FINAL REPORT

Investigating Relationship between Driving Patterns and Traffic Safety using Smartphones Based Mobile Sensor Data

Date: May 2016

Olcay Sahin, Graduate Research Assistant, Old Dominion University
Rajesh Paleti, PhD, Assistant Professor Old Dominion University
Mecit Cetin, PhD, Associate Professor, Old Dominion University

Prepared by:
Transportation Research Institute
Old Dominion University
135 Kaufman Hall
Norfolk, VA, 23529

Prepared for:
Virginia Transportation Research Council
530 Edgemont Road
Charlottesville, VA 22903

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle	5. Report Date
Investigating Relationship between Driving Patterns and Traffic Safety using Smartphones Based Mobile Sensor Data	5/23/2016
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Olcay Sahin, Rajesh Paleti, and Mecit Cetin	

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS
Old Dominion University 135 Kaufman Hall Norfolk, VA, 23529	**11. Contract or Grant No.** DTRT13-G-UTC33

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
US Department of Transportation Office of the Secretary-Research UTC Program, RDT-30 1200 New Jersey Ave., SE Washington, DC 20590	Final 10/1/2015 – 5/22/2016
	14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

In spite of various advancements in vehicle safety technologies and improved roadway design practices, roadway crashes remain a major challenge. While certain hotspots may be unsafe primarily due to the geometric features of these locations, in many cases the safety risk seems to be an outcome of the unsafe driving patterns along the roadway stretching downstream and/or upstream of the actual crash locations. Even though there is plenty of research on correlating safety measures to roadway characteristics and some elements of traffic flow (e.g., exposure, speed), there is no significant literature on analyzing the correlation between high-resolution speed and acceleration data and crash risks along highway segments. Collecting such high-resolution data is now feasible with the mobile consumer devices such as smartphones. Smartphones are now equipped with sensors capable of recording vehicle performance data at a very fine temporal resolution in a cost-effective way. The current project used this mobile sensor data to identify unsafe driving patterns and quantified the relationship between these driving patterns and traffic crash incidences. The models with microscopic traffic measures were shown to be statistically better than traditional models that only control for roadway geometry and traffic exposure variables. Also, from a methodological standpoint, generalized count models that provide more flexibility through spatial dependency, heterogeneous dispersion, and random parameter heterogeneity were found to perform better than standard Poisson and Negative Binomial models.

17. Key Words	18. Distribution Statement
Mobile sensors, crash frequency, speed, acceleration, count models, Poisson, Negative Binomial, generalized ordered response	No restrictions. This document is available from the National Technical Information Service, Springfield, VA 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	33	

ACKNOWLEDGMENTS

This research project is funded by Virginia Transportation Research Council (VTRC) of Virginia Department of Transportation (VDOT) as part of VDOT's cost-share agreement to support MATS UTC. The authors would like to thank Catherine McGhee and Michael Fontaine at VTRC for their help in assembling the data sources and their valuable feedback over the course of this research project. The authors would also like to acknowledge Mr. Kenneth Wynne who helped with the literature synthesis and preliminary descriptive analysis of the data.

DISCLAIMER

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

PROBLEM

In the United States, there were 32,675 fatalities as a result of motor vehicle crashes in 2014 and current trends show that an increase of about 8.1 percent is expected in 2015 (NHTSA 2015). In the year 2014, in Virginia alone, 700 people were killed and 63,384 people were injured in a total of 120,282 motor vehicle accidents (DMV 2014). This combined with the fact that recent vehicle miles travelled (VMT) estimate of a compound annual growth rate of about 1% through the year 2033 makes traffic safety a matter of great concern (FHWA 2015). These crashes not only cause injury and loss of life, but they also cost a considerable amount to the people involved. For instance, in 2010, the economic costs of motor vehicle crashes in the nation totaled $242 billion. These costs come not only from the damage to vehicles and the medical bills of the injured but also include items such as $28 billion due to congestion (Blincoe *et al.* 2015).

Safety engineers have relied on crash frequency modeling to inform safety policy making concerning prioritization and implementation of countermeasures to improve safety. Crash frequency modelling is an attempt to quantify the expected number of crashes in a certain period (e.g., one year) at a specific location (e.g., roadway segment or intersection) as function of variables describing the location and the traffic conditions at the location. These models are referred to as the Safety Performance Functions (SPFs) in the Highway Safety Manual (HSM). In the past, most of these SPFs for roadways only used geometry (e.g., presence of shoulder, median width *etc*) and aggregate traffic measures (*e.g.*, traffic volume) as explanatory variables. However, there is limited literature on analyzing the correlation between microscopic traffic measures (e.g., high-resolution speed and acceleration) and crash risk. Lack of microscopic traffic data has been the primary impediment for limited research in this direction. In the absence of these microscopic measures, the parameter estimates in the SPFs can be biased and lead to wrong policy implications. For instance, it is possible that in the absence of microscopic traffic measures, the SPFs overestimate the impact of roadway improvements on safety because they confound the effect of driving patterns and the roadway characteristics. Also, the SPFs that lack microscopic traffic measures are not sensitive to countermeasures that are focused on changing the driving patterns (e.g., speed harmonization) rather than geometric features.

APPROACH

Smartphones are now equipped with sensors that are capable of recording vehicle performance data at a fine temporal resolution in a cost-effective way (Zhen and Qiang 2014). In fact, several auto insurance firms (e.g., Progressive's Snapshot) have been experimenting with monitoring driving activity (e.g., hard-brakes per mile) through on-board diagnostic (OBD) devices to assess and valuate the crash risk of individual drivers. However, there is no significant research on investigating the potential use of high-resolution data from mobile sensors of smartphones in understanding crash risks and safety measures for highway sections. The current study aims to make use of smartphone sensors to extract microscopic traffic measures that can serve as indicators of driving patterns and test the relationship between these microscopic traffic measures and crash frequency along freeway segments. To start-off, mobile sensor data was collected by driving along major roadways in the Hampton Roads region. Next, this data was overlaid on the transportation network to map probe data and the roadway segments. Then, several acceleration and deceleration

metrics were calculated for each roadway using the mobile sensor data. Subsequently, these metrics were appended to the Virginia Department of Transportation (VDOT) crash data for the past one year. Supplementary data sources were used to assemble information regarding roadway inventory data and traffic exposure information. Next, statistical model estimation was undertaken to quantify the relationship between microscopic traffic measures and crash incidences along major interstates in Hampton Roads.

METHODOLOGY

The research methodology adopted to accomplish the project goals comprises of four keys components. A brief description of these components follows.

Identification of Spatial and Temporal Resolution

One of the first steps to crash frequency modeling is selecting the spatial unit of analysis, *i.e.* the geographical extent of region over which the expected crash frequency is modeled. The current study focusses on crash frequency along major interstates in Hampton Roads. So, the empirical context implies that the interstates must be split into smaller segments that constitute the unit of analysis. However, this decision cannot be made arbitrarily because the availability of roadway inventory data and the homogeneity of resulting segments are critical to developing an accurate crash frequency model. So, several segment definitions were explored prior to choosing the spatial unit of analysis. Based on the relative merits of three different segment definitions, this study adopted the VDOT segment definition as the spatial unit of analysis (more details in the section titled 'Complete Documentation'. There were 513 unique roadway segments along major freeways in Hampton Roads (see Figure 1).

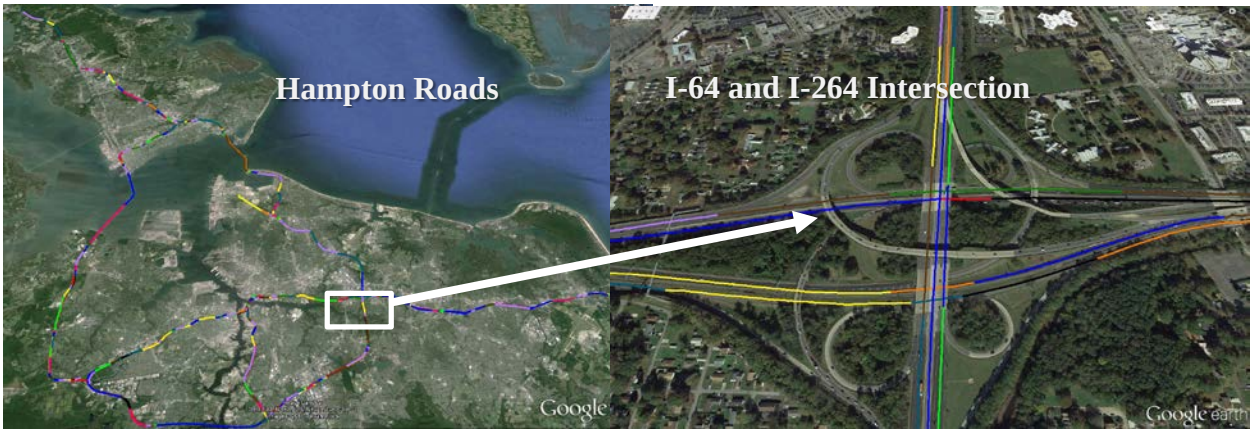

Figure 1 Spatial Unit of Analysis: Roadway Segments Definition

With regard to the temporal resolution, weekend crashes were omitted due to travel patterns being inconsistent with other travel days. Also, the evening peak period between 4:00 pm and 6:00 pm had the highest crash rate with nearly 18% of all crashes occurring during the two hour. This observation coupled with the constraint that it is not feasible to collect probe vehicle data using smartphones along all interstates during all hours of the day, the two hour time period between 4 and 6 pm was chosen as the temporal unit of analysis. So, the dependent variable of analysis is

crash frequency between 4 pm and 6 pm during weekdays in one year along each of the 513 freeway segments in Hampton Roads.

