Accident Analysis and Prevention 101 (2017) 107–116

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

A preliminary investigation of the relationships between historical crash and naturalistic driving

Anurag Pande [a], Sai Chand [b], Neeraj Saxena [b], Vinayak Dixit [b,*], James Loy [a], Brian Wolshon [c], Joshua D. Kent [d]

[a] Civil & Environmental Engineering, Cal Poly State University, San Luis Obispo, CA 93407, United States
[b] Research Centre for Integrated Transport Innovation, School of Civil & Environmental Engineering, University of New South Wales, Sydney, Australia
[c] Civil & Environmental Engineering, Louisiana State University, Baton Rouge, LA 70803, United States
[d] Louisiana State University, Baton Rouge, LA 70803, United States

ARTICLE INFO

Article history:
Received 16 September 2014
Received in revised form 30 January 2017
Accepted 31 January 2017
Available online 16 February 2017

Keywords:
Naturalistic driving data
Crash frequency
Negative binomial model
Random parameter negative binomial model
Traffic safety

ABSTRACT

This paper describes a project that was undertaken using naturalistic driving data collected via Global Positioning System (GPS) devices to demonstrate a proof-of-concept for proactive safety assessments of crash-prone locations. The main hypothesis for the study is that the segments where drivers have to apply hard braking (higher jerks) more frequently might be the "unsafe" segments with more crashes over a long-term. The linear referencing methodology in ArcMap was used to link the GPS data with roadway characteristic data of US Highway 101 northbound (NB) and southbound (SB) in San Luis Obispo, California. The process used to merge GPS data with quarter-mile freeway segments for traditional crash frequency analysis is also discussed in the paper. A negative binomial regression analyses showed that proportion of high magnitude jerks while decelerating on freeway segments (from the driving data) was significantly related with the long-term crash frequency of those segments. A random parameter negative binomial model with uniformly distributed parameter for ADT and a fixed parameter for jerk provided a statistically significant estimate for quarter-mile segments. The results also indicated that roadway curvature and the presence of auxiliary lane are not significantly related with crash frequency for the highway segments under consideration. The results from this exploration are promising since the data used to derive the explanatory variable(s) can be collected using most off-the-shelf GPS devices, including many smartphones.

1. Introduction

The emergence of geographic information systems (GIS) and global positioning systems (GPS) technologies have made it possible to collect, process, and graphically represent driving data at levels of detail and accuracy not possible a decade ago. The applications of these technologies span many facets of transportation such as turn by turn navigation, transit fleet tracking, etc. Prior research has suggested that traffic safety analyses can also be conducted with this technology; which complements the traditional methods typically based on the analysis of long-term crash data. This research sought to correlate the jerk (hard braking) data collected using GPS-based naturalistic driving experiments with long-term crash frequency at freeway segments. The occurrences of hard braking are not as rare as crashes or even "near-misses". When compared with the instances of normal braking, the rate at which drivers undertook the hard braking on segments helps in understanding the crash potential at the study locations. Since these measures can be derived using data from many commonly available devices (including many smartphones), the promising results from this research may lead to the ability to proactively flag locations with higher crash potential (i.e., sites with promise as defined by Hauer (1996)).

2. Background

Although there are a wide variety of methods and performance metrics to quantify "traffic safety" (Dixit et al., 2011, 2014; Dixit, 2013; Pande et al., 2015; Gettman et al., 2008), most road agencies use a single measure, i.e., traffic crashes. More specifically, the

* Corresponding author.
E-mail addresses: apande@calpoly.edu (A. Pande), saichand.chakka@unsw.edu.au (S. Chand), n.saxena@unsw.edu.au (N. Saxena), v.dixit@unsw.edu.au (V. Dixit), jmloy@calpoly.edu (J. Loy), brian@rsip.lsu.edu (B. Wolshon).

http://dx.doi.org/10.1016/j.aap.2017.01.023
0001-4575/

frequency, rate, severity, and cost estimates for crashes are all metrics used to identify high-hazard locations and used as the primary criteria to assess the need for and associated benefits of various road improvements (Hauer, 1996; Tarko and Kanodia, 2004).

Researchers have used different statistical methods to estimate model parameters and thereby to understand the roadway crash characteristics in a better way. For example, count data models were extensively used to estimate crash frequencies at selected locations (Anastasopoulos and Mannering, 2009; Bhat et al., 2014); Tobit model was used to study the accident rates (Anastasopoulos et al., 2012); the mixed logit (Moore et al., 2011), the ordered logit and the ordered probit models (O'Donnell and Connor, 1996; Kockelman and Kweon, 2002) were used to predict injury severity. In all these studies, the locations with the highest rates, frequencies, severities, and combinations thereof are regarded to be "less safe" or more hazardous than other locations. Analysis of crash data involves many issues (such as over and under-dispersion, under-reporting, low sample mean and small sample size, etc. to name few) associated with crash frequency data as put forward by Lord and Mannering (2010).

Few of the research efforts have recognized the potential power of identifying abnormal driving events and conflicts that caused near-collisions or have the potential to cause a collision. These events, otherwise known as crash surrogates have the ability to strengthen traditional traffic analysis methods by identifying "crash potential" (Guo et al., 2010; Klauer et al., 2006). Almost two decades ago, Hummer (1994) suggested that conflicts could be used as a supplement to traffic accident studies when estimating the traffic accident potential at an intersection or other locations. These conditions can be observed and identified much sooner than the results of traffic crash studies, which can often require several years for adequate data collection and analysis. Hauer (1996) proposed the use of "deviations from norm" to evaluate surrogate safety measures and identified locations that should receive the most attention and allocation of improvement resources.

The concept of crash surrogates is easy to comprehend, but it is often hard to identify these events within larger data sources. The most recent studies seek to proactively address unsafe conditions and behavior through the collection and analysis of naturalistic driving data with the extensive instrumentation of the subject vehicles (Dingus et al., 2006; Wu and Jovanis, 2012; Einink et al., 2014; Blatt et al., 2015). These studies help in understanding the way drivers interact with their own vehicle, in-vehicle and portable technologies, vehicle occupants and other road users, and road infrastructure, in different driving environments (Australian Naturalistic Driving Study, ANDS, 2015). The data from naturalistic studies have been analyzed to understand driver inattention and crash risk (Victor et al., 2015), prediction of crashes and near-crash events (McLaughlin et al., 2008), alcohol consumption (Smith et al., 2015), cell phone usage (Cook et al., 2015), etc. Some research studies have even attempted to identify drivers' safety critical maneuvers through detection of jerks (Bagdadi and Várhelyi, 2013) and braking maneuvers (Bagdadi, 2013). In those studies, it was critical to differentiate between controlled powerful braking versus "critical braking" that might lead to a crash or near-miss in real-time.

In this research, global positioning system (GPS) sensors were used to record detailed time and position information which, in turn, was used to measure a variety of movement parameters including speed, acceleration, travel direction, etc. Further, the GPS data were linked with the roadway network information and crash databases. The analysis facilitated the search for locations of recurrent atypical maneuvering suggestive of driving conflicts and determine if these locations also show patterns of crashes that are related, directly or indirectly, to such conflicts. Interested readers are recommended to refer McLaughlin and Hankey (2015) for an excellent overview of the linking procedure of GPS data with road data.

