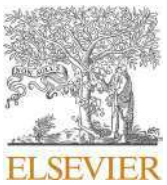
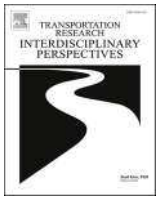
Transportation Research Interdisciplinary Perspectives 24 (2024) 101051



Contents lists available at ScienceDirect

Transportation Research Interdisciplinary Perspectives

journal homepage: www.sciencedirect.com/journal/transportation-research-interdisciplinary-perspectives



Extracting driving volatility from connected vehicle data in exploring Space-Time relationships with crashes in the city of Saint Louis

Abdul Rashid Mussah *, Yaw Adu-Gyamfi

Civil and Environmental Engineering, University of Missouri, W1024 Lafferre Hall, Columbia 65211, MO, United States

A R T I C L E I N F O

Keywords:
Connected Vehicle
Driving Volatility
Traffic Safety
Human Centered Analysis
Space-time analysis

A B S T R A C T

The analysis of factors that influence the occurrence of roadway crashes within a specified locality have historically been reliant on the assessment of physical infrastructure, historical crash frequency, environmental factors and driver characteristics. The consensus over the years has been drawn to the idea that human factors, specifically regarding driving behaviors, account for the majority of crash outcomes on our roadways. With the emergence of connected vehicle data in the last few years, the capacity to analyze real time driving behavior has become a possibility for safety analysts. Driving volatility has emerged as a valuable proxy for driving behavior and indicator of safety. In this study, evidence of the spatial relationship between driving volatility and historical crash hotspots is uncovered. Utilizing an entropy-based analysis, this study discovered generally strong positive spatial relationships between locations of volatile driving events and historical crashes, with R^2 values ranging from 0.015 to 0.970 and a mean of 0.612 for hard accelerations, and 0.048 to 0.996 and a mean of 0.678 for hard decelerations. Including temporal context presented insights showing that the relationships are significant for over 60 % of the coverage area usually between the hours of 7 am to 7 pm, with average R^2 values of 0.594 for hard accelerations, and 0.629 for hard decelerations.

Introduction

Roadway travel is still by far one of the most preferred modes of transport as reported by the National Household Travel Survey, as Americans take 1.1 billion trips per day of which 87 percent is carried out using personal vehicles (Santos et al., 2011). Effective and efficient roadway travel options are thus the lifeblood of the social and economic landscape for the average American. Assessing the crash risk implication of the roadway network in real time presents a valuable framework for making sure that roadway users are protected during their daily commute. This framework can include, but is not limited to, a more proactive and rapid response to crashes and the situations that allow them to occur or escalate. Missourians are estimated to incur a cost of 3.3 billion dollars each year as the hidden cost to operating and maintaining a roadway vehicle (TRIP, 2010). Another 4.8 billion dollars are incurred yearly as the safety cost for roadway travel as a result of motor vehicle crashes, which includes medical costs, lost workplace and household productivity, property damage, congestion costs and legal costs (Blincoe et al., 2015, TRIP, 2010). This high cost to travel is a great burden to the economic situation of the state and the people. Dealing with the issues that degrade the safety and efficiency of the state's roadway network remains paramount. The field of transportation safety has evolved over the years into a more systems approach domain, which not only investigates seeking safety through vehicle or roadway design and operational changes, but also relying on behavioral changes of roadway users in their daily interactions within the transportation network system. The National Highway Traffic Safety Administration (NHTSA) with support from many state departments of transportation, have moved towards adopting the Vision Zero initiative which aims to achieve a highway system with no fatalities or serious injuries involving road traffic. At the heart of the Vizion Zero initiative is the human centered systems approach, with efforts being made over the years to collect comprehensive data on driver behavior. Programs such as the second Strategic Highway Research Program (SHRP-2) and the Ann Arbor Connected Vehicle Test Environment (AACVTE) have provided an incredible amount of data on roadway user interactions, which include but is not limited to, kinematic vehicle data, driver perspective video data, as well as driver-facing video feed. With this renewed focus on exploring the intricacies of the different forms of interactions between roadway users and the safety implications thereof, many researchers

* Corresponding author.
E-mail addresses: akm2fx@mail.missouri.edu (A. Rashid Mussah), adugyamfiy@missouri.edu (Y. Adu-Gyamfi).

https://doi.org/10.1016/j.trip.2024.101051
Received 23 June 2022; Received in revised form 24 December 2022; Accepted 21 February 2024
Available online 1 March 2024

have defined and explored new metrics which can be used in identifying the needs and areas of improvements of our transportation networks. This study will begin with a literature review which will investigate the existing work carried out in this new domain of traffic and transportation safety, and then follow up with the research contributions.

Literature review

Driving volatility, which describes rapid changes in driving kinematics, as a proxy indicator of driving behavior has been extensively researched for its real-world safety implication (Wang et al., 2014, Liu et al., 2017, Wali et al., 2019). With the understanding that minute details in the way individuals drive can be extracted in real time and measured against a baseline to determine the possible consequence of any sort of deviations from the norm due to the emergence of connected vehicle (CV) data, a major direction of research has focused on advancing vehicle to vehicle (V2V) and vehicle to infrastructure (V2I) technologies (Osman and Ishak, 2015). Other areas of research have focused primarily on accounting for these deviations, by introducing safety systems that can counteract them, such as lane keep assist (Guo et al., 2021), brake assist and cooperative adaptive cruise controls (Milanés et al., 2013). The most prevalent systems designed to assess and characterize driver behavior usually factor for over speeding, although much focus has now been drawn to other behaviors such as hard accelerations and decelerations, both in the longitudinal and lateral directions of motion, as well as jerk, a measure of the rate of change of acceleration with time (Kamrani et al., 2017).

Given the advancements of V2I and V2V technologies in the last decade, CV data is now available in real time to aid roadway management and safety agencies to make real world, instantaneous decisions on developing situations on the roadway network. This has opened the way for researchers to not only be able to look at the specific situations which lead to safety critical outcomes at the macro-level, but also micro-level analysis of individual behaviors that contribute to a wider range of issues within the transportation system, such as vehicle emissions (Ferreira et al., 2021). The importance of the collective consequence of driving behaviors on the transportation network system cannot be overstated as many studies have drawn correlates between definable and quantifiable metrics of driving instances and crash outcomes (Wang et al., 2014, Liu et al., 2017, Wali et al., 2019, Wali and Khattak, 2020).

