DRIVER BEHAVIOR INDICES FROM LARGE-SCALE FLEET TELEMATICS DATA AS SURROGATE SAFETY MEASURES

Patrick Alrassy[1], Andrew W. Smyth[2], Jinwoo Jang[3]

Abstract

With the rapid proliferation of connected vehicle technologies, large-scale telematics data enable a high-resolution inference of road network's safety conditions and driver behavior. Although many researchers have investigated how to define meaningful safety surrogates and crash predictors from telematics, no comprehensive study analyzes the driver behavior derived from large-scale telematics data and relates them to crash data and the road networks in metropolitan cities. This study extracts driver behavior indices (e.g., speed, speed variation, hard braking rate, and hard acceleration rate) from large-scale telematics data, collected from 4,000 vehicles in New York City five boroughs. These indices are compared to collision frequencies and collision rates at the street level. Moderate correlations were found between the safety surrogate measures and collision rates, summarized as follows: (i) When normalizing crash frequencies with traffic volume, using a traffic AADT model, safety-critical regions almost remain the same. (ii) The correlation magnitude of hard braking and hard acceleration varies by road types: hard braking clusters are more indicative of higher collision rates on highways, whereas hard acceleration is a stronger hazard indicator on non-highway urban roads. (iii) Locations with higher travel times coincide with locations of high crash incidence on non-highway roads. (iv) However, speeding on highways is indicative of collision risks. After establishing the spatial correlation between the driver behavior indices and crash data, two prototype safety metrics are proposed: *speed corridor maps* and *hard braking and hard acceleration hot-spots*. Overall, this paper shows that data-driven network screening enabled by connected vehicle technology has great potential to advance our understanding of road safety assessment.

Keywords: Safety surrogate measures, Collisions data, Telematics, Connected Vehicles Technology, Smart Cities

[1]P. Alrassy holds a Ph.D. degree from the Department of Civil Engineering and Engineering Mechanics, Columbia University, New York, NY, 10027 USA, and currently works as a Data Engineer at Meta Platforms, Inc. (e-mail: patrick.alrassy@columbia.edu; smyth@civil.columbia.edu).

[2]A.W Smyth is a Professor of Civil Engineering and Engineering Mechanics, Columbia University, New York, NY, 10027 USA, (e-mail: smyth@civil.columbia.edu).

[3]J. Jang is an Assistant Professor of the Department of Civil, Environmental and Geomatics Engineering, Florida Atlantic

1. Introduction

Road infrastructure in metropolitan cities has dynamically developed to meet the needs of growing traffic and mobility. However, a substantial increase in traffic has resulted in a growing number of accidents and fatalities (Rodrigue, 2017). Therefore, city planners and policy makers have endeavored to monitor unsafe areas for motor vehicles, pedestrians and bikers. According to the U.S. Department of Transportation Federal Highway Administration (2016), the first step in the roadway safety management program is network screening, which mostly cuts down the list of hot-spots to a manageable list. Planners select a performance measure for analyzing the safety performance of each site. The most common safety performance measure is historical collision data. Improvement interventions are classically implemented at locations that historically had a relatively high frequency of collisions. However, existing collision databases are prone to have errors, omissions (Stipancic et al., 2018b) and underreporting (Kockelman & Kweon, 2002). Moreover, collision data are relatively of small sample sizes (Lord, 2006). Therefore, city planners are shifting towards adopting larger objective performance data when making improvements to road design and signalization. Mobile sensing has expanded to many application domains, such as intelligent transportation systems that offer major implications to the traffic engineering community (Amin et al., 2019). Traffic sensing infrastructure evolved from the use of fixed radars and manual collection to the use of inductive loop detectors, video cameras and on-board diagnostic devices mounted on the vehicles. New technologies have made the data collection task simpler and cost efficient (Eren et al., 2012).