The contribution offered by this research differs from previous efforts in two critical ways:

(1) First, since real-time classification of crashes/near-misses from 'normal' driving is not the goal of this effort; events of interest for this research (such as braking, etc.) are not necessarily near-misses. As discussed later, crash rates on the segments in this study were expressed in terms of the rate at which all study participants cumulatively had to make an abrupt deceleration while braking (either due to intentional powerful braking or a safety critical near-miss). In other words, the hypothesis is that the segments where drivers have to brake hard more often might be the "unsafe" segments with more crashes over the long-term. These occurrences in the crash data are not going to be as rare as crashes or even "near-misses". Hence, it is necessary to account for not the just occurrence of these events, but the rate at which drivers undertook these maneuvers relative to normal braking.
(2) The variable(s) used for the analysis can also be obtained from a simple instrumentation set up.

The analysis is set up as a crash frequency estimation problem with data from the driving experiment serving as the independent variables alongside roadway characteristics and traffic information. The negative binomial regression modeling framework was used in the study followed by random parameter count model estimation to address potential concerns related to Omitted Variable Bias (OVB).

3. Data collection

Driving data utilized for this research were collected from 33 staff members of California Polytechnic State University, San Luis Obispo. Participants in the study were selected using a screening questionnaire which solicited both personal and driving information. Using the information provided, participants were segregated based on age, typical route selections, and driving frequency. The test subjects all commuted from various cities and communities within the counties of San Luis Obispo and Santa Barbara. However, all of the participants pass through the same US101 corridor in San Luis Obispo as part of their daily commute.

This enabled us to get more information of the drivers along this corridor than other corridors. Participants were assigned a random identification number that was later linked to the GPS data collection unit (described later) to maintain driver confidentiality. All drivers were between the ages of 25 and 55 years old and included 23 female drivers and 10 male drivers. Data collection was undertaken during the period from July 2012 to January 2013 with each driver providing about 10 days of driving data.

Each study participant was given a GPS Data Logger V3.15 (data logger) to keep in their personal commute vehicle for a period of two weeks. GPS data loggers were programmed to record data in a comma-separated format from parsed sentences that followed the National Marine Electronics Association (NEMA) standards. GPS positional values were recorded as per the WGS 84 standard. The information recorded by the data loggers included latitude; longitude; altitude; heading; speed; number of satellites used; position dilution of precision (PDOP); horizontal dilution of precision (HDOP); vertical dilution of precision (VDOP); fix; universal time code (UTC); year; month; and day. Data was nominally recorded at an average rate of 3 hertz (i.e., 3 readings per second). To preserve battery life, a "sleep" mode was implemented to stop the data recording when the vehicle remained stationary for a period

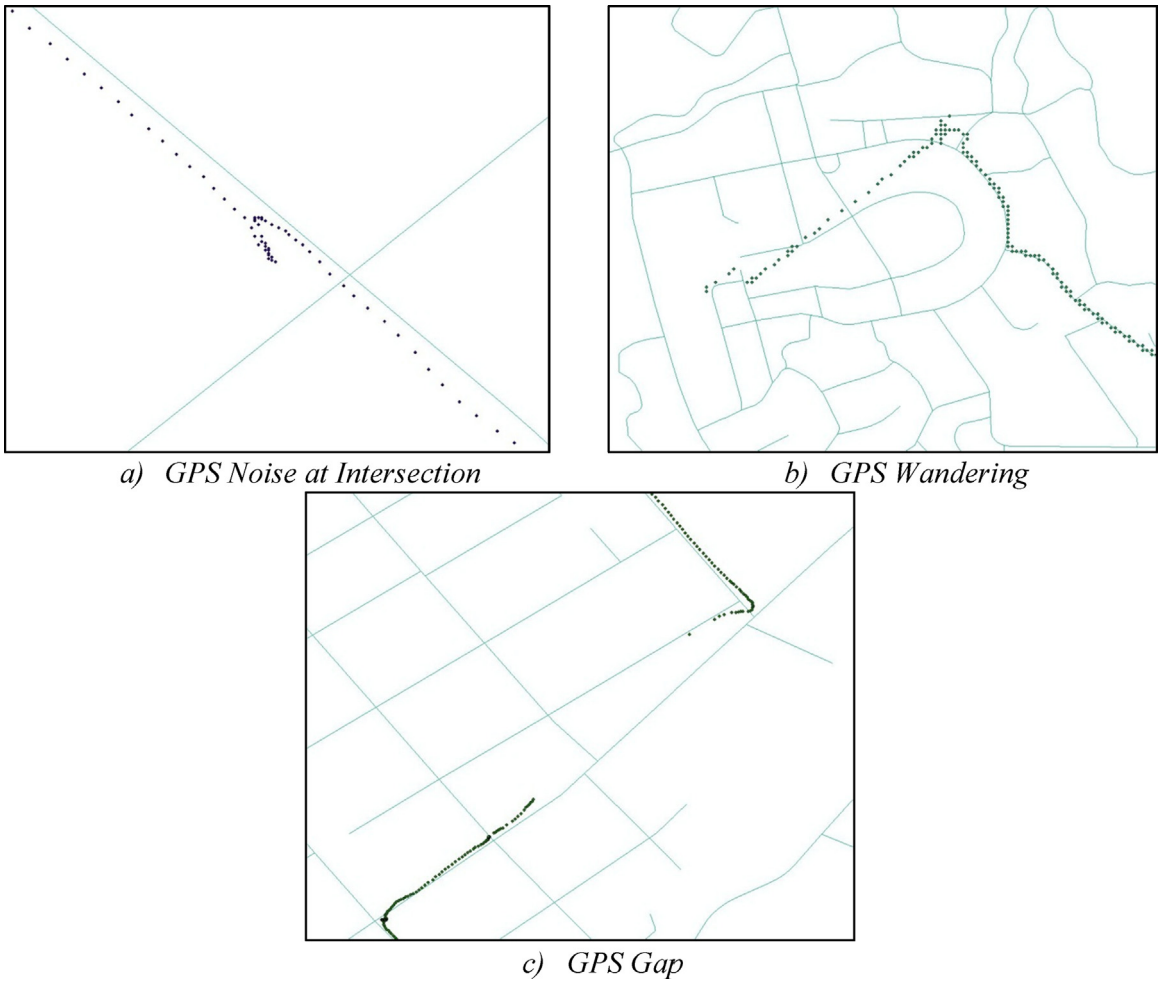

Fig. 1. Illustration of GPS errors on base map.

longer than 300 s. The fully charged battery for the devices typically lasted for the data collection period, depending on driving frequency.

During the experimentation period, the data loggers were strategically placed in participant's glove boxes or vehicle center consoles to prevent the data loggers from sliding and collecting erroneous data. Through extensive testing, it was also found that placing the logger in either location did not significantly affect the GPS reception or the data. Participants were also asked to refrain from allowing other relatives or friends to drive their vehicle while the logger was collecting data. However, we cannot disregard the possibility of others driving the car. As there was no video capturing mechanism in the cars, it was practically impossible to cross-check this instruction. As we show later, irrespective of the characteristics of individual participants, the jerk rate was correlated with high crash locations.