With the emergence of large quantities of vehicle kinematic data, driving video data and driver characteristic data, insights into abnormal driving can be separated into two main fields. The first approach deals with information on driver actions based on video feeds and other forms of sensors that collect information which describe their behaviors during their trips. One such study is the distracted driving behavior study carried out by and expanded on by Halley Sutton (2017), analyzed data from over 3 million drivers, who carried out up to 570 million trips, covering 5.6 billion miles of travel. Findings indicated that drivers used their phones in up to 88 percent of the trips analyzed, a worrying statistic. Other studies such as Papazikou et al. (2017, 2019) having extensively investigated a combination of both video data and the corresponding vehicle kinematics from naturalistic driving data, were able to look into how specific behaviors reflected in the signal from the kinematic information. The other approach to driver behavior analysis involves analyzing only kinematic information to quantify instantaneous driving decisions. With a large enough pool of driver kinematic information, it then becomes possible to define thresholds of deviations from normal driving. Wang et al. (2014), and Liu et al. (2017) have explored how vehicle trajectory information can be harnessed at both the individual trip level, and over the whole transportation network to define volatile driving behaviors, which have huge safety implications. The conclusions drawn from these studies provide insight into how specific forms of connected vehicle data can be harnessed in the push to improve on the safety of transportation network system. In considering driving volatility from streaming data as a viable surrogate of crash detection and prediction signals, it is paramount that the spatial and temporal nature of the assessed variables under consideration to be taken into account.

Analysis of spatial approaches in road safety has been carried out in recent years given that spatial heterogeneity and the transferability of defined model have become important. Ziakopoulos and Yannis (2020) carried out an extensive study looking into factors that influence the outcome of spatial analysis such as the boundary problem, the modifiable areal unit problem and spatial proximity. This investigation also presented the advantages and disadvantages of the different spatial modeling approaches including statistical, bayesian and machine learning approaches. Ziakopoulos (2021) followed up with an investigation of the spatial analysis of hard braking (deceleration) events in urban environments, concluding that spatial autocorrelation is detected in harsh event frequencies. Unobserved heterogeneity and spatial autocorrelation have thus been established as an important factor to investigate in the analysis of crash data and crash surrogates (Cai et al., 2021, Ziakopoulos et al., 2022, Wei et al., 2021).

Consequently, temporal correlation cannot be discounted when spatial analysis of crashes is being investigated. analysis of crash data, where variables allow for, have also provided much useful insight into the changing dynamics of the influence of contributory factors to crash events, with respect to when and where they occur. Wen et al. (2019) understanding the effects of the dynamic nature of weather factors, explored the use of models to measure their associations with crash events. The study showed significant correlations in the interactions of many weather-related variables and geometric factors on the outcome of crashes on freeways. The implications of the findings of all the studies presented above draw a vivid presentation of the importance of investigating the spatial, as well as temporal nature of crash risk factors on crash outcome to ensure accuracy in the outcomes of potential future analyses.

Research objectives and contribution

Like most states, the state of Missouri collects and manages a comprehensive roadway crash database. Roadway crash data is invaluable in the analysis and assessment of the safety implications of the roadway network, and the decision-making process to improve on it. The usual historic precedent with roadway safety analysis is reliant on modeling variables that dwell with the interaction of the physical infrastructure, environmental factors and human behavior. Safety treatments thus tend to be usually reactive and applied when crash data is gathered over a long period of time. The Turner-Fairbank Research Center reports that in over 90 percent of roadway crashes, driver behavior serves as the critical reason (FHWA, 2021). Driving behaviors which have been labeled as aggressive constitute a large proportion of these situations that lead to potentially dangerous outcomes on our roadway. Aggressive driving consists of, following too closely, driving at excessive speeds, weaving through traffic, and running stop lights and signs, among other acts. With the advent of connected vehicle (CV) data that provides real time driving metrics, driving volatility has emerged as a useful proxy for driving behavior that deviates from the norm. Volatility can be quantified in multiple ways from speeding, aggressive or hard acceleration and deceleration, as well as other dispersion from generally acceptable driving behaviors. The challenge therefore in the pursuit of a proactive safety system, which can identify and mitigate latent risks in the transportation system before they occur, rather than reacting afterwards, begins with exploring the different available real time variables and selecting which proxy indicators of driving behavior correlate best with crash occurrences. The focus of this paper is to explore the spatial and temporal correlations of volatility as a proxy of driving behavior, and crashes, within the transportation network that overlays Saint Louis City of the state of Missouri. By analyzing the number of historical crash events in conjunction with the number of observed volatile events and specific location, insights into the

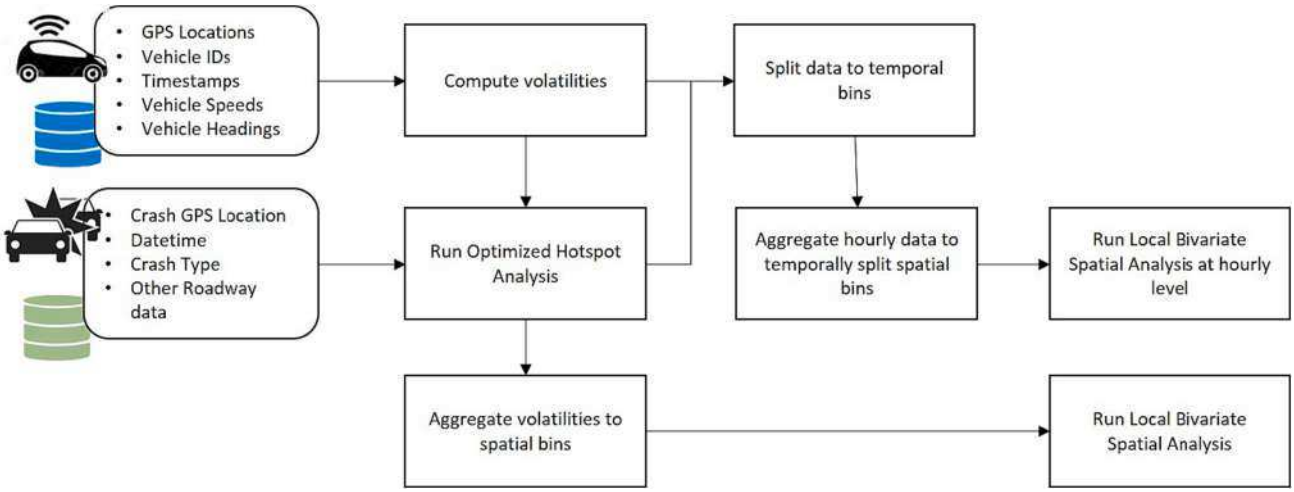


Fig. 1. Conceptual Framework of Analysis.