Data

The data for the empirical analysis was compiled from four different data sources (see Figure 2). *First*, crash information was obtained through a VDOT database that includes all reported crashes from October 2014 to October 2015 for the entire Hampton Roads Region. Next, each crash was geocoded to one of the roadway segments (*i.e.*, spatial unit of analysis). Lastly, all crash occurrences on each roadway segment in the past year were aggregated to obtain the crash frequency that serves as the dependent variable of analysis. *Second*, the roadway inventory information was obtained from a VDOT maintained database that contains information regarding the physical characteristics of the roadway including number of lanes, surface type (plant mix *versus* Portland Cement Concrete), presence of shoulder, presence of median and its type, HOV status, reversible lane status, and type of facility (one way, two-way divided, and two-way non-divided). *Third*, the traffic exposure information was obtained from 222 Wavetronix sensors maintained by VDOT (see Figure 3). This information was accessed from the Hampton Roads Traffic Operations Center (HRTOC) data repository. *Lastly*, mobile sensor data was collected by driving vehicles equipped with smartphones which were linked to on board diagnostic (OBD) devices through Bluetooth. The smartphone runs an Android application named GoGreen which has the capability of recording information from the GPS, accelerometer and gyroscope sensors in the smartphone along with the data recorded by the OBD device (speed and rotations per minute of the engine) (see Figure 4). The GPS feature in the smartphone was enabled to track vehicles as they drive along the interstates. The GPS coordinates were also used to map the probe vehicle onto the roadway segments that constitute the spatial unit of analysis.

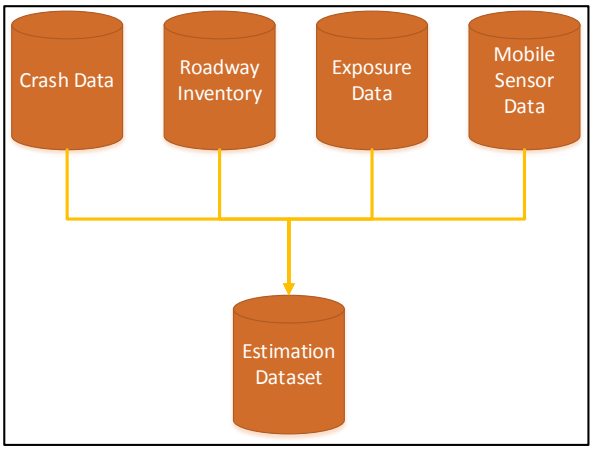

Figure 2 Data Components

Figure 3 Wavetronix Sensor Locations

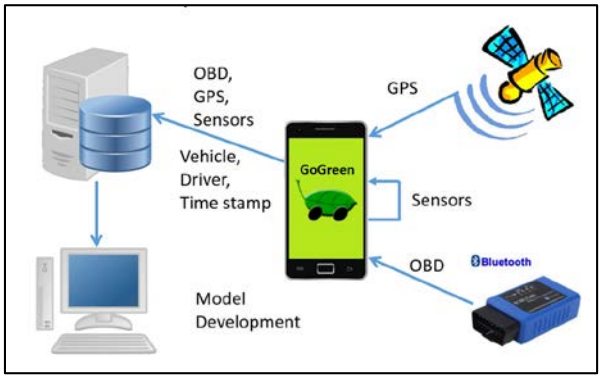

Figure 4 Mobile Sensor Data Collection

The mobile sensor data was collected by using the floating car technique in which the probe vehicle drove at a speed very close to that of the surrounding traffic in the right lane and passing slower traffic when feasible to mimic the "average" commuter. After data collection, the recorded data by the GoGreen app was downloaded to an offline server where all subsequent analysis was undertaken including computation of speed and acceleration based microscopic traffic metrics.

Statistical Model Development

Crashes are rare and random events. So, the observed crash frequency at a location can vary year to year. This is the reason safety analysts use '*expected crash frequency*' which is the long term average crash frequency as the safety metric for all policy analysis. From a statistical standpoint, crash frequency is a count variable starting at 0 and extending without a predetermined upper bound. So, linear regression techniques that deal with continuous data are not suited for modeling crash frequency data. The Poisson and Negative Binomial (NB) models are the two most commonly used count models in the literature. The Poisson model can only handle count data with equidispersion, *i.e.*, the mean is equal to the variance of count data. However, crash frequency data is typically over-dispersed, *i.e.* variance is greater than mean. In such cases, the NB model is better suited. However, recently generalized count models were developed in the literature that improves the statistical fit considerably by relaxing the restrictive assumptions of standard count models. These generalized variants considered in this research include:

NB model with Heterogeneous Dispersion (NB HD): The dispersion parameter in the NB model that accounts for over-dispersion in crash frequency data is assumed to be the same for all roadway segments. However, this is a restrictive assumption because the degree of variation about the mean can vary for different types of segments. The NB HD model relaxes this assumption by parameterizing the dispersion parameter as a function of different segment attributes (Hariharan *et al.* 2016, Narayan *et al.* 2016).

Zero-Inflated Models: There were zero crashes for 44% of the 513 roadway segments in Hampton Roads in the past one year during the evening peak period. This over-representation of zeroes in the count data is referred to as the *excess zeroes* problem. Zero-inflated models that assume two different data generation processes for count outcomes are found to be capable of accounting for the excess zeroes problem (Lord *et al.* 2005).

Generalized Ordered Response (GOR) Models: Recently, researchers have shown that generalized ordered (GOR) models subsume standard count models including Poisson and NB models as special cases. These models have an additional risk propensity component that provides more flexibility for count data modeling (Castro *et al.* 2012).

Random Parameter Heterogeneity: There may be several unobserved factors that are not controlled in crash frequency modeling. These unobserved factors can moderate the influence of explanatory variables in the model leading to random parameter heterogeneity. For example, the marginal effect of presence of shoulder on crash frequency can be -0.5 for segment 'x' and -0.7 for another segment 'y'. To capture this heterogeneity, each parameter estimate in the count model is assumed to be a realization from a normal distribution and both the mean and standard deviation of the distribution are estimated as opposed to simple mean in standard models (Mannering *et al.* 2016).

Spatial Dependency: In addition to the driving patterns along a roadway segment, the driving patterns in the neighboring segments can also influence the segment's crash risk propensity. Moreover, it is expected that the spatial dependency weakens with distance. These effects can be captured by adding spatially lagged explanatory variables to the model (Castro *et al.* 2012, Narayanamoorthy *et al.* 2013).

Count models addressing all the aforementioned issues were developed in this research. More details of these models are provided in the 'Complete Documentation" section. The models were developed incrementally at each step verifying the statistical and behavioral validity of the model results.

Post-Estimation Analysis

The statistical fit of a model may be improved by adding several variables that make little or no intuitive sense. To ensure that the improvement in model fit is statistically significant, model fit comparisons were performed using log-likelihood ratio (LR) tests for nested models and Bayesian Information Criterion (BIC) for non-nested models[*]. Also, traditional models without microscopic traffic measures were developed and compared with final models to demonstrate the utility of including microscopic measures in crash frequency modeling. Lastly, it is difficult to understand the magnitude of impact of different variables by directly looking at the parameter estimates in the count models. So, elasticity analysis that computes the percentage change in crash frequency for a unit change in explanatory variables was undertaken to quantify the impact of different factors on crash occurrences.

FINDINGS

All the estimation results are provided in the "Complete Documentation" section. A brief over of key findings from these models is provided below:

1. There is significant over-dispersion in the crash frequency data. This is also reflected in the log-likelihood (LL) improvement from -1126.2 in Poisson model to -884.84 in the NB model which is significant at any reasonable confidence level.
2. In the absence of the dispersion parameter, the Zero-Inflated Poisson (ZI Poisson) model was found to fit the data better than simple Poisson model. However, there was no evidence in support of zero-inflation after accounting for the over-dispersion in the NB model. Moreover, the NB model also fit the data better than the ZI Poisson model as indicated by their BIC values of 1825.8 and 2210.9, respectively.
3. As opposed to homogenous dispersion, roadway segments with less than or equal to two lanes were found to have greater variation in crash frequencies whereas segments with median were found to have lower variation in crash frequencies compared to all other segments. Also, the LR test statistic of comparison between the NB HD and NB models was 10.95 which is greater than the critical squared value of 5.99 for two degrees of freedom at 95% confidence level. This suggests superior data fit in the NB HD model.
4. The Generalized Ordered Response (GOR) variant of the NB model proved to be a better model that standard NB model. Two variables were found to influence risk propensity

[*] A model with lower BIC value is preferred over a model with higher BIC value.

component but not the expected count component – surface type and indicator variable for whether a segment is at an interchange or not. To be specific, roads with plant mix surface (*i.e.*, bituminous) were found to have a marginally higher crash risk propensity compared to PCC roads in a fixed effects model without random heterogeneity. Also, interchange segments have a higher risk propensity compared to other regular segments of freeways consistent with the complex driving patterns associated with these segments. The LR test statistic of comparison between the GOR and NB HD models was 5.32 which was greater than critical chi-squared value of 4.65 for two degrees of freedom at 90% confidence level.

