In Canada, the number of road traffic fatalities has declined by about 62.0 percent from 1990 to 2014, and the number of injuries has declined by about 68.0 percent during the same period [3]. However, traffic fatalities have increased in developing countries from 1990 to 2014 (*i.e.* 44.0 percent in Malaysia and about 243.0 percent in China) [1]. Developing countries bear a large share of the burden, accounting for 85.0 percent of annual deaths and 90.0 percent of the disability-adjusted life years. More than one-half of all road traffic deaths globally involve people ages 15 to 44, during their most productive earning years. Moreover, the disability burden for this age group accounts for about 60.0 percent of all disability-adjusted life years. The costs and consequences of these losses are significant. Three-quarters of all poor families who lost a member in a traffic crash reported a decrease in their standard of living, and about 61.0 percent reported having to borrow money to cover expenses following their loss [4]. The World Bank estimates that road traffic injuries cost 2.0 percent to 3.0 percent of the Gross National Product of developing countries, or twice the total amount of development aid received worldwide by developing countries [5]. Crash-related fatalities and injuries can be prevented or at least minimized by a joint involvement from multiple sectors (*i.e.* transportation agencies, police, health departments, education institutions) that oversee road safety, vehicles, and the drivers themselves. Effective interventions include design of safer infrastructure and incorporation of road safety features into land-use and transport planning; improvement of vehicle safety features; improvement of post-crash care for victims of road crashes, and improvement of driver behavior, such as setting and enforcing laws relating to key risk factors, and raising public awareness [6]. In addition, vehicular crash data can assist with the development of generalized theories concerning road safety. A range of basic laws have been put forth to help explain the relationship between the occurrence of road crashes and potential risk factors, such as: the universal law of learning, which implies that the crash rate tends to decline as the number of kilometers travelled increases; the law of rare events, which states that rare events, such as environmental hazards, would have more effect on crash rates than regular events; and the law of complexity, which implies that the more complex the traf-

fic situation road users encounter, the higher the probability of crash occurrence [7]. Although transportation agencies often try to identify the most dangerous road sites, and put great efforts into preventive measures, such as illumination and policy enforcement, the annual number of traffic crashes has not yet significantly decreased. For instance, 35,092 traffic fatalities were recorded in the US during 2015, an increase of 7.2% as compared to the previous year [8]. The fatality rate per 100 million vehicle miles traveled increased 3.7% from 2014-2015. Thirty-five States had more motor vehicle fatalities in 2015 than in 2014. Given this trend, it is imperative to gain a better understanding of crash data sources, trends and problems.

## **2. The Importance of Collecting Vehicular Crash Data**

Vehicular crash data are used to respond to requests from the congress, federal agencies, state and local governments, universities and research organizations, highway safety communities, the media, and private citizens. Accurate data are required to support the development, implementation, and assessment of highway safety programs aimed at reducing crash tolls. An example of the practical importance of collecting and maintaining vehicular crash data is the recent emerging of the crash data retrieval tools, commonly referred to as the vehicle black boxes. Based upon a rule imposed by the National Highway Traffic Safety Administration (NHTSA), most vehicles manufactured and sold in North America after 2012 are equipped with Event Data Recorders (EDRs) that collect, store, and retrieve vehicle crash event data. The EDRs can help law enforcement investigating vehicle crashes to recover crucial crash data parameters from a vehicle that has been involved in a crash, including pre-crash data that will help better understand important factors that led to the crash occurrence [9]. Another practical example is the use of the Crash Outcome Data Evaluation System (CODES), which is a program managed by NHTSA, to link crash records to injury outcome records collected at the scene by emergency medical services. CODES data has been utilized to improve traffic safety issues in different ways, such as examining whether the increased crash rates for teen drivers have resulted in an increased injury to their passengers, and exploring the seat belt usage in preventing injuries and fatalities. CODES data has also been used to inform and educate traffic safety decision-makers at federal, state, and local levels in many circumstances, for instance, providing federal and state legislators with CODES reports on the importance of seat belt use in preventing injuries and fatalities; delivering data to the state highway administrations to develop long-term, statewide strategic plans for traffic and highway safety; and publishing CODES fact sheets that can help educate the public [10].