4. Data processing methodology

4.1. Diagnostic analysis

GPS data collected from each study participant was processed by combining multiple data files from the data logger into a single file named according to its identification number. Driver attributes (including gender, age, commuting behavior, etc.) collected in the initial survey were linked with collected GPS data. Each observation in the database was assigned a "trip number" attribute based on the relative time difference between successive GPS data point readings to separate different trips. A new trip number was assigned to an observation if the time difference between consecutive data points was greater than 30 min. Individual participant's driving data was then imported into ESRI ArcMap and overlaid onto the San Luis Obispo County GIS base map. GIS-based symbology options were also used to visually inspect speed and heading attributes to accurately identify the streets to which the data belonged.

It is worth mentioning that the data reported by these devices was not without errors. Fig. 1 illustrates the most common errors: (a) GPS noise at intersections, (b) GPS wandering, and (c) GPS gaps. GPS noise, which often occurred when the vehicle was stopped or driving at low speeds, was the most common error observed in the data. This error was encountered when the GPS positional data changed while the observed speeds and heading remained constant, resulting in a large "cluster" of data points observed at the actual stopping location. GPS noise also occurred during very short braking periods in which GPS noise resulted in inaccurately recording the vehicle speed. GPS wandering was also observed in some participant's data and was identified when the GPS positional data differed from the true location of the vehicle. Wandering typically was a random error and was identified by observing vehicles seemingly traveling on roadways that did not physically exist. Large gaps in the GPS data also were seen for few participants. These errors were caused by a sudden lack of GPS satellite availability due to some impedance or communication malfunction. This lack of accuracy in the data precluded the ability to detect lane changing events or which lane a participant was in during a given reading.

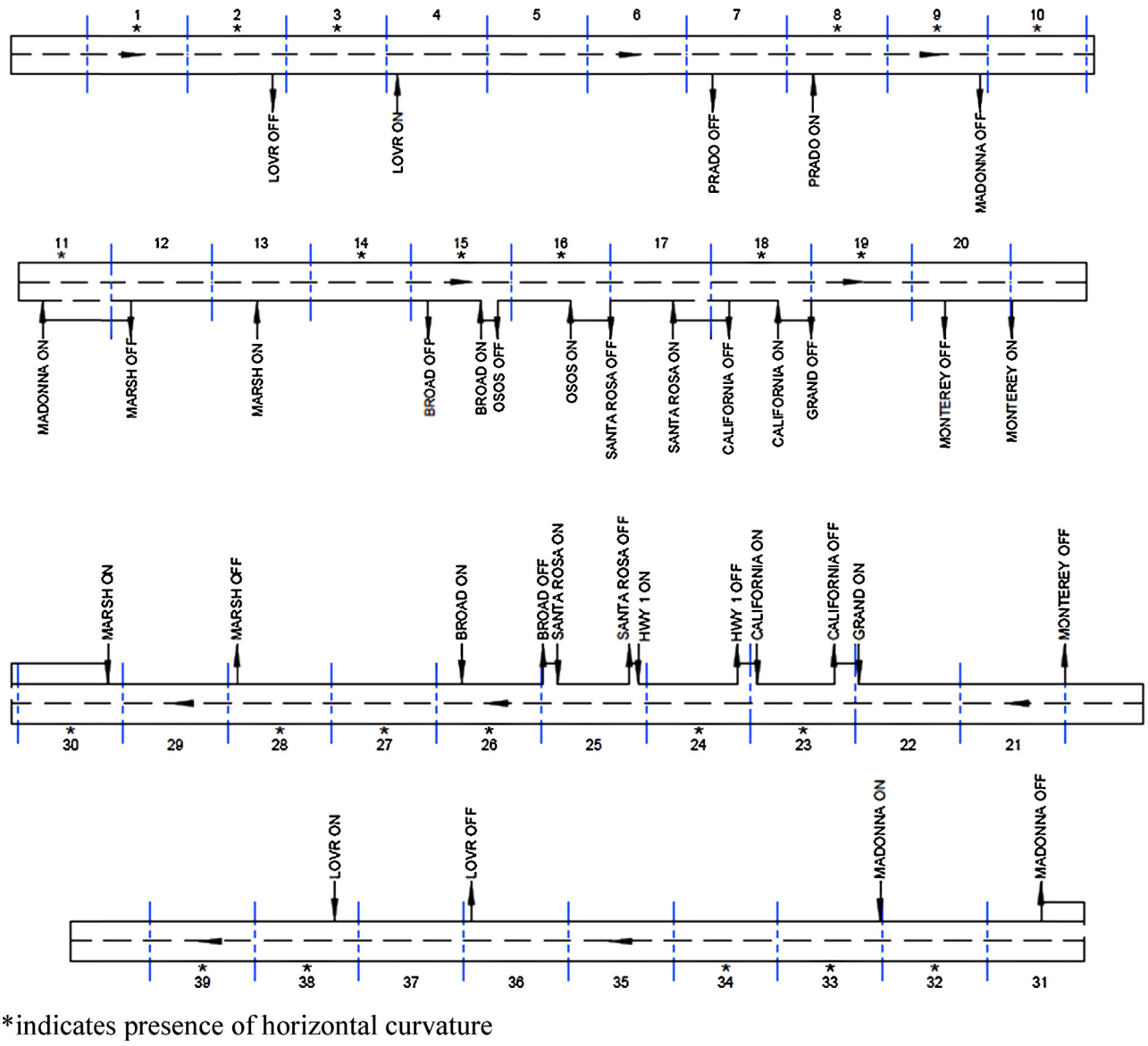

Fig. 2. Straight line diagrams for US 101 NB (top; 1–20) and SB (bottom 21–39) segments with on and off-ramp locations and direction of traffic movements. *indicates presence of horizontal curvature.

4.2. GIS linear referencing

A consistent methodology needed to be implemented to ensure proper GPS data was being utilized for analysis relating to individual roadways, due to spatial anomalies in the data. The process of linear referencing available in ArcMAP was selected to approach this problem. Linear referencing is a spatial-analytical technique for storing and referencing point events relative to their positions along a measured route. The process of linear referencing creates a linear axis for which GPS data points could be spatially compared with one another. In order to correlate the GPS data collected in this study with historical crash occurrences, both datasets were referenced to the San Luis Obispo road network. Using the road network maintained within the San Luis Obispo County GIS base map, individual route segments were merged together to form a single street network. The routes of interest in this research were the northbound (NB) and southbound (SB) lanes of US Highway 101; a four-lane freeway with a posted speed of 65 mph located within San Luis Obispo. Using the "*Create Routes*" tool in the ArcGIS Desktop software package, the highway sections were dynamically segmented and assigned a unique identifier based on an alpha-numeric reference scheme. The resulting output included two unique polyline features which contained spatial information along the length of the freeway.