Telematics sensors are currently being installed in many vehicle fleets, enabling, through their real-time and wide coverage, data-assisted traffic management and safety assessement. they include important vehicle-centric data, collected from extra sensor modules (e.g., On-Board Diagnostics (OBD-II)), able to obtain engine speed, Time-to-Collision (TTC), hard braking and acceleration events, in addition to GPS data. The data are then uploaded to storage servers (**?**) via continuous cellular connection (Rémy et al., 2012) or delayed Wi-Fi connection (Ramamoorthy et al., 2014). Several driver behavior indices can be derived from the OBD-collected data and have been used in the literature as safety surrogates. But to the best of the authors' knowledge, there has been no comprehensive study of driver behavior analytics and their spatial relationship with crashes, based on large-scale telematics data, crash data, and complex urban road network. Cai et al. (2020) studied the correlations between high-incidence locations for aberrant driving behaviors and locations of road traffic accidents based on vehicle OBD data. However only selected arterials and avenues from a mountainous city are considered. Also, the driving behavior data is recorded as a string,

University, FL, 33431, (e-mail: jangj@fau.edu).

for instance a hard acceleration event is reported as "Rapid acceleration" instead of a g-value. Yannis et al. (2016) propose monitoring driver traffic and safety behavior through OBD data, by focusing on the causality between harsh driving and probability of an accident, rather than on the spatial relationship between the two variables. Similarly, Ellison et al. (2015) introduce driver behavior profiles as an approach for evaluating driver behavior as a function of the risk of the casualty crash. The data is collected using GPS devices, which are prone to excessive noise.

For a performance measure to serve as a safety surrogate, it should be correlated with the outcome: collision frequency and collision rate in our context (Tarko et al., 2009). The main focus of this work is to understand the existence or absence of spatial correlation between the proposed safety surrogate measures and crashes, and not on inferring the causality between the two. Thus, this research work examines the direction and magnitude of the spatial correlation between driver behavior indices (speed, speed variation, hard braking rate, and hard acceleration rate), and collision data (i.e., absolute collision counts, and collision rate normalized with respect to traffic volume). The driver behavior indices are derived from real-world big telematics data, for the 2015-2016-year period, collected from in-vehicle sensing devices, mounted on 4,500 city-owned vehicles, in the New York City area. The vehicles are managed by the New York City Department of Citywide Administrative Services (NYC DCAS). Only the light-duty vehicles were considered in this study, to better represent the normal public driving population. A map-matching engine developed by the authors in Alrassy et al. (2019) was used to match the telematics data to the road segments. This study validates that the telematics data represent the general population traffic patterns, by comparing the OBD-II speeds with spot speeds radar data.

This work contributes to the existing literature from four angles: methodology, data quality, application, and findings. The behavioral big data are unique in terms of the road network type (New York city dense road network) and road network wide coverage (analysis carried across the entire city). Unlike the current research work, which mostly relies on GPS data to estimate speed, hard acceleration and hard deceleration parameters are recorded directly from the CAN bus through the OBD-II connection. To the best of the authors' knowledge, no comprehensive study analyzes the differences of driver behavior indices as safety surrogate measures both on highways and in dense urban network. Few studies focus on understanding the spatial relationship between harsh driving and collisions, but rather focus on the causality between the two variables. This paper highlights several key findings: (i) When normalizing crash frequencies at intersections with exposure, hot regions almost remain the same. (ii) The correlation magnitude of hard braking and acceleration varies by road type: hard braking is more indicative of collision rates on highways than hard

acceleration, whereas hard acceleration is a stronger safety indicator than hard braking in dense urban roads. (iii) The correlation direction of speed changes also by road type. Longer travel times (i.e., lower mean speed values) may be linked to crashes in dense urban roads. However, high speeding on highways is more indicative of collision risks.

The remainder of this paper is structured as follows: Section 2 provides an overview of the network screening methods and correlation studies of safety surrogate measures with observed crash data. Section 3 describes the data and processing methods, as well as extraction of driver behavior indices, and presents two safety metrics derived from the safety surrogate measures. Section 4 presents the correlation results. Section 5 discusses the results and future research directions. Finally, conclusions are provided in Sections 6.

2. Literature Review

Johnsson et al. (2018) list the requirements for an "ideal" surrogate safety measure (SSM): An SSM should reflect collision and injury risks in different settings, should have robust validity through measuring the correlation magnitude for instance, and should be reliable and replicable to produce an accurate result irrespective of the setting. A regression analysis has been the most common approach to study the validity of SSM (Zheng et al., 2014). However, Davis et al. (2011) outline the SSM-crash relationship as a probabilistic model that computes the probability of a crash given a set of non-crash events. Zheng et al. (2014) state the need for more sophisticated approaches, such as extreme value theory (EVT) for road safety analysis because of its power to identify the likelihood of extreme events from a short period of observations and that is the intent of SSMs. Machine learning techniques have also been used for network screening and collision prediction tools. Moosavi et al. (2019) implemented a deep neural network model using traffic events data (congestion, lane-blocked, accident), weather data (visibility, temperature, rain, snow), and point-of-interest annotation tags (roundabout, bump, traffic signal, etc.). Yuan et al. (2018) used a convolutional long short-term memory neural network model (LSTM) to predict crash frequency using traffic volume, road condition, rainfall, temperature, and satellite images.