## **3. Road Traffic Data Collection Methods**

Most studies of traffic related problems begin with the collection of data. Generally, traffic data collection methods can be classified as one of two categories: intrusive and non-intrusive methods. Intrusive methods typically involve a data

recorder and a sensor placing on or in the road [11]. The most common intrusive devices are:

- Pneumatic road tubes: rubber tubes placed across the road lanes to detect vehicles from pressure changes that are produced when a vehicle tire passes over the tube. The pulse of air that is created is recorded and processed by a counter located on the side of the road. The main drawback of this technology is that it has limited lane coverage and its efficiency is subject to weather, temperature and traffic conditions.
- Piezoelectric sensors: sensors are placed in a groove along roadway surface of the lane(s) monitored. The principle is to convert mechanical energy into electrical energy. The amplitude and frequency of the signal is directly proportional to the degree of deformation.
- Magnetic loops: this is the most conventional technology used to collect traffic data. The loops are embedded in roadways in a square formation that generates a magnetic field. The information is then transmitted to a counting device placed on the side of the road. This has a generally short life expectancy because it can be damaged by heavy vehicles, but is not affected by bad weather conditions.

Non-intrusive techniques are based on remote observations ranging from human observation to those based on new technologies [12]:

- Manual counts: Trained observers gather traffic data such as vehicle occupancy rate, pedestrians and vehicle classifications that cannot be efficiently obtained through automated counts. Equipment needs are rather basic with the observers usually requiring only a tally sheet, mechanical and/or electronic counting devices.
- Passive and active infra-red sensors: the presence, speed and type of vehicles can be detected based on the infrared energy radiating from the detection area. The main drawbacks of this method are the sensor's performance during bad weather, and limited lane coverage.
- Passive magnetic sensors: magnetic sensors can be fixed under or on top of the roadbed. The sensors record the number of vehicles, their type and speed. However, in some operating conditions, the sensors have difficulty differentiating between closely spaced vehicles.
- Microwave radar sensors: these sensors can detect moving vehicles and record vehicle counts, speed and vehicle classification and are not usually compromised by weather conditions.
- Ultrasonic and passive acoustic sensors: these devices emit sound waves to detect vehicles by measuring the time for the signal to return to the device. The ultrasonic sensors can be placed directly over the lane or alongside the road to collect vehicle counts, speed and classification data. However, the collection ability of these sensors can be adversely affected by temperature or bad weather.
- Video image detection: video cameras can be used to record vehicle numbers, type and speed by means of different video techniques e.g. trip line and

tracking. Video detection systems can be sensitive to weather conditions.

The Floating Car Data (FCD) can be used to collect traffic data by locating the vehicle via mobile phones or GPS over the entire road network. Data such as car location, speed and direction of travel can then be sent anonymously to a central processing center. After being collected and extracted, useful information can be redistributed to the drivers on the road [13].

There are two important traffic measures that are widely used in modeling traffic data, namely: the average annual daily traffic (AADT); and the vehicle miles travelled (VMT). These two traffic variables, usually derived from fixed sensors measurements, play a key role in traffic crash analysis and policy decisions [14]. AADT is the average (calculated over a year) number of vehicles passing a point along a particular counting section each day. Thus, AADT represents the vehicle flow over a road section (e.g. highway segment) on an average day of the year. Methods for calculating AADT are generally based on data from two types of counts: permanent automatic traffic counts and short-period traffic counts. A combination of these two measurements is generally used to obtain an AADT estimate over a larger road network. In the US, the factoring method is a common methodology used to estimate AADT. This method has been adopted by many transportation agencies as a standard protocol corresponding with federal guidelines. The 2013 Traffic Monitoring Guide serves as a reference document that provides general guidance on the development of traffic monitoring programs for highway agencies. In particular, the TMG provides guidance on the collection of traffic volume, vehicle classification, and weight information [15]. VMT refers to the distance travelled by vehicles. It is often used as an indicator of traffic demand and for analyzing mobility patterns and travel trends. It plays a key role in various important decision-makings such as air quality compliance, roadway pavement maintenance, and crash analysis. There are four methods commonly used to calculate VMT [16]:

- Odometer readings (vehicle-based method) at regular vehicle inspections, the average distance travelled by the vehicles is determined and then multiplied by the number of road vehicles.
- Traffic counts (road-based method) for one considered link, the VMT is calculated by multiplying the AADT by the length of the link. VMT for a roadway can then be obtained by summing the VMT of each segment.
- Driver survey questionnaires sent to households with one or more cars soliciting information such as the number of miles driven by each vehicle during the whole year and unit consumption.
- Fuel consumption the volume of road traffic is estimated from information about fuel supply and fuel consumption as derived from estimates of miles driven per fuel gallon for typical types of vehicles.

