Accident Analysis and Prevention 132 (2019) 105253

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Improving freeway segment crash prediction models by including disaggregate speed data from different sources

Nancy Dutta[a,*], Michael D. Fontaine[b]

[a] T3 Design Corporation, 10340 Democracy Ln, Fairfax VA 22030, United States
[b] Virginia Transportation Research Council, 530 Edgemont Rd, Charlottesville, VA 22903, United States

ARTICLE INFO

Keywords:
Crash Analysis
Safety Performance
Disaggregated Data

ABSTRACT

Traditional traffic safety analyses use highly aggregated data, typically annual average daily traffic (AADT) and annual crash counts. This approach neglects the time-varying nature of critical factors such as traffic speed, volume, and density, and their effects on traffic safety. This paper evaluated the relationship between crashes and quality of flow at different levels of temporal aggregation using continuous count station data and probe data from 4 lane rural freeway and 6 lane urban freeway segments in Virginia. The performance of crash prediction models using traffic and geometric information at 15-minute, hourly, and annual aggregation intervals were contrasted. This study also assessed whether inclusion of speed data improved model performance and examined the effects of using speeds from physical sensors versus speed estimates from private-sector probe speed data. The results showed that using average hourly volume along with average speed and selected geometric variables improved predictions compared to annual models that did not use speed information. When comparing an AADT-based model to an average hourly volume model for total crashes, the mean absolute prediction error improved by 11% for rural models and 20% for urban models. This result was based on volume and speed data from continuous count stations. When private sector probe speed data was used, the rural model performance improved by 10% and urban models by 20%. This trend was consistent for all crash types irrespective of level of injury or number of vehicles involved. Even though models using private sector data performed slightly worse than the ones based on continuous count data, they were still far better than AADT based models. These results indicate that probe based data can be used in developing crash models without harming prediction capability.

1. Introduction

Transportation safety research has sought to gain a better understanding of how different variables such as roadway geometry, driver behavior, traffic conditions, and environmental factors affect crash occurrence. The influence of those factors on traffic crashes cannot be fully understood without detailed information not only on crash itself, but also on its surrounding circumstances.

The Highway Safety Manual (HSM) provides standard scientific techniques and knowledge to help transportation officials make educated decisions regarding road safety (AASHTO, 2010). The safety performance functions (SPF) recommended in the HSM relate crash occurrence to annual average daily traffic (AADT). A difficulty with this approach is that a freeway with an intense flow during rush periods would clearly have a different crash potential than a freeway with the same AADT but with flow more evenly spread throughout the day. The customary means of using AADT in safety analysis may be too aggregate to capture all the variation in traffic flow that occurs throughout the day and could mask safety effects of operational improvements on a roadway. When considering the flow of traffic along a freeway, three parameters are of considerable significance. Speed and density describe the quality of service experienced by the stream, while volume measures the quantity of the stream and the demand on the highway facility. Drivers change or adapt their driving behavior according to the level of traffic present on the road. They may become more alert as traffic increases, but small driver errors may be more likely to result in collisions if traffic is congested. As traffic flow increases, the vehicles may travel at a lower speed, which could reduce crash severity during those conditions. Likewise, several past studies have indicated that speed variance may play a role in crash likelihood (Choudhary et al., 2018;Garber and Gadiraju, 1989;Garber and Ehrhart, 2000; Quddus, 2013; Solomon, 1964; Tanishita and Van Wee, 2017).

* Corresponding author.
E-mail addresses: dutta.nan@gmail.com (N. Dutta), Michael.Fontaine@VDOT.Virginia.Gov (M.D. Fontaine).

https://doi.org/10.1016/j.aap.2019.07.029
Received 3 January 2019; Received in revised form 17 June 2019; Accepted 27 July 2019
Available online 05 August 2019

There has been considerable research conducted in recent years into establishing predictive crash relationships for freeway segments. Despite overall progress, there is still no clear understanding about how different traffic flow characteristics that represent quality of flow affect safety. This paper addresses the limitations of current SPFs by developing crash prediction models using traffic and geometric information that is provided at sub-daily aggregation intervals for urban and rural freeway segments.

2. Research objective

As a practical matter, relationships between traffic crashes and traffic flow parameters are inherently difficult to establish due to limitations in matching crash data with available traffic data sources. It is further restricted by the random nature of crash occurrence and the quality of available crash and traffic data. For this reason, there is considerable interest in surrogate measures such as speed, standard deviation of speed, density, or volume to capacity (*v/c)* ratios that may help in identifying problems.

This paper seeks to evaluate the relationship between crashes and quality of flow both at 15 min and hourly levels of aggregation for freeway segments using data from Virginia. Using different geometric and traffic variables, predictive models were developed for both urban and rural freeway segments and for different crash types. The goal is to identify how traffic flow variables are related to crash occurrence and if disaggregate traffic data could improve the quality of crash predictions over AADT-based models. This paper also evaluates the possibility of using private sector probe speed data as an alternate source for speed data. Models were first developed using speed data from continuous count stations, and then these models were repeated using the probe data from INRIX. A comparison between these two data sources in terms of prediction accuracy is one of the major objectives this paper wanted to address since probe data is widely available on freeways at a much lower per-mile cost than sensor data.

3. Literature review

Crashes are complex events and are influenced by many factors such as road geometric design, traffic volume and composition, speed differentials between vehicles, and so on. This paper addresses two major topics that have been of interest in the field of traffic safety research: (a) geometric and traffic variables influencing the frequency of crashes and (b) disaggregated analysis, where exposure is defined by sub-daily data instead of AADT.

Horizontal curvature, grade, median width, lane width, and shoulder width are some of the significant factors influencing road crashes on freeway segments. Crash frequency has been shown to decrease as curve radius increases (Khan et al., 2013; Shaw-pin, 1994; Tegge et al., 2010). Prior research has also demonstrated that steeper vertical grades are associated with higher crash rates (Geedipally et al., 2017; Tegge et al., 2010). These effects of horizontal curve radius, horizontal curve length, and percent grade are included in the HSM in the form of crash modification functions (CMF) based on studies by Zegeer et al. (Zegeer et al., 1990) and Harwood et al. (Graham et al., 2014). Crashes also tend to increase with wider medians, even though it largely depends on crash type. Cross median crashes tend to decrease with increasing median width, whereas rollover crashes tend to increase (Khan et al., 2013; Shaw-pin, 1994).

Speed and speed variation are widely believed to be key issues in the understanding of traffic crashes. In 1964 the Federal Highway Administration (FHWA) published a report by Solomon that studied the relationship between crashes on 2-lane and 4-lane roadways and a number of factors (Solomon, 1964). From an analysis of 10,000 crashes, it was concluded that crash rates were lowest for travel speeds near the mean speed of traffic and increased with greater deviations above and below the mean. Solomon's work is often cited as the source of the 85th percentile speed rule for setting speeds. Imprialou et al. re-examined crash–speed relationships by creating a new crash data aggregation approach that enables improved representation of the road conditions just before crash occurrences (Imprialou et al., 2016). Crashes from Strategic Road Network of England in 2012 were aggregated according to the similarity of their pre-crash traffic and geometric conditions, forming an alternative crash count dataset termed as a condition-based approach. The results showed that high speeds trigger crash frequency. But the speed–crash relationship is negative regardless of crash severity. Empirical examination of the relationship between flow–density, speed, and crash rate on selected freeways in Colorado by Kononov et. al. suggested that as flow–density increases, the crash rate initially remains constant until a certain critical threshold combination of speed and density is reached (Kononov et al., 2012). Once this threshold is exceeded, the crash rate rises rapidly. Lord et. al. developed predictive models from data collected on freeway segments from Montreal, Quebec. For rural segments, as density and V/C increased, the number of single-vehicle crashes decreased, and the number of multi-vehicle crashes increased. The data showed that crashes become less severe with an increasing v/c ratio but did not seem to be affected by density (Lord et al., 2005).

Persaud and Dzbik developed crash prediction models at both the macro level (in crashes per unit length per year), and micro level (in crashes per unit length per hour) using the generalized linear modeling approach with a negative binomial error structure (Persaud and Dzbik, 1993). Microscopic models showed a decreasing slope in regression lines as hourly volume increased, perhaps capturing the influence of decreasing speed. This is in contrast to the macroscopic model, which showed increasing slopes. Evaluation of freeway safety as a function of traffic flow by Golob et al. revealed that the highest crash rates (6.3 crashes per million vehicle miles traveled (VMT)) occurred during the morning peak period with heavily congested flow, corresponding to low mean speeds, low speed variation, low flows, and low flow variation. In contrast, the lowest crash rates (0.6 per million VMT) were characterized by high speeds and low speed variation (Golob et al., 2004). Ivan et.al concluded that there is evidence that the hourly volume explains much of the variation in highway crash rates. They focused on actual hourly exposure values of seventeen rural, two-lane highway segments in Connecticut, with varying land-use patterns (Ivan et al., 2000). Single-vehicle crashes occurred most often in the evening and at night. On the other hand, multi-vehicle crashes were more likely to occur under daylight conditions at midday and during the evening peak period. Yu et al. investigated the impacts of data aggregation approaches based on traffic data from Shanghai's urban expressway system (Yu et al., 2018). Crash frequency analyses with a segment-based approach and a scenario-based approach were conducted first, and then crash risk analyses were developed at the individual crash level. It was found that during the congested period, an increase in operating speed would reduce crash likelihood. For medium operating speeds, the changes in operating speed do not have substantial effects on crash occurrence probability. For free-flow periods, increases in operating speed would further increase the probability of crashes.

The study of the relationship of crashes and traffic flow state is largely constrained by the difficulty in acquiring widespread information on quality of flow. Each agency has a different strategy regarding whether the roads are maintained at the state level or city/county level, which can also limit the ability to collect consistent crash and traffic detector data. Limited research focuses on different levels of data disaggregation, considers temporal traffic flow characteristics, compares different model forms and modeling techniques, and goes through vigorous model validation. A recent study by Wang et al. shed some light on this area (Wang et al., 2018). They developed different models to estimate crash frequency using annual daily traffic and annual hourly traffic. The study segments were from three expressways in Orlando, Florida and included basic freeway segments, merging segments and weaving segments. It was found the logarithm of volume, the standard deviation of speed, the logarithm of segment length, and the existence of a diverge segment were significant in the models. Weaving segments experienced higher daily and hourly crash frequencies than merge and basic segments.

This work discussed in this paper addresses similar concerns as those described by Wang et al., but the scope of these two papers are different in terms of dataset used and research objective. In this study, focus was on establishing a relationship between traffic flow variables and crashes using disaggregate traffic data over a broad statewide network. This paper also evaluates whether widely available probe data could serve as a substitute for loop detector data, which could broadly expand the applicability of crash prediction models that use speed as an input factor.

4. Data collection and preparation

Volume and speed data were collected for 2-lane directional rural freeway segments and 3-lane directional urban freeway segments in 15-minute increments from 2011 to 2017 using the Virginia Department of Transportation (VDOT) Traffic Management System. A total of 31 continuous count station were identified from rural 2 lane segments and 24 from urban 3 lane segments as shown in Fig. 1.