5. There was evidence for significant random heterogeneity on the effect of surface type on risk propensity component of the GOR model. In fact, the mean effect of surface type was found to be zero in the final model with random heterogeneity. This does not mean surface type does not have any effect on crash occurrences. Instead, this implies that 50% of segments with plant mix surface have higher crash risk propensity whereas the other 50% have lower risk propensity compared to PCC roads. Moreover, the standard deviation parameter estimate of 0.8120 implies that the effect of surface type on crash risk propensity is between [-0.8120, 0.8120], [-1.624, 1.624], and [-2.436, 2.436] for 68%, 95%, and 99.7% of the segments, respectively.
6. The results also indicate evidence for the presence of spatial dependency effects of driving patterns. Specifically, acceleration patterns downstream and upstream of a segment were found to influence crash incidences on that segment. Also, the structure of spatial dependency was found to be inverse distance squared, *i.e.* the influence of a neighboring segment B on a segment A is inversely proportional to the square of the distance between the segments A and B. In fact, the LR test statistic of comparison between the final model with spatial effects and a model without spatial effects was 17.9 which is much greater than critical chi squared value of 3.84 for one degree of freedom at 95% confidence level.
7. Among all the models tested the GOR variant of the NB model with heterogeneous dispersion (HD), random parameter heterogeneity, and spatial effects was found to be the best model with the highest LL value (-874.4) and least BIC value (1817.5).
8. The final model was re-estimated by dropping all microscopic traffic measures. The LL of this restricted model without microscopic traffic measures was -926.5. The model with microscopic traffic measures has four additional parameters and LL of -874.4. The LR test statistic of comparison between these two models was 104.2 which is much greater than 9.48 which is critical chi squared value for four degrees of freedom at 95% confidence level. This demonstrates the capability of microscopic measures to improve crash frequency models
9. Higher traffic exposure levels, lower vehicle speeds, interchange segments, and greater variation in the acceleration profile along a segment were found to be associated to higher crash frequencies. On the other hand, presence of extreme positive accelerations without any extreme decelerations was associated with fewer crash occurrences. These results suggest that stop-and-go movements and lower speeds in congested conditions lead to more crashes compared to free flow conditions that are associated with higher speeds and accelerations.
10. Interestingly, none of the roadway geometry variables were found to directly affect expected crash frequency in the final model. These variables were, however, significant in Poisson model that ignores over-dispersion in the data. So, this suggests that Poisson model can over-

estimate the effect of otherwise unimportant variables to compensate for over-dispersion in the crash frequency data.

11. On average, everything else being same, a 10% increase in traffic exposure (*i.e.*, traffic volume during evening peak period) was found to increase crash frequency by 2.1%.
12. A 10% increase in speed was found to decrease the crash frequency by 5.5%. It is important to note this result does not suggest that 'speeding' reduces crashes because the model results may be interpreted only within the range of speeds observed in the estimation dataset. All that this result indicates is that there is higher likelihood of crashes in low speed congestion conditions compared to free flow conditions.
13. On average, interchange segments have 15.7% more crashes than regular segments probably indicative of the fact that complex driving patterns of exiting, merging and weaving at an interchange increase the likelihood of a crash.
14. A 10% increase in the standard deviation of acceleration along a segment increases crash frequency by about 3.4%. Importantly, a 10% increase in standard deviation of acceleration of all neighboring segments (other than the current segment) was found to increase crash frequency at the current segment by 9%. This is interesting result as it indicates that the direct effect (3.4%) is less than the spatial dependency effect (9%) and underscores the importance of accounting for spatial effects in crash frequency models.
15. Segments that only have extreme accelerations but not extreme decelerations were found to have 40% fewer crashes compared to other segments. Again this result suggests that free flow conditions are safer compared to stop-and-go movements.
16. The elasticity effects of microscopic traffic measures including speed and acceleration are higher than standard variables such as traffic exposure that are typically the only controlled variables in earlier studies. So, not only is it important to include these variables in the crash frequency models from a statistical fit standpoint but also from a policy standpoint.

CONCLUSIONS

Microscopic traffic measures describing speed and acceleration patterns enhance the crash frequencies models considerably. In fact, not only does the statistical fit improve but also the magnitude of elasticity effects of these microscopic measures was found to be larger than standard variables such as traffic exposure. Moreover, the results also suggest strong spatial dependency effects whereby the crash risk on a segment depends on the acceleration patterns of segments in close proximity in addition to acceleration patterns on the same segment. Also, recent advancements in the count modeling literature including Heterogeneous Dispersion (HD) and Generalized Ordered Response (GOR) modeling methods are better suited for analyzing crash frequency data compared to standard Poisson and Negative Binomial (NB) models.

RECOMMENDATIONS

Based on the study findings, the following two recommendations are provided for safety engineers and policy makers:

1. Data from mobile sensors (e.g., smartphones) enables monitoring vehicle dynamics and traffic flow at high resolution (e.g., second-by-second). This data can be used to develop microscopic traffic measures that serve as better indicators of actual driving patterns. There are

opportunities to expand the mobile sensor data collection on a larger scale (e.g., statewide or nationwide) by partnering with probe data providers (e.g., INRIX, HERE). Currently, most of the Safety Performance Functions (SPFs) are only sensitive to aggregate variables such as traffic exposure and geometric attributes (e.g.: presence of shoulder). Federal safety agencies and state DOTs would benefit immensely by updating these SPFs using microscopic traffic measures so that the predictions are more accurate and these models can also be used to evaluate countermeasures that primarily affect driving behavior (*e.g.,* variable speed limits).

2. Currently, most of the SPFs in the HSM are either Poisson or Negative Binomial models. However, there is considerable scope for improving these models without adding significant computational complexity. Specifically, the Heterogeneous Dispersion (HD) and Generalized Ordered Response (GOR) variants of the NB model are relatively easy to estimate and were found to improve the statistical fit significantly. To start-off, state DOTs can test and evaluate the relative merits (better prediction accuracy and policy sensitive) and difficulties (data collection) of these models for specific locations (e.g., freeways, intersections *etc*.).

COMPLETE DOCUMENTATION

This section provides a detailed overview of all the work undertaken in this project including literature synthesis, data collection process, data description, mathematical formulation of different models, estimation results, and post-estimation analysis.

Past Literature

Crashes are rare and random events. So, the number of observed crashes at any given location can fluctuate year-to-year even if all the observable crash causation conditions remain the same between the two years. If the observed crash frequency is very high in one year, then it is more likely to be followed by relatively lower crash frequency in the next year, and vice-versa. This effect is referred to as the 'Regression-To-Mean Bias'. This inherent variation in observed crash frequency poses a challenge to evaluating the effectiveness of different safety countermeasures. For instance, it is unclear if the reduction (or increase) in crash occurrences is due to random fluctuation or the safety countermeasure. To address this problem, safety analysts rely on estimates of the long term average crash frequency, also referred to as 'Expected Crash Frequency', as a proxy for crash risk. The observed crash frequency across several locations is used to statistically estimate the expected crash frequency. Expected crash frequency modelling is a reliable method for determining the safety of a segment of roadway.

Previous studies have looked at explanatory variables primarily in two categories, physical characteristics of the location (e.g., roadway or intersection) and aggregate traffic characteristics at that location (e.g., AADT, % of left turning traffic, % of heavy vehicle traffic *etc.*) (Shankar *et al.* 1997, Qin *et al.* 2005, Lord and Mannering 2010). A majority of these early studies focused on physical characteristics of the roadway due to a lack of consistent and accurate data collection means (Ogle 2005). Unfortunately, data such as this is unable to capture the actual driving patterns (*i.e.,* flow and movement of individual vehicles). It is difficult to develop an accurate representation of expected crash frequencies when the characteristics of the actual vehicles travelling the corridor are not considered. For instance, the overall congested crash rate in the state of Indiana is 24.1 times greater than the uncongested crash rate (Mekker *et al.* 2016). In addition

to higher traffic volumes, there are most likely unique driving patterns that contributed to high crash rates during congested period. Simple aggregate measures (average daily traffic and truck volumes) cannot capture these differences between congested and uncongested conditions.

A potential source for speed data could be crash reports that were completed at the scene of an accident by the police. This would appear to be a simple way to obtain a piece of driving patterns. But, obtaining speed from a police crash report is not recommended because the police may be under a lot of stress during incident investigations and may not be able to accurately determine the speed at which the driver was going. Also, the driver may underreport the estimated speed which they were travelling in an attempt to lessen the likelihood of receiving additional infractions for an incident. Alternatively, several researchers have used speed limit as a proxy for traffic speed. Probe vehicle data, on the other hand, can be used to capture the speed and acceleration profiles that serve as reliable indictors of congested traffic conditions. Some of the previous studies have relied on simulation models to capture naturalistic driving data regarding the movements of the actual vehicle itself through space and time (Gettman and Head 2003). This method of data collection allows the researcher to control for every aspect of the simulation while being able to alter the simulation to fit different scenarios. Multiple simulation inputs may be evaluated in a short period of time to get the most accurate results. A limitation of these methods, however, is that it is based on simulation and not driving behavior in reality.