Multiple studies have attempted to identify possible driver behavior surrogate safety measures for network screening. TTC is defined by Hayward in Hayward (1972) as the time needed for two vehicles to collide with each other when they keep the same speed and the same travel path. El-Basyouny and Sayed (2013) develop a two-phase model. The first model of the two-phase model predicts conflicts using a log-normal model with the aid of traffic volume, road type and geometry features as covariates. The TTC is used to define a traffic conflict, and two trained observers were stationed at intersections to observe traffic conflicts.

The second model uses the predicted conflicts to compute a Negative Binomial safety performance function. This model might not be scalable given that labor work is needed to observe conflicts, although computer vision techniques can automate some of this work. Mi et al. (2020) mention that TTC is inappropriate for intersection safety assessment since the characteristics of vehicle movements at intersections, such as frequent acceleration and non-lane-based vehicle movements, are not considered. Thus, they propose instead a modified TTC (MTTC) value, calculated from the video record, and derived from the relative speed and acceleration of the interacting vehicles.

PET is another safety surrogate measure developed in the literature. The post-encroachment time is calculated as the time between the instance when the first vehicle leaves the path of the second and the instance when the second reaches the path of the first (Johnsson et al., 2018). Zheng et al. (2014) found, using an extreme value modeling approach, a correlation between post-encroachment time measures from 4189 lane change maneuvers recorded at 29 directional freeway segments, and crash data collected over four years.

Agerholm & Lahrmann (2012) built a predictive model to identify hazardous road locations based on GPS jerk data (the time derivative of acceleration data). However, they mention that large-scale studies are needed to test the reliability of the jerk-based model, as other parameters (speed prior to jerks, deceleration start and end time) need to be identifiable to avoid erroneous results. Tageldin et al. (2015) use an automated video-based analysis technique to detect jerk rates, in order to measure traffic conflicts as indicators of safety. Tageldin and Sayed (2016) suggest that evasive action-based indicators, which represent variations in the spatio-temporal gait parameters (i.e., step length, step frequency and walk ratio), are possible indicators of pedestrian conflicts.

Speed is an essential safety surrogate measure in the road safety analysis. It is generally believed that an increase in speed threatens road safety (Rolison et al., 2018). Though, it can also be argued that driving at high speed reduces the length of time exposure and thus the likelihood of a crash (Pei et al., 2012). Inconsistent findings were reported in the literature. Some researches show that there is a negative or insignificant relationship between speed and crashes (Quddus, 2013; Stipancic et al., 2017), while others suggest a positive relationship (Taylor et al., 2000). In the work of Stipancic et al. (2017), congestion index, average speed, and the coefficient of variation of speed, were compared with crash data collected over an 11-year period in Quebec City. Driver behavior indices were derived from smartphone GPS data. The correlations with crash frequencies were found to be weak to moderate. The congestion index was shown to be positively correlated with crash frequency. Higher congestion levels were related to crashes with major

injuries, whereas low congestion levels were related to crashes with minor injuries. The average speed was found to be negatively correlated with crash frequency. However the coefficient of variation of speed was positively correlated. On the other side, Quddus et al. (2013) used a random-effects Negative Binomial and a mixed-effects spatial model to explore the effects of speeds on minor-injured and major-injured collisions using segment-based traffic, road geometry, and accident data from 266 road segments including 13 different motorways in London. Both models indicated a negative, yet statistically insignificant, relationship between average speed and collisions. The results of Quddus et al. (2013) suggest that average speeds are not correlated with accident rates when controlling for other factors affecting accidents such as traffic volume and road geometry. Wang et.al (2009) conducted a precise congestion measurement in the M25 London orbital motorway and concluded that traffic congestion had little or no impact on the frequency of road accidents, but was negatively correlated to collision severity. Martin (2002) claimed based on observations made on 2000 km of French interurban motorways over two years, that light traffic was a safety problem in terms of frequency and severity of accidents. Wang et al. (2018) studied this relationship on urban arterials using taxi-based high-frequency GPS data and concluded that higher average speeds were associated with higher crash frequencies, but the sample size was formed only of eight arterials in downtown Shanghai. Urban roads in Canada were studied by Gargoum & El-Basyouny (2016) to explore their speed-safety relationship. The authors reported that a 1% increase in average speed was associated with a 0.018% increase in collision frequency.