4.1. Volume and speed data

In Virginia, traffic volumes are determined using both continuous count stations and short-term counts conducted throughout the state on a rotating basis. This study relied on data from the continuous count stations only because of their high level of quality control and their ability to produce accurate volume counts over the entire study period. Only time periods where both volume and speed data meet a quality threshold set by VDOT were included in the dataset used for this research.

As an alternate data source, speed data was also obtained from the private sector travel time data provider INRIX for both 15-minute and hourly interval. INRIX is a private company that processes GPS probe data to estimate speeds, which are reported spatially using traffic message channel (TMC) links. TMC links are spatial representations developed by digital mapping companies for reporting traffic data and consist of homogeneous segments of roadways. VDOT currently uses INRIX data to support a variety of performance measurement and traveler information applications, and several external and internal evaluations have supported the accuracy of the travel time data for freeways (Haghani et al., 2009). INRIX provides confidence scores for each 1-minute interval travel time, with a confidence score of 30 representing real-time data and scores of 10 and 20 representing historic data during overnight and daytime periods, respectively. For the purposes of this analysis, no threshold was set for the confidence scores and both real time and historic speed data was used in model development.

While continuous volume data is available only at a discrete number of locations with sensors installed, INRIX speed data is broadly available across the roadway network in Virginia. Use of INRIX data in crash modeling will help to overcome the difficulties associated with using only continuous count station data.

4.2. Geometry data

The VDOT Highway Traffic Records Information System (HTRIS) (2019) was used to extract all geometric and traffic control information used for this analysis. Using this database, information such as number of lanes, speed limit, shoulder width, median type, rural/urban designation, etc. was gathered for the study segments. The vertical curvature (VC) data are expressed in the form of percent grade, with positive grades indicating uphill segments and negative grades indicating downhill segments. Horizontal curvature (HC) was expressed using a variety of variables, including length of the curve, presence of curve as

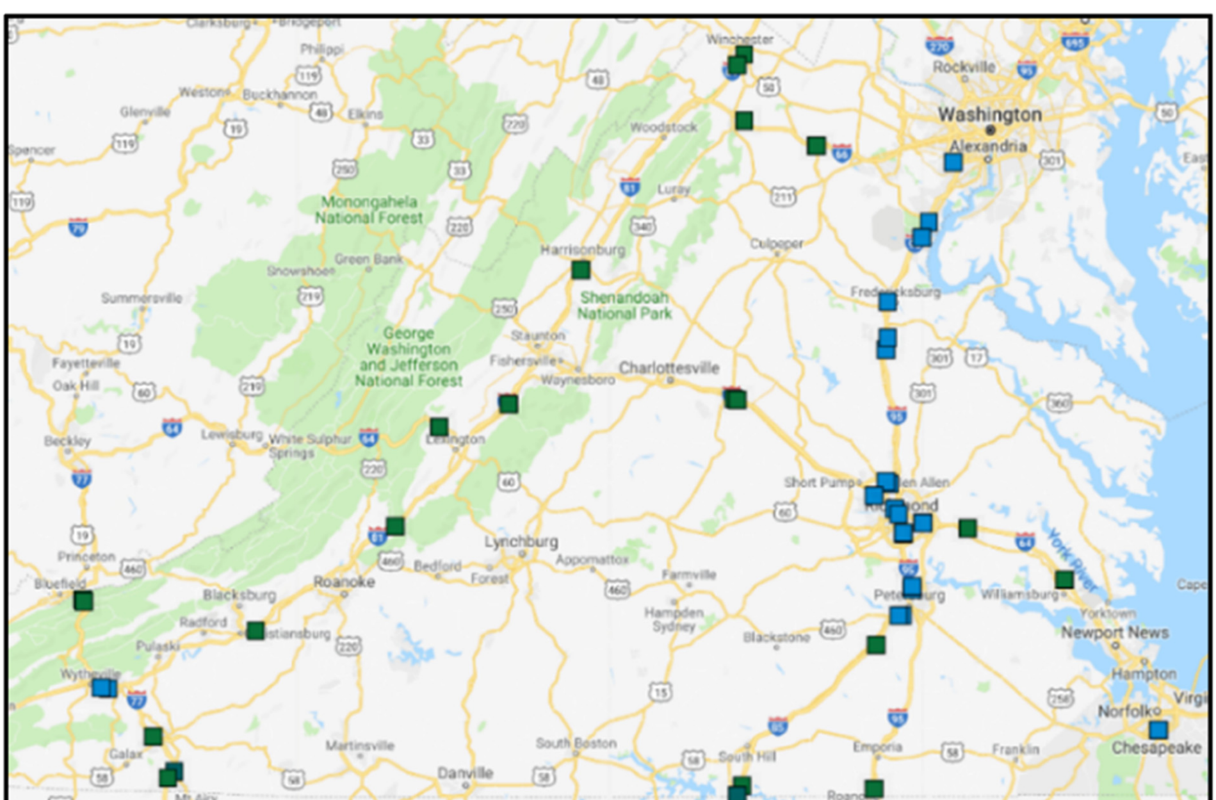

Fig. 1. Continuous Count Stations on Freeway Segments Used in This Research (Green squares represent rural stations; blue station represents urban stations) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

Table 1
Summary of data.

Type of Data	Source (Maintaining Agency)	Data Format	Data Elements
Traffic Volume	Continuous Count Stations on Virginia Freeways (VDOT)	Data was extracted for every 15 minute interval for the entire study period. This raw data was then converted to an average 15-minute volume and average hourly volume for each year. AADT data was directly available from the source.	Average 15-minute Volume Average Hourly Volume AADT
Speed	Continuous Count Stations on Virginia Freeways (VDOT) Probe data from private data source (INRIX)	Data was extracted for every 15 minute interval for the entire study period. The raw data was then converted to average 15-minute speed and average hourly speed for each year using the two different data sets.	Average 15-minute Speed (Count Station and INRIX) Average Hourly Speed (Count Station and INRIX)
Roadway Geometry	Highway Traffic Records Information System (VDOT)	Comprehensive inventory on detailed geometric data for Virginia. The data is provided for each roadway link in the state.	Number of lanes Horizontal & Vertical Curvature Median Type & Width Shoulder Width Speed Limit
Crashes	Roadway Network System (VDOT)	Detailed information on time, location, crash type, injury and road condition. The crash models predicted either crashes/15 min or crashes/hour depending on the time resolution under study. During validation, predicted number of crashes were summed to create yearly numbers to compare with AADT based crash models that predict crashes/year.	Total Crashes Fatal and Injury Crashes Property Damage Crashes Single Vehicle Crashes Multiple Vehicle Crashes

a percentage of segment length, and radius of curve. Length and radius of curve for each segment were directly available in the dataset.

4.3. Crash data

Crash data for all the sections were obtained from VDOT as well (Roadway Network System, 2019 Virginia Department of Transportation.). The data included detailed information on crash location and date, crash type, severity, number of vehicles involved, etc. For all the segments, crash information was also collected between 2011 and 2017.

A summary of all the data sources are included in Table 1. Fig. 2(a) shows a sample of the data format for a segment and Fig. 2(b) provides explanation of the data format.

4.4. Selection of freeway segment

For this research, only basic freeway segments free from ramps or interchanges were considered. This was done using the detector database maintained by the Traffic Engineering Division, 2019 at VDOT and the VDOT GIS Integrator, 2019. The GIS integrator stores layers of different elements such as mile markers, exits and traffic count stations. All these elements contain direction and location information that helped to define the segments in a way that there is no entry/exit ramp within 0.5 miles of start/end of the segment.

It was important for this analysis to define a segment surrounding each count station where it could be assumed that homogeneous flow conditions were present for the entire length. If the station was on a link with homogeneous geometric characteristics that was greater than 2 miles in length, a buffer of a maximum 2 miles around the actual location of the detector (1 mile upstream and downstream) was created.

The number of lanes, lane and shoulder width, speed limit, median type, and median width were used to define the geometric homogeneity of segment. Generally speaking, horizontal and vertical curvature was not significant on these segments since they were located on interstates with high geometric design standards. Since this research focuses on interaction between geometry and flow parameters and how they define safety instead of a design focused approach, horizontal and vertical curvature was not used to define the segment, instead they were used as variables to identify their interaction with flow. Table 2 summarizes the properties of the study segments.

5. Methodology

A series of crash prediction models were developed using a variety of variables including volume, segment length, heavy vehicle percentage, horizontal curvature, vertical curvature, median width, median type, and speed. Speed was expressed in a variety of ways, including average speed, standard deviation of speed, and difference between speed limit and average speed. Volume and segment length are already used in the HSM SPFs.

Previous research by the authors indicated that crash prediction models using raw hourly volume and speed data as observed on each site perform worse than the AADT based model (Dutta and Fontaine, 2018) With raw hourly data, data errors and availability can create issues and imputation of missing values can be problematic and increase errors. The same research also showed that use of the average volume calculation helped to smooth out the discrepancies created by missing raw hourly data. Based on this previous experience, volumes in this research were expressed in the form of AADT, average 15-min volume, and average hourly volume for each site over each year. Quality of flow variables were summarized using a similar definition in each case. The disaggregated models were compared to each other and with the AADT model to determine how the predictions vary from typical HSM-like models. Three different regression model forms were evaluated as part of this research:

1 Models using volume and segment length only
2 Models using volume, segment length, and geometric variables
3 Models using volume, segment length, geometric variables, and traffic flow parameters

For each model, volumes were expressed as AADT (to be consistent with the current HSM SPFs), average 15-min volume, and average hourly volume. Both negative binomial and zero inflated negative binomial regression methods were evaluated. Models using AADT data were created for the first two model forms so that model performance could be compared using the same datasets. To be consistent with the HSM, length was used as an offset variable in the models. One additional step for the third model was that it was developed twice. First a model was selected for each crash type using volume, geometry and flow parameters based on data from continuous count station. Once the models are finalized, they were regenerated by keeping the same form but using speed data from INRIX. For this iteration, volume data was still generated by the continuous count stations, but speed components were coming from probe based INRIX data.