Recent studies have focused on obtaining and using data collected directly in the field to develop more accurate crash frequency models. GPS sensors and OBD devices are now regularly used in transportation research to obtain the aforementioned naturalistic driving behavior data (Ogle 2005, Jun 2006). Another option when considering probe vehicle data is using data that is crowd-sourced, collected, and combined into a dataset by a third party source (Mekker *et al.* 2016). This data source has the benefit of allowing the researchers to have a more robust dataset that encompasses a greater length of time. The data can be collected and stored for multiple years rather than only being available for the duration of research period. This allows the researcher to have access to probe vehicle data that was collected around the time that actual accidents occurred. (Wåhlberg 2004) looked at the acceleration profiles of busses as a potential indicator of crash frequency and study concluded that driver acceleration behavior could be used as a predictor of accidents. But, due to some discrepancies between samples it was difficult to determine the validity of this finding. Also, in this study, the acceleration data was recorded on-board using a g-analyst which measured the acceleration at 10 Hertz to 100th of 1g (9.81 m/s2) accuracy. This tool did not measure the acceleration from the vehicle directly but, simply measured the g-force felt by the bus's start and stop motions. This may have resulted in errors due to the vehicle not producing the data itself.

To summarize, past literature highlighted the importance of considering microscopic traffic measures in crash prediction models but there is relatively little research in this area primarily due to the challenges associated with data collection. The current study addresses this limitation by using a smartphone application that can record data collected by mobile sensors including GPS, accelerometer, and gyroscope. The next section describes the data collection effort and provides a description of the final dataset.

Data Assembly and Description

Crash Database

Vehicle crash data was obtained through a Virginia Department of Transportation (VDOT) database that includes all reported crashes from October 2014 to October 2015 for the entire Hampton Roads Region. There were many records in the database which were affiliated with disabled vehicles. These records were omitted because the study is only interested in actual vehicle crashes. This raw data contained 111 characteristics for each crash. Some of this information is administrative in nature such as who recorded the crash, how it was recorded, and who last modified the report; these variables were not beneficial in the analysis. The database also recorded the type of crash (vehicle accident, multi-vehicle accident, or tractor trailer accident) and time impact severity of the crash (< 30 min., 30 min. to 2 hours, or > 2 hours). However, this study focused only on total crash frequency. So, these variables were not used in the analysis. One variable of particular importance in the crash database was the location (*i.e.,* latitude and longitude) of crash occurrence. The location of the crash was used to overlay the crash data onto the transportation network. Next, each crash was geocoded to one of the roadway segments (*i.e.*, spatial unit of analysis). Lastly, all crash occurrences on each roadway segment in the past year were aggregated to obtain the crash frequency that serves as the dependent variable of analysis.

Identifying Spatial Unit of Analysis

One of the first steps to crash frequency modeling is selecting the spatial unit of analysis, *i.e.* the geographical extent of region over which the expected crash frequency is modeled. The current study focusses on crash frequency along major interstates in Hampton Roads. So, the empirical context implies that the interstates must be split into smaller segments that constitute the unit of analysis. However, this decision cannot be made arbitrarily because the availability of roadway inventory data and the homogeneity of resulting segments are critical to developing an accurate crash frequency model. So, several segment definitions were explored prior to choosing the spatial unit of analysis. For instance, the easiest and straightforward segment definition is uniform one-mile segments starting from the first mile marker of each interstate. However, such segmentation can result in non-homogenous segments, *i.e.* the roadway geometric characteristics and traffic conditions can vary considerably within each segment. For instance, a portion of the one mile stretch may correspond to the freeway portion and the remaining portion corresponds to ramp area. Another alternative was the publicly available Census Bureau's TIGER (Topologically Integrated Geographic Encoding and Referencing) database that divides each roadway into a contiguous stretch of several smaller segments. It is important to note that these segments are homogenous but not uniform. However, one of the limitations of using the TIGER segments was unavailability of extensive roadway inventory data. Barring a few important variables such as number of lanes and segment length, other key attributes such as shoulder and median presence were missing. The third alternative was using the segment definition in the VDOT's roadway inventory database that provided detailed information characterizing each segment. However, just as the TIGER segments, the VDOT segments were also not uniform. Based on the relative merits of the three segment definitions (uniform, TIGER, and VDOT), this study adopted the VDOT segment definition as the spatial unit of analysis.

Identifying the Time Period of Analysis

Weekend crashes were omitted due to travel patterns being inconsistent with other travel days. Also, a histogram of the crash data, seen in Figure 5, shows that nearly 18% of all crashes in the past year occurred during the two hour PM peak period. This observation coupled with the constraint that it is not feasible to collect probe vehicle data using smartphones along all interstates during all hours of the day, the two hour time period between 4 and 6 pm was chosen as the temporal unit of analysis. Figure 6 displays the frequency distribution for the number of crashes per segment during the two hour PM peak window. It can be seen that there were no crashes in the entire year during the evening peak period in nearly 44% of the segments. At the same time, there are several segments with more than crash during the same time period. The average number of crashes per segment was 2.16 and the variance across all segments was 14.6. This preliminary descriptive analysis suggests over-dispersion in the dataset.

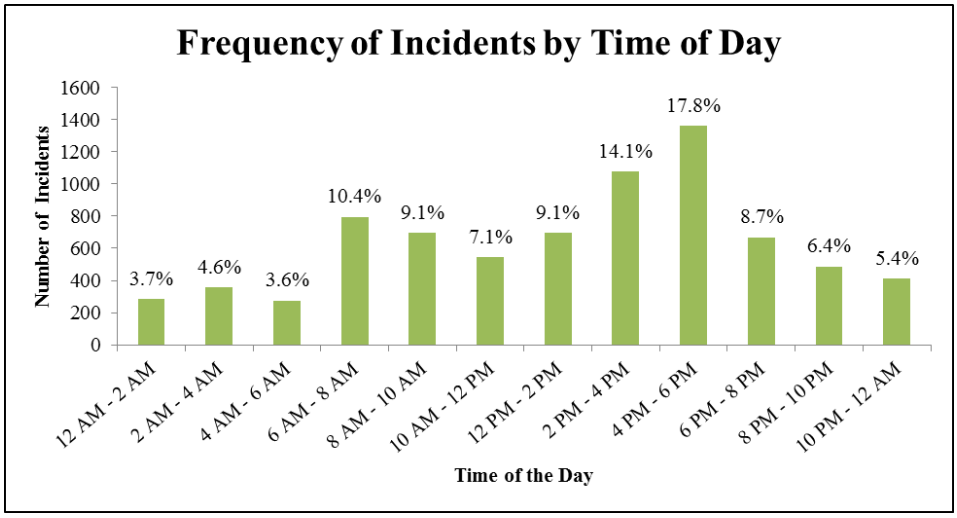

Figure 5 Total Incidents by Time-of-Day

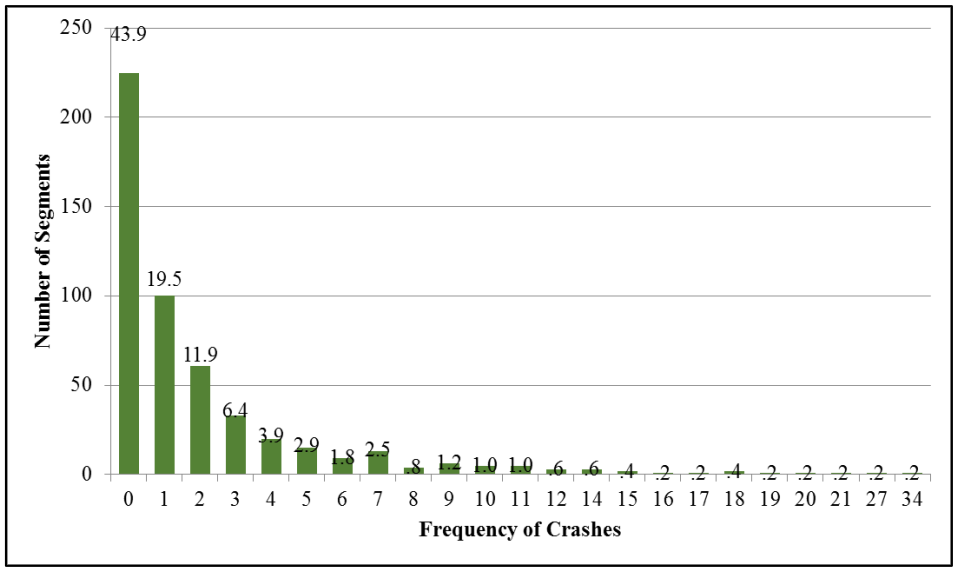

Figure 6 Frequency of Crashes per Roadway Segment

Traffic Exposure Data

Traffic exposure refers to the average traffic volume during the two hour peak period during weekdays in the past one year. This traffic volume was obtained from 222 Wavetronix sensors that are maintained by VDOT. The 222 sensors were distributed among 513 segments. There were some segments with multiple sensors as well as segments without any sensors. In cases where multiple sensors were located along a segment, the average of traffic counts across the corresponding sensors was used as the exposure variable. In cases where there were no sensors within the segment, the traffic counts for nearest segment were used as representative exposure variables. Table 1 displays the mean, 5^{th} percentile, 95^{th} percentile, and standard deviation for the traffic exposure variable used in the final dataset.