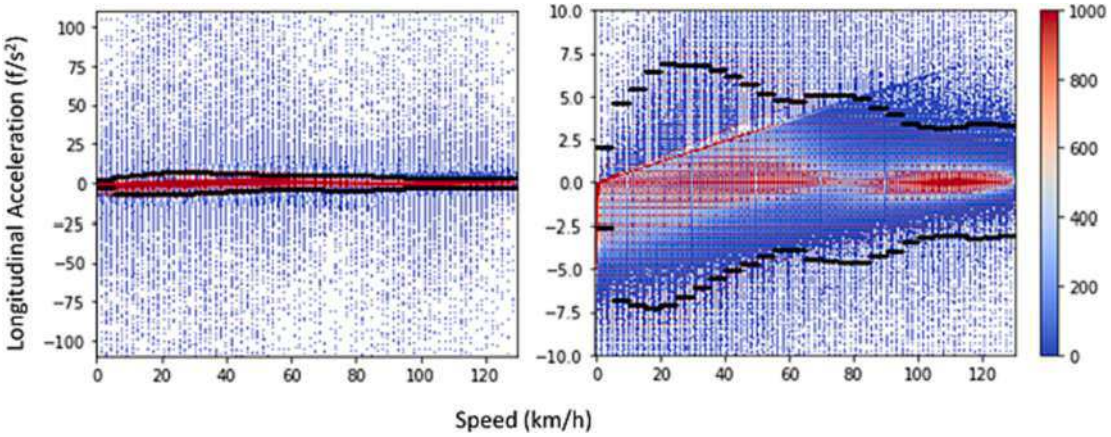


Fig. 2. Scatter density plots (at two levels of magnification) of calculated acceleration and deceleration 2-sigma thresholds defined by 5 km/h speed bins (black bars).

relationship between the two occurrences can be gleamed. Expanding the analysis further within an hourly context also allows us to observe the changes in the development of the relationships and their magnitudes at different times of the day. The potential benefit of this study will help researchers by presenting variables which will aid in developing proactive safety systems and a more efficient response to safety critical outcomes on the roadways, towards an effective allocation of risk reduction and alleviation resources.

Methodology and data

Conceptual framework

With ten years' worth of geocoded historical crash data provided by the Missouri Department of Transportation (MoDOT) between the years of 2010 to 2019, and 2 months' worth of CV data for roughly 373,000 unique vehicle IDs at a resolution of 3 s, sourced from Otonomo technologies, a car data service platform, the general spatial and temporal context of where and when crashes usually occur, as well as where and when specific driving behaviors are observed on the roadway network can be scrutinized in deeper context to explore areas for potential necessary safety modifications, as well as proactive countermeasures.

This study begins with an initial optimized analysis to define areas of statistically significant spatial clusters of high crash counts, as a means of narrowing down the area of observation if necessary. Similarly, the local analysis tool in ArcGIS Pro software (v. 2.8.3) is utilized to quantify the relationship between areas of where crash occur and whether it is dependent on or influenced by areas where different volatility measures occur as well as if these relationships vary over geographic space within the Saint Louis City boundary.

With the initial spatial exploratory analysis carried out, the next stage of this study investigates exploring the additional temporal context of when these behaviors are most prevalent whilst keeping their spatial context of where they happen in mind. It is hypothesized that certain confounding factors, such as the changing density of traffic at different times of the day (exposure) and lighting conditions, have a relatively large effect on the outcome of the number of observed volatile driving events, as well as crash outcomes, which can be accounted for to some degree with an inclusion of the temporal context. Exploring the spatial and temporal correlation of these variables under observation will allow us to be more exact in analyzing the dynamic relationship of their influence and how that can be leveraged to improve the safety of our roadways. The conceptual framework of how the analysis was carried out in this paper is presented in Fig. 1.

Data

With ten years' worth of geocoded historical crash data, numbering 122,151 crashes, provided by the Missouri Department of Transportation (MoDOT) between the years of 2010 to 2019, and 2 months' worth of CV data, numbering over 36 million data points, from roughly

Table 1
Descriptive Statistics of Key Variables from Connected Vehicle (CV) data.

	All CV Data			Hard Accelerations			Hard Decelerations		
N =	36,522,722			281,876			804,901		
Variable	Mean	SD	Min/Max	Mean	SD	Min/Max	Mean	SD	Min/Max
Speed (km/hr)	39.34	40.41	0/129.98	58.85	26.46	1/129.94	17.05*	25.86	0/129.64*
Longitudinal Acc/Dec (ft/sec)	–23.14	2.2	−109.83/109.83	8.25	8.37	2.05/109.83	5.64*	5.4	2.55*/109.82*
Lateral Acc/Dec (ft/sec)	2.62	18.63	−29.07/134.38	0.75	5.42	–23.88/134.38	0.32	5.57	−26.25/138.38

NB: negative values of lateral acceleration denote movement from right to left
* Computed absolute values of deceleration magnitudes

Table 2
Crash Data for Saint Louis City for 2010–2019.

	All Crashes	Crashes by Year								
		2010	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**
N =	122,151	10,626	10,165	8,530	8,847	10,290	13,969	15,577	14,774	13,992	15,385
Crash Type											
Rear End	29,829	2,452	2,380	2,157	2,358	2,575	3,537	3,774	3,620	3,326	3,620
Parking or Parked Car	18,127	1,675	1,619	1,171	1,169	1,611	2,125	2,305	2,179	1,986	2,263
Out of Control	15,606	1,565	1,396	988	1,123	1,381	1,723	1,861	1,824	1,817	1,909
Passing	13,200	827	814	931	925	1,015	1,487	1,718	1,828	1,652	1,988
Right Angle	13,034	1,029	1,062	887	915	1,018	1,416	1,677	1,648	1,683	1,681
Roadway Geometry Type											
Straight	114,425	9,902	9,571	7,901	8,169	9,552	13,089	14,623	13,876	13,180	14,562
Curve	7,545	702	577	622	648	719	854	927	881	804	811

373,000 unique vehicle IDs at a resolution of 3 s, sourced from Otonomo technologies, a car data service platform, the general spatial and temporal context of where and when crashes usually occur, as well as where and when specific driving behaviors are observed on the roadway network can be scrutinized in deeper context to explore areas for potential necessary safety modifications, as well as proactive countermeasures. As the data streamed from the connected vehicles are highly granular and do not provide information specifically on the volatility metrics, large scale processing of the data is carried out as the first step.

Exploratory spatial and temporal analysis

The analysis begins with exploring the spatial and temporal nature of the crashes and volatility metrics. Understanding that different hours of the day account for different volumes of traffic on the roadway, and as such different driving behaviors is necessary to our analysis. The different environmental factors, such as lighting conditions, also influence the behaviors of drivers (Yannis et al., 2013). As such, plotting the locations of where and when crashes and volatilities happen will allow us to find a unique perspective about the roadway network.

Driving volatility

With the advantage of the real time stream of connected vehicle data, instantaneous driving volatility can be defined and measured. In this study, as speed data in the longitudinal direction is available at a resolution of every 3 s, allowing us to calculate the rate of change of speed (acceleration/deceleration) as well in real time, by finding the difference between the current state speed s_t and previous state speed s_{t-1} over the time interval t for every unique vehicle, during every unique trip carried out by the vehicle.

A threshold for normal acceleration and deceleration behavior is then computed for different speed bins within 2 standard deviations of the observed values. Value points that exceed the threshold values are thus labeled as highly volatile events, with positive values defined as hard accelerations and negative values defined as hard decelerations. Fig. 2 below shows the scatter density plot of the values of acceleration or decelerations against the respective vehicle speed value and bounded by the defined thresholds set to separate normal driving from volatile driving.