5.1. Selection of model form

There are a wide variety of statistical methods that researchers have

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	LinkID	Year	Hour	Volume	HV	Total	Speed	Std	Length	Percent HC	L_HC	Radius_HC	Grade_VC	Mwidth	Speed Limit	Delta
2	10148	2011	0	203.13	39.80	0	48.88	4.21	1.78	12.92	0.23	0.78	0.32	40	70	21.12
3	10148	2011	1	157.27	42.23	0	48.31	3.98	1.78	12.92	0.23	0.78	0.32	40	70	21.69
4	10148	2011	2	144.88	41.00	0	48.69	3.43	1.78	12.92	0.23	0.78	0.32	40	70	21.31
5	10148	2011	3	124.67	43.19	1	48.44	3.71	1.78	12.92	0.23	0.78	0.32	40	70	21.56
6	10148	2011	4	135.88	41.30	0	48.93	3.2	1.78	12.92	0.23	0.78	0.32	40	70	21.07
7	10148	2011	5	181.82	38.49	0	50.03	3.44	1.78	12.92	0.23	0.78	0.32	40	70	19.97
8	10148	2011	6	289.33	30.31	0	52.2	3.1	1.78	12.92	0.23	0.78	0.32	40	70	17.8
9	10148	2011	7	438.81	24.57	0	53.23	3.13	1.78	12.92	0.23	0.78	0.32	40	70	16.77
10	10148	2011	8	590.53	21.40	1	53.24	3.56	1.78	12.92	0.23	0.78	0.32	40	70	16.76
11	10148	2011	9	738.24	18.93	0	52.93	3.98	1.78	12.92	0.23	0.78	0.32	40	70	17.07
12	10148	2011	10	892.68	17.21	1	52.58	4.17	1.78	12.92	0.23	0.78	0.32	40	70	17.42

(a)

Column	Explanation
LinkID	Identifier for Segments
Year	Year
Hour	Hour
Volume	Average Hourly Volume (vph)
HV	Heavy Vehicle Percentage (%)
Total	Hourly Total Crashes
Speed	Average Hourly Speed (mph)
Std	Standard Deviation of Speed
Length	Length of Segment (mile)
Percent HC	Percent of Horizontal Curve Presence (%)
L_HC	Length of Horizontal Curve (mile)
Radius_HC	Radius of Horizontal Curve (mile)
Grade_VC	Grade of Vertical Curve (%)
Mwidth	Median Width (ft)
SL	Speed Limit (mph)
Delta	Difference between Speed Limit & Speed (mph)

(b)

Fig. 2. Sample Data (a) Data Format, (b) Data Explanation.

been using to model crash frequency over the years. Although Poisson models have served as a starting point for crash analysis, they are often criticized for its inability to handle over- and under-dispersed data (Lord and Mannering, 2010). The negative binomial regression model is an extension of the Poisson model that helps overcome possible over dispersion in the data. Negative binomial regression has become the most common method for developing SPFs, and is also the recommended modeling approach in the HSM (Highway Safety Manual, 2010). In a negative binomial regression model, the probability of roadway entity i having yi crashes per time period is defined as:

$$\boldsymbol{P}(\boldsymbol{y_i} = \frac{\boldsymbol{exp}(-\boldsymbol{\lambda_i}) * \boldsymbol{\lambda_i}^{\boldsymbol{y_i}}}{\boldsymbol{y_i}\,!} \quad (1)$$

$$\boldsymbol{\lambda_i} = \boldsymbol{exp}(\boldsymbol{\beta X_i} + \varepsilon_i) \quad (2)$$

Where exp (ε_i) is a gamma-distributed error term with mean 1 and variance α (Simon et al., 2010). The addition of this term allows the variance to differ from the mean as:

$$\text{VAR}(yi) = \text{E}(yi)\,[1 + \alpha \text{E}(yi)] = \text{E}(yi) + \alpha\, \text{E}(yi)^2 \quad (3)$$

Since crashes are random events, researchers are often left with a dataset that is characterized by a significant number of zeros. As the data becomes more disaggregated, zero crashes become more common for the selected interval (hour or 15-minute). Zero inflated models have been developed to handle data characterized by a significant number of zeros or more zeros than the one would expect in a traditional Poisson or negative binomial/Poisson-gamma model. These models operate on the principle that the excess zero density that cannot be accommodated by a traditional count structure is accounted for by a splitting regime that models a crash-free versus a crash prone propensity of a roadway segment (Lord and Mannering, 2010; Simon et al., 2010). If the probability of a data point being zero is π and probability of it being non-zero is (1 – π), then, the probability distribution of the ZINB random variable y_i can be written as:

$$\boldsymbol{P_r}(\boldsymbol{y_i} = \boldsymbol{j}) = \begin{cases} \pi_i + (1 - \pi_i) g(\boldsymbol{y_i} = 0) \; if\ j = 0 \\ (1 - \pi_i) g(\boldsymbol{y_i}) \; if\ j > 0 \end{cases} \quad (4)$$

Where πi is the logistic link function and g(yi) is the negative binomial distribution given by:

$$\boldsymbol{g}(\boldsymbol{y_i}) = \boldsymbol{P_r}(\boldsymbol{Y} = \boldsymbol{y_i} | \boldsymbol{\mu_i},\ \boldsymbol{\alpha}) = \frac{\Gamma(\boldsymbol{y_i} + \alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(\boldsymbol{y_i} + 1)} \left(\frac{1}{1 + \alpha\mu_i}\right)^{\alpha^{-1}} \left(\frac{\alpha\mu_i}{1 + \alpha\mu_i}\right)^{y_i} \quad (5)$$

5.2. Selection of modeling technique

Generalized linear models are extensions of traditional regression models that allow the mean to depend on the explanatory variables through a link function, and the response variable to be any member of a set of distributions called the exponential family (e.g., Normal, Poisson, Binomial) (McCullagh and Nelder, 1989). In a generalized

Table 2
Summary of the descriptive statistics of freeway study segments.

Type of Segment	Total Mileage (mile)	Variable	Mean	Std. Deviation	Min	Max
Rural 4 Lane Segments	57.21	AADT	21360	7926	4420	34200
		Average Hourly Volume (vph)	855	600	29	2754
		Average Hourly Speed (mph)	69.36	4.19	48.31	75.72
		Segment Length (mile)	1.85	0.41	1.06	2.00
		Lane Width (ft)	12.00	0.00	12.00	12.00
		Right Shoulder Width (ft)	6.47	4.93	0.00	10.00
		Left Shoulder Width (ft)	3.88	4.77	0.00	10.00
		Median Width (ft)	115	54	34	220
		Horizontal Curvature Radius (mile)	2.10	1.47	0.00	5.31
		Horizontal Curvature Length(mile)	0.37	0.23	0.00	0.99
		Grade (%)	−0.27	0.96	−1.67	2.69
		Speed Limit (mph)	69	1.70	65	70
		Annual Total Crashes	7.00	7.00	1.00	13.00
		Annual Fatal & Injury Crashes	2.00	3.00	0.00	5.00
		Annual Property Damage Crashes	5.00	5.00	0.00	8.00
		Annual Single Vehicle Crashes	4.00	2.00	0.00	9.00
		Annual Multiple Vehicle Crashes	3.00	4.00	0.00	4.00
Urban 6 Lane Segments	38.67	AADT	43840	15754	20137	80656
		Average Hourly Volume (vph)	1717	1209	69	5243
		Average Hourly Speed (mph)	63.96	7.31	19.92	74.71
		Segment Length (mile)	1.59	0.41	0.81	2.21
		Lane Width (ft)	12.00	0.00	12.00	12.00
		Right Shoulder Width (ft)	5.38	5.19	0.00	12.00
		Left Shoulder Width (ft)	4.77	5.39	0.00	12.00
		Median Width (ft)	66.35	51.62	5	220
		Horizontal Curvature Radius (mile)	1.51	0.88	0	3.61
		Horizontal Curvature Length(mile)	0.35	0.19	0	0.78
		Grade (%)	−0.18	1.09	−2.58	2.20
		Speed Limit (mph)	63.05	4.19	55	70
		Annual Total Crashes	16.00	17.00	0.00	74.00
		Annual Fatal & Injury Crashes	4.00	5.00	0.00	34.00
		Annual Property Damage Crashes	12.00	13.00	0.00	60.00
		Annual Single Vehicle Crashes	5.00	4.00	0.00	32.00
		Annual Multiple Vehicle Crashes	11.00	15.00	0.00	64.00

linear model (GLM), each outcome Y of the dependent variables is assumed to be generated from the exponential family. The mean, $\boldsymbol{\mu}$, of the distribution depends on the independent variables, X, through:

$$\boldsymbol{E}(\boldsymbol{Y}) = \boldsymbol{\mu} = \boldsymbol{g}^{-(X\beta)} \tag{6}$$

Where E(Y) is the expected value of Y; $X\beta$ is the *linear predictor*, a linear combination of unknown parameters $\boldsymbol{\beta}$; g is the link function. The unknown parameters, β, are typically estimated with maximum likelihood. This method estimates model parameters by selecting those that maximize a likelihood function that describes the underlying statistical distribution assumed for the regression model. For a negative binomial regression model, the likelihood function can be described as -

$$\boldsymbol{L}\,(\boldsymbol{\lambda_i}) = \prod_{\boldsymbol{i}} \frac{\boldsymbol{\Gamma}\,(\boldsymbol{y_i} + \left(\frac{1}{\alpha}\right))}{\boldsymbol{y_i}\,!\,\boldsymbol{\Gamma}\left(\frac{1}{\alpha}\right)} \cdot \left[\frac{\boldsymbol{\alpha\lambda_i}}{1 + \boldsymbol{\alpha\lambda_i}}\right]^{y_i} \cdot \left[\frac{1}{1 + \boldsymbol{\alpha\lambda_i}}\right]^{1/\alpha} \tag{7}$$

Where $\Gamma(x)$ is the gamma function, variance is α, λ is the mean and y_i is number of crashes per period for roadway segment i.

5.3. Vuong test

The Vuong test statistic (V) has been proposed for non-nested models to compare the fitness of zero inflated models versus regular count models (Vuong, 1989) :

$$\boldsymbol{V} = \frac{\boldsymbol{\bar{m}} \ * \sqrt{\boldsymbol{N}}}{\boldsymbol{S_m}} \tag{8}$$

Where, $m_i = \log[\frac{f_1\,(y_i)}{f_2\,(y_i)}]$
N = number of observations
$\bar{m}$ = Mean of m_i
S_m = Standard deviation of m_i
f_1, f_2 = Two competing models

V has a standard normal distribution, and has three possible outcomes:

- If the absolute value of V is less than 1.96 for a 0.95 confidence level, then neither model is preferred by the test result.
- V is a large positive value, then model 1 is preferred.
- V is a large negative value, then model 2 is preferred.

This test was used to select which model form is appropriate for the dataset.

5.4. Model selection and validation

While comparing the models, it is important to have a consistent methodology to select a model from a series of models that has been developed for each technique. A popular method for model selection is the Akaike information criterion (AIC)(Akaike, 1974). AIC offers an estimate of the relative information lost when a given model is used to represent the process that generated the data.

$$\mathbf{AIC} = \mathbf{-2LL} + \mathbf{2p} \tag{9}$$

Where p is the number of estimated parameters included in the model. A lower value of AIC indicates a better model. It should be noted that the values of AIC are only relevant to that particular disaggregation level. Different levels of data aggregation lead to very different total numbers of data points in all these models, and the interaction among variables changes for different levels of disaggregation as well. As a result, AIC values should not be compared across different data aggregation levels.

It is important to note that an objective assessment of the predictive performance of a particular model can be made only through the evaluation of several goodness of fit (GOF) criteria. The GOF measures used to conduct external model validation included mean prediction bias (MPB), mean absolute deviation (MAD), and mean squared prediction error (MSPE) (Washington et al., 2010.)

Since AADT based models predict annual crashes while hourly volume models predicted hourly crashes, the summation of hourly predictions was used to generate annual predicted numbers of crashes for the GOF calculations. The average hourly volume data was computed by averaging data for each available hour for each site, so there were always 24 h of data available for each year and each site for validation. A similar methodology was followed for average 15 min data as well. Data from the years 2011 to 2015 was used to build the models, and data from 2016 and 2017 were used for validation. The calculation of these measures was based on the following equations:

$$\textbf{\textit{Mean Absolute Deviation (MAD)}} = \frac{\sum_{i=1}^{n} |Y_{model} - Y_{observed}|}{n} \quad (10)$$

$$\textbf{\textit{Mean Absolute Prediction Error (MAPE)}} = \frac{100}{n}\sum_{i=1}^{n} \left| \frac{Y_{observed} - Y_{model}}{Y_{observed}} \right| \quad (11)$$

$$\textbf{\textit{Mean Squared Prediction Error (MSPE)}} = \frac{\sum_{i=1}^{n} (Y_{model} - Y_{observed})^2}{n} \quad (12)$$

Where –
Y_{model} = Predicted Crash Frequency
$Y_{observed}$ = Observed Crash Frequency
n = Sample Size

Fig. 3 below provides an overview of the methodology followed in this research. Fig. 4 provides a sample flowchart for all the tasks under a particular model. Other models were developed by performing same tasks with added variables.