#### **4. Sources of Vehicular Crash Data**

In the U.S., a variety of efforts to collect, maintain and/or distribute information on vehicular crash data have been utilized. Some of the crash data sources that

are publicly available are listed below:

#### **4.1. Fatality Analysis Reporting System (FARS)**

FARS is an online database of fatal motor vehicle crashes that documents all fatalities that occurred within the 50 States since 1975. FARS qualifying crashes had to involve a motor vehicle traveling on a public traffic way, and must have resulted in the death of a motorist or a non-motorist within 30 days of the crash. FARS is administered by the National Center for Statistics and Analysis (NCSA) within the National Highway Traffic Safety Administration (NHTSA). FARS data are collected from each State's government by trained state employees, who are responsible for gathering, and transmitting their state's data to NCSA in a standard format. After the data file is created, quality checks are performed on the data, and the electronic data are made available online to the public in Statistical Analysis System (SAS) data files as well as Database Files (DBF). The main SAS data files include: the Accident file, which contains information about crash characteristics and environmental conditions at the time of the crash; the Vehicle file, which contains information describing the in-transport motor vehicles and the drivers of in-transport motor vehicle who are involved in the crash; the Person file, which contains information describing all persons involved in the crash including motorists and non-motorists (e.g., pedestrians); the Damage file, which contains information about all areas on the vehicle that were damaged in the crash; the Drimpair file, which contains information about physical impairments of drivers of motor vehicles; the Factor file, which contains information about vehicle circumstances that may have contributed to the crash; the Violatn file, which contains information about violations that were charged to drivers; and the Vindecode file, which contains vehicle descriptors based on the vehicle's VIN. The temporal coverage of FARS data includes some variables such as, the time of the crash, the date, the month, and the year. The spatial coverage of FARS data includes the latitude and longitude coordinates of each crash location. The FARS data are generally complete, reliable, and publicly available online [17]. However, one of the FARS data weaknesses is that FARS data cannot be downloaded for multiple years at a time due to the system complexities, and when data is downloaded from FARS website, the user can obtain data by only one variable at a time. In addition, as mentioned above, the FARS data does not provide the injury-severity only crashes, and property-damage only crashes.

#### **4.2. The NASS-GES**

The National Automotive Sampling System (NASS)-General Estimates System (GES) obtains its data from a representative crash sample selected from more than five million police-reported crashes annually in the US. These crashes include those that result in a fatality or injury and those involving major property damage as well. The data are obtained by NASS-GES data collectors in 60 geographic sites across the United States. These data collectors make visits to approximately 400 police agencies within the 60 sites, where they randomly sample

about 50,000 crashes per year. NASS-GES data are made available to the public in Statistical Analysis System (SAS) data files as well as Database Files (DBF). The main SAS data files of NASS-GES include similar FARS files mentioned above. The temporal coverage of the NASS-GES data includes variables such as, time of the crash, the date, the month, and the year. The spatial coverage only includes the land use of the crash location without providing the latitude and longitude of the crash location or the x, y coordinates. One weakness in NASS-GES data is that it uses a weighted data element that produces the overall national estimates that may differ from the true state-level values because they are based on a probability sample of crashes among the country, and this cannot give the accurate state-level estimates, which decreases the reliability of the data. Another weakness is that the NASS-GES data are obtained either directly from the police accident report (PAR) or by interpreting the information provided in the PAR through reviewing the crash diagram, or combinations of data elements on the PAR. Because of this interpretation, an important portion of data can be missing in the system [18].