Table 1: Traffic Exposure Data

Continuous Variable	Units	Mean	5th Percentile	95th Percentile	Standard Deviation
Average Annual Weekday Peak Period Traffic	Vehicles	7,323.37	2,104.00	12,413.10	7,431.37

Figure 7 Segments with Multiple and No Sensors

Roadway Inventory Data

The roadway inventory information was obtained from a VDOT maintained database that contains information regarding to the physical characteristics of the roadway. The length of each segment was recorded to account for varying lengths between segments. The number of lanes varied from one lane to five lanes. This variable was broken down into three separate categories: less than or equal to two lanes, three lanes, and greater than or equal to 4 lanes. The next variable used from this dataset was the surface type. This category was only broken down into two types within the roadway segments considered: plant mix and Portland Cement Concrete (PCC). The plant mix category is a typical asphalt roadway and PCC is a concrete surface. Surface width was taken from the database and broken into three categories: less than or equal to 24', 24' to 48', and greater than or equal to 48'. The presence of shoulder on both the right and left side of the roadway was also included in this database and recorded for analysis. If a shoulder is present, the width of the shoulder was also recorded and broken into three categories: Less than or equal to 8', 8'-12', and greater than or equal to 12'. The database provided information as to whether or not the roadway segment was a high occupancy vehicle (HOV) lane or a regular lane. Along with HOV lanes, the database considered whether or not the lane was a reversible lane. Median presence was also

considered and if there was a median, its type and size was considered. Types of median were split between grass median and a combination of positive barrier and curbed median for the analysis. The width of these medians was also considered. This category was broken into median widths which are less than 20', widths that are greater than or equal to 20' and less than or equal to 40', and widths that are greater than 40'. The final variable considered from the roadway inventory was the type of facility. This variable was broken into two categories: two-way divided roadway and a combination of roadways which were one-way, and two-way non-divided roadways. Table 2 and Table 3 provide an overview of all of the previously mentioned roadway inventory explanatory variables, along with their frequency and percentage distributions used in the final dataset.

Table 2: Categorical Roadway Inventory Data

Explanatory Variable	Frequency	Percentage
Number of Lanes		
Less Than or Equal To 2	259	50.4%
3	143	27.8%
Greater Than or Equal to 4	111	21.6%
Total	513	100.0%
Surface Width		
Less Than or Equal To 24'	255	49.7%
24'-48'	147	28.6%
Greater Than or Equal to 48'	111	21.6%
Total	513	100.0%
Surface Type		
Plant Mix	231	45.0%
Portland Cement Concrete (PCC)	282	55.0%
Total	513	100.0%
Presence of Right Shoulder		
Shoulder Present	288	56.1%
No Shoulder	225	43.9%
Total	513	100.0%
Right Shoulder Width		
Less Than or Equal To 8'	242	47.1%
8'-12'	265	51.6%
Greater Than or Equal to 12'	6	1.1%
Total	513	100.0%
Presence of Left Shoulder		
Shoulder Present	298	58.0%
No Shoulder	215	42.0%
Total	513	100.0%
Left Shoulder Width		
Less Than or Equal To 8'	287	55.9%
8'-12'	220	42.8%

Explanatory Variable	Frequency	Percentage
Greater Than or Equal to 12'	6	1.1%
Total	513	100.0%
HOV Lane		
Lane is an HOV Lane	36	7.0%
Lane is not an HOV Lane	477	93.0%
Total	513	100.0%
Reversible Lane		
Lane is Reversible	4	0.8%
Lane is Non-Reversible	509	99.2%
Total	513	100.0%
Median Type		
Grass/Unprotected	205	39.9%
Positive Barrier or Curbed	148	28.8%
No Median	160	31.1%
Total	513	100.0%
Median Width Minimum		
Less than 20'	420	81.8%
Greater Than or Equal to 20' and Less Than or Equal to 40'	8	1.5%
Greater than 40'	85	16.5%
Total	513	100.0%
Facility Type		
One Way or Two-Way Non-Divided	70	13.6%
Two-way Divided	443	86.4%
Total	513	100.0%

Table 3: Continuous Roadway Inventory Data

Continuous Variable	Units	Mean	5th Percentile	95th Percentile	Standard Deviation
Segment Length	Miles	0.47	0.05	1.56	0.57

Mobile Sensor Data Collection

Mobile sensor data was collected by driving vehicles equipped with cellular devices that were linked to an OBD device through Bluetooth. The OBD device interfaces with the computer system within the vehicle itself. This device records information such as the velocity of the vehicle and the rotations per minute of the engine. The cellular device runs an Android application named 'GoGreen' that has the capability of recording data from sensors located inside the phone along with the data recoded by the OBD device. The vehicles equipped with the data collection system were driven on interstate roadways within Hampton Roads during the 4pm to 6 pm time period. The data was collected by using the car following technique in which the probe vehicle drove at a speed very close the surrounding traffic in the right hand lane and passing slower traffic when feasible to mimic the "average" commuter. The GPS feature in the smartphone was also enabled

to track vehicles as they drive along the interstates and map the probe vehicle onto the roadway segments that constitute the spatial unit of analysis. Figure 8 shows the distribution of probe vehicle trips across segments. On average, 11 probe vehicle trips were made per roadway segment to collect mobile sensor data. Also, at least five probe vehicle trips were made along nearly 90% of the segments. Next, several metrics were calculated for each segment using the mobile sensor data to capture driving behavior.

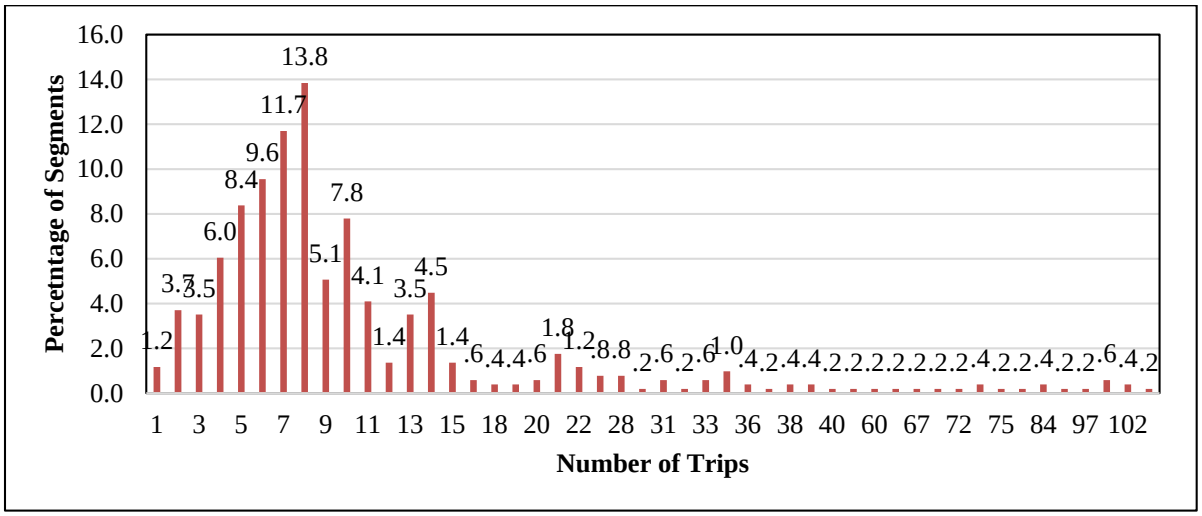

Figure 8 Number of Probe-Vehicle Trips per Segment

First, the mean traffic speed for each segment was obtained by averaging the speed recordings for all trips contained within that single segment. Both linear and non-linear speed effects were tested during model estimation. For the non-linear effects, traffic speed was broken down into three distinct categories: <45 mph, >= 45 mph and <60 mph, and >=60 mph. Next, speed data taken from the OBD device was used in order to calculate acceleration values for the model. The OBD device recorded speed values at one second frequency. The difference between two consecutive velocity readings over a one second time period was considered the acceleration for that data point. This acceleration value was then converted to feet per second^2 for analysis. The acceleration data was also divided into two separate categories: accelerations and decelerations. All positive acceleration recordings were considered as 'accelerations' and all negative acceleration recordings were considered as 'decelerations'. Average, minimum, maximum, and standard deviation metrics of all acceleration and deceleration recordings were computed for each segment. To capture extreme driving patterns, two additional metrics were calculated. The 5th and 95th percentile accelerations were calculated and if a segment had a deceleration recording below the 5th percentile it was considered to have an extreme deceleration. Similarly, if a segment had acceleration recording above the 95th percentile it was considered to have an extreme acceleration. The average speed was also calculated from probe vehicle recordings and included in this database. Table 4 displays the mean, 5th percentile, 95th percentile, and standard deviation for continuous microscopic traffic measures computed. While estimating the count models, care was taken so that highly correlated continuous acceleration metrics (see Table 5) from the mobile sensor data were not used simultaneously in the model. Table 6 displays the frequency and percentage distributions of the categorical traffic measures computed.