6. Results and discussion

6.1. Vuong test

The Vuong test results showed that, in general, negative binomial models performed better than the zero inflated ones with respect to AIC value, variable significance, and sign of estimated coefficients. For the 15 min volume dataset, the volume and geometry model for single vehicle crashes on urban segments and fatal and injury crashes for rural segments were the only two categories where the Vuong test results preferred the zero inflated model over the negative binomial form. The results supported negative binomial model for all other cases as documented in Tables 3 and 4. For hourly data, negative binomial models outperformed the zero inflated models for both rural and urban segments, irrespective of crash type. To maintain consistency in model form, negative binomial models were used for both total and injury crashes, and those results are documented in the remainder of this paper.

6.2. Volume and length model

For the first set of models, volume was significant for all levels of aggregation for all types of crashes. This model is consistent with the current HSM SPFs in terms of variables used but differs in how volume is being used. This serves as the basic model, and more variables are added in subsequent steps to increase the complexity of the models. Sample results are summarized in Tables 5 and 6 for the volume only models for rural and urban segments respectively.

6.3. Volume, length and geometry model

Next, geometric variables were added to the volume only model discussed previously. Due to limited variability in lane width and shoulder width for this particular data set, they were not significant in the modeling process. Other geometric variables such as median width, horizontal curvature, vertical curvature were evaluated. Horizontal curvature was included in the models as length of curvature and percentage of curvature (ratio of curve length and segment length) variables. Vertical curvature was categorized as positive and negative grades.

For urban segments, the only statistically significant geometric variable was median width for all levels of aggregation and crash types. The segments with curve presence were mostly comprised of long, gentle horizontal curves that almost resemble a tangent section. There were little variability in vertical grades for these segments as well. Median width was negatively associated with crash frequency, indicating wider medians in urban segments reduce the total number of crashes. Previous research indicated that median width between 20 and 30 ft generally shows a mixed effect on crashes and median width of 60–80 ft has decreasing effect on crashes (Gang-Len and Xiang, 2003; Knuiman et al., 1993). About 55% of the urban dataset had median widths within this range so the negative relationship between median width and crashes is intuitive. Table 7 summarizes some sample models for this step. The geometric variables that were significant are a

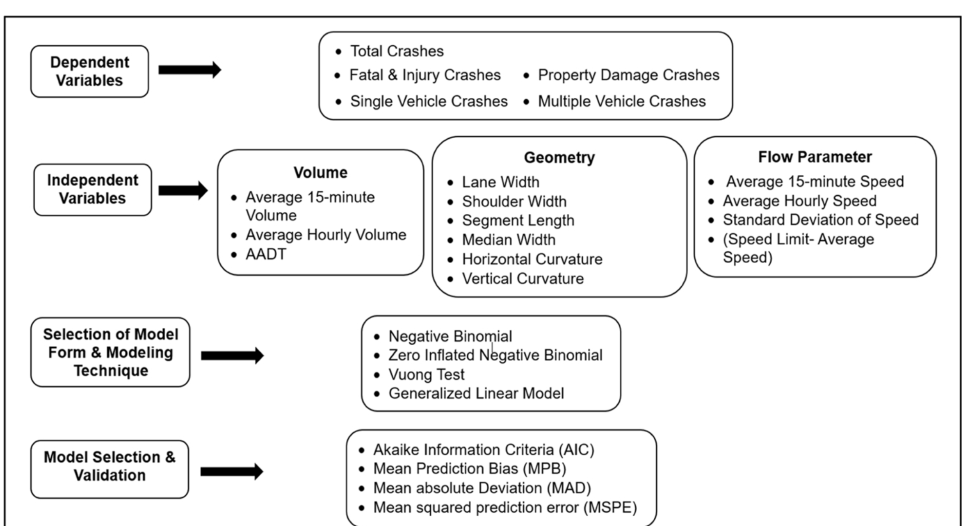

Fig. 3. Summary of Methodology.

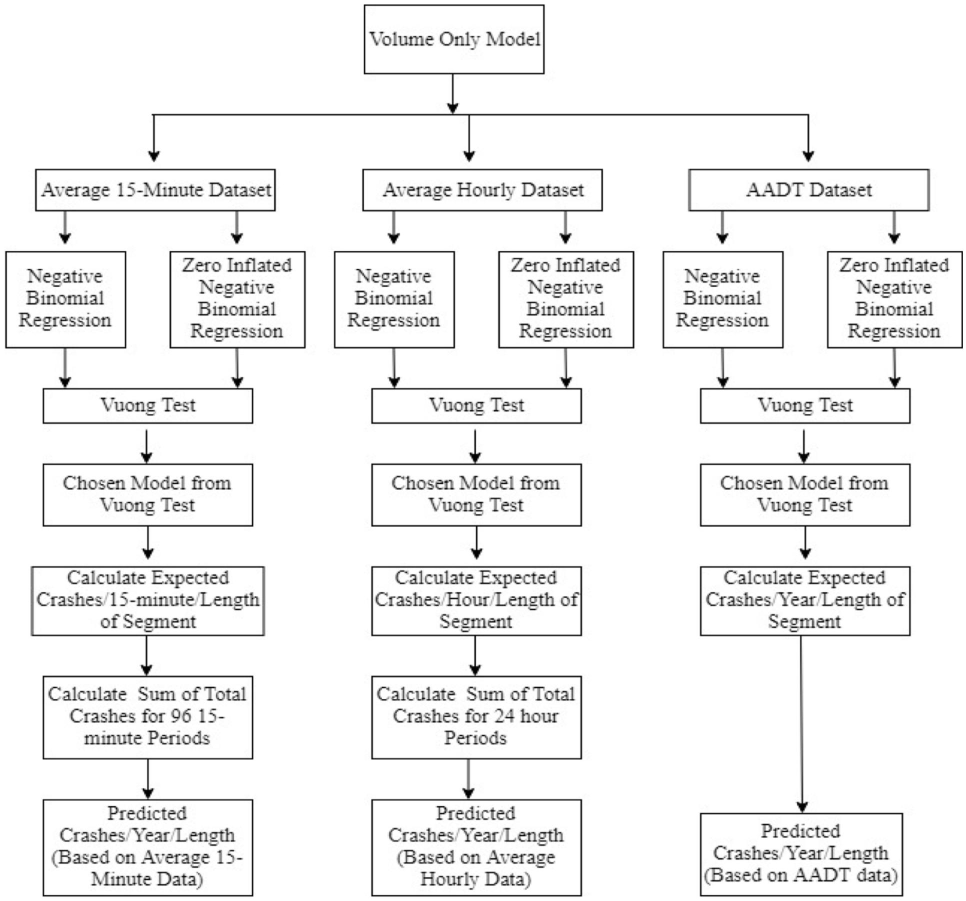

Fig. 4. Flowchart of Modeling Tasks.

Table 3
Vuong Test Results for Rural 4 Lane Freeway Segments.

Total Crashes				
	Model	AIC Corrected	BIC Corrected	Result
Average 15-minute Data	Volume, length, and geometry models	3.273	4.009	model 1 > model 2
	Volume, length, geometry, and flow state models.	4.093	5.092	model 1 > model 2
Average Hourly Data	Volume, length, and geometry models	2.292	3.399	model 1 > model 2
	Volume, length, geometry, and flow state models.	2.867	3.816	model 1 > model 2
Injury Crashes				
	Model	AIC Corrected	BIC Corrected	Result
Average 15-minute Data	Volume, length, and geometry models	−2.667	−1.693	model 2 > model 1
	Volume, length, geometry, and flow state models.	1.233	2.013	model 1 > model 2
Average Hourly Data	Volume, length, and geometry models	1.261	1.925	model 1 > model 2
	Volume, length, geometry, and flow state models.	0.962	1.474	model 1 > model 2
Single Vehicle Crashes				
	Model	AIC Corrected	BIC Corrected	Result
Average 15-minute Data	Volume, length, and geometry models	1.465	2.368	model 1 > model 2
	Volume, length, geometry, and flow state models.	1.693	2.667	model 1 > model 2
Average Hourly Data	Volume, length, and geometry models	1.983	2.138	model 1 > model 2
	Volume, length, geometry, and flow state models.	1.798	2.126	model 1 > model 2

function of the data set available for modeling, and these models may not reflect variation that would be seen across a broader cross section of sites.

For rural segments, 71% of the data came from segments with median widths greater than 80 ft and no median barrier. The results indicated that wider medians generally had more crashes. This is contradictory to the urban segments, but consistent with previous research (Shankar et al., 2004; Graham et al., 2014). The relationship between median width and crashes largely depend on type of facility, crash type, and also presence and type of median barrier. Cross median crashes tend to decrease with increasing median width, whereas rollover crashes tend to increase. For vertical curvature, presence of grade (both positive and negative) increases the probability of any types of crash. For injury crashes and single vehicle crashes, only negative grades had a statistically significant affect. These findings were similar irrespective of the volume disaggregation level, and also align with the results from previous research (Graham et al., 2014; Shankar et al., 2004; Watson et al., 2014). Table 8 includes final selected models from this step.

Table 4
Vuong Test Results for Urban 6 Lane Freeway Segments.

Total Crashes				
	Model	AIC Corrected	BIC Corrected	Result
Average 15-minute Data	Volume, length, and geometry models	4.775	5.734	model 1 > model 2
	Volume, length, geometry, and flow state models.	6.879	7.592	model 1 > model 2
Average Hourly Data	Volume, length, and geometry models	5.321	6.135	model 1 > model 2
	Volume, length, geometry, and flow state models.	6.499	7.107	model 1 > model 2
Injury Crashes				
	Model	AIC Corrected	BIC Corrected	Result
Average 15-minute Data	Volume, length, and geometry models	5.292	5.781	model 1 > model 2
	Volume, length, geometry, and flow state models.	6.005	6.658	model 1 > model 2
Average Hourly Data	Volume, length, and geometry models	7.435	8.358	model 1 > model 2
	Volume, length, geometry, and flow state models.	7.603	7.407	model 1 > model 2
Single Vehicle Crashes				
	Model	AIC Corrected	BIC Corrected	Result
Average 15-minute Data	Volume, length, and geometry models	−6.076	−4.962	model 2 > model 1
	Volume, length, geometry, and flow state models.	5.061	6.135	model 1 > model 2
Average Hourly Data	Volume, length, and geometry models	−5.859	−3.981	model 1 > model 2
	Volume, length, geometry, and flow state models.	4.711	5.251	model 1 > model 2

Table 5
Parameter Estimates for Volume Only Models for Rural 4 Lane Freeway Segments.