After processing the GIS base map data, the next step was to assign each GPS data point's attributes to the corresponding location along the linearly referenced route. Assignment of GPS data points to the individual routes yielded the ability to determine the relative distance traveled between GPS data points. A large point radius distance of 300 feet was adopted to capture all data points on the route bypassing any possible GPS noise or wandering. These linear referencing steps were repeated for each individual participant set driving data and then combined into a linear referenced dataset.

4.3. Analysis variables and freeway segment information

U.S. 101 Highway was divided into 39 one-quarter-mile segments within the city limits of San Luis Obispo. While the segments were of equal length, they varied based on Average Daily Traffic

(ADT), the presence of horizontal curves, and the presence of auxiliary lanes near on- and off- ramps. Of the 39 freeway segments, 20 of the ¼-mile segments were in the NB direction and 19 were in the SB direction. These segments are shown in Fig. 2. The figure also illustrates the locations of on- and off-ramps. The sections of highway with an auxiliary lane are represented with an additional straight line drawn between the off- and on-ramp arrows. The freeway sections that contained horizontal curves were marked with an asterisk (*).

The GPS data was used to calculate accelerations (a) and jerks (j) at each time step and for every participant based on the following equations:

$$a = \frac{dv}{dt} \tag{1}$$

$$j = \frac{da}{dt} \tag{2}$$

where a, Acceleration (ft/s^2); dv, Change in velocity between successive observations (ft/s); dt, Change in time between successive observations (s); j, Jerk (ft/s^3); da, Change in acceleration between successive observations (ft/s^2).

With the complete linear referenced dataset, the GPS driving data was then filtered to further remove data points that were either erroneous and/or in reality belonged to other nearby surface streets. GPS data points with lower values of a total number of satellites in communication ($\leq$5 total satellites; about 1% of all data) were excluded from the analysis. Additionally, GPS data points with heading values that did not fall within an expected heading range for individual freeway segments were also eliminated. This filtering ensured that only accurate data with relatively reliable position values remained in the sample.

Histograms of acceleration and jerk (variables defined in Eqs. (1) and (2), respectively) were examined for each of the 39 segments to understand spatial distributions of these variables. Based on this heuristic examination, the rate of change in acceleration, i.e., jerk (Eq. (2)) was further evaluated. If a vehicle was decelerating gradually, one would expect the jerk value to be closer to zero compared to a situation where brakes were applied forcefully. In the case of the later, one would expect the magnitude of the jerk to be high. It was hypothesized that on unsafe freeway segments characterized by high long-term crash rates the need for sudden forceful braking would be encountered more often by the drivers.

Next, a set of 10 binary variables based on threshold value "X" was defined as follows:

High_Jerk_X = 1; If a vehicle was decelerating and magnitude of the rate of change in deceleration is greater than "X" ft/s^3

High_Jerk_X = 0; Otherwise.

The threshold value X was varied from 0.50 ft/s^3 to 2.75 ft/s^3 with an increment of .25 ft/s^3. This changing value of X produced 10 different binary variables defined as "High_Jerk_.50" through "High_Jerk_2.75". This classification of jerk values is similar to the kinematic search criteria proposed by Wu and Jovanis (2013). Relative frequencies of the levels of the 10 binary variables were then examined for each of the 39 freeway segments.

Of course, higher threshold values resulted in fewer instances of the corresponding binary variable equal to 1. In other words, at lower thresholds (e.g., 0.50 ft/s^3), High_Jerk_X = 1 was a more common occurrence; while at higher thresholds (e.g., 2.75 ft/s^3), it was a much rarer occurrence. Based on the relative frequencies of the two levels we estimated the percentage of observations for each segment where binary variables were equal to 1. In the next section, the correlation between this percentage (at various thresholds) and a measure of crash rate is examined for the 39 segments.

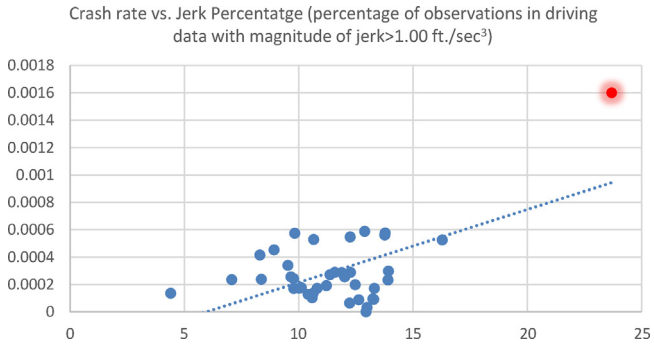

Fig. 3. The relationship between crash rate and jerk percentage (jerk >1 ft/s^3).

4.4. Crash data and correlation analysis

In terms of identifying safe and unsafe crash sites, the long-term crash counts/rates are a commonly accepted measure (Hauer, 1996). Therefore, in addition to the GPS data, traffic crash data for a 10-year period (2002 through 2011) were gathered from UC Berkeley's Traffic Injury Mapping System (TIMS) database ("Transportation Injury Mapping System (TIMS)," 2015). This crash data was also linear referenced to the same reference network as the GPS data. The long-term crash counts were extracted for all 39 US 101 segments. Utilizing Caltrans traffic data, an estimate of ADT for each segment was applied to normalize the total number of crashes by the amount of daily traffic (a measure of exposure). This provided a measure akin to crash rate to assess its relationship with the GPS driving data. Also, note that while crash counts were collected over a 10-year period, we normalized them with the 2011 ADT values (latest available) for the segments. San Luis Obispo is not a high growth city and the ADT values have experienced only gradual changes over the period since 2002. Hence, 2011 ADT could be used for normalization. Also, while crash rates have been used in this preliminary analysis, the analysis presented in the next section models crash counts and used ADT as a covariate in the model. This is based on the work by Hauer and Ezra (1995) that explored the issue of exposure and crash rate in detail.

Fig. 3 shows the plot of the estimates of crash rate and "jerk rate". The rate of events with jerk magnitude higher than 1 ft/s^3 is considered for this plot. The jerk percentage in the figure is the percentage of observations in the driving data with the magnitude of jerk >1 ft/s^3. A clear correlation between jerk percentage and crash rate for a segment can be seen from the Figure. One may recognize that differently marked point in Fig. 3 could potentially be an outlier. Even excluding the outlier, we observed a similar relationship between jerk rate and crash rate.

Fig. 4 shows the Pearson's correlation coefficient between crash rates and percentage of observations with High_Jerk_X = 1 based on four different thresholds (i.e., X = 0.50, X = 1.00, X = 2.00, and X = 2.75). All four of these correlation coefficients are statistically significant. In addition, it may be observed that the correlation coefficient rises as the threshold value, X, was increased from 0.5 to 2.75 even as the rise was most noteworthy when the threshold was changed from 0.5 to 1.0. The increase in correlation coefficient was modest when the threshold was changed from 2.00 to 2.75. Hence, the variable from GPS driving data used for crash frequency modeling in the next section was based on a threshold X = 2.00 ft/s^3. The variable for each freeway segment was the corresponding percentage of observations where the magnitude of jerk greater than 2.00 while decelerating.