#### **4.3. The NASS-CDS**

The National Automotive Sampling System (NASS)-Crashworthiness Data System (CDS) obtains its data from 24 geographic sites in the US. These data are weighted to represent all police reported motor vehicle crashes occurring in the USA during the year including light vehicles, such as, passenger cars, SUVs, and vans. The NASS-CDS files are available in a Statistical Analysis System (SAS) dataset, and contain similar FARS files. The NASS-CDE system provides temporal coverage of data through variables such as, time of the crash, the date, the month, and the year. There is no spatial coverage within the NASS-CDS data, as it does not provide the latitude and longitude of the crash location nor the x, y coordinates. One weakness of the NASS-CDS data is that the data from these crashes are weighted to produce national estimates, and cannot give the state-level estimates, which decreases the reliability of data [19].

#### **4.4. The State Data System (SDS)**

The State Data System (SDS) is maintained by NHTSA's National Center for Statistics and Analysis (NCSA), and only thirty-two states are participating in the system, including the state of Missouri. While the (FARS) only has fatal crash data, SDS provides data on injury and property-damage-only crashes as well. In contrast to the data in (NASS-GES), the SDS consists of census data taken directly from police accident reports. The law enforcement agencies within a state are the primary source of information on crashes occurring within a state. All states have requirements for documenting fatal, injury or property damage crashes (with damage above a certain dollar threshold). Each participating state has its own reporting system, for instance, in the state of Missouri, the Missouri Statewide Traffic Accident Records System (STARS) is managed by the Missouri State Highway Patrol (MSHP), and all Missouri law enforcement agencies are



required by law to submit a Missouri Uniform Traffic Crash Report to STARS if a traffic crash occurred that involves a death, a personal injury, or a property damage. STARS involves many recording files, such as, the Crash and Personal Severity, which includes fatal, personal injury, and property damage; the Crash Circumstances file, which includes motorcycles crashes by year; Speed Involved Traffic Crash file; Alcohol Involved Traffic Crash file; Young Driver Involved Traffic Crash file; and Mature Driver Involved Traffic Crash file. All files are provided in excel and PDF format, complete, reliable, and available online for the public (MSHP 2016). The temporal coverage of the SDS data includes variables such as, time of the crash, the date, the month, and the year. The spatial coverage only includes the x, y coordinates of the crash locations in only some spots. One weakness of the SDS data is that it does not provide a comprehensive list of risk variables and details that exist in the FARS and NASS-GES systems [20].

#### 4.5. The Highway Safety Information System (HSIS)

The Highway Safety Information System (HSIS) is a highway data system funded by the U.S. Federal Highway Administration (FHWA), with data voluntarily provided to HSIS by the participating states, which are California, Washington, Minnesota, Illinois, Ohio, Maine, and North Carolina. HSIS began operation in 1987, and the participating states were selected based on their data availability, quantity, and quality of data. HSIS supports the FHWA safety research program, and can be accessed online by researchers, universities, and safety professionals. The HSIS files are available in a (SAS) format, and the main files include four basic files namely; the Accident file, the Vehicle file, the Occupant file, and the Roadway file. The temporal coverage of the HSIS data includes variables such as, time of the crash, date, month, and the year. The spatial coverage only includes the section length, and the milepost of the crash location without providing the latitude and longitude of the crash location nor the x, y coordinates. The HSIS data are generally complete with very few missing data, reliable, and publicly available. One weakness of the HSIS data is that it does not cover all states within the US, and also their main files should be merged in order to get the required information [21].

#### 4.6. Data.Gov

The Data.gov is a federal open US government online database that includes all states, and local government's metadata describing their open data resources. Data.gov began operation in 2009, and is managed and hosted by the U.S. General Services Administration, Office of Citizen Services and Innovative Technologies, and follows the Project Open Data schema that includes fields, such as title, description, tags, publisher, etc. for every data set displayed on the website. Different data topics are available, such as Agriculture, Health, Business, Climate, Energy, Finance, and Science. The transportation statistics series consists of analyzed statistical information on motor fuel, vehicle crashes, motor vehicle



registrations, driver licenses, highway user taxation, highway mileage, travel, and highway finance. The files are available in CSV format, and can be freely downloaded without registration [22].

#### **4.7. The U.S. Census Bureau**

The U.S. Census Bureau is part of the Department of Commerce, and is overseen by the Economics and Statistics Administration. The transportation section within the online database provides data on civil air transportation, water transportation, revenues, passenger and freight traffic volume, trains, highway mileage and finances, highway crash data, characteristics of public transit, and railroads. Data are available in excel format for public use [23].