	Total Crashes								
	Average 15 Minute Volume			Average Hourly Volume			AADT		
	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)
Intercept	−7.59	0.255	< 2e-16	−7.21	0.338	< 2e-16	−6.32	1.120	2E-08
log (Volume)	0.55	0.047	< 2e-16	0.59	0.049	< 2e-16	0.65	0.114	1E-08
AIC	6027			3924			853		
	Injury Crashes								
	Average 15 Minute Volume			**Average Hourly Volume**			**AADT**		
	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)
Intercept	−8.68	0.448	< 2e-16	−8.06	0.564	< 2e-16	−9.38	1.734	6E-08
log (Volume)	0.54	0.083	8E-11	0.54	0.083	5E-11	0.83	0.175	2E-06
AIC	2425			1764			553		
	Single Vehicle Crashes								
	Average 15 Minute Volume			**Average Hourly Volume**			**AADT**		
	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)
Intercept	−6.3	0.288	< 2e-16	−5.45	0.363	< 2e-16	−2.47	1.085	0.0023
log (Volume)	0.20	0.055	3E-04	0.23	0.055	2E-05	0.18	0.110	0.001
AIC	3953			2741			656		

Table 6
Parameter Estimates for Volume Only Models for Urban 6 Lane Freeway Segments.

	Total Crashes								
	Average 15 Minute Volume			Average Hourly Volume			AADT		
	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)
Intercept	−10.05	0.301	< 2e-16	−7.51	0.359	< 2e-16	−19.71	1.951	< 2e-16
log (Volume)	1.06	0.047	< 2e-16	0.74	0.048	< 2e-16	1.95	0.184	< 2e-16
AIC	8941			5571			825		
	Injury Crashes								
	Average 15 Minute Volume			**Average Hourly Volume**			**AADT**		
	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)
Intercept	−10.43	0.499	< 2e-16	−9.04	0.556	< 2e-16	−23.06	2.420	< 2e-16
log (Volume)	0.92	0.078	< 2e-16	0.76	0.073	< 2e-16	2.14	0.225	< 2e-16
AIC	3687			2594			548		
	Single Vehicle Crashes								
	Average 15 Minute Volume			**Average Hourly Volume**			**AADT**		
	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)
Intercept	−5.65	0.326	< 2e-16	−4.31	0.406	< 2e-16	−9.41	2.125	10E-06
log (Volume)	0.15	0.054	8E-03	0.13	0.056	0.0177	0.89	0.199	9E-06
AIC	3927			2763			621		

Table 7
Parameter Estimates for Volume and Geometry Based Models for Urban Segments.

	Total Crashes								
	Average 15 Minute Volume			Average Hourly Volume			AADT		
	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)
Intercept	−9.65	0.301	< 2e-16	−6.91	0.359	< 2e-16	−17.36	2.401	4.81E-13
log (Volume)	0.97	0.051	< 2e-16	0.70	0.048	< 2e-16	1.75	0.236	1.07E-13
Median Width	−0.13	0.032	4E-09	−0.11	0.063	< 2e-16	−0.21	0.001	0.0043
AIC	8868			5520			742		
	Injury Crashes								
	Average 15 Minute Volume			**Average Hourly Volume**			**AADT**		
	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)
Intercept	−9.99	0.499	< 2e-16	−8.56	0.563	< 2e-16	−23.12	2.805	< 2e-16
log (Volume)	0.82	0.083	< 2e-16	0.75	0.073	< 2e-16	2.19	0.272	7.55E-16
Median Width	−0.11	0.541	6E-06	−0.32	0.423	1E-09	−0.15	0.001	5.97E-07
AIC	3964			2559			469		
	Single Vehicle Crashes								
	Average 15 Minute Volume			**Average Hourly Volume**			**AADT**		
	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)
Intercept	−5.32	0.335	< 2e-16	−3.74	0.427	< 2e-16	−6.87	2.604	0.00833
log (Volume)	0.09	0.059	3E-04	0.11	0.057	0.0571	0.67	0.256	8.59E-03
Median Width	−0.12	0.025	4E-04	−0.21	0.001	9E-11	−0.01	0.321	0.00272
AIC	3909			2721			550		

6.4. Volume, geometry and flow parameter models

The final sets of models were created by adding flow parameters to the models selected in the previous step. Average speed, standard deviation of speed, and the difference between speed limit and average speed (called the delta speed hereafter) were selected to represent traffic flow. AADT based models were not developed for this alternative since average speed over a year showed little variability. These models were developed twice; first with speed data from the continuous count stations and then repeating the same model with speed data from INRIX. This was done to compare the quality of data between these sources and how they affect the model fit. Data from 2016 and 2017

Table 8
Parameter Estimates for Volume and Geometry Based Models for Rural Segments.

	Total Crashes								
	Average 15 Minute Volume			Average Hourly Volume			AADT		
	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|	Estimate	Std. Error	Pr(> |z|)
Intercept	−8.41	0.307	< 2e-16	−8.04	0.385	< 2e-16	−9.04	1.289	2E-12
log (Volume)	0.56	0.051	< 2e-16	0.57	0.0524	< 2e-16	0.87	0.131	2E-11
Median Width	0.43	0.138	5E-06	0.51	0.315	3E-09	0.65	0.902	0.0153
Grade of VC									
Negative	0.42	0.092	5.E-06	0.48	0.098	8.E-07	0.55	0.106	3E-07
Positive	0.38	0.133	0.00465	0.33	0.143	2E-02	0.41	0.161	0.0102
Percent of HC	0.05	0.001	6E-11	0.06	0.009	9E-11	0.06	0.008	3E-14
Length of HC	−2.74	0.472	4E-09	−3.15	0.515	8E-10	−3.82	0.476	1E-15
AIC	5977			3864			796		
	Injury Crashes								
	Average 15 Minute Volume			**Average Hourly Volume**			**AADT**		
	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)
Intercept	−9.29	0.532	< 2e-16	−8.46	0.629	< 2e-16	−11.07	2.125	2E-07
log (Volume)	0.51	0.087	8E-09	0.47	0.086	3E-08	0.95	0.216	1E-05
Median Width	0.24	0.711	1E-03	0.36	0.667	2E-05	0.29	0.201	0.0255
Grade of VC									
Negative	0.35	0.161	3E-02	0.36	0.161	0.0248	0.49	0.167	3E-03
Positive	0.16	0.243	0.4989	0.03	0.255	0.8909	0.36	0.267	0.1784
Percent of HC	0.05	0.015	8E-04	0.04	0.016	9E-03	0.041	0.014	4E-03
Length of HC	−2.59	0.855	2E-03	−2.61	0.936	5E-03	−2.70	0.807	8E-04
AIC	2416			1745			544		
	Single Vehicle Crashes								
	Average 15 Minute Volume			**Average Hourly Volume**			**AADT**		
	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)
Intercept	−7.09	0.374	< 2e-16	−6.11	0.433	< 2e-16	−4.03	1.309	2.E-03
log (Volume)	0.23	0.059	1E-04	0.22	0.058	2E-04	0.28	0.135	3E-02
Median Width	0.33	0.612	4E-03	0.41	0.125	2E-06	0.22	0.311	0.0122
Grade of VC									
Negative	0.54	0.131	3E-05	0.44	0.124	4E-04	0.61	0.125	1E-06
Positive	0.37	0.175	0.0351	0.14	0.179	0.4419	0.36	0.192	0.0602
Percent of HC	0.04	0.012	6E-04	0.06	0.012	9E-07	0.05	0.009	3E-07
Length of HC	−2.32	0.654	4E-04	−3.47	0.672	2E-07	−2.99	0.556	7E-08
AIC	3934			2699			624		

Table 9
Parameter Estimates for Volume, Geometry and Flow Parameter Based Models for Rural Segments.

	Models with Detector Speed						Models with INRIX Speed					
Total Crashes	Average 15 Minute Volume			Average Hourly Volume			Average 15 Minute Volume			Average Hourly Volume		
	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)
Intercept	−1.31	1.770	< 2e-16	−6.91	0.394	< 2e-16	−1.66	1.760	< 2e-16	−5.95	0.413	< 2e-16
log (Volume)	0.42	0.052	6E-16	0.45	0.054	< 2e-16	0.36	0.058	3E-10	0.34	0.061	2E-08
Grade of VC												
Negative	0.31	0.098	0.0016	0.33	0.098	0.0003	0.24	0.096	0.014	0.36	0.094	0.0002
Positive	0.17	0.137	0.2164	0.07	0.149	0.3004	0.09	0.139	0.476	0.03	0.135	0.8272
Percent of HC	0.05	0.008	2E-10	0.04	0.009	1E-09	0.05	0.007	3E-09	0.04	0.007	2E-06
Length of HC	−2.66	0.496	8E-08	−2.21	0.463	7E-09	−2.21	0.471	2E-06	−2.13	0.439	1E-06
Average Speed	−0.09	0.026	0.0001	0.06	0.012	3E-05	−0.09	0.026	8E-04	0.07	0.017	2E-05
Std of Speed	0.13	0.022	2E-09	0.17	0.024	3E-10	0.15	0.022	2E-11	0.17	0.021	3E-07
AIC	5805			3708			5903			3826		
Injury Crashes	**Average 15 Minute Volume**			**Average Hourly Volume**			**Average 15 Minute Volume**			**Average Hourly Volume**		
	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)
Intercept	−5.15	0.678	< 2e-16	−7.25	0.612	0.0002	−8.19	0.902	< 2e-16	−6.33	0.691	< 2e-16
log (Volume)	0.37	0.086	2E-05	0.36	0.093	0.0001	0.32	0.097	9E-04	0.26	0.101	0.001
Percent of HC	0.04	0.015	0.0033	0.17	0.057	0.0618	0.04	0.014	0.006	0.21	0.053	0.001
Length of HC	−1.88	0.826	0.0023	−0.59	0.334	0.0066	−1.60	0.797	0.045	−0.68	0.329	0.037
Std of Speed	0.11	0.039	0.0035	0.19	0.042	7E-07	0.12	0.041	0.004	0.17	0.035	3E-06
Delta	−0.15	0.044	0.0002	−0.07	0.024	0.0003	−0.16	0.039	6E-05	−0.07	0.029	0.011
AIC	2368			1681			2399			1741		
Property Damage Crashes	**Average 15 Minute Volume**			**Average Hourly Volume**			**Average 15 Minute Volume**			**Average Hourly Volume**		
	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)
Intercept	−2.73	0.132	0.008	−7.77	0.463	< 2e-16	−8.04	0.406	< 2e-16	−6.72	0.492	< 2e-16
log (Volume)	0.44	0.063	2E-12	0.52	0.07	< 2e-16	0.37	0.069	1E-08	0.38	0.071	1E-07
Grade of VC												
Negative	0.38	0.122	0.0017	0.42	0.156	1E-05	0.35	0.112	2E-04	0.43	0.111	0.0001
Positive	0.26	0.168	0.1052	0.12	0.173	0.0504	0.23	0.163	0.158	0.19	0.156	0.2272
Percent of HC	0.06	0.011	2E-08	0.06	0.011	5E-08	0.05	0.009	2E-08	0.05	0.008	3E-07
Length of HC	−2.83	0.595	2E-06	−2.51	0.592	6E-08	−2.66	0.542	1E-06	−2.41	0.501	1E-06
Std of Speed	0.14	0.027	1E-07	0.16	0.027	7E-08	0.18	0.026	2E-12	0.17	0.023	7E-12
Delta	−0.07	0.028	0.008	−0.05	0.015	3E-06	−0.03	0.018	0.058	−0.09	0.019	1E-08
AIC	4406			2910			4493			2995		
Single Vehicle Crashes	**Average 15 Minute Volume**			**Average Hourly Volume**			**Average 15 Minute Volume**			**Average Hourly Volume**		
	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)
Intercept	−7.04	0.374	< 2e-16	−12.23	1.47	< 2e-16	−9.69	1.38	2E-12	−13.28	1.53	< 2e-16
log (Volume)	0.21	0.062	0.0015	0.11	0.06	0.0008	0.14	0.069	0.052	0.07	0.069	0.0005
Grade of VC												
Negative	0.55	0.132	3E-05	0.36	0.126	0.0004	0.53	0.129	6E-05	0.42	0.121	0.0005
Positive	0.33	0.177	0.0731	0.06	0.186	0.7654	0.36	0.181	0.047	0.08	0.173	0.658
Percent of HC	0.04	0.012	0.0003	0.04	0.012	2E-06	0.04	0.011	5E-04	0.04	0.009	8E-05
Length of HC	−2.66	0.682	9E-05	−2.93	0.678	6E-05	−2.42	0.642	2E-04	−2.91	0.589	6E-07
Average Speed	0.03	0.015	0.0263	0.11	0.021	4E-06	0.06	0.021	0.006	0.13	0.024	3E-08
Std of Speed	0.06	0.031	0.0073	0.13	0.035	0.0022	0.07	0.035	0.039	0.12	0.031	0.0002
AIC	3855			2575			3918			2684		
Multiple Vehicle Crashes	**Average 15 Minute Volume**			**Average Hourly Volume**			**Average 15 Minute Volume**			**Average Hourly Volume**		
	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)
Intercept	−3.53	0.609	< 2e-16	−3.94	0.893	< 2e-16	−1.29	2.066	0.005	−1.42	0.283	0.0005
log (Volume)	0.99	0.103	< 2e-16	1.05	0.107	< 2e-16	0.92	0.105	< 2e-16	0.95	0.111	< 2e-16
Percent of HC	0.05	0.011	1E-05	0.04	0.013	0.001	0.04	0.01	2E-05	0.02	0.011	0.0506
Length of HC	−2.15	0.701	0.0021	−1.76	0.806	0.0029	−1.73	0.623	0.006	−1.57	0.681	0.004
Average Speed	−0.12	0.033	0.0003	−0.12	0.041	8E-05	−0.15	0.031	2E-06	−0.14	0.034	3E-05
Std of Speed	0.22	0.029	8E-15	0.19	0.034	4E-06	0.22	0.028	4E-15	0.19	0.029	1E-10
Delta	−0.14	0.031	1E-05	−0.16	0.037	9E-06	−0.17	0.029	5E-09	−0.18	0.033	4E-07
AIC	3020			2005			3071			2053		