Table 4: Continuous Metrics Computed Using Mobile Sensor Data

Continuous Metric	Units	Mean	5th Percentile	95th Percentile	Standard Deviation
Maximum Deceleration (MAXDEC)	ft/sec^2	-4.15	-9.35	-0.75	2.81
Maximum Acceleration (MAXACC)	ft/sec^2	3.29	0.73	8.26	2.33
Average across all Acceleration and deceleration recordings (AVGACC)	ft/sec^2	0.72	0.26	1.43	0.39
Standard Deviation across all acceleration and Deceleration recordings (SDACC)	ft/sec^2	1.05	0.42	2.01	0.52
Average Speed (AVGSPD)	mph	46.76	18.80	62.92	13.23

Table 5 Correlation Among Acceleration Metrics

	MAXDEC	MAXACC	TAVGACC	TSDACC
MAXDEC	1.000	-.636	-.618	-.757
MAXACC	-.636	1.000	.535	.665
AVGACC	-.618	.535	1.000	.931
SDACC	-.757	.665	.931	1.000

Table 6: Categorical Metrics Computed using Mobile Sensor Data

Categorical Metrics	Frequency	Percentage
Presence of Extreme Accelerations		
Yes	197	38.4%
No	316	61.6%
Total	513	100.0%
Presence of Extreme Decelerations		
Yes	166	32.4%
No	347	67.6%
Total	513	100.0%
Presence of Extreme Accelerations without Extreme Decelerations		
Yes	49	9.6.0%
No	464	90.4%
Total	513	100.0%
Presence of Extreme Decelerations without Extreme Accelerations		
Yes	80	15.6%
No	433	84.4%
Total	513	100.0%
Presence of Both Extreme Accelerations And Extreme Decelerations		
Yes	267	52.0%
No	246	48.0%

Categorical Metrics	Frequency	Percentage
Total	513	100.0%
Average Speed		
<45 mph	193	37.6%
>=45 mph and <60 mph	235	45.8%
>=60 mph	85	16.5%
Total	513	100.0%

Spatial Weight Matrix

One hypothesis that this research intends to test is the presence of significant spatial dependency of driving patterns across segments. To do this, a network distance matrix $\boldsymbol{D}$ was created in which each cell was populated with the distance along the roadway network between the midpoints of two segments in the corresponding row and column. The size of this matrix is 513×513 because there are a total of 513 segments in the dataset. Distance between roadway segments in opposite directions as well as distance two separate roads was coded as infinity (some large number) because it is unlikely that there will be any interactions between traffic on these roads. Next, different spatial weight $\boldsymbol{W}$ matrices were computed by populating cell elements with different distance-based measures including distance, inverse distance squared, and inverse distance cubed. The diagonal entries of these spatial weight matrices are recoded as 0. Next, each row of the spatial weight matrix was normalized by dividing its cell elements by the sum of all entries in that row. The distance measure used in the spatial weight matrix calculation controls the structure of spatial dependency. For instance, an inverse distance structure indicates stronger spatial dependency compared to inverse distance squared structure. If we denote each cell element in row *s* and column *s'* of the spatial weight matrix by $w_{s,s'}$, then $\sum_{s'=1}^{513} w_{ss'} = 1 \; and \; w_{ss} = 0 \; \forall \; s \in [1,513]$.

Interchange Segments

Segments that are completely within an interchange are expected to have more crashes because of complex traffic movements including exiting, merging, diverging, and weaving. To quantify this additional risk associated with interchange segments, an indicator variable for whether a segment is at an interchange or not was created. For instance, the segments marked in red color in Figure 8 are identified as interchange segments. Please note that these interchange segments are different from ramp segments which were excluded from this analysis.

Figure 9 Interchange Segments

Methodological Framework

A brief discussion of modeling methods follows. Let *s* be the index for the roadway segment.

Poisson Model

Assuming that crash data are realizations from a Poisson distribution, the probability of observing a count outcome y_s conditional on the expected mean parameter $\lambda_s$is given by:

$$P(Y = y_s) = \frac{e^{-\lambda_s} \times {\lambda_s}^{y_s}}{y_s!} \quad \text{Equation (1)}$$

As indicated earlier, the Poisson model has the *equidispersion* property which implies that the variance of the Poisson distribution is equal to the expected mean parameter λ_s. So, to ensure that the λ_s parameter is always greater than 0 during model estimation, it is parameterized as $e^{LOG(\lambda_s)}$ and $LOG(\lambda_s)$ is specified as a linear function of different exogenous variables as follows: $LOG(\lambda_s) = \boldsymbol{\beta}'\boldsymbol{X}_s$ where $\boldsymbol{X}_s$ is the vector of exogenous variables and $\boldsymbol{\beta}$ is the corresponding vector of coefficients that were estimated using maximum likelihood approach.

Negative Binomial Model

In the NB model, the probability of observing count outcome y_s conditional on the expected mean parameter λ_s and dispersion parameter $r > 0$ is given by:

$$P(Y = y_s) = \left(\frac{r}{r + \lambda_s}\right)^r \times \frac{\Gamma(r + y_s)}{\Gamma(y_s + 1)\Gamma(r)} \times \left(\frac{\lambda_s}{r + \lambda_s}\right)^{y_s} \quad \text{Equation (2)}$$

Where Γ is the gamma function defined as follows:

$$\Gamma(t) = \begin{cases} \int_{x=0}^{\infty} x^{t-1}e^{-x}dx \text{ for positive non} - \text{integer } t \\ (t-1)! \text{ for positive integer } t \end{cases} \quad \text{Equation (3)}$$

The expected mean of the NB model is λ_s whereas the variance is $\lambda_s + \frac{\lambda_s^2}{r}$ making the model particularly suited for handling over-dispersion. In the NB model, the dispersion parameter *r* must also be estimated in addition to the β parameters in the $LOG(\lambda_s)$ specification.

Negative Binomial Model with Heterogeneous Dispersion

Standard negative binomial (NB) model assumes that the dispersion parameter is the same for all segments. However, this is a restrictive assumption because crashes along different groups of segments can have varying degrees of variance in crash occurrences. Recently, NB models with heterogeneous dispersion were developed for modeling crash frequency (Hariharan *et al.* 2016, Narayan *et al.* 2016). This model parameterizes the over-dispersion parameter in the NB model component as follows: $r_s = e^{\boldsymbol{\delta}'\boldsymbol{W}_s}$, where $\boldsymbol{W_s}$ is the vector of segment characteristics. This model will be referred to as 'NB HD' model in this report.

Zero-Inflated Modeling Framework

In the Zero-Inflated (ZI) framework, the data is assumed to be generated from two different states – a zero count state and a normal count process state. So, zero crashes in any given segment can result either because of the zero state or the normal count state resulting in over-representation of zeroes in the crash dataset (Lord *et al.* 2005).

Zero-Inflated Poisson (ZI Poisson) model

$$P(Y = 0) = ZIP[0] + ZIP[1]e^{-\lambda_s}$$

$$P(Y = y_s) = ZIP[1]\frac{e^{-\lambda_s} \times {\lambda_s}^{y_s}}{y_s!} \ \forall\ y_s > 1 \qquad \text{Equation (4)}$$

Zero Inflated Negative Binomial (ZI NB) model

$$P(Y = 0) = ZIP[0] + ZIP[1]\left(\frac{r}{r+\lambda}\right)^r$$

$$P(Y = y_s) = ZIP[1]\left(\frac{r}{r+\lambda_s}\right)^r \frac{\Gamma(r+y_s)}{\Gamma(y_s+1)\Gamma(r)}\left(\frac{\lambda_s}{r+\lambda_s}\right)^{y_s} \ \forall\ y_s > 1 \qquad \text{Equation (5)}$$

$ZIP[0]$ and $ZIP[1]$ are the probabilities associated with the zero inflation component and

$ZI[0] + \ ZI[1] = 1 \ and \ ZI[0] = \frac{e^{\boldsymbol{\pi}' \boldsymbol{D}_S}}{1+e^{\boldsymbol{\pi}' \boldsymbol{D}_S}}$ where $\boldsymbol{D}_s$ is the vector of segment characteristics and $\boldsymbol{\pi}$ is the corresponding vector of coefficients.

Generalized Ordered Response Probit (GORP) Framework

In the GORP framework, a latent risk propensity y_s^* is mapped into observed count outcomes y_s by threshold parameters ψ_s^k where k is the index for all possible count outcomes. Assuming specific functional forms for these threshold parameters will result in the GORP framework replicating standard count models. The latent risk propensity y_s^* in the standard ordered response framework can be written as:

$$y_s^* = \boldsymbol{\gamma}' \boldsymbol{Z}_s + \varepsilon_s \qquad \text{Equation (6)}$$

Where $\boldsymbol{Z}_s$ is a vector of all exogenous variables and $\boldsymbol{\gamma}$ is the corresponding vector of coefficients; ε_s is the stochastic error term that represents all unobserved factors (not captured in the exogenous variables) that can impact y_s^* and is assumed to be an independent realization from a standard normal distribution, *i.e.*, $\varepsilon_s \sim N(0,1)$. In the GORP framework, the probability that the observed outcome is y_s is given by:

$$P(Y = y_s) = P\left(\psi_s^{y_s-1} < y_s^* < \psi_s^{y_s}\right) = P\left(\psi_s^{y_s-1} < \boldsymbol{\gamma}' \boldsymbol{Z}_s + \varepsilon < \psi_s^{y_s}\right) \qquad \text{Equation (6a)}$$

$$= P\left(\psi_s^{y_s-1} - \boldsymbol{\gamma}' \boldsymbol{Z}_s < \varepsilon < \psi_s^{y_s} - \boldsymbol{\gamma}' \boldsymbol{Z}_s\right)$$

$$= \Phi\left(\psi_s^{y_s-1} - \boldsymbol{\gamma}' \boldsymbol{Z}_s\right) - \Phi\left(\psi_s^{y_s} - \boldsymbol{\gamma}' \boldsymbol{Z}_s\right) \qquad \text{Equation (6b)}$$

Where $\Phi(.)$ is the cumulative distribution function of standard normal random variable

Standard count models including the Poisson and NB models can be obtained by imposing certain constraints on the GORP model, *i.e.*, the implied probability expressions for different count outcomes would be identical for the GORP (Equation (5) and standard count models (Equations (3) and (4)). To see this, consider the constraints and functional forms imposed on ψ_s^k parameters below:

Generalized Poisson Model (GORP Poisson)

$$\psi_s^k = \Phi^{-1}\left(\sum_{p=0}^{k} \frac{e^{-\lambda_s} \times {\lambda_s}^p}{p!}\right) + \alpha_k \; \forall \; k \geq 0 \qquad \text{Equation (7)}$$

If (1) ψ_s^k is parameterized as shown in Equation (6), (2) all γ parameters in the propensity equation are equal to 0, and (3) all α_k parameters are equal to 0, then the GORP model collapses to the standard Poisson model.