Exploring the plot in detail, we notice that for most vehicles driving at speeds of up to 40 km/h (~25 mph), acceleration and deceleration rates of up to 7.5 feet per second squared are observed, this falls within reasonable range of vehicles traveling on low-speed roadway segments such as local and minor roads/arterials. As the vehicle speeds increase, the general acceleration and deceleration rates are seen to be within a lowered range of 5 feet per second squared, indicating that progressive braking is more preferred approach for safety and comfort reasons.

Hot Spot analysis

The initial spatial exploratory analysis carried out in this study involves identifying statistically significant hot spots and cold spots using the Getis-Ord Gi* statistic. Hot spots and cold spots are defined as areas of significant clusters of high values and low values respectively. The local sum for a feature and its neighbors is compared proportionally to the sum of all features, a statistically significant z-score results when the local sum is very different from the expected local sum, and when that difference is too large to be the result of random chance (Getis and Ord, 1992, Anselin, 1995). False Discovery Rate (FDR) correction is applied to account for multiple testing and spatial dependency and control the proportion of false positive results (type 1 errors). The general equation for the Getis-Ord local statistic is presented below by equation (1).

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j}x_j - \bar{X}\sum_{j=1}^{n} w_{i,j}}{S\sqrt{\frac{n\sum_{j=1}^{n} w^2_{i,j} - \left(\sum_{j=1}^{n} w_{i,j}\right)^2}{n-1}}}$$

where x_j is the attribute value for feature j, w_{ij} is the spatial weight between feature i and j, n is equal to the total number of features such that

$$\bar{X} = \frac{\sum_{j=1}^{n} x_j}{n}$$

and

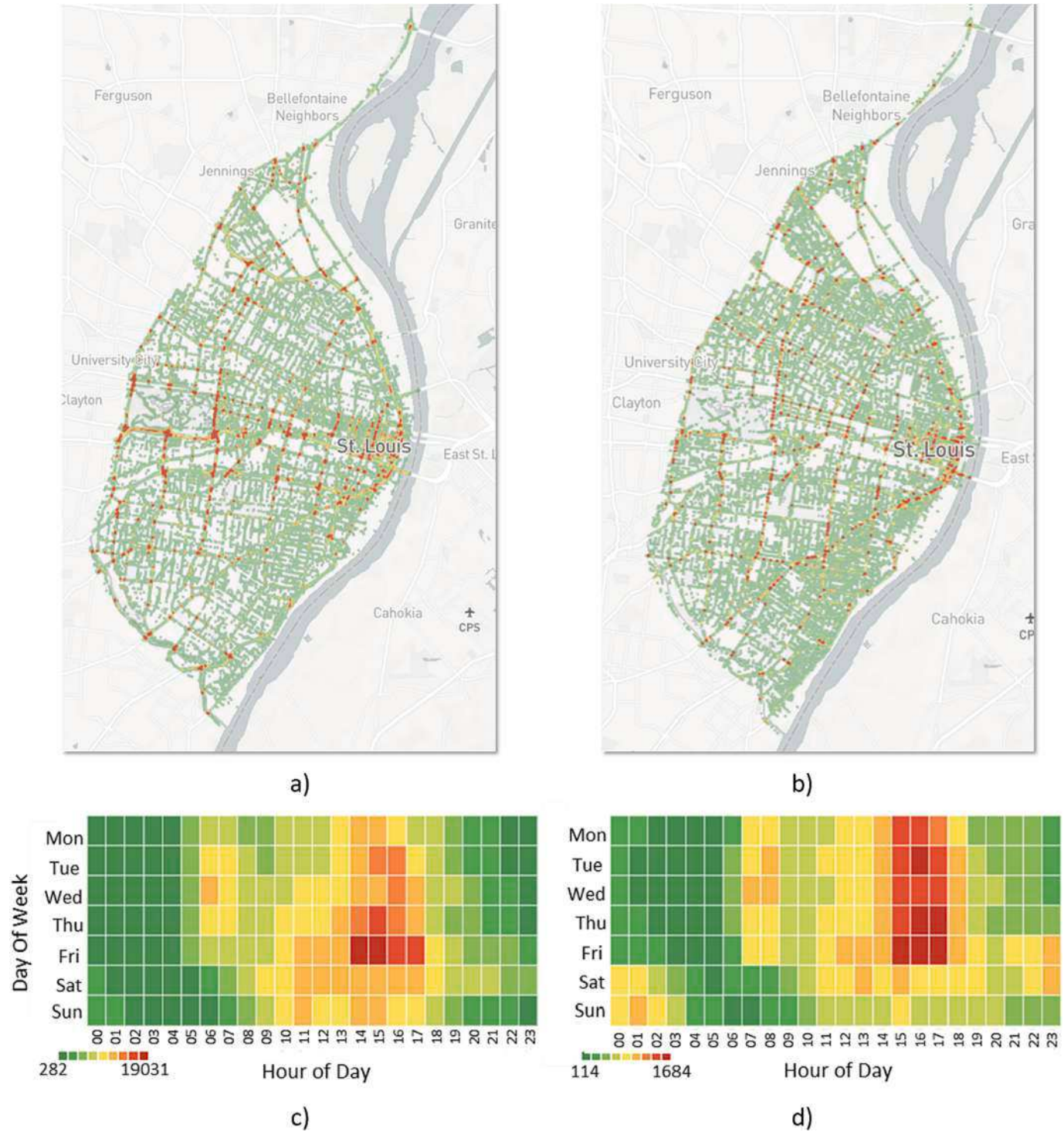


Fig. 3. Spatial Density plot of a) volatility locations and b) crash locations. As well as corresponding heatmaps showing the temporal distribution of both c) volatile driving and d) crashes at different hours of the day for different days of the week. Note a strong spatial and temporal relationship between crashes and highly volatile events.

$$S = \sqrt{\frac{n\sum_{j=1}^{n} x^2{}_j}{n} - (\bar{X})^2}$$

Given that the Gi* statistic is a z-score, no further calculations are thus required.

Entropy based local spatial analysis

Given that the relationship between two variables may change over geographical space, either in regression coefficients or relational form, it begs that specific local relationships at a given location need to be analyzed independently over defined spatial boundaries. The approach carried out in this study is explained in detail by Guo (2010), capitalizing on an approximation of the Renyi entropy for the multivariate data in each locally defined geographical region. According to the work done

Table 3
Crash Optimized Hotspot Analysis Summary.

Initial Data Assessment	
There are 122,151 valid input features.	
There were 1432 outlier locations; these will not be used to compute the hexagon size. The total study area is 171441572.1358 Meters.	
Incident Aggregation	
Using a hexagon of width 308.3050 Meters and height 267.0000 Meters, the aggregation process resulted in 4745 weighted polygons.	
Incident Count Properties	
Min	0
Max	650
Mean	25.74
Std. Dev.	51.73
Scale of Analysis	
The optimal fixed distance band is based on peak clustering found at 1254.1454 Meters	
Hot Spot Analysis	
There are 1887 output features statistically significant based on an FDR correction for multiple testing and spatial dependence. 0 % of features had less than 8 neighbors based on the distance band of 1254.1454 Meters	

by Guo, each value of local entropy is converted to a p-value by comparing to a distribution of permutation entropy values for the same region. The final few steps in the process involve processing the p-values for each region using several statistical tests to control for the multiple-testing problem, and then mapping locations of statistical significance with the definition of the relationship form between the variables under observation.