#### **4.8. The SHRP2-NDS**

The Strategic Highway Research Program 2-Naturalistic Driving Study (SHRP2-NDS) is an online database related to the Transportation Research Board (TRB)'s second safety project for an in-vehicle driving behavior field study collected from naturalistic driving data and associated participant, vehicle, and crash-related data. The project was conducted by six site contractors located at geographically distributed data collection sites throughout the United States and more than 3000 individuals participated in the study. Given that the SHRP 2-NDS is a federally funded study that involves human subjects, the collection of the data and its use in analysis are subject to the approval of institutional review boards. The SHRP 2-NDS database is managed by the Virginia Tech Transportation Institute, and researchers interested in accessing the data must demonstrate that they are qualified researchers seeking the data for research purposes [24].

#### **4.9. The Center for Advanced Public Safety (CAPS)**

The Center for Advanced Public Safety (CAPS) is a research center at the University of Alabama that deals with vehicular crash data, and traffic safety improvements, among other research areas. CAPS has developed a tool for crash data analysis called the Critical Analysis Reporting Environment (CARE), which has many useful analytical functions such as, frequency distributions, cross-tabulations, and statistical significance tests. CARE can compare the performance of one subset of data against another in terms of all potential variables that could demonstrate performance differentials. CARE analysis software is free to download and is required to analyze and visualize the electronic data contained within CAPS datasets. The CAPS online crash datasets are free to download, and contains a variety of crash data files that mainly belongs to the state of Alabama, such as the vehicle crash files, the driver data file, the person data file, and the road data file [25].

### **5. Count Data**

When discussing traffic crash data, it is important to differentiate between a count, and count data. The term count typically refers to an enumeration of events.

Count data, on the other hand, refers to the observations made about events that are enumerated [26]. A common quality of count data is that (0.0) is the most frequently observed value, (1.0) is the next most observed, (2.0) the next, and so on. Use of count data is widespread in many disciplines, including transportation engineering. Examples of count data applications in transportation include the number of driver route changes per day, the number of trip departure changes per week, number of vehicles waiting in a queue, and the number of crashes observed on road segments per some time period, such as a year, or five years. Count data are often described as random events, sporadic (*i.e.* isolated or scattered), rare, discrete, not continuous, and non-negative integers [27]. One frequent pitfall is to model count data as continuous data by applying an ordinary least square regression [28]. This approach is inappropriate because regression models can produce predicted values that are non-integers and can also predict values that are negative, both of which are inconsistent with count data. In addition, many distributions of count data are positively skewed with many observations in the data set having a value of 0.0. The high number of zeros in the data set prevents the transformation of a skewed distribution into a normal one, which is a requirement of normal distribution. An alternative is to use a Poisson distribution or one of its variants. Poisson distributions have a number of advantages over an ordinary normal distribution, including a skew, discrete distribution, and the restriction of predicted values to non-negative numbers [28].

## 6. Common Problems with Crash Data

Crash data suffer from some problems or issues that have been identified over the years. These problems are a potential source of error in modeling crash data that may cause incorrect estimates and inferences. These issues are summarized below:

- **Over-dispersion:** over-dispersion occurs when the observed variance exceeds the theoretical variance of the crash counts, which violates the assumption of the most common count-data modeling approach. Over-dispersion in crash data can result from a variety of factors, such as the clustering of data, unaccounted temporal correlation, and model miss-specification (Cameron and Trivedi 1998). When data are over-dispersed, estimation of a crash model can lead to biased parameter estimates, which in turn could lead to incorrect inferences regarding the factors that determine crash-frequencies [29] [30] [31] [32].
- **Under-dispersion:** under-dispersion occurs when the observed variance of the crash counts is smaller than the assumed (*i.e.* theoretical) variance, and most likely to occur with small sample sizes. Although rare, however, under-dispersion can lead to incorrect parameter estimates and crash prediction [29] [31] [32] [33].
- **Small Sample Size:** crash data collection process may be expensive, therefore crash data are sometimes characterized by a small number of observations

(*i.e.* small sample size), which can produce low sample-mean. Small samples can cause estimation problems in traditional count prediction models. For example, with small sample sizes, the maximum likelihood estimation of parameters could produce insufficient results [34] [35]. Also, it was shown that the dispersion parameter of the negative binomial model can be incorrectly estimated when using data characterized by a small sample size and low sample mean [36].