Generalized Negative Binomial Model (GORP NB)

$$\psi_s^k = \Phi^{-1}\left(\sum_{p=0}^{k} \left(\frac{r}{r+\lambda_s}\right)^r \times \frac{\Gamma(r+p)}{\Gamma(p+1)\Gamma(r)} \times \left(\frac{\lambda_s}{r+\lambda_s}\right)^p\right) + \alpha_k \; \forall \; k \geq 0 \qquad \text{Equation (8)}$$

If (1) ψ_s^k is parameterized as shown in Equation (8), (2) all γ parameters in the propensity equation are equal to 0, and (3) all α_k parameters are equal to 0, then the GORP model collapses to the standard NB model.

Although theoretically one could estimate one α_k parameter specific to each count outcome *k*, from a practical standpoint, α_k can be fixed as α_K where *K* is a pre-determined count outcome depending on the empirical context, *i.e.*, $\alpha_k = \alpha_K \; \forall \; k \geq K$. Also, the α_k parameters control for any additional probability mass that is not captured by the parameters in the λ_s and y_s^* specifications. So, the GORP versions of Poisson and NB models can easily handle over or under-representation of multiple count outcomes without necessitating a hurdle or inflated model set-up (Castro *et al.* 2012). In the GORP versions of Poisson and NB models, the analyst must also estimate the γ parameters in propensity y_s^* and the α_k parameters in thresholds ψ_k in addition to the ψ_s^k parameters in $LOG(\lambda_s)$ specification and dispersion parameter *r* (in case of NB models).

<u>Generalized Ordered Response Probit (GORP) Framework with Random Heterogeneity</u>

This model is also referred to as the mixed GORP (MGORP) model in the literature. There may be several unobserved factors that influence crash occurrences along a segment. These unobserved factors can moderate the influence of different exogenous variables considered in our study. This can lead to unobserved heterogeneity in the parameter estimates both in the expected mean λ_s specification as well as latent propensity. Ignoring the presence of this random heterogeneity can lead to biased parameter estimates. The fixed parameters GORP framework can be extended to capture random heterogeneity by allowing random variation in the parameters in λ_s and y_s^* as follows:

$LOG(\lambda_s) = \boldsymbol{\beta_s}'\boldsymbol{X_s}$ where $\boldsymbol{X_s}$ is the vector of exogenous variables and $\boldsymbol{\beta_s}$ is the corresponding vector of coefficients for segment *s*. $\boldsymbol{\beta_s}$ is assumed to be a random realization from a multivariate normal distribution with mean $\boldsymbol{\beta}$ and variance $\boldsymbol{\Omega}$.

$y_s^* = \boldsymbol{\gamma_s}'\boldsymbol{Z_s} + \varepsilon_s$ where $\boldsymbol{X_s}$ is the vector of exogenous variables and $\boldsymbol{\gamma_s}$ is the corresponding vector of coefficients for segment *s*. $\boldsymbol{\gamma_s}$ is assumed to be a random realization from a multivariate normal distribution with mean $\boldsymbol{\gamma}$ and variance $\boldsymbol{\Sigma}$.

The analyst must also estimate the elements of the $\boldsymbol{\Omega}$ and $\boldsymbol{\Sigma}$ in addition to the parameters in the fixed parameters GORP models. The resulting model was estimated using the maximum simulated likelihood (MSL) approach using 200 Halton draws (Bhat 2003). However, if there is random parameter heterogeneity only in the risk propensity component y_s^*, then the model can be estimated using traditional maximum likelihood inference method without simulation.

Spatial Effects

Spatially weighted explanatory variables can be added to both in the λ_s and y_s^* components of the model as follows:

$$LOG(\lambda_s) = \boldsymbol{\beta_s}'\boldsymbol{X_s} + \widetilde{\boldsymbol{\beta_s}}' \sum_{s'=1}^{513} w_{ss'}\boldsymbol{X_s}$$

$$y_s^* = \boldsymbol{\gamma_s}'\boldsymbol{Z_s} + \widetilde{\boldsymbol{\gamma_s}}' \sum_{s'=1}^{513} w_{ss'}\boldsymbol{Z_s} + \varepsilon_s \qquad \text{Equation (9)}$$

where $w_{s,s'}$ is an element of the spatial weight matrix $\boldsymbol{W}$ and $\sum_{s'=1}^{513} w_{ss'} = 1 \; \forall \; s \in [1,513]$. The parameters $\widetilde{\beta_s}$ and $\widetilde{\gamma_s}$ are coefficients on spatially lagged explanatory variables that must be estimated in addition to all the other parameters.

Estimation Results

Table 7 presents the results of NB model and its variants. The results of all other intermediate models are shown in the Appendix. For brevity, the results of only the best model (last column of the table), *i.e.* GOR variant of NB model with heterogeneous dispersion, random parameter heterogeneity, and spatial dependency effects are discussed in this report.

Positive signs on parameter estimates in the expected count component λ_s indicate that higher levels of traffic exposure and greater variation in acceleration patterns both in the current and proximal segments (due to spatial effects) are associated with higher average crash frequencies. Similarly, negative parameter estimates in λ_s suggest that higher speeds and presence of extreme accelerations without any extreme deceleration are associated with lower crash occurrences. The parameter estimate on LOG(Segment Length) was fixed to one because of one to one relationship between crash frequency and segment length. *i.e.*, everything else remaining the same, increasing the segment length by 10% must increase crash frequency also by 10%.

The dispersion parameter r_s was found to vary across segments. Specifically , roadway segments with less or than or equal to two lanes were found to have larger variation in crash frequency whereas segments with median were found to have lower variation in crash frequencies compared to all other segments. Please note that the dispersion parameter r_s is inversely proportional to variance and the interpretation is in the opposite direction of the parameter sign.

Table 7 Estimation Results of Negative Binomial Model and its Variants

Variables	***NB***	***NB HD***	***GORP NB HD***	***MGORP NB HD with Spatial Effects***
	Coefficient (t-stat)	*Coefficient (t-stat)*	*Coefficient (t-stat)*	*Coefficient (t-stat)*
Expected Count $LOG(\lambda_s)$				
Constant	-1.2145 (-1.16)	-1.198 (-1.15)	-1.7233 (-1.63)	-1.6183 (-1.55)
LOG(Traffic Exposure)	0.2602 (2.21)	0.2497 (2.12)	0.3119 (2.56)	0.2967 (2.45)
Average speed (mph)	-0.016 (-3.10)	-0.0161 (-2.99)	-0.0191 (-3.39)	-0.0200 (-3.62)
SDACC (ft/sec^2)	0.3111 (1.84)	0.3759 (2.11)	0.3265 (1.79)	0.3332 (1.92)
Spatially Weighted SDACC (ft/sec^2)	0.8055 (3.84)	0.8334 (3.80)	0.8966 (4.00)	0.9198 (4.22)
Presence of extreme accel. w/o extreme decel.	-0.6916 (-2.78)	-0.6399 (-2.56)	-0.6468 (-2.54)	-0.6286 (-2.48)
Intersection segment	0.2696 (2.16)	0.217 (1.73)		
LOG(Segment Length)	1.0000	1.0000	1.0000	1.0000
Dispersion $LOG(r_s)$				
Constant		-0.262 (-1.17)	-0.4284 (-1.77)	0.0000[†]
Lanes ≤ 2		-0.5414 (-2.18)	-0.577 (-2.22)	-0.4100 (-1.49)
Presence of median		0.6339 (2.63)	0.7371 (2.88)	0.4715 (2.33)
Propensity				
Plant mix roadway surface			0.2092 (2.09)	0.0000[†]
Standard Deviation				0.8120 (5.05)
Intersection segment			0.1793 (1.89)	0.1994 (1.93)
Number of Observations	513	513	513	513
Number of Parameters Estimated	9	10	12	12
Log-Likelihood at convergence	-884.8	-879.4	-876.7	-874.4

The propensity component of the final model showed some interesting results. The estimates corresponding to the surface type variable are a mean effect of 0 and a standard deviation of 0.8120. This indicates that although roadway segments with plant mix surface, on average across

[†] This parameter was fixed to zero because it turned out to be statistically equal to zero.

all segments, have no difference in crash frequency compared to PCC roads, 50% of plant mix segments have higher crash risk propensity compared to PCC roads and *vice versa*. Furthermore, the standard deviation captures the probability distribution of this effect across all the segments. Lastly, interchange segments have a higher risk propensity compared to other regular segments of freeways consistent with the complex driving patterns associated with these segments.