were used to validate and compare the models.

Table 9 shows the rural models that include speed parameters. For total crashes, speed was negatively related to crashes, meaning that lower average speed is correlated with higher crash frequency. Lower average speeds indicate the presence of congestion, so this relationship is intuitive for rural sites. These models also show that as standard deviation of hourly average speeds or 15-minute average speed increases, probability of crashes also increases.

For models based on level of injury, it was found that standard deviation is positively related to crashes for both injury and PDO crashes. The variable delta speed had a negative relationship for both types of crashes. Delta speed is defined as the difference between posted speed limit and average speed. A positive value of delta would mean average speed is lower than the speed limit, indicating congestion. A negative value, on the other hand, would represent free flow conditions. A negative relationship between this variable and property damage crashes means that this type of crashes increases when congestion increases. This is a logical relationship since during congestion, speed is lower so the probability of the crash being an injury crash is lower. The negative relationship between delta speed and injury crashes seems counter intuitive since higher speeds are generally associated with more severe injuries. This result could be due to how injury was defined, and the type of data used for modeling. Fatal and injury crashes were combined in this category and range from a crash being fatal to a minor injury that does not require any doctor or hospital visit. Separating fatal and severe injury crashes from minor injury crashes might shed some

Table 10
Parameter Estimates for Volume, Geometry and Flow Parameter Based Models for Urban Segments.

Total Crashes	Models with Detector Speed						Models with INRIX Speed					
	Average 15 Minute Volume			Average Hourly Volume			Average 15 Minute Volume			Average Hourly Volume		
	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)
Intercept	−10.59	0.755	< 2e-16	−4.87	0.347	< 2e-16	−7.64	0.665	< 2e-16	−4.81	0.338	< 2e-16
log (Volume)	0.55	0.051	< 2e-16	0.34	0.048	9E-13	0.58	0.052	< 2e-16	0.32	0.048	9E-12
Median Width	−0.21	0.167	< 2e-16	−0.22	0.366	< 2e-16	−0.15	0.208	< 2e-16	−0.28	0.278	< 2e-16
Std of Speed	0.08	0.009	< 2e-16	0.13	0.012	< 2e-16	0.11	0.013	< 2e-16	0.16	0.111	< 2e-16
Delta	0.11	0.013	< 2e-16	0.05	0.006	< 2e-16	0.03	0.011	0.0002	0.02	0.009	1.4E-05
AIC	7709			4264			8407			4887		
Injury Crashes	**Average 15 Minute Volume**			**Average Hourly Volume**			**Average 15 Minute Volume**			**Average Hourly Volume**		
	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)
Intercept	−11.14	1.34	< 2e-16	−6.62	0.594	< 2e-16	−8.07	1.16	< 2e-16	−6.07	0.539	< 2e-16
log (Volume)	0.51	0.089	1E-08	0.36	0.079	6E-07	0.44	0.088	6E-07	0.32	0.072	8E-06
Median Width	−0.13	0.128	1E-10	−0.25	0.123	< 2e-16	−0.17	0.146	2E-06	−0.19	0.135	< 2e-16
Std of Speed	0.07	0.017	2E-05	0.17	0.013	< 2e-16	0.12	0.018	3E-10	0.15	0.009	< 2e-16
AIC	3274			2059			3515			2328		
Property Damage Crashes	**Average 15 Minute Volume**			**Average Hourly Volume**			**Average 15 Minute Volume**			**Average Hourly Volume**		
	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)
Intercept	−11.07	0.857	< 2e-16	−3.22	0.497	< 2e-16	−8.59	0.764	< 2e-16	−3.58	0.617	5E-09
log (Volume)	0.57	0.059	< 2e-16	0.31	0.055	2E-08	0.64	0.061	< 2e-16	0.33	0.052	3E-10
Median Width	−0.17	0.112	< 2e-16	−0.16	0.011	1E-14	−0.13	0.011	< 2e-16	−0.22	0.068	< 2e-16
Average Speed	−0.06	0.013	6E-13	−0.07	0.052	5E-10	−0.11	0.012	< 2e-16	−0.03	0.085	2E-06
Std of Speed	0.08	0.011	9E-14	0.15	0.523	5E-10	0.09	0.012	0.0001	0.16	0.0118	< 2e-16
AIC	6288			3556			6869			4159		
Single Vehicle Crashes	**Average 15 Minute Volume**			**Average Hourly Volume**			**Average 15 Minute Volume**			**Average Hourly Volume**		
	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)
Intercept	−8.75	1.131	< 2e-16	−3.46	0.467	< 2e-16	−7.19	1.05	7E-13	−2.94	0.449	< 2e-16
log (Volume)	0.11	0.062	0.0007	0.47	0.066	5E-05	0.13	0.056	2E-05	0.49	0.065	8E-07
Std of Speed	0.06	0.018	0.0006	0.05	0.022	2E-06	0.08	0.016	0.0003	0.07	0.018	8E-06
Average Speed	0.09	0.022	2E-06	0.02	0.008	0.0003	0.07	0.018	3E-06	0.03	0.015	0.00009
AIC	3694			2411			3898			2681		
Multiple Vehicle Crashes	**Average 15 Minute Volume**			**Average Hourly Volume**			**Average 15 Minute Volume**			**Average Hourly Volume**		
	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)
Intercept	−13.94	0.988	< 2e-16	−7.79	0.486	< 2e-16	−11.06	0.867	< 2e-16	−7.74	0.464	< 2e-16
log (Volume)	1.02	0.077	< 2e-16	0.64	0.065	< 2e-16	1.07	0.574	< 2e-16	0.63	0.064	< 2e-16
Std of Speed	0.09	0.107	< 2e-16	0.15	0.014	3E-08	0.13	0.118	4E-14	0.19	0.013	< 2e-16
Delta	0.11	0.602	4E-11	0.15	0.006	6E-06	0.15	0.472	< 2e-16	0.11	0.011	2E-06
AIC	5619			3092			6204			3689		

light on the relationship. Even though this issue was not explicitly addressed in this paper, it could be an interesting area for future research. This relationship also might be specific to this particular dataset. This analysis was based on rural continuous count station data where the maximum hourly volume observed was 3822 vph across two lanes. Thus, these results may be driven by the fact that this dataset is dominated by locations that are often traveling near free flow and a broader variation in traffic speed is not expected.

For crashes involving single vehicle, average speed and standard deviation both were significant and followed an intuitive relation showing that as speed increases, single vehicle crashes increase. Increase in average speed means there is no congestion, which also means fewer vehicles on the road. As the number of vehicles on the road increases and speed decreases, the probability of multiple vehicles being involved in a crash also increases. Multiple vehicle crash models for rural segments showed this relationship as well.

The AIC value for models using INRIX speeds were worse than the continuous count station based models. This is expected since continuous count station data is based on speeds of all traffic, whereas the probe data estimates link speed based on a sample of vehicles less than the entire population. While the AIC values for the INRIX models were lower than the count station models, the goal here was to evaluate if inclusion of the INRIX data could improve crash prediction models as compared to models without traffic speed parameters. The prediction accuracy is discussed in the model validation section below. For all the models, parameters for speed related variables didn't vary much between two data sources. Since these two models essentially had the same data other than the speed component, this is an indication that the speed data from these two sources are not significantly different than each other in terms of their effect on the model.

Table 10 shows that the speed parameters showed consistent results for urban segments as well. Standard deviation of speed always had an increasing effect on crash frequency for all crash types. For total crashes, another significant flow parameter was delta speed. The models showed that crashes on urban segments increase as congestion increases.

For injury models, only standard deviation of speed was a statistically significant flow parameter. For property damage crashes, models indicated that crashes increase as average speed decreases, which is again a logical finding.

For crashes involving a single vehicle, urban segments showed similar results as rural segments where single vehicle crashes showed increasing trends with increases in average speed. A reverse trend was observed between congestion and multi vehicle crashes, where crash frequency increased with increasing congestion.

Similar to the rural models, these relationships were consistent irrespective of the level of aggregation. The model parameters for volume, geometry, and flow models based on continuous count data and the corresponding models based on INRIX data were very close to each other. The AIC value for INRIX models were worse than the detector data based models, which is again consistent with the findings from rural segments. In this case also, the main focus was to identify whether this drop in quality in models make a significantly lower prediction quality or not.

Table 11
Model Comparison for Urban Segments[*].