For a real-valued d-dimensional data space R^d, the entropy is defined by equation (4) as:

$$H_\lambda = \frac{1}{1-\lambda}\log\left(\int_{R^d} f(x)^\lambda dx\right), \quad \lambda \geq 0, \lambda \neq 1$$

where x is a d-dimensional vector, $f(x)$ is the probability density function, and $\lambda \geq 0$ is the order of the Renyi entropy.

The outcome of the process as defined by Guo (2010) is a non-parametric approach that defines spatially varying multivariate relations which can take many different functional forms. A 'good' multivariate relationship exists when one variable is dependent on the other, and as such one may be used to predict the other. Some common relationship expected from the entropy based local bivariate spatial analysis, are linear relationships, quadratic polynomials, and in some cases, complex functions (Guo, 2010).

Results and discussion

Descriptive statistics

Table 1 represents the descriptive statistics of the CV data with the key variables employed in the analysis. The general statistics such as the mean, standard deviation, minimum and maximum values of the variables help in presenting a context to the general nature of the distributions of the selected variables. From the table it can be inferred that hard decelerations tend to occur more frequently on the roadways, this is consistent with literature (Wang et al., 2014, Kamrani et al., 2017). The standard deviation, a measure of variability, is smaller for the hard decelerations than hard accelerations. Several factors may account for this observation as deceleration occur more frequently on roadways, especially with regards to interactions with traffic control measures such as stop signs and traffic signals. From the context of driver safety, the instinctive response to a safety critical situation for most drivers is to slow down or come to a stop, in order to avoid serious bodily injury or harm. Further exploration of the data shows that of the 372,721 vehicle IDs present in the data set, 40,391 (10.84 %) were involved in hard acceleration events and 146,580 (39.33 %) were involved in hard deceleration events. A total of 150,952 (40.5 %) of the vehicle IDs were involved in a highly volatile driving event. Table 2 presents the general overview of the nature of the roadway crash events for the last 10 years in Saint Louis City. Most of the crashes happen on straight roadway segments, and of the top five most frequent crash types, rear end crashes were the most prevalent.

Analysis results

To practically examine whether any type of relationship exists between the two variables under observation, spatial bins are defined to aggregate the events of crashes and volatilities to a quantifiable context. The process carried out begins first with plotting the locations of all crash events and volatility events (Fig. 3), and as observed, there exists strong spatial correlation between the two variables. Optimized Hotspot analysis is carried out using the ArcGIS Pro software (v. 2.8.3) in order to create spatial hexagons for which the crash and volatility events will then be aggregated into. The Optimized Hotspot analysis also allows us

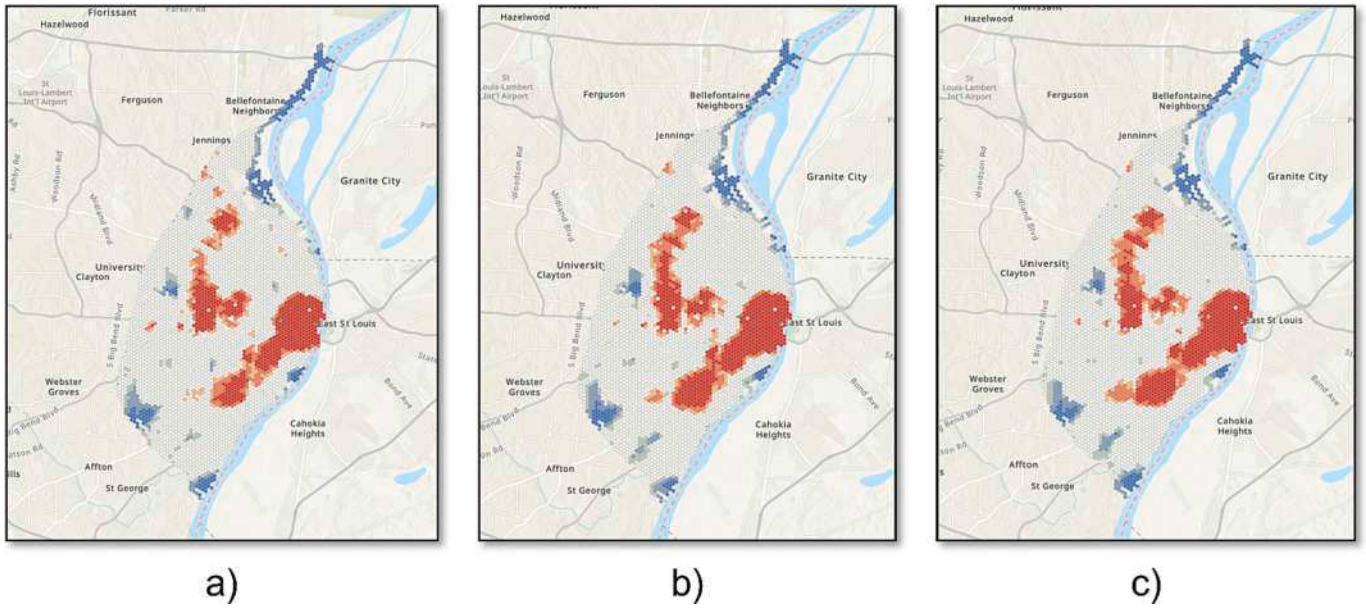


Fig. 4. Comparison of Optimized Hotspot Analysis for (a) 1-year of crash data (b) 5-years of crash data and (c) 10-years of crash data, for the City of St. Louis.