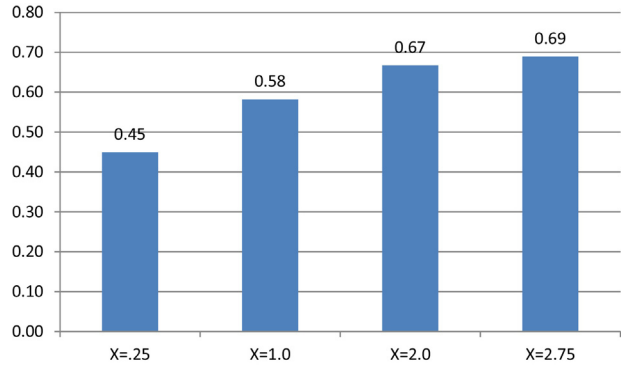

Fig. 4. Correlation coefficient for US 101 freeway segments between crash rates and percentage of observations with High_Jerk_X = 1 (based on four different threshold, i.e., X, values).

Table 1
Crash frequency estimation model with ADT and geometric variables.

Analysis of maximum likelihood parameter estimates			
Parameter	Estimate	Std. error	Wald chi-square
ADT/10,000	2.2306	2.4387	0.84
Curve	0.013	0.249	0.00
Auxiliary Lane	0.3284	0.3145	1.09
Dispersion	0.3921***	0.1174	
AIC (Akaike information criterion): 242.14			

*Significant at 90%.
** Significant at 95%.
*** Significant at 99%.

Table 2
Crash frequency estimation model with percentage of high magnitude jerk values while braking.

Analysis of maximum likelihood parameter estimates			
Parameter	Estimate	Std. error	Wald chi-square
Intercept	1.1616***	0.2686	18.71
Percentage of observations with High_Jerk_2.00 =1	0.1271***	0.0441	8.30
Dispersion	0.2908	0.1001	
AIC (Akaike information criterion): 230.94			

* Significant at 90%.
** Significant at 95%.
*** Significant at 99%.

Table 3
Crash frequency estimation model with ADT, geometric variables, and percentage of high magnitude jerk values while braking.

Analysis of maximum likelihood parameter estimates			
Parameter	Estimate	Std. error	Wald chi-square
Intercept	0.3798	0.6841	0.31
Percentage of observations with High_Jerk_2.00 = 1	0.1503***	0.0461	10.62
ADT/10,000	3.3005	2.1623	2.33
Auxiliary Lane	−0.0493	0.2852	0.03
Curve	−0.0370	0.2162	0.03
Dispersion	0.2637***	0.0937	
AIC (Akaike information criterion): 234.14			

* Significant at 90%.
** Significant at 95%.
*** Significant at 99%.

4.5. Crash frequency modeling

The next step in the analysis was to estimate long-term crash frequency using the negative binomial model for the 39 segments. Negative binomial models are widely used for crash frequency estimation (e.g., Abdel-Aty and Radwan, 2000; Milton and Mannering, 1998; Shankar et al., 1995). Three different specifications of the negative binomial model were estimated in this analysis. First, model estimation was undertaken with traditional measures as independent variables, i.e., 2011 ADT and binary variables representing the presence of curvature and auxiliary lane. It is worth noting here that the normalization of ADT as discussed in the previous section is limited to the preliminary exploration of correlations between jerk rates and crash rates. In the negative binomial formulation, however, ADT of the year 2011 is used. The results of the model are shown in Table 1. It is apparent from Table 1 that none of these variables were significantly related to long-term crash counts.

Next, the crash frequency was modeled as a function of *Percentage of observations with High_Jerk_2.00 = 1* (i.e., the percentage of observations with the magnitude of jerk >2.00 ft/s^3 while decelerating) for each segment. This variable was significant in the crash frequency model (See Table 2). Similar results were obtained with corresponding percentages based on higher threshold values (i.e., X). Next, the model using *ADT*, *Curvature*, and *Auxiliary lane presence* along with *Percentage of observations with High_Jerk_2.00 = 1* was estimated (See Table 3). However, the AIC (Akaike Information Criterion; the measure of model fit) did not improve (i.e., it increased) and *Percentage of observations with High_Jerk_2.00 = 1* was still the only significant variable. These results show that crash frequency on US 101 freeway was explained by the percentage of observations with a high jerk (model shown in Table 2) while it cannot be better explained based on the variables used traditionally in the crash frequency analysis.[1] Finally, the model presented in Table 2 is the suggested one as it is parsimonious and at the same time has a better model fit.

While these models suggest that locations with a higher rate of atypical braking events are also the locations with higher long-term crash frequency; the estimates from the negative binomial model may not be reliable as pointed out by Washington et al. (2010). Traditional statistical model, such as the NB model shown in Tables 1–3, do not allow parameter estimates to vary across observations (Lord and Mannering, 2010) meaning that the effect of explanatory variables on the frequency of crashes is constrained to be the same for all observations. However, due to unobserved variations from one roadway segment to the next (i.e., unobserved heterogeneity), it may not be an appropriate constraint. In this scenario, the fixed parameter model (e.g., the models shown in Tables 1–3) parameter estimates may be biased. Specifically, the absence of such important data can potentially cause serious specification problems that can lead to biased and inconsistent parameter estimates and erroneous crash predictions (Mannering et al., 2016). In other words, there might be some unobserved but relevant variables that are correlated with the observed variables.

The negative binomial model assumes fixed attribute effects across segments under consideration. The problem can be addressed by different methods:

[1] Though the subjects were widely distributed across age and gender, they are unlikely to be a representative sample of drivers using the US 101 corridor leading to selectivity bias. However, no demographic impacts were found on their jerk rates and therefore the trajectory data provides useful traffic probe data information (Dixit et al., 2011).