Another safety surrogate addressed in the literature is speed variation. Speed variations are used to represent the inconsistency of vehicle speed along a segment (Wang et al., 2018). Most of the related studies in the literature converge to an idea that speed variations are positively correlated with crash occurrence. However, this conclusion is based to a great extent on research on rural roads and freeways (Wang et al., 2018). Boonsiripant et al. (2011) derived the speed variation from on-board vehicle speed sensor data in Atlanta metropolitan area, and found no significant relationship with the crash frequency under likely free-flow conditions. Pei et al. (2012) evaluated the relationship between speed and crash risk using disaggregated crash and speed data collected from 112 road segments in Hong Kong, and stated that there was no evidence in their Bayesian crash model that the standard deviation of speed was significantly associated with the likelihood of crash occurrence or crash severity. Wang et al. (2018) analyzed speed variation data derived from taxi-based GPS data on eight urban arterials and found that speed variation was significantly positively associated with crash frequencies. Oh et al. (2001) have looked at the speed variation parameter from real-time traffic data instead of the speed quantity itself when estimating the likelihood of an accident and proved

that reducing speed variation increased safety and reduced the accident likelihood.

Deceleration and acceleration-based indicators have been investigated to act as safety surrogate measures of collisions. Hard braking and acceleration are measures of how fast the speed of a vehicle changes (N. Algerholm, 2012). Most of the related research work focus on the hard braking and acceleration correlation with collision risk from the driver's perspective. In other words, researchers examined whether the drivers that are involved in aggressive maneuvers also have the highest crash records (Bagdadi, 2013; Johnson & Trivedi, 2011; Laureshyn et al., 2009). A few studies examined the spatial correlation between the two quantities and whether the road intersections with the most dangerous maneuvers are associated with higher crash rates. Jun. et al. (2007) found that the frequency of hard deceleration events was strongly related to the crash involvement rate of individual drivers location-wise, but it was not clear if both quantities incorporated traffic volume or not. Stipancic et al. (2018a) explained in an empirical study, based on GPS-enabled mobile devices, that locations with more hard braking and hard acceleration counts also tended to have more collisions. Li et al. (2021) utilize critical bus driving events extracted from GPS trajectory data as pedestrian and bicycle surrogate safety measures for bus stops. Correlations were then examined using Spearman's rank correlation coefficient. The study concludes that bus stops with more hard acceleration events are expected to have more pedestrian and bicycle crashes.

This study attempts to address several shortcomings that are apparent in the existing literature: when studying the spatial correlation between hard braking and acceleration and crash frequencies, normalization by traffic volume is taken into account since the latter is a confounding variable that influences both quantities. Also, as mentioned earlier, driver behavior data (maximum speed, hard braking, and hard acceleration) are extracted directly from the vehicle's CAN bus, instead of being calculated from GPS measurements that can be noisy. Moreover, highway and non-highway roads are considered separately as both the driving conditions and settings change, whereas many previous studies considered highways only or individual neighborhoods, corridors or links only. Lastly, there has been no similar comprehensive study of driver behavior analytics based on large-scale telematics data, crash data, and road network.

3. Methodology

The work of this paper can be divided into three steps: pre-processing, processing and post-processing, as shown in Figure 1. In the pre-processing step, the telematics data are assigned to the road segments using the map matching engine. Then, the OBD-II speeds are compared to spot speeds radar data for validation. In the processing step, driver behavior indices are extracted and grouped by road segments.