- Time Interval Variations: crash data are typically collected over some time period, such as one year, three years, and five years. Over the collection period, some explanatory variables and their relationship to the crash incidents may change a reality that is not usually considered due to the lack of detailed data within the collection period. Ignoring within-period variation in explanatory variables may result in biased estimation of parameters and incorrect prediction of crashes as a result of unobserved heterogeneity [28] [35].
- Temporal and Spatial Autocorrelations: the prediction of crash models can be improved when several years of crash data are utilized, such as a period of three years instead of one year [37]. However, this means that the same roadway entity will generate multiple observations, which will be correlated over time because many of the unobserved effects associated with a specific roadway entity will remain the same over time. This phenomenon is termed temporal autocorrelation, which can adversely affect the precision of parameter estimates. Similarly, correlation of observations over space can exist given that roadway entities may be in close proximity and may share unobserved effects. This phenomenon is termed spatial autocorrelation and if not appropriately addressed, can also lead to incorrect parameter estimates [38] [39] [40] [41] [42].
- Omitted-Variables Bias: modeling crash prediction with few explanatory variables could produce simplified models with omitted-variables bias. Leaving out important explanatory variables can result in biased parameter estimates and incorrect inferences, especially if the omitted variable is correlated with variables included in the model, which is often the case [43] [44] [45] [46].
- Under-Reporting: traffic crash data may suffer from under-reporting effects, especially for minor and less severe crashes. The unknown parameters in the models are generally estimated assuming random sampling from the population, therefore, if under-reporting is not accounted for, then it could result in biased samples that are likely to produce incorrect parameters in the model-estimation process [28] [44] [47].
- Non-Linear Relationships Bias: many crash prediction models assume that explanatory variables influence the dependent variables in linear manner. However, it has been shown that non-linear functions can often better characterize the relationships between crash frequencies and explanatory variables. For example, using traffic flow as a measure of exposure, some have found that the crash prediction per unit of exposure becomes smaller as traf-

fic flow increases pointing to unobserved heterogeneity and possible other specification problems in the functional form of the model [39] [48].

## 7. Conclusion

Road traffic crash data are useful tools to support highway safety programs that tend to reduce road traffic crashes. They can be used by many authorities such as: law enforcements to identify persons at fault in road traffic crashes; insurers seeking facts about traffic crash claims; road safety researchers to access crash reliable database; decision makers to develop long-term, statewide strategic plans for traffic and highway safety; and highway safety administrators to help educate the public. Given such trends, this paper presented a general overview of the sources, collection methods, and problems associated with crash data to better gaining an understanding of road traffic operational problems, locating hazardous road sections, identifying risk factors, developing accurate diagnosis and remedial measures, and evaluating the effectiveness of road safety programs.