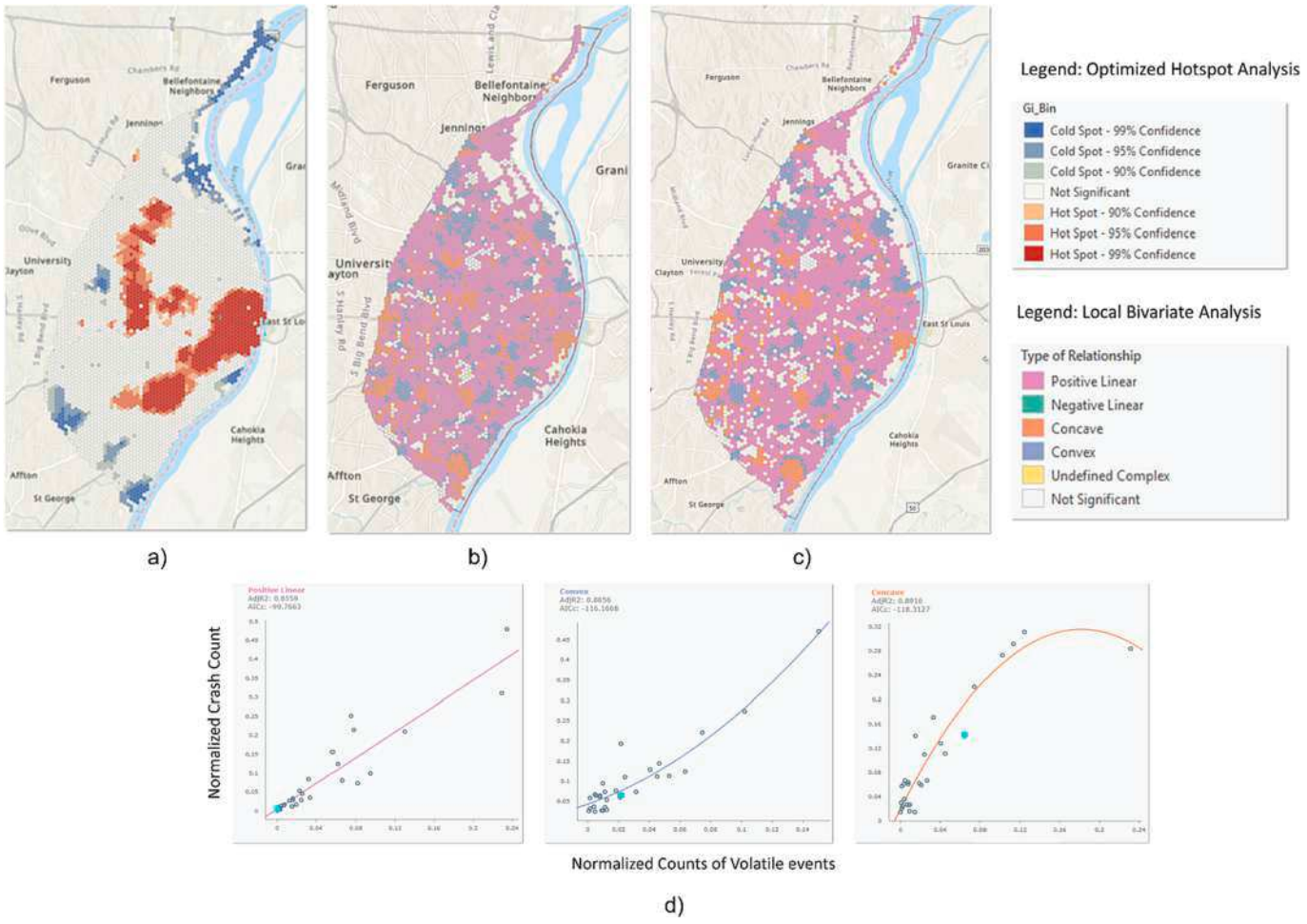


Fig. 5. Analysis Results. Locations of (a) statistically significant crash hotspots and coldspots; and bivariate relationships of crash counts with (b) volatile decelerations and (c) volatile accelerations from CV data with (d) examples of underlying observed linear and quadratic relationships.

Table 4
Local Bivariate Analysis Results for Hard Accelerations and Hard Deceleration with Crash counts.

Hard Accelerations and Crash Count					Hard Decelerations and Crash Count				
CategoricalSummary	# features	% features			CategoricalSummary	# features	% features		
Positive Linear	2333	67.64			Positive Linear	2368	63.21		
Negative Linear	0	0			Negative Linear	0	0		
Concave	417	12.09			Concave	763	20.37		
Convex	523	15.16			Convex	533	14.23		
Undefined Complex	5	0.14			Undefined Complex	0	0		
Not Significant	171	4.96			Not Significant	82	2.19		
Total	3449	100			Total	3746	100		
Entropy Results Summary	Min	Max	Mean	Median	Entropy Results Summary	Min	Max	Mean	Median
Entropy	0.26	1.19	0.66	0.64	Entropy	0.22	1.12	0.54	0.52
p-value	0.005	0.875	0.0223	0.005	p-value	0.005	0.605	0.0126	0.005
R^2	0.015	0.970	0.612	0.629	R^2	0.048	0.996	0.678	0.708
FDR Comparison	# significant	% significant			FDR Comparison	# significant	% significant		
Without FDR	3288	95.33			Without FDR	3664	97.81		
FDR	3278	95.04			FDR	3664	97.81		

to explore the spatial relationship of crash hotspots within the area of interest. The results of the analysis are presented in Table 3 below.

Given that the crash data utilized in this study is a multiyear data, in order to account for temporal variations in the travel patterns, geometric improvements, operational improvements as well as vehicle technology, the Optimized Hotspot Analysis is carried out on 1 year and 5 years' worth of crash data to explore whether changing patterns in the general hotspot locations will be observed as influenced by the aforementioned factors. Fig. 4 below is a side-by-side comparison of the Optimized Hotspot Analysis for the different crash data collection ranges discussed. It can be observed that the general locations of the hotspots and coldspots do not change much irrespective of how many years of data is used. The difference in intensity and quantity of the hotspots and coldspots can be attributed to the difference in quantity of crashes. The number of

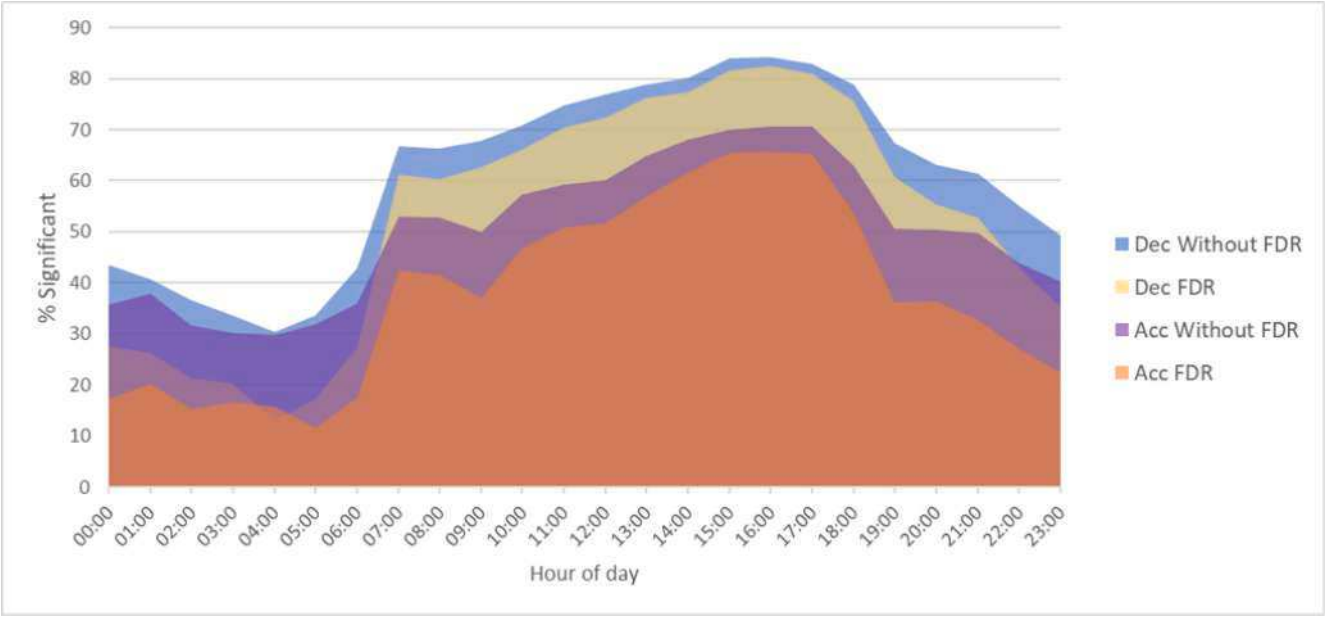


Fig. 6. Space time analysis percentage significant plot. Note the observed change in percentage significance of spatial association of the local bivariate analysis results at different hours of the day.