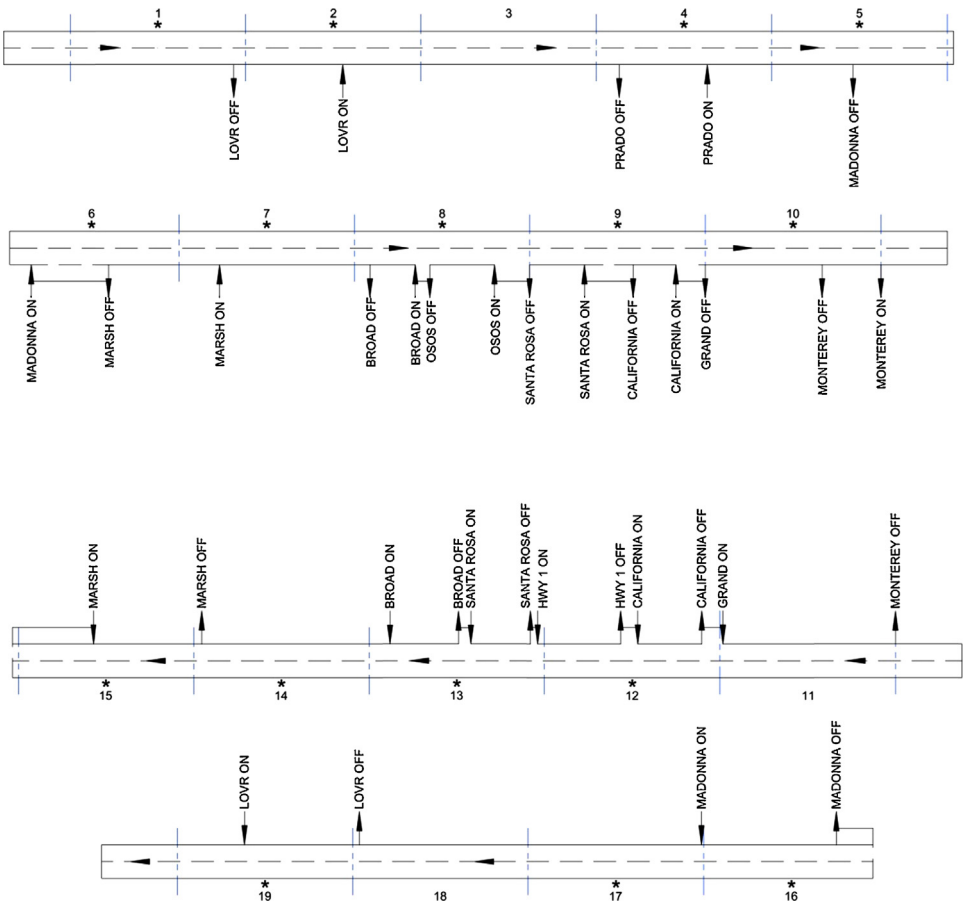

Fig. 5. Straight line diagrams for US NB (top; 1–10) and SB (bottom, 11–20) half-mile segments.

- Parametric approaches such as random parameter models (Anastasopoulos and Mannering, 2009; Mitra and Washington, 2012; Venkataraman et al., 2014; Bhat et al., 2014),
- Semi-parametric approaches such as finite-mixture latent class models (Park and Lord, 2009; Peng and Lord, 2011),
- Markov switching models (Malyshkina and Mannering, 2010), etc.

In this study, we used random parameter negative binomial model formulation discussed in the next section.

4.6. Random parameter count models

Random parameter approaches have been developed and widely used by the researchers to account for unobserved heterogeneity. The analyst can test for random parameters using a specific mixing distribution (such as normal, log-normal, uniform, etc.), across all observations for each explanatory variable included (Mannering et al., 2016).

In the current study, there were four explanatory variables under consideration. Of these four variables, jerk and ADT are continuous while the presence of horizontal curvature and auxiliary lane are binary. A variety of random-parameters negative binomial regression model specifications were tested, by altering the distributions of the parameters. LIMDEP (Greene, 2016), a widely used econometric software package among the research community was used and the random parameters were simulated using 1000 Standard Halton draws (Train, 2009). All the specifications explored had the nearly equal statistical goodness of fit measures, i.e., log-likelihood, AIC and BIC. Table 4 shows the final recommended model based on the goodness of fit measures. This model has the parameter of ADT variable as uniformly distributed, with the mean and scale parameter being statistically significant (p values of 0.06 and <0.001 respectively). The mean jerk effect was found to be highly significant and positive. It is important to note that all the other model specifications with various mixing distributions for random parameters yielded jerk as the only significant variable, with the scale parameter being insignificant, resulting in a point estimate. The following section shows the NB model resulting from segmentation of corridor into half-mile segments.

Table 4
Results from random parameter negative binomial model on the observed data.

Attribute	Estimated value
Mean of random parameters	
ADT	2.81**
Standard deviation of random parameters	
ADT	3.18753***
Fixed parameters	
Constant	0.53159
Jerk	0.22038***
Over-dispersion parameter	0.15977**
Log-likelihood at convergence	−112.089
AIC	234.2
BIC	242.49

* Significant at 90%.
** Significant at 95%.
*** Significant at 99%.

Table 5
Crash estimation utilizing negative binomial regression models for half-mile segments.

Half mile segment model 1
Negative binomial regression model with High Jerk
AIC (Akaike information criterion): 125.16
Analysis of maximum likelihood parameter estimates

Parameter	Estimate	Std. error	Wald chi square
Intercept	1.6340***	0.4899	11.12
High Jerk percentage	0.0944**	0.0410	5.31
Dispersion	0.0795*	0.0457	

Half mile segment model 2
Negative binomial regression model with ADT and geometric variables
AIC (Akaike information criterion): 127.50
Analysis of maximum likelihood parameter estimates

Parameter	Estimate	Std. error	Wald chi square
Intercept	1.9638***	0.6291	9.74
Curve	0.5724**	0.2906	3.88
Weaving	0.2136	0.2189	0.95
Average daily traffic	0.7622	2.2424	0.12
Dispersion	0.0759*	0.0421	

Half mile segment model 3
Negative binomial regression model with High Jerk, ADT and geometric variables
AIC (Akaike information criterion): 123.90
Analysis of maximum likelihood parameter estimates

Parameter	Estimate	Std. error	Wald chi square
Intercept	0.6163	0.7729	0.64
High Jerk percentage	0.0929***	0.0363	6.55
Curve	0.5084**	0.2626	3.75
Weaving	0.1925	0.1908	1.02
Average daily traffic	1.9171	2.0402	0.88
Dispersion	0.0412	0.0315	

Half mile segment model 4
Negative binomial regression model with High Jerk and ADT
AIC (Akaike information criterion): 126.46
Analysis of maximum likelihood parameter estimates

Parameter	Estimate	Std. error	Wald chi square
Intercept	1.0643	0.8253	1.66
High Jerk percentage	0.1041***	0.0416	6.25
Average daily traffic	1.6744	1.9802	0.71
Dispersion	0.0746*	0.0440	

* Significant at 90%.
** Significant at 95%.
*** Significant at 99%.

4.7. Half mile analysis

The fixed-effect negative binomial analysis was repeated utilizing ½-mile long segments to verify that trends observed in the first analysis remained true with larger analysis segment. ½-mile segments were constructed by combining two ¼-mile segments. A total of 19 ½-mile long segments were constructed from the first 38 ¼-mile long segment. From these 19 segments, 10 are from the NB direction of US Highway 101, and 9 of them are from the SB direction of US Highway 101. Fig. 5 shows the straight-line diagrams for these 19 ½-mile long segments.

Four negative binomial regression model specifications were evaluated using the ½-mile long analysis segments. Table 5 shows the results of each model specification discussing the statistical significance of explanatory variables at different confidence levels. Results of the three models for ½-mile segments were similar to the results seen in the ¼-mile segment analysis. Model 1 for the ½-mile long segments shows that the high jerk percentage has a significant relationship with total crashes observed. This fact holds true in Model 3 and Model 4 as well. However, Model 2 showed that the feature "curve", which is a binary variable assigned to the segment if a curve is present in the segment, appears to be significant as well. This variable was not significant in the ¼-mile analysis. However, when observing the data, it can be noted that there were only 3 out of the total 19 considered ½-mile segments that did not contain curvature. This lack of sampling size for non-curve sections resulted in a very high standard error in model coefficient. Because of this high standard error, it is believed that this significant correlation with curvature is a function of the small sample and may not be a reliable indicator of crash frequency. Finally, the Model 1 presented in Table 5 is the suggested one for crash prediction because of its parsimony and a better model fit compared to other specifications.