## References

- [1] WHO (2015) WHO Global Report on Road Safety 2015.  
[http://www.who.int/violence\\_injury\\_prevention/road\\_safety\\_status/2015/en/](http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/)
- [2] NCSA (2015) NHTSA-National Center for Statistics and Analysis.  
<http://www.nhtsa.gov/NCSA>
- [3] Transport Canada (2016) Road Safety in Canada.  
<http://www.tc.gc.ca/eng/motorvehiclesafety/tp-tp15145-1201.htm>
- [4] Beirness, D.J. and Beasley, E. (2011) A Comparison of Drug- and Alcohol-Involved Motor Vehicle Driver Fatalities. Canadian Centre on Substance Abuse, Ottawa.
- [5] World Bank (2015) The World Bank—Transport for Development.  
[http://blogs.worldbank.org/transport/why-vehicle-safety-matters-crash-related-deaths?cid=EXT\\_WBBlogSocialShare\\_D\\_EXT](http://blogs.worldbank.org/transport/why-vehicle-safety-matters-crash-related-deaths?cid=EXT_WBBlogSocialShare_D_EXT)
- [6] Mohan, D. (2002) Road Safety in Less-Motorized Environments: Future Concerns. *International Journal of Epidemiology*, **31**, 527-532.  
<https://doi.org/10.1093/ije/31.3.527>
- [7] Elvik, R. (2006) Laws of Accident Causation. *Accident Analysis and Prevention*, **38**, 742-747.
- [8] NCSA (2016) NHTSA-National Center for Statistics and Analysis.  
<http://www.nhtsa.gov/NCSA>
- [9] NHTSA-Ruling (2010) The Crash Data Services.  
<http://www.crashdataservices.net/NHTSAruling.html>
- [10] NHTSA-CODES (2011) The Crash Data Outcome Evaluation System.  
<http://www.nrd.nhtsa.dot.gov/cats/listpublications.aspx?Id=219&ShowBy=Category>
- [11] Bar-Gera, H. (2007) Evaluation of a Cellular Phone-Based System for Measurements of Traffic Speeds and Travel Times: A Case Study from Israel. *Transportation Research Part C*, **15**, 380-391.
- [12] Fraser, S. (2007) The Use of Floating Cellular Telephone Data for Real-Time Transportation Incident Management. McMaster University, Hamilton.
- [13] Robichaud, K. and Gordon, M. (2003) Assessment of Data-Collection Techniques for Highway Agencies. *Transportation Research Record*, **1855**, 129-135.

- <https://doi.org/10.3141/1855-16>
- [14] Sliupas, T. (2006) Annual Average Daily Traffic Forecasting Using Different Techniques. *Transport*, **21**, 38-43.
- [15] Ehlert, A., Bell, M.G.H. and Grosso, S. (2006) The Optimization of Traffic Count Locations in Road Networks. *Transportation Research Part B*, **40**, 460-479.
- [16] Fricker, J. and Kumapley, R. (2002) Updating Procedures to Estimate and Forecast Vehicle-Miles Traveled. Purdue University, West Lafayette.  
<https://doi.org/10.5703/1288284313337>
- [17] NHTSA-FARS (2016) The Fatality Analysis Reporting System.  
<http://www.nhtsa.gov/FARS>
- [18] NASS-GES (2016) The General Estimate System.  
<https://www.nhtsa.gov/national-automotive-sampling-system-nass/nass-general-estimates-system#11381>
- [19] NASS-CDS (2016) The Crash Worthiness Data System.  
<https://www.nhtsa.gov/national-automotive-sampling-system-nass/crashworthiness-data-system>
- [20] NHTSA-SDS (2016) NHTSA-The State Data System.  
<https://www.nhtsa.gov/state-data-programs/sds-overview>
- [21] HSIS (2016) The Highway Safety Information System.  
<http://www.hsisinfo.org/index.cfm>
- [22] Data.gov. (2016) US Government Data.  
<https://catalog.data.gov/dataset>
- [23] The US Census Bureau (2016) The Data of the US Census Bureau.  
<http://www.census.gov/en.html>
- [24] SHRP2-NDS (2016) The Strategic Highway Research Program 2—Naturalistic Driving Study. <https://insight.shrp2nds.us/>
- [25] CAPS (2016) The Center for Advanced Public Safety.  
<http://www.caps.ua.edu/analytics/downloads/datasets/public/>
- [26] Hilbe, J. (2014) Modeling Count Data. Cambridge University Press, Cambridge.  
<https://doi.org/10.1017/cbo9781139236065>
- [27] Hauer, E. (1992) Traffic Conflicts and Exposure. *Accident Analysis and Prevention*, **14**, 359-364.
- [28] Glenberg, A. (1996) Learning from Data: An Introduction to Statistical Reasoning. 2nd Edition, Lawrence Erlbaum Associates, Mahwah.
- [29] Cameron, A.C. and Trivedi, P.K. (1998) Regression Analysis of Count Data. Cambridge University Press, Cambridge. <https://doi.org/10.1017/CBO9780511814365>
- [30] Miaou, P. (1994) The Relationship between Truck Accidents and Geometric Design of Road Sections: Poisson versus Negative Binomial Regressions. *Accident Analysis and Prevention*, **26**, 471-482.
- [31] Park, S. and Lord, D. (2007) Multivariate Poisson-Lognormal Models for Jointly Modeling Crash Frequency by Severity. *Transportation Research Record*, **2019**, 1-6.  
<https://doi.org/10.3141/2019-01>
- [32] Abdulhafedh, A. (2016) Crash Frequency Analysis. *Journal of Transportation Technologies*, **6**, 169-180. <https://doi.org/10.4236/jtts.2016.64017>
- [33] Oh, J., Washington, S. and Nam, D. (2006) Accident Prediction Model for Railway-Highway Interfaces. *Accident Analysis and Prevention*, **38**, 346-356.
- [34] Wood, G.R. (2002) Generalized Linear Accident Models and Goodness of Fit Testing. *Accident Analysis and Prevention*, **34**, 417-427.