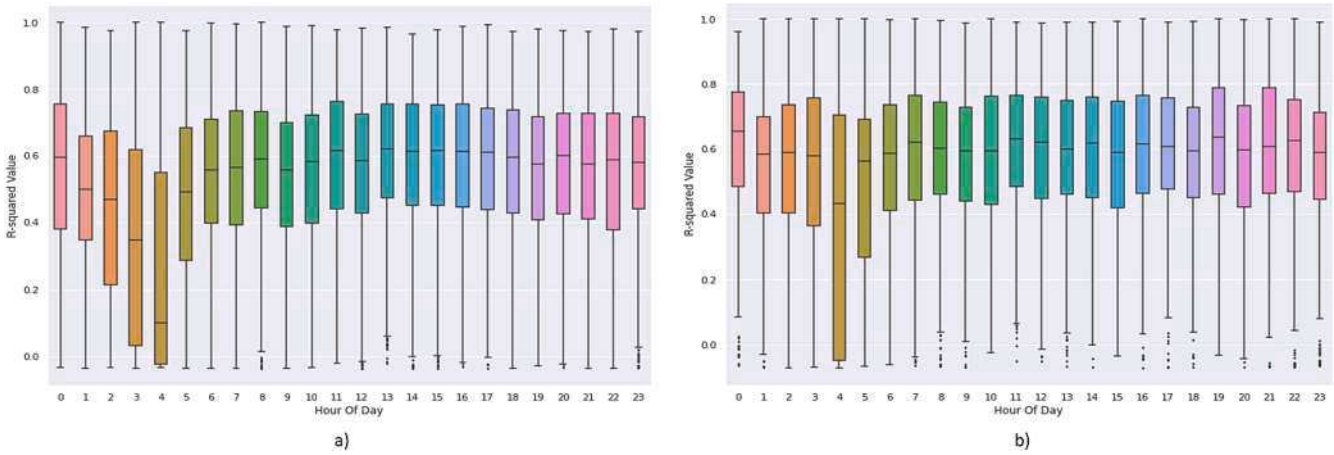


Fig. 7. Boxplots of R^2 value distribution of defined spatial associations between crashes and (a) hard accelerations, as well as (b) hard decelerations at different hours of the day.

crashes observed within the 10-year period is 122,151, whilst that for the 5-year period is 73,697 and for the 1-year period we had a total of 15,385 crashes.

The outcome of the Optimized Hotspot analysis created 4745 spatial hexagons with crash data aggregated within raging from values of a minimum of 0 crashes to a maximum of 650 crashes for the 10 years' worth of crash data utilized in the final analysis. Statistically significant clusters of hotspots and coldspots are observed in 1887 of the spatial hexagons, accounting for 39.77 % of the area of interest and this can be observed in Fig. 5 (a) below. The next step, as discussed, is to employ the local bivariate analysis of counts of hard accelerations, as well as hard decelerations, with crash counts within the defined spatial hexagons. The results of these analyses are observed in Fig. 5 (b) and (c) and summarized in Table 4 below. Hard decelerations show higher statistically significant relationships with crash counts, with a coverage area of 97.81 % of locations within the study area of observed data.

The relationships observed within the spatial bins are mostly positive linear in 63.21 % for hard decelerations within 3746 spatial bins, and 67.64 % for hard accelerations within 3449 spatial bins. High values of the coefficient of determination (bivariate correlation) for the observed relationships, denoted by R^2, which provide a measure of how well observed outcomes are replicated based on the proportion of total variation of outcomes explained by the defined relationship were observed. Further observation of the output Fig. 5(b) and 5(c), for both analyses, show that the relationships where they exist, are similar in the spatial bins for both volatility metrics. Finally, the analysis involved including temporal context in exploring how these observed relationships change at different hours of the day, as influenced by assumed confounders such as exposure (higher traffic) and lighting conditions. Fig. 6 below defines the changing values of the percentage significance of the analysis as carried out for data for different hours of the day. As expected, significant relationships between the variables and crash counts are highest during hours of higher economic activity and traffic volumes. Fig. 7 presents a visual representation of the change in R^2 value for different hours of the day, and as speculated, the association between locations of volatile driving and crash events are not only more prevalent during hours of high activity, but also show significantly greater values in the range of correlation coefficients.

In summary, the analysis presented in this study shows that during daylight hours, a larger percentage of the study area shows a significant defined relationship between volatility and crashes with very strong correlation, and this is important in helping define and develop models to access the safety situation of roadways more efficiently. Studies that have employed volatility metrics, especially hard braking/deceleration in safety analyses conclude with high certainty the positive correlation these metrics and crash outcomes (Wen et al., 2019, Desai et al., 2021, Hunter et al., 2021, Ziakopoulos et al., 2022, Mussah et al., 2022). Comparatively, Desai et al. (2021) concluded with an R^2 of 0.85 and significant positive correlation between hard braking and crash outcomes. Hunter et al. (2021) utilized 4.5 years of rear-end crash data and one month of hard braking data at 8 intersections to conclude that strong correlations occurred in specific locations at particular times of day. Unlike the earlier mentioned studies, this study explores both spatial and temporal correlation across the whole expanse of the defined study area.

Conclusion

Connected vehicle data presents a new era for traffic safety analysis, with realtime information from vehicles in motion. Driving volatility has emerged as a useful indicator of driver behavior and given that driver behavior accounts for the greater majority of causes of roadway crashes, it is important to explore how we can capitalize on CV data to draw on developing proactive safety countermeasure to developing situation on our roadways. This study contributes by analyzing and exploring the importance of the spatially varying multivariate relationships between defined driving volatility metrics and roadway crashes, as well as how these relationships vary in intensity over different hours of the day.