Heat charts showing relative values of total historical crashes, high jerk percentages, and ADT values for each bin are shown in Fig. 6. Like the plots from the ¼-mile analyses, the trends in the high jerk percentage follow the trends in crash data more closely than the ADT.

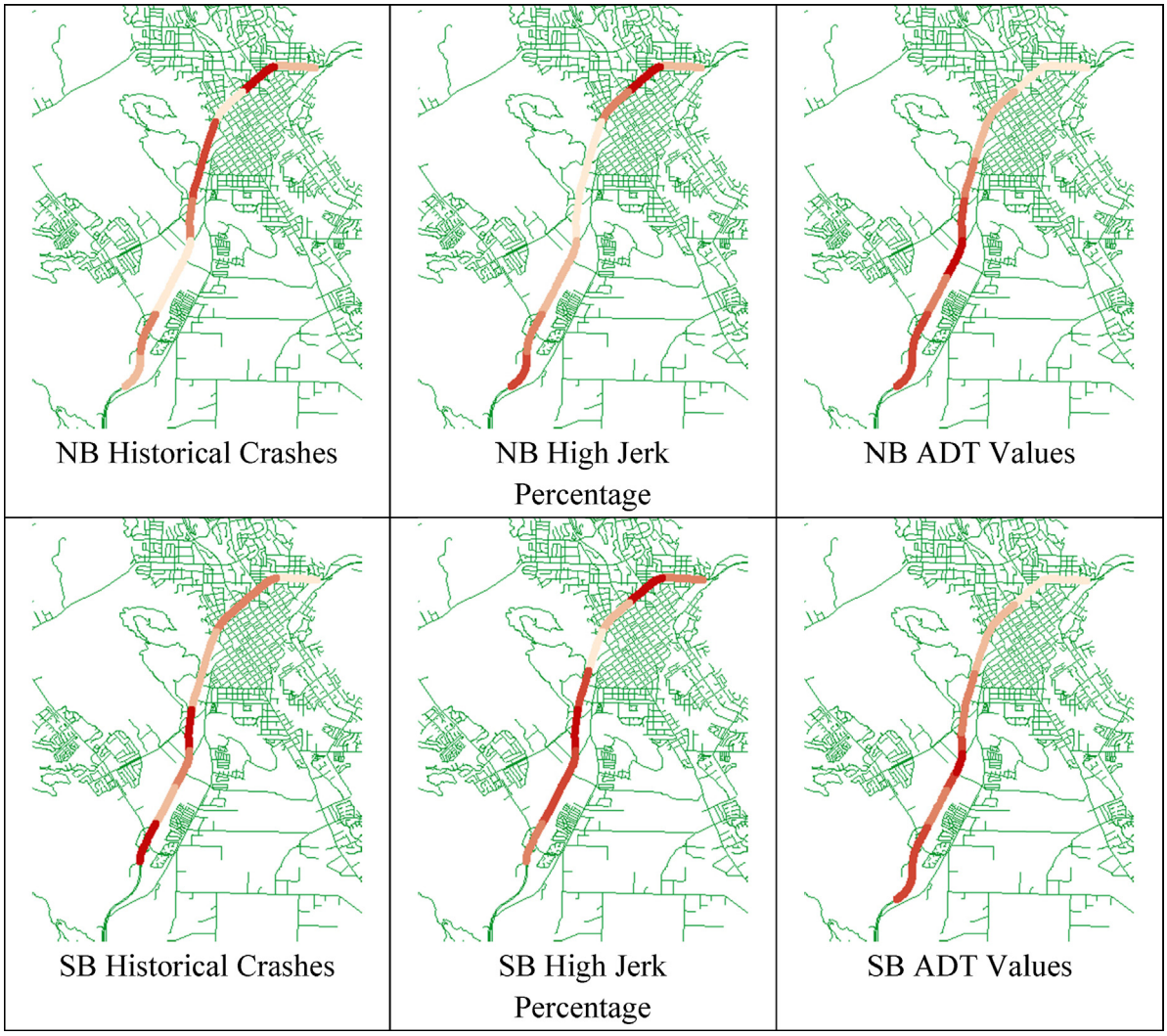

Fig. 6. Linear heat maps for half-mile analysis.

5. Conclusions

GPS driving data divided into trips was linear referenced to quarter-mile segments of NB and SB US 101 within the city limits of San Luis Obispo. The data was used to calculate acceleration and rate of change of acceleration (jerk). Analysis of GPS driving data in conjunction with the crash data revealed that freeway segments' long-term crash rates were correlated with the rate at which drivers had to exert high magnitude of jerk while decelerating. In a negative binomial crash frequency model estimation framework, the *percentage of observations with high magnitude of jerk while decelerating* was the only significant independent variable; whereas more traditional measures were found to be statistically insignificant. Furthermore, a random parameter negative binomial model with uniformly distributed parameter for ADT and fixed parameter for jerk provided a statistically significant estimate. The model results suggest that the frequency and rates of sudden deceleration events on a freeway segment can also be used as a surrogate safety measure in addition to other measures. Although the current study serves as a proof-of-concept, additional investigation through a larger scale study and different statistical models is required to generalize the findings in this paper. As all the road sections considered in the current analysis are on a single highway and within a close vicinity, some of the traditional explanatory variables such as percentage of heavy vehicles, weather and pavement conditions are not considered in the current study. The possibility of omitted variable bias cannot be ruled out and may be considered as a limitation of the current study.

Today, with the increasing availability of GPS-enabled smartphones; future research may be able to leverage "crowdsourcing" techniques for collecting data required for safety estimates. In addition, a smaller sample available from commercial fleet can also be used to gather this information. This study provides a foundation for these explorations in terms of merging the GPS data with GIS through linear referencing. Using the parameters available from the GPS data may help in determining whether certain safety measures need to be undertaken without having to wait for sufficient crash data to be accumulated. The procedure described here is expected to support future researchers and practitioners in crowdsourcing the data collection thereby making an estimation of the relevant variables easier and more cost-effective.

Acknowledgements

The authors want to thank National Science Foundation (NSF) for supporting the work presented herein through Grant #0927123 and Grant #0927358. The authors also thank two anonymous reviewers for their insightful comments that helped to improve the paper significantly.

References

Abdel-Aty, M.A., Radwan, A.E., 2000. Modeling traffic accident occurrence and involvement. Accid. Anal. Prev. 32 (5), 633–642.

Anastasopoulos, P.C., Mannering, F.L., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. Accid. Anal. Prev. 41, 153–159.

Anastasopoulos, P.C., Mannering, F.L., Shankar, V.N., Haddock, J.E., 2012. A study of factors affecting highway accident rates using the random-parameters tobit model. Accid. Anal. Prev. 45, 628–633.

Australian Naturalistic Driving Study (ANDS 2015). Retrieved from http://www.ands.unsw.edu.au/.

Bagdadi, O., 2013. Assessing safety critical braking events in naturalistic driving studies. Transp. Res. Part F: Traffic Psychol. Behav. 16, 117–126.