- [35] Lord, D. and Bonneson, A. (2007) Development of Accident Modification Factors for Rural Frontage Road Segments in Texas. *Transportation Research Record*, **2023**, 20-27. <https://doi.org/10.3141/2023-03>
- [36] Lord, D. (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 and Prevention*, **46**, 751-766.
- [37] Mohammadi, M., Samaranayake, V. and Bham, G. (2014) Crash Frequency Modeling Using Negative Binomial Models: An Application of Generalized Estimating Equation to Longitudinal Data. *Accident Analysis and Prevention*, **2**, 52-69.
- [38] Gujarati, D. (1992) Essentials of Econometrics. McGraw-Hill, New York.
- [39] Lord, D. and Persaud, N. (2000) Accident Prediction Models with and without Trend: Application of the Generalized Estimating Equations Procedure. *Transportation Research Record*, **1717**, 102-108. <https://doi.org/10.3141/1717-13>
- [40] Washington, P., Karlaftis, G. and Mannering, F. (2010) Statistical and Econometric Methods for Transportation Data Analysis. 2nd Edition, Chapman Hall, CRC, Boca Raton.
- [41] Lord, D. and Mannering, F. (2010) The Statistical Analysis of Crash Frequency Data: A Review and Assessment of Methodological Alternatives. *Accident Analysis and Prevention*, **44**, 291-305.
- [42] Savolainen, P., Mannering, F., Lord, D. and Quddus, M. (2011) The Statistical Analysis of Highway Crash-Injury Severities: A Review and Assessment of Methodological Alternatives. *Accident Analysis and Prevention*, **43**, 1666-1676.
- [43] Arminger, G., Clogg, C. and Sobel, M. (1995) Handbook of Statistical Modeling for the Social and Behavioral Sciences. Plenum Press, New York. <https://doi.org/10.1007/978-1-4899-1292-3>
- [44] Caliendo, C., Guida, M. and Parisi, A. (2007) A Crash-Prediction Model for Multi-lane Roads. *Accident Analysis and Prevention*, **39**, 657-670.
- [45] El-Basyouny, K. and Sayed, T. (2009) Collision Prediction Models Using Multivariate Poisson-Lognormal Regression. *Accident Analysis and Prevention*, **41**, 820-828.
- [46] Geedipally, R., Lord, D. and Dhavala, S. (2012) The Negative-Binomial Lindley Generalized Linear Model: Characteristics and Application Using Crash Data. *Accident Analysis and Prevention*, **45**, 258-265.
- [47] Anastasopoulos, P.C. and Mannering, F. (2009) A Note on Modeling Vehicle Accident Frequencies with Random-Parameters Count Models. *Accident Analysis and Prevention*, **41**, 153-159.
- [48] Shankar, N., Milton, J.C. and Mannering, F. (1997) Modeling Accident Frequencies as Zero-Altered Probability Process: An Empirical Enquiry. *Accident Analysis and Prevention*, **29**, 829-837.

**Submit or recommend next manuscript to SCIRP and we will provide best service for you:**

Accepting pre-submission inquiries through Email, Facebook, LinkedIn, Twitter, etc.

A wide selection of journals (inclusive of 9 subjects, more than 200 journals)

Providing 24-hour high-quality service

User-friendly online submission system

Fair and swift peer-review system

Efficient typesetting and proofreading procedure

Display of the result of downloads and visits, as well as the number of cited articles

Maximum dissemination of your research work

Submit your manuscript at: <http://papersubmission.scirp.org/>

Or contact [jts@scirp.org](mailto:jts@scirp.org)