Then, the l variation, hard braki ıd collision rate, norm ety metrics derived fro

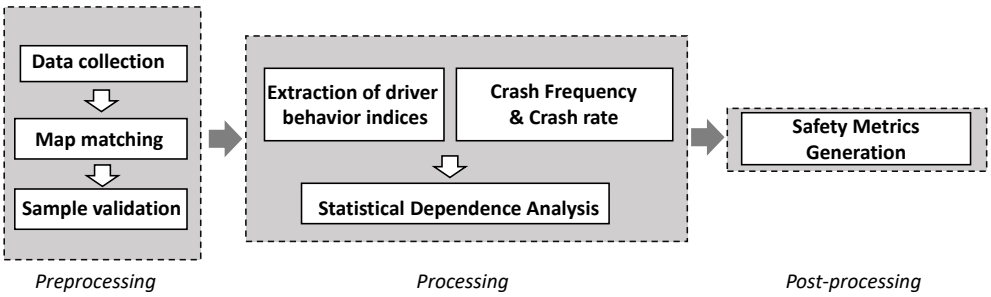

Figure 1: Flowchart explaining the methodology for generating safety metrics

3.1. New York City Fleet Raw Data Structure

The driver behavior data consist of 1.5 years of raw *GPS trajectories* sampled from GPS devices over the 2015-2016 period, installed on 4,500 distinct vehicles which are classified as "light utility" within the NYC fleet database, and generally would be considered to be cars or SUVs. Trucks, trash collection vehicles, or other heavy vehicles, which would not be representative of traffic at large, were disregarded to be as close as possible to the driving population. The fleet-based driver behavior data comprise OBD data, GPS trajectory, and acceleration data. A raw GPS trajectory is a sequence of N noisy data points $P = (p_i|i = 1, \ \ldots, \ N)$ in a chronological order. A time interval between two consecutive points does not exceed a certain threshold Δt, which is the sampling rate. The regular sampling rate of the NYC fleet data considered in this study is 30 seconds. Each data point p_i has the following parameters: 1) longitude, latitude and altitude values, 2) timestamp, 3) the number of visible satellites, 4) maximum speed from OBD-II communication, and 5) hard braking and hard acceleration data from the OBD-II port. The OBD-II speed data point $v_{\max}$ represents a *maximum speed* value between the previous data point and the current data point (typically within the last 30 second time interval). The illustration of the GPS trajectory data structure is shown in Figure 2. In addition, hard braking and hard acceleration events are also logged in the database. These events are logged in addition to the regular 30-second interval maximum speed samples.

Smith et al. 2003 analyzed the last-second braking response of drivers based on experimental data collected from test track studies in car-following situations, and found that the 50-percentile average decel-

	Timestamp,	Latitude,	Longitude,	Altitude,	Value ,	N_satellite
p_1:	2016/01/01 13:12:00,	42.7128,	-74.0000,	33.1253,	15 mph	7
p_2:	2016/01/01 13:12:30,	42.7138,	-74.0220,	33.1293,	23 mph	4
	••••• ••••• •••••					
P_n:	2016/01/01 15:17:45,	42.7138,	-74.0345,	33.1323,	0.33 g	5

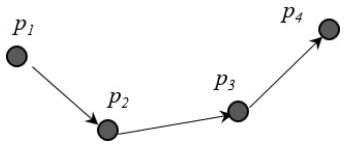

Figure 2: Illustration of GPS trajectory raw data structure.

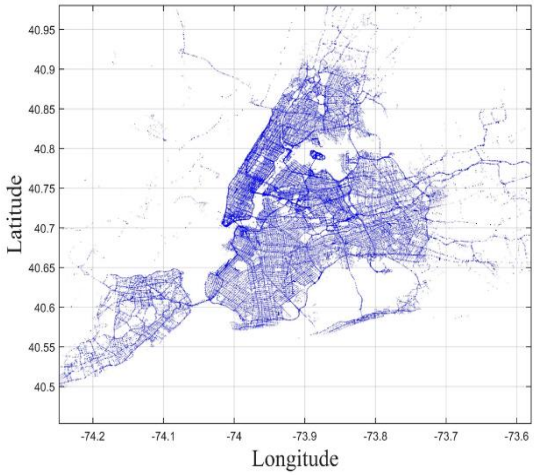

Figure 3: A sample of one day's recorded data points from the NYC vehicle fleet.

eration values range from 0.17g to 0.27g in the "hard braking" conditions as the speed increases from 3 to 20 meters per second. We use in this study a threshold value of 0.18 g to define a hard braking or a hard acceleration event as values below 0.18 g are not logged in the database. Figure 3 shows the spatial coverage of a one day sample of recorded data points from the NYC vehicle fleet.