Creating a unique data set that provides realtime driving context on the roadway, with regards to driving behavior, from over 36 million data points from 373,000 unique vehicle IDs, the potential of exploring the landscape of how these behaviors relate with areas of crash occurrences is explored in this study. A simple measure of driving volatility is quantified as the 2-sigma rate of change of driving speed over time both in the positive and negative domains and are labeled as hard accelerations and hard decelerations respectively. The provided spatial information from both historically archived crash data collected over ten years, and the generated volatility data set, reveals an observable spatial correlation between the locations of where both sets of events occur. Employing an entropy based spatial analytical framework, the underlying spatially varying bivariate relationships between the defined driving volatility metric and roadway crashes is computed. To partially account for unobserved confounding influence of some safety critical variables, particularly the effect of exposure (intensity of traffic) and lighting conditions, the temporal context of the dynamism of the bivariate relationships is also explored. The analysis shows that driving volatility has a strong association with crash location occurrences and can serve as a great indicator of realtime crash risk. With the application of statistical, bayesian and machine learning approaches that can factor spatial and temporal influence into the analysis, hard decelerations are becoming a very useful and preferred proxy for exploring realtime crash risks of roadway segments (Ziakopoulos et al., 2022, and Adu-Gyamfi 2022).

The implication of this study draws on the need to provide proactive safety warnings to drivers about potential safety critical situations in specific areas of the transportation network, through advanced warning systems transmitted via connected vehicle technologies. Future studies will look into defining crash risk at a roadway segment level capitalizing on the insights of this study to develop a realtime crash risk dashboard that can help drivers and traffic management organizations respond to situations on the roadways as they develop, in a more efficient and effective manner.

CRediT authorship contribution statement

Abdul Rashid Mussah: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Visualization. **Yaw Adu-Gyamfi:** Conceptualization, Writing – review & editing, Software, Supervision, Project administration, Funding acquisition.

Declaration of competing 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.

Data availability

The authors do not have permission to share data.

Acknowledgments

This study done in this paper is supported by the Missouri Department of Transportation, United States through their collaboration with the Civil and Environmental Engineering Department of the University of Missouri, Columbia, United States.

References

Anselin, L., 1995. Local indicators of spatial association - LISA. Geogr. Anal. 27, 93–115.
Blincoe, J. L., Miller, T. R., Zaloshnja, E., Lawrence, B. A., and Washington, DC: National Highway Traffic Safety Administration. 2015. The economic and societal impact of motor vehicle crashes, 2010. (Revised).
Cai, Z., Wei, F., Wang, Z., Guo, Y., Chen, L., Li, X., 2021. Modeling of low visibility-related rural single-vehicle crashes considering unobserved heterogeneity and spatial correlation. Sustainability 13, 7438.
Desai, J., Li, H., Mathew, J.K., 2021. Correlating hard-braking activity with crash occurrences on interstate construction projects in Indiana. J. Big Data Anal. Transp 3, 27–41.
Ferreira, E., Fernandes, P., Bahmankhah, B., Coelho, M.C., 2021. Micro-analysis of a single vehicle driving volatility and impacts on emissions for intercity corridors. Int. J. Sustain. Transp. 1–23.
(FHWA), F. H. A. 2021. Office Of Research, Development, And Technology at The Turner-Fairbank Highway Research Center.
Getis, A., Ord, J.K., 1992. The analysis of spatial association by use of distance statistics. Geogr. Anal. 24, 189–206.
Guo, D., 2010. Local entropy map: a nonparametric approach to detecting spatially varying multivariate relationships. Int. J. Geogr. Inf. Sci. 24, 1367–1389.
Guo, Y., Zhang, H., Wang, C., Sun, Q., Li, W., 2021. Driver lane change intention recognition in the connected environment. Physica A 575, 126057.
Hunter, M., Saldivar-Carranza, E., Desai, J., 2021. A proactive approach to evaluating intersection safety using hard-braking data. J. Big Data Anal. Transp 3, 81–94.
Kamrani, M., Wali, B., Khattak, A.J., 2017. Can data generated by connected vehicles enhance safety? Proactive approach to intersection safety management. Transp. Res. Rec. 2659, 80–90.
Liu, J., Khattak, A., Wang, X., 2017. A comparative study of driving performance in metropolitan regions using large-scale vehicle trajectory data: implications for sustainable cities. Int. J. Sustain. Transp. 11, 170–185.
Milanés, V., Shladover, S.E., Spring, J., Nowakowski, C., Kawazoe, H., Nakamura, M., 2013. Cooperative adaptive cruise control in real traffic situations. IEEE Trans. Intell. Transp. Syst. 15, 296–305.
Mussah, A., Rashid, Y., Adu-Gyamfi, 2022. Machine learning framework for real-time assessment of traffic safety utilizing connected vehicle data. Sustainability 14, 15348.
Osman, O.A., Ishak, S., 2015. A network level connectivity robustness measure for connected vehicle environments. Transportation Research Part c: Emerging Technologies 53, 48–58.
Papazikou, E., Quddus, M., and Thomas, P. 2017. Detecting deviation from normal driving using SHRP2 NDS data.
Papazikou, E., Quddus, M., Thomas, P., Kidd, D., 2019. What came before the crash? An investigation through SHRP2 NDS data. Saf. Sci. 119, 150–161.
Santos, A., Mcguckin, N., Nakamoto, H. Y., Gray, D., Liss, S., and United States. Federal Highway Administration 2011. Summary of travel trends: 2009 national household travel survey (No. FHWA-PL-11-022).
Sutton, H., 2017. Study finds unsafe driving practices run rampant near campuses. Campus Security Report 14, 9.
The Road Information Program (TRIP) 2010. Missouri Transportation by the Numbers.
Wali, B., Khattak, A.J., 2020. Harnessing ambient sensing & naturalistic driving systems to understand links between driving volatility and crash propensity in school zones-a generalized hierarchical mixed logit framework. Transportation Research Part c: Emerging Technologies 114, 405–424.

Wali, B., Khattak, A.J., Karnowski, T., 2019. Exploring microscopic driving volatility in naturalistic driving environment prior to involvement in safety critical events-concept of event-based driving volatility. Accid. Anal. Prev. 132, 105277.
Wang, X., Khattak, A.J., Liu, J., Masghati-Amoli, G., Son, S., 2014. What is the level of. instantaneous driving decisions? Transportation Research Part c: Emerging Technologies 58, 413–427.
Wei, F., Cai, Z., Wang, Z., Guo, Y., Li, X., Wu, X., 2021. Investigating rural single-vehicle crash severity by vehicle types using full bayesian spatial random parameters logit model. Appl. Sci. 11, 7819.
Wen, H., Zhang, X., Zeng, Q., Sze, N.N., 2019. Bayesian spatial-temporal model for the main and interaction effects of roadway and weather characteristics on freeway crash incidence. Accid. Anal. Prev. 132, 105249.
Yannis, G., Kondyli, A., Mitzalis, N., 2013. Effect of lighting on frequency and severity of road accidents. Proceedings of the Institution of Civil Engineers-Transport 166, 271–281.
Ziakopoulos, A., 2021. Spatial analysis of harsh driving behavior events in urban networks using high resolution smartphone and geometric data. Accid. Anal. Prev 157, 106189.
Ziakopoulos, A., Vlahogianni, E., Antoniou, C., Yannis, G., 2022. Spatial predictions of harsh driving events using statistical and machine learning methods. Saf. Sci. 150, 105722.
Ziakopoulos, A., Yannis, G., 2020. A review of spatial approaches in road safety. Accid. Anal. Prev. 135, 105323.