	Total Crashes								
	Average 15 min Volume			Average Hourly Volume			AADT		
	MAD	MAPE	MSPE	MAD	MAPE	MSPE	MAD	MAPE	MSPE
Volume and length models	8.22	58%	152.42	8.19	54%	140.59	8.65	59%	185.38
	(-5%)	(-1%)	(-18%)	(-5%)	(-5%)	(-24%)			
Volume, length, and geometry models	8.13	52%	140.56	7.72	40%	120.92	8.52	52%	181.91
	(-5%)	(0%)	(-23%)	(-9%)	(-12%)	(-34%)			
Volume, length, geometry, and flow state models **	7.87	45%	129.97	6.81	30%	112.47	–	–	–
	(-8%)	(-7%)	(-29%)	(-20%)	(-22%)	(-38%)			
Volume, length, geometry, and flow state models (INRIX) **	8.11	40%	138.81	6.82	33%	124.35	–	–	–
	(-5%)	(-12%)	(-24%)	(-20%)	(-19%)	(-32%)			
	Fatal & Injury Crashes								
	Average 15 min Volume			**Average Hourly Volume**			**AADT**		
	MAD	MAPE	MSPE	MAD	MAPE	MSPE	MAD	MAPE	MSPE
Volume and length models	2.81	39%	18.85	2.65	28%	12.13	2.79	31%	18.21
	(+1%)	(+8%)	(+4%)	(-5%)	(-3%)	(-33%)			
Volume, length, and geometry models	2.77	36%	15.14	2.56	21%	10.83	2.68	29%	14.94
	(+3%)	(+7%)	(+1%)	(-4%)	(-8%)	(-28%)			
Volume, length, geometry, and flow state models **	2.53	29%	12.68	2.44	19%	8.61	–	–	–
	(-6%)	(0%)	(-15%)	(-9%)	(-10%)	(-42%)			
Volume, length, geometry, and flow state models (INRIX) **	2.67	31%	12.73	2.34	24%	9.71	–	–	–
	(0%)	(+2%)	(-15%)	(-13%)	(-5%)	(-35%)			
	PDO Crashes								
	Average 15 min Volume			**Average Hourly Volume**			**AADT**		
	MAD	MAPE	MSPE	MAD	MAPE	MSPE	MAD	MAPE	MSPE
Volume and length models	6.53	47%	90.4	6.36	49%	80.97	6.85	49%	106.72
	(-5%)	(-2%)	(-15%)	(-7%)	(0%)	(-24%)			
Volume, length, and geometry models	6.41	44%	86.56	6.18	42%	71.72	6.78	47%	104.29
	(-5%)	(-3%)	(-17%)	(-9%)	(-5%)	(-31%)			
Volume, length, geometry, and flow state models **	6.23	41%	77.76	5.83	37%	63.14	–	–	–
	(-8%)	(-6%)	(-25%)	(-14%)	(-10%)	(-39%)			
Volume, length, geometry, and flow state models (INRIX) **	6.11	37%	81.09	5.95	38%	68.41	–	–	–
	(-10%)	(-10%)	(-22%)	(-12%)	(-9%)	(-34%)			
	Single Vehicle Crashes								
	Average 15 min Volume			**Average Hourly Volume**			**AADT**		
	MAD	MAPE	MSPE	MAD	MAPE	MSPE	MAD	MAPE	MSPE
Volume and length models	2.79	59%	13.79	2.71	55%	11.3	2.72	57%	12.19
	(+3%)	(+2%)	(+13%)	(0%)	(-2%)	(-7%)			
Volume, length, and geometry models	2.63	53%	11.31	2.59	47%	9.94	2.64	55%	11.72
	(+0%)	(-2%)	(-3%)	(-2%)	(-8%)	(-15%)			
Volume, length, geometry, and flow state models **	2.46	42%	10.88	2.44	39%	6.64	–	–	–
	(-7%)	(-13%)	(-7%)	(-8%)	(-16%)	(-43%)			
Volume, length, geometry, and flow state models (INRIX) **	2.61	51%	11.00	2.47	43%	9.01	–	–	–
	(-1%)	(-4%)	(-6%)	(-6%)	(-12%)	(-23%)			
	Multiple Vehicle Crashes								
	Average 15 min Volume			**Average Hourly Volume**			**AADT**		
	MAD	MAPE	MSPE	MAD	MAPE	MSPE	MAD	MAPE	MSPE
Volume and length models	6.66	44%	95.51	6.63	42%	82.56	6.76	44%	137.63
	(-1%)	(0%)	(-31%)	(-2%)	(-2%)	(-40%)			
Volume, length, and geometry models	5.58	42%	63.57	5.66	33%	71.08	6.37	44%	133.06
	(-12%)	(-2%)	(-52%)	(-11%)	(-11%)	(-47%)			
Volume, length, geometry, and flow state models **	5.51	35%	60.17	4.94	26%	57.81	–	–	–
	(-14%)	(-9%)	(-55%)	(-22%)	(-18%)	(-57%)			
Volume, length, geometry, and flow state models (INRIX) **	5.54	38%	63.01	5.37	33%	69.18	–	–	–
	(-13%)	(-6%)	(-53%)	(-16%)	(-11%)	(-48%)			

* Value in the parentheses represents the change compared to respective AADT based models.
** These models were compared to the AADT based volume, length, and geometry models.

6.5. Model comparison

Tables 11 and 12 shows the comparison of performance among the models developed. The AADT models didn't include speed as a variable because averaging hourly speed over a year did not capture the effect of speed on traffic conditions and crashes. For comparison purposes, the volume, flow, and geometry model was compared to the AADT based volume and geometry models. Model comparison provided a check on whether adding geometric and then flow parameters really improve model performance as expected. It also showed how different levels of data aggregation affect the performance. Finally, it shows whether the model performance is significantly different depending on the source of speed data. For all models, irrespective of type of facility, type of crash, or level of data aggregation, prediction accuracy consistently improved as geometric and then speed variables were added. This improvement was higher in magnitude across all validation MOEs when the speed component was added to the model.

The average 15 min volume models gave a mixed result in comparison to the AADT model. Even though it did perform better than the AADT based model most of the time, there were certain models (injury models for urban segments, single vehicle models for rural segments) when the model could not outperform the AADT models. It is possible that at a 15-min level data is too noisy to capture the true relationship between crashes and flow parameters. Likewise, inaccuracies in time

Table 12
Model Comparison for Rural Segments*.

	Total Crashes								
	Average 15 min Volume			Average Hourly Volume			AADT		
	MAD	MAPE	MSPE	MAD	MAPE	MSPE	MAD	MAPE	MSPE
Volume and length models	3.58	58%	30.63	3.52	70%	29.89	3.92	85%	34.06
	(-9%)	(-27%)	(-10%)	(-10%)	(-15%)	(-12%)			
Volume, length, and geometry models	3.47	58%	27.88	3.45	58%	25.24	3.62	76%	28.18
	(-4%)	(-18%)	(-1%)	(-5%)	(-18%)	(-10%)			
Volume, length, geometry, and flow state models **	3.37	54%	23.77	3.21	43%	20.11	–	–	–
	(-7%)	(-22%)	(-16%)	(-11%)	(-33%)	(-29%)			
Volume, length, geometry, and flow state models (INRIX) **	3.35	59%	22.07	3.24	48%	23.5	–	–	–
	(-7%)	(-17%)	(-22%)	(-10%)	(-28%)	(-17%)			
	Fatal & Injury Crashes								
	Average 15 min Volume			**Average Hourly Volume**			**AADT**		
	MAD	MAPE	MSPE	MAD	MAPE	MSPE	MAD	MAPE	MSPE
Volume and length models	1.31	47%	2.97	1.15	50%	2.83	1.22	62%	3.17
	(+7%)	(-15%)	(-6%)	(-6%)	(-12%)	(-11%)			
Volume, length, and geometry models	1.31	41%	2.86	1.13	40%	2.23	1.2	55%	2.94
	(+9%)	(-14%)	(-3%)	(-6%)	(-15%)	(-24%)			
Volume, length, geometry, and flow state models **	1.17	39%	2.47	1.09	33%	1.72	–	–	–
	(-3%)	(-16%)	(-16%)	(-9%)	(-22%)	(-41%)			
Volume, length, geometry, and flow state models (INRIX) **	1.17	37%	2.83	1.1	38%	1.85	–	–	–
	(-3%)	(-18%)	(-4%)	(-8%)	(-17%)	(-37%)			
	PDO Crashes								
	Average 15 min Volume			**Average Hourly Volume**			**AADT**		
	MAD	MAPE	MSPE	MAD	MAPE	MSPE	MAD	MAPE	MSPE
Volume and length models	2.89	67%	19.2	2.99	59%	22.45	3.07	93%	23.79
	(-6%)	(-26%)	(-19%)	(-3%)	(-34%)	(-6%)			
Volume, length, and geometry models	2.75	64%	17.46	2.87	55%	18.4	2.92	87%	22.7
	(-6%)	(-23%)	(-23%)	(-2%)	(-32%)	(-19%)			
Volume, length, geometry, and flow state models **	2.65	58%	12.84	2.77	52%	17.45	–	–	–
	(-9%)	(-29%)	(-43%)	(-5%)	(-35%)	(-23%)			
Volume, length, geometry, and flow state models (INRIX) **	2.7	60%	16.23	2.74	58%	17.1	–	–	–
	(-8%)	(-27%)	(-29%)	(-6%)	(-29%)	(-25%)			
	Single Vehicle Crashes								
	Average 15 min Volume			**Average Hourly Volume**			**AADT**		
	MAD	MAPE	MSPE	MAD	MAPE	MSPE	MAD	MAPE	MSPE
Volume and length models	2.15	85%	9.79	2.09	71%	6.08	2.05	83%	6.68
	(+5%)	(+2%)	(+47%)	(+2%)	(-12%)	(-9%)			
Volume, length, and geometry models	1.92	72%	6.43	1.88	62%	5.85	1.98	77%	6.42
	(-3%)	(-5%)	(0%)	(-5%)	(-15%)	(-9%)			
Volume, length, geometry, and flow state models **	1.89	65%	6.27	1.79	57%	5.8	–	–	–
	(-5%)	(-12%)	(-2%)	(-10%)	(-20%)	(-10%)			
Volume, length, geometry, and flow state models (INRIX) **	1.92	58%	6.38	1.83	60%	5.64	–	–	–
	(-3%)	(-19%)	(-1%)	(-8%)	(-17%)	(-12%)			
	Multiple Vehicle Crashes								
	Average 15 min Volume			**Average Hourly Volume**			**AADT**		
	MAD	MAPE	MSPE	MAD	MAPE	MSPE	MAD	MAPE	MSPE
Volume and length models	2.44	49%	15.13	2.51	53%	14.31	2.53	61%	15.05
	(-4%)	(-12%)	(+1%)	(-1%)	(-8%)	(-5%)			
Volume, length, and geometry models	2.31	47%	13.26	2.23	43%	12.87	2.41	54%	12.93
	(-4%)	(-7%)	(+3%)	(-7%)	(-11%)	(0%)			
Volume, length, geometry, and flow state models **	2.19	44%	11.25	2.03	34%	11.31	–	–	–
	(-9%)	(-10%)	(-13%)	(-16%)	(-20%)	(-13%)			
Volume, length, geometry, and flow state models (INRIX) **	2.23	41%	10.03	2.19	39%	12.09	–	–	–
	(-7%)	(-13%)	(-22%)	(-9%)	(-15%)	(-6%)			

* Value in the parentheses represents the change compared to respective AADT based models.
** These models were compared to the AADT based volume, length, and geometry models.

stamps of crash reports could influence results at that level. The prediction accuracy improved significantly for all models when average hourly data was used. In this case, the aggregation interval was not too disaggregated to capture the random nature of crashes, also not too aggregated to lose the variation in traffic. These models consistently performed better than the AADT based model for all MOEs.