Model Comparison

Two models are said to be nested if one of the two models (restricted model) can be obtained by placing restrictions on the other model (unrestricted model). Such nested models can be compared using the log-likelihood ratio (LR) test. In this test, the LR test statistic computed as $2 \times (LL\ of\ Unrestricted\ Model - LL\ of\ Restricted\ Model)$ is compared with the critical chi-squared value corresponding to the number of degrees of freedom equal to the number of additional parameters in the unrestricted model. The NB (restricted) and NB HD (unrestricted) models are nested models where there are two additional parameters in the NB HD model. The LR test statistic of comparison between these two models is 10.94 which is greater than the critical chi squared value of 5.99 for two degrees of freedom at 95% confidence level. So, the NB HD model is statistically better than the NB model. However, this test is not applicable to non-nested models. In such cases, Bayesian Information Criterion (BIC) is used to determine the better model and is computed as $-2 * LL + K \times LN(N)$ where K is the number of parameters in the model and N is the number of observations in the dataset. The BIC statistic penalizes models that attain higher LL values using more parameters. Between two non-nested models, a model with lower BIC value is the preferred model. For instance, the NB (BIC value = 1825.8) is better than the zero-inflated (ZI) Poisson ((BIC value = 2210.9). Among all the models considered, the MGOR NB HD model with spatial effects has the highest LL value and the least BIC value suggesting superior data fit.

Elasticity Effects

The parameter estimates in Table 7 do not directly indicate the magnitude of impact of different variables on expected crash frequency. To do this, elasticity effects that indicate percentage change in the dependent variable for a unit change in the explanatory variables were computed. For indicator variables, pseudo-elasticity effects were computed as follows. First, expected crash frequency was computed assuming the indicator variable assumes a value of zero for all segments. Next, the indicator variable was changed to one for all segments and expected crash frequency was recomputed. Next, percentage change in expected crash frequency was computed for each segment which is then averaged across all segments to get the average elasticity effect. For continuous variables, the same procedure was used except that a 10% increase in the variable was assumed. Table 8 presents the results of this elasticity results for the best model (last column of Table 7). The first entry in the table is 2.12% for traffic exposure which indicates that, on average if everything else remains the same, a 10% increase in traffic volumes during evening peak period will result in 2.12% more crashes on a roadway segment. A -40.36% elasticity effect for presence of extreme accelerations suggests that segments with extreme accelerations and no extreme decelerations have 40% fewer crashes, on average, compared to other segments. Other numbers in the table can be interpreted similarly.

Table 8 Elasticity Effects

Explanatory Variable	Elasticity Effect (%)
Traffic Exposure	2.12
Average speed	-5.49
SDACC	3.36
Spatially Weighted SDACC	9.05
Presence of extreme accel. w/o extreme decel.	-40.36
Intersection segment	15.67
Less than or Equal to Two Lanes	-0.76
Presence of Median	0.48

REFERENCE LIST

Bhat, C. R. (2003). "Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences." Transportation Research Part B: Methodological 37(9): 837-855.

Blincoe, L., T. R. Miller, E. Zaloshnja and B. A. Lawrence (2015). The economic and societal impact of motor vehicle crashes, 2010 (Revised).

Castro, M., R. Paleti and C. R. Bhat (2012). "A latent variable representation of count data models to accommodate spatial and temporal dependence: Application to predicting crash frequency at intersections." Transportation Research Part B: Methodological 46(1): 253-272.

DMV (2014). Virginia Traffic Crash Facts. D. o. M. Vehicles. 2014.

FHWA (2015). FHWA Forecasts of Vehicle Miles Traveled (VMT): May 2015. F. H. Administration.

Gettman, D. and L. Head (2003). "Surrogate Safety Measures from Traffic Simulation Models." Transportation Research Record: Journal of the Transportation Research Board 1840: 104-115.

Hariharan, B., J. Hong, V. Shankar, N. Venkataraman, J. C. Milton and I. Van Schalkwyk (2016). Roadside Geometry Effects on the Overdispersion Parameter. Transportation Research Board 95th Annual Meeting.

Jun, J. (2006). Potential Crash Exposure Measures Based on GPS-Observed Driving Behavior Activity Metrics, Citeseer.

Lord, D. and F. Mannering (2010). "The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives." Transportation Research Part A: Policy and Practice 44(5): 291-305.

Lord, D., S. P. Washington and J. N. Ivan (2005). "Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory." Accident Analysis & Prevention 37(1): 35-46.

Mannering, F. L., V. Shankar and C. R. Bhat (2016). "Unobserved heterogeneity and the statistical analysis of highway accident data." Analytic Methods in Accident Research 11: 1-16.

Mekker, M. M., S. M. Remias, M. L. McNamara and D. M. Bullock (2016). Characterizing Interstate Crash Rates Based on Traffic Congestion Using Probe Vehicle Data. Transportation Research Board 95th Annual Meeting.

Narayan, V., V. Shankar, J. Blum, B. Hariharan and J. Hong (2016). Transferability Analysis of Heterogeneous Overdispersion Parameter Negative Binomial Safety Performance Functions: A Case Study from California. Transportation Research Board 95th Annual Meeting.

Narayanamoorthy, S., R. Paleti and C. R. Bhat (2013). "On accommodating spatial dependence in bicycle and pedestrian injury counts by severity level." Transportation Research Part B: Methodological 55: 245-264.

NHTSA (2015). Traffic Safety Facts Crash-Stats. N. H. T. S. Administration.

Ogle, J. H. (2005). Quantitative assessment of driver speeding behavior using instrumented vehicles, Citeseer.

Qin, X., J. N. Ivan, N. Ravishanker and J. Liu (2005). "Hierarchical Bayesian Estimation of Safety Performance Functions for Two-Lane Highways Using Markov Chain Monte Carlo Modeling." Journal of Transportation Engineering 131(5): 345-351.

Shankar, V., J. Milton and F. Mannering (1997). "Modeling accident frequencies as zero-altered probability processes: An empirical inquiry." Accident Analysis & Prevention 29(6): 829-837.

Wåhlberg, A. E. (2004). "The stability of driver acceleration behavior, and a replication of its relation to bus accidents." Accident Analysis & Prevention 36(1): 83-92.

Zhen, C. and G. Qiang (2014). Mobile Sensor Data Collecting System Based on Smart Phone. Pervasive Computing and the Networked World: Joint International Conference, ICPCA/SWS 2013, Vina del Mar, Chile, December 5-7, 2013. Revised Selected Papers. Q. Zu, M. Vargas-Vera and B. Hu. Cham, Springer International Publishing: 8-14.

APPENDIX

Table 9: Poisson Model Parameter Estimates

Variables	***Coefficient***	***t-stat***
Roadway Inventory Parameters		
Constant	-2.8928	-5.10
LOG(Segment Length)	1.0000	-
Presence of Left Shoulder		
(Base: No Shoulder Present)		
Left Shoulder is Present	-0.1986	-2.10
Presence of Right Shoulder		
(Base: No Shoulder Present)		
Right Shoulder is Present	-0.2177	-2.30
Presence of Interchange		
(Base: No interchange Present)		
Interchange Segment	0.2911	4.50
Probe Vehicle Data Parameters		
Average speed	-0.0154	-5.74
Weighted Ave. of s.d. of Neighbor Segments	0.8190	8.75
S.D. of Accel. and Decel.	0.5650	7.15
Presence of Extreme Accel. w/o Extreme Decel	-0.4944	-3.08
Exposure Parameter		
LOG(AADT)	0.4197	7.00
Number of Cases	513	
Log Likelihood	-1126.22	

Table 10: Zero Inflation Poisson Model Parameter Estimates

Variables	***Coefficient***	***t-stat***
Zero Inflation Component		
Constant	-1.9484	-6.06
Less than or Equal to Two Lanes	0.9855	2.66
Expected Count $\boldsymbol{LOG(\lambda_s)}$		
Constant	-2.1928	-3.36
LOG(Segment Length)	1.0000	-
Presence of Left Shoulder	-0.2421	-2.51
Presence of Right Shoulder	-0.2558	-2.66
Interchange Segment	0.4262	6.12
Speed Less than or Equal to 45 mph	0.2449	3.15
SDACC(ft/sec^2)	0.6651	6.28
Spatially Weighed SDACC (ft/sec^2)	0.7650	8.53
Presence of Extreme Accel. w/o Extreme Decel	-0.3905	-2.23
LOG(Traffic Exposure)	0.2619	3.87
Number of Cases	513	
Log Likelihood	-1068.05	

Table 11: GOR NB HD Model without Spatial Variables

Variables	*Coefficient*	*t-stat*
Expected Count $LOG(\lambda_s)$		
Constant	-0.9282	-0.86
LOG(AADT)	0.2989	2.42
Average speed (mph)	-0.0208	-3.70
SDACC (ft/sec^2)	0.7220	4.67
Presence of extreme accel. w/o extreme decel.	-0.6444	-2.57
LOG(Segment Length)	1.0000	.
Dispersion $LOG(r_s)$		
Constant	0.0000†	
Lanes ≤ 2	-0.5245	-2.07
Presence of median	0.4059	2.10
Propensity		
Plant mix roadway surface	0.0000†	
Standard Deviation	0.7295	4.42
Intersection segment	0.1459	1.45
Number of Observations	513	513
Log-composite likelihood at convergence	-883.41	

Table 12 MGOR NB HD Model without Microscopic Traffic Measures

Variables	*Coefficient*	*t-stat*
Expected Count $LOG(\lambda_s)$		
Constant	0.1784	0.40
LOG(AADT)	0.1644	3.08
LOG(Segment Length)	1.0000	.
Dispersion $LOG(r_s)$		
Constant	0.6500	4.47
Lanes ≤ 2	-0.2358	-1.68
Presence of median	0.2005	1.38
Propensity		
Plant mix roadway surface	0.0000†	
Standard Deviation	0.7295	4.42
Intersection segment	0.1459	1.45
Number of Observations	513	513
Log-composite likelihood at convergence	-926.5	