Bagdadi, O., Várhelyi, A., 2013. Development of a method for detecting jerks in safety-critical events. Accid. Anal. Prev. 50, 83–91.

Bhat, C.R., Born, K., Sidharthan, R., Bhat, P.C., 2014. A count data model with endogenous covariates: formulation and application to roadway crash frequency at intersections. Anal. Methods Accid. Res. 1, 53–71.

Blatt, A., et al., 2015. Naturalistic Driving Study: Field Data Collection. The Second Strategic Highway Research Program. Transportation Research Board, Washington, D.C.

Cook, J.K., Antin, J.F., Atkins, W.M., Hankey, J.M., 2015. Naturalistic Driving Study: Collecting Data on Cell Phone Use. The Second Strategic Highway Research Program (SHRP). Transportation Research Board, Washington, D.C.

Dingus, T.A., Klauer, S.G., Neale, V.L., Petersen, A., Lee, S.E., Sudweeks, J.D., et al., 2006. The 100-car Naturalistic Driving Study, Phase II-Results of the 100-car Field Experiment (No. DOT HS 810 593), Retrieved from http://trid.trb.org/view.aspx?id=783477.

Dixit, V.V., 2013. Behavioural foundations of two-fluid model for urban traffic. Transp. Res. Part C: Emerg. Technol. 35 (October), 115–126.

Dixit, V.V., Harrison, G.W., Rutstrom, E.E., 2014. Estimating the subjective risks of driving simulator accidents. Accid. Anal. Prev. 62 (2014), 63–78.

Dixit, V., Pande, A., Abdel-Aty, M., Das, A., Radwan, E., 2011. Quality of traffic flow on urban arterial streets and its relationship with safety. Accid. Anal. Prev. 43 (5), 1610–1616.

Einink, R., Barnard, Y., Baumann, M., Augros, X., Utesch, F., 2014. UDRIVE: The European Naturalistic Driving Study. Transport Research Arena, Paris, France.

Gettman, D., Pu, L., Sayed, T., Shelby, S.G., 2008. Surrogate Safety Assessment Model and Validation: Final Report, Retrieved from http://trid.trb.org/view.aspx?id=864039.

Greene, W.H., 2016. Limdep, Version 11. Econometric Software Inc., Plainview, NY.

Guo, F., Klauer, S.G., Hankey, J.M., Dingus, T.A., 2010. Near crashes as crash surrogate for naturalistic driving studies. Transp. Res. Rec.: J. Transp. Res. Board 2147, 66–74.

Hauer, Ezra, 1995. On exposure and accident rate. Traffic Eng. Control 36 (3), 134–138.

Hauer, E., 1996. Identification of sites with promise. Transp. Res. Rec.: J. Transp. Res. Board 1542 (1), 54–60.

Hummer, J.E., 1994. Traffic Conflict Studies, Retrieved from http://trid.trb.org/view.aspx?id=759177.

Klauer, S.G., Dingus, T.A., Neale, V.L., Sudweeks, J.D., Ramsey, D.J., 2006. The Impact of Driver Inattention on Near-Crash/Crash Risk: An Analysis Using the 100-Car Naturalistic Driving Study data [Electronic Version] (Retrieved January 7, 2007).

Kockelman, K.M., Kweon, Y.J., 2002. Driver injury severity: an application of ordered probit models. Accid. Anal. Prev. 34, 313–321.

Lord, D., Mannering, F.L., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transp. Res. Part A 44, 291–305.

Malyshkina, N., Mannering, F., 2010. Zero-state Markov switching count-data models: an empirical assessment. Accid. Anal. Prev. 42 (1), 122–130.

Mannering, F., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Anal. Methods Accid. Res. 11, 1–16.

McLaughlin, S.B., Hankey, J.M., 2015. Naturalistic Driving Study: Linking the Study Data to the Roadway Information Database. The Second Strategic Highway Research Program (SHRP). Transportation Research Board, Washington, D.C.

McLaughlin, S.B., Hankey, J.M., Dingus, T.A., 2008. A method for evaluating collision avoidance systems using naturalistic driving data. Accid. Anal. Prev. 40 (1), 8–16.

Milton, J., Mannering, F., 1998. The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies. Transportation 25 (4), 395–413.

Mitra, S., Washington, S., 2012. On the significance of omitted variables in intersection crash modeling. Accid. Anal. Prev. 49, 439–448.

Moore, D.N., Schneider IV, W.H., Savolainen, P.T., Farzaneh, M., 2011. Mixed logit analysis of bicyclist injury severity resulting from motor vehicle crashes at intersection and non-intersection locations. Accid. Anal. Prev. 43, 621–630.

O'Donnell, C.J., Connor, D.H., 1996. Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice. Accid. Anal. Prev. 28 (6), 739–753.

Pande, A., Loy, J., Dixit, V.V., Spansel, K., Wolshon, B., 2015. Integrity of estimates of the two-fluid model and gender impacts. Transp. Res. Part C: Emerg. Technol. 50, 141–149.

Park, B.-J., Lord, D., 2009. Application of finite mixture models for vehicle crash data analysis. Accid. Anal. Prev. 41 (4), 683–691.

Peng, Y., Lord, D., 2011. Application of latent class growth model to longitudinal analysis of traffic crashes. Transp. Res. Rec. 2236, 102–109.

Shankar, V., Mannering, F., Barfield, W., 1995. Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. Accid. Anal. Prev. 27 (3), 371–389.

Smith, R.C., Doerzaph, Z., Hankey, J., 2015. Naturalistic Driving Study: Alcohol Sensor Performance. The Second Strategic Highway Research Program (SHRP), Report S2-S31-RW-2. Transportation Research Board, Washington, D.C.

Tarko, A.P., Kanodia, M., 2004. Effective and fair identification of hazardous locations. Transp. Res. Rec.: J. Transp. Res. Board 1897 (1), 64–70.

Train, K.E., 2009. Discrete Choice Methods with Simulation. Cambridge University Press, Cambridge, New York.

Transportation Injury Mapping System (TIMS) (2015). Retrieved April 28, 2012, from http://www.tims.berkeley.edu/.

Washington, S.P., Karlaftis, M.G., Mannering, F., 2010. Statistical and Econometric Methods for Transportation Data Analysis. CRC Press.

Venkataraman, N., Shankar, V., Ulfarsson, G., Deptuch, D., 2014. Modeling the effects of interchange configuration on heterogeneous influences of interstate geometrics on crash frequencies. Anal. Methods Accid. Res. 2, 12–20.

Victor, T., et al., 2015. Analysis of Naturalistic Driving Study Data: Safer Glances, Driver Inattention, and Crash Risk. The Second Strategic Highway Research Program (SHRP). Transportation Research Board, Washington, D.C.

Wu, K.-F., Jovanis, P.P., 2012. Crashes and crash-surrogate events: exploratory modeling with naturalistic driving data. Accid. Anal. Prev. 45, 507–516.

Wu, K.-F., Jovanis, P.P., 2013. Defining and screening crash surrogate events using naturalistic driving data. Accid. Anal. Prev. 61, 10–22.