For the rural hourly volume, geometry, and flow models, MAD, MAPE, and MSPE improved by 11%, 33% and 29% respectively while using continuous count station as speed data source and 10%, 28%, and 17% respectively when INRIX speed data was used. For the urban models, similar trends was observed where MAD, MAPE and MSPE improved by 20%, 22%, and 38% respectively for detector data and 20%, 19% and 32% for INRIX data. In both cases, these models were compared to AADT based volume and geometry models. This trend was consistent for all other models as well. Even though models using INRIX data performed slightly worse than the ones based on continuous count data, they were still far better than AADT based models.

The comparison results reinforce the importance of selecting an appropriate disaggregation level. Due to the random nature of crash occurrence, the 15 min data had too much variability to generate useful models. Similarly, aggregated models that rely on AADT may fail to capture variations traffic flow that could influence safety. Finding a proper disaggregation level as well as significant variables that influence crash frequency is one of the major concerns in the area of crash prediction modeling. The improved model accuracy came from a combination of both the choice of time interval and also the use of traffic flow information. Another very important finding is that speed variables played a significant role in model performance irrespective of their source. This essentially opens up the possibility to extend the analysis to sections without a continuous count station. Since current models only rely on volume, quality of volume data dictates the quality of model. This research showed that INRIX data can be used as an alternate source for speed data without reducing the quality of crash prediction models. Models developed using INRIX data performed very similarly to the continuous count station data during model validation. For all crash types and model categories, INRIX models consistently outperformed AADT models. This indicates that it may be possible to use INRIX data along with historic volume distributions from short duration count stations (where only volume data is collected intermittently) to run this analysis on a larger scale.

7. Conclusions and future research

Crash prediction models have been a major focus for researchers in the field of traffic safety. Past research examining the influence of traffic speed relied on data from point detectors, hence limiting the coverage. This study developed a general relationship that accounts for both hourly speed and volume on freeway segments in Virginia. The results indicated that inclusion of variables, like speed, standard deviation, and difference in speed could create improvements in the quality of crash prediction models. As availability and reliability of observed traffic data significantly affect the accuracy of the study, using probe data, which has better network coverage, might be useful to improve the availability and quality of data. This research showed that average hourly volume profiles could be coupled with hourly speed to generate better crash predictions even when the speed data does not come from a continuous count station. This finding is important since INRIX data has been used in numerous studies related to traffic operations so far, but their application in safety research has been limited.

Future research could use average hourly volume distributions derived from short-term counts in combination with probe data to make the methodology developed in this paper more broadly applicable. Future work in this direction would consider year to year correlation in the data that was not addressed in this paper.

Additionally, models that assess safety when traffic control and cross section are changed dynamically have not been estimated previously. Currently, there is no existing methodology for safety assessment of facilities with dynamic traffic control or geometry such as part time shoulder use or variable speed limits (Dutta et al., 2019; Gonzales and Fontaine, 2018). Another gap in current research is in the area of work zone safety. Work zones are only active for a portion of the day, and it is important to know how the timing of lane closures impacts safety. The HSM provides crash modification functions that account for the effects of project length and duration on crash frequency but do not allow for explicit comparisons of safety effects of daily lane closures (Kweon et al., 2014). Further extension of the approaches developed in this paper could enable more proactive analysis of work zone impacts while the traffic management plans are in the planning stage.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Akaike, H.A., 1974. New look at the statistical model identification. IEEE Trans. Automat. Contr. 19 (6), 716–723.
American Association of State Highway Transportation Officials, 2010. Highway Safety Manual.
Choudhary, P., Imprialou, M., Velaga, N.R., Choudhary, A., 2018. Impacts of speed variations on freeway crashes by severity and vehicle type. Accid. Anal. Prev. 121, 213–222. https://doi.org/10.1016/j.aap.2018.09.015.
Dutta, Nancy, Fontaine, M., 2018. Developing Rural Four Lane Freeway Crash Prediction Models Using Hourly Flow Parameters. 2018 (*Under Review for IATSS Research*). .
Dutta, N., Boateng, R., Fontaine, M.D., 2019. Safety and operational effects of the interstate 66 active traffic management system. J. Transp. Eng. Part A Syst. 145. https://doi.org/10.1061/JTEPBS.0000220.
Gang-Len, C., Xiang, H., 2003. The Relationship Between Congestion Levels and Accidents. University of Maryland-College Park Publication MD-03-SP 208B46.
Garber, N.J., Gadiraju, R., 1989. Factors affecting speed variance and its influence on accidents. Transportation Research Record: Journal of the Transportation Research Board 1239, 11–19.
Garber, N.J., Ehrhart, A., 2000. The Effect of Speed, Flow, and Geometric Characteristics on Crash Rates for Different Types of Virginia Highways. Virginia Transportation research Council Publication VTRC 00-R15.
Geedipally, S.R., Pratt, M.P., Lord, D., 2017. Effects of geometry and pavement friction on horizontal curve crash frequency. J. Transp. Saf. Secur. https://doi.org/10.1080/19439962.2017.1365317.
GIS Integrator, Virginia Department of Transportation http://integrator/. Accessed on July 2017.
Golob, T.F., Recker, W.W., Alvarez, V.M., 2004. Freeway safety as a function of traffic flow. Accid. Anal. Prev. 36, 933–946. https://doi.org/10.1016/j.aap.2003.09.006.
Gonzales, D., Fontaine, M.D., 2018. Evaluation of the Safety Effects of the I-77 Active Traffic and Safety Management System: Phase 2.FHWA/VTRC 19-R6. Virginia Department of Transportation.
Graham, J.L., Harwood, D.W., O'Laughlin, M.K., Donnell, E.T., Brennan, S., 2014. Median Cross-Section Design for Rural Divided Highways. NCHRP Report 794. Transportation Research Board of the National Academies.
Haghani, A., Hamedi, M., Sadabadi, K.F., 2009. I-95 Corridor Coalition Vehicle Probe Project: Validation of INRIX Data. I-95 Corridor Coalition.
Highway Traffic Records Information System (HTRIS), 2019. Traffic Engineering Division, Virginia Department of Transportation.
Imprialou, M.I.M., Quddus, M., Pitfield, D.E., Lord, D., 2016. Re-visiting crash-speed relationships: a new perspective in crash modelling. Accid. Anal. Prev. 86, 173–185. https://doi.org/10.1016/j.aap.2015.10.001.
Ivan, J.N., Wang, C., Bernardo, N.R., 2000. Explaining two-lane highway crash rates using land use and hourly exposure. Accid. Anal. Prev. 32, 787–795.
Khan, G., Bill, A., Chitturi, M., Noyce, D., 2013. Safety evaluation of horizontal curves on rural undivided roads. Transportation Research Record: Journal of the Transportation Research Board 2386, 147–157. https://doi.org/10.3141/2386-17.
Knuiman, M., Council, F., Reinfurt, D., 1993. Association of median width and highway accident rates. Transportation Research Record: Journal of the Transportation Research Board 1401, 70–82.
Kononov, J., Hersey, S., Reeves, D., Allery, B., 2012. Relationship Between Freeway Flow Parameters and Safety and Its Implications for Hard Shoulder Running. Transportation Research Record: Journal of the Transportation Research Board 2279, 118–123. https://doi.org/10.3141/2280-02.
Kweon, Y.-J., Lim, I.K., Fontaine, M.D., 2014. Work Zone Safety Performance Measures for Virginia. Publication FHWA/VTRC 16-R10. Virginia Transportation research Council.
Lord, D., Manar, A., Vizioli, A., 2005. Modeling crash-flow-density and crash-flow-V/C ratio relationships for rural and urban freeway segments. Accid. Anal. Prev. 37, 185–199. https://doi.org/10.1016/j.aap.2004.07.003.
Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transp. Res. Part A Policy Pract. 44, 291–305. https://doi.org/10.1016/j.tra.2010.02.001.
McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, second edition. Chapman and Hall, Ltd., London.
Persaud, B., Dzbik, L., 1993. Accident prediction models for freeways. Journal of the Transportation Research Board: Transportation Research Record 1401, 55–60.
Quddus, M., 2013. Exploring the relationship between average speed, speed variation, and accident rates using spatial statistical models and GIS. J. Transp. Saf. Secur. 5 (1). https://doi.org/10.1080/19439962.2012.705232.
Roadway Network System, Virginia Department of Transportation http://rns/app/. Accessed on July 2017.
Shankar, V., Chayanan, S., Sittkariya, S., Shyu, M., Ulfarsson, G., Juvva, N.K., 2004. Median Crossover Accident Analyses and Effectiveness of Median Barriers. Publication WA-RD 591.1. Washington State Transportation Center (TRAC).
Shaw-pin, M., 1994. The relationship between truck accidents and geometric design of road sections: poisson versus negative binomial regressions. Accid. Anal. Prev. 26, 471–482.
Solomon, David, 1964. Accidents on Main Rural Highways Related to Speed, Driver, and Vehicle. U.S. Department of Transportation, Federal Highway Administration.
Tanishita, M., Van Wee, B., 2017. Impact of vehicle speeds and changes in mean speeds on per vehicle-kilometer traffic accident rates in Japan. Iatss Res. 41, 107–112. https://doi.org/10.1016/j.iatssr.2016.09.003.
Tegge, R.A., Jo, J.-H., Ouyang, Y., 2010. Development Ans Application of Safety Performance Function in Illinois. Publication No ICT-R27-20. Illinois Center for Transportation.
Traffic Engineering Division, Virginia Department of Transportation. http://tedweb/tms/jsp/. Accessed on July 2017.
Vuong, Q.H., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57, 307. https://doi.org/10.2307/1912557.
Wang, L., Abdel-aty, M., Wang, X., Yu, R., 2018. Analysis and comparison of safety

models using average daily, average hourly, and microscopic tra fic. Accid. Anal. Prev. 111, 271–279. https://doi.org/10.1016/j.aap.2017.12.007.

Washington, S.P., Matthew, G., Karlaftis, F.M., 2010. Statistical and Econometric Methods for Transportation Data Analysis, second edition. Chapman and Hall/CRC.

Watson, D.C., Al-kaisy, A., Anderson, N.D., 2014. Examining the effect of speed, roadside features, and roadway geometry on crash experience along a rural corridor. J. Mod. Transp. 22, 84–95. https://doi.org/10.1007/s40534-014-0040-4.

Yu, R., Quddus, M., Wang, X., Yang, K., 2018. Impact of data aggregation approaches on the relationships between operating speed and traffic safety. Accid. Anal. Prev. 120, 304–310. https://doi.org/10.1016/j.aap.2018.06.007.

Zegeer, C., Stewart, R., Reinfurt, D., Council, F., Neuman, T., Hamilton, E., Miller, T., Hunter, W., 1990. Cost Effective Geometric Improvements for Safety Upgrading of Horizontal Curves. FHWA-RD-90-021, Federal Highway Administration.