3.2. Map Matching

The raw GPS trajectories are noisy and do not fall on the road network segments where they actually traveled. The team has built a core data processing engine, detailed in Alrassy et al. (2019). The outcome of the map-matching algorithm, shown in Figure 4, is a sequence of projected points c_i $(i = 1, \ldots, N)$ on the road segments of the digital map that represent a reconstructed path that a driver has taken in a chronological order, given the N noisy data points $P = (p_i | i = 1, \ \ldots, \ N)$. The candidate point c_i of data point p_i is projected on the road segment e_i using Equation 1 below:

$$c_i = \arg\min \text{dist}(p_i, c_i), \qquad \forall c_i \in e_i. \tag{1}$$

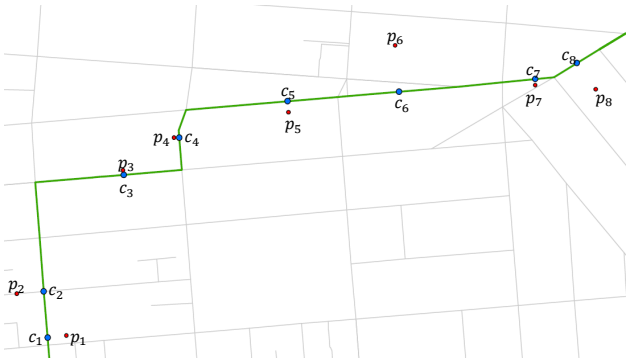

Figure 4: Reconstructed path from a GPS trajectory generated by the map-matching engine

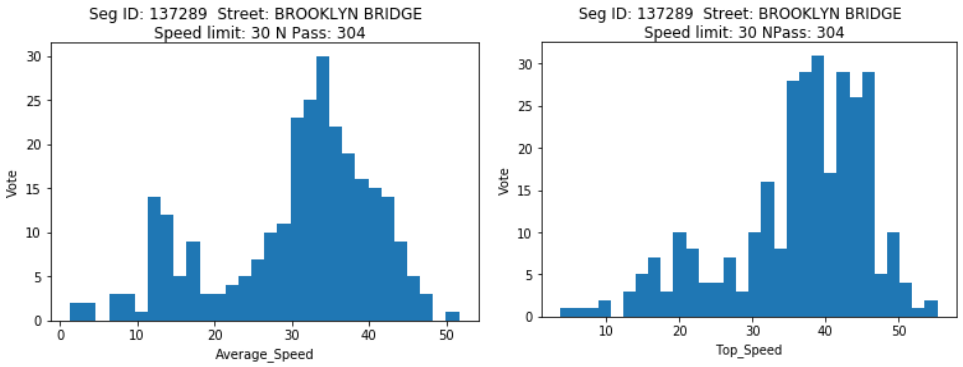

Figure 5: Average and Maximum speed profiles for a given segment

The core engine uses the LION geographic base map of New York City streets (NYC Department of City Planning, 2020) as the digital map, adopted by the New York City Department of Transportation for mapping collision data.

Between two intermediate GPS points (p_{i-1}, p_i), the sampled vehicle OBD-II speed $v_{\max,i}$, associated with a data point (p_i), is assigned to the set of intermediate segments forming the actual path $P(c_{i-1}, c_i)$ from (c_{i-1}, c_i). The intermediate path $P(c_{i-1}, c_i)$ is computed using the Dijkstra algorithm described in (Dreyfus, 1969). A travel speed parameter between each pair of GPS points (p_{i-1}, p_i) is also determined as the distance traveled $P(c_{i-1}, c_i)$ divided by the sampling time interval Δt as shown in Equation 2 below :

$$v_{\text{travel speed}} = \frac{P(c_{i-1}, c_i)}{\Delta t}. \tag{2}$$

Analytical methods, described in Section 3 of this paper, are performed to compute the statistics of the driver behavior indices, and to locate specific clusters within road segments of high rates of hard braking,

Vehicle Model, Year, Classification
FORD F350, 2011, HEAVY DUTY
FORD ESCAPE, 2013, LIGHT DUTY
••••• ••••• •••••
TOYOTA CAMRY,2013, LIGHT DUTY

Figure 6: Fleet classification by model provided by the New York City Department of Citywide Administrative Services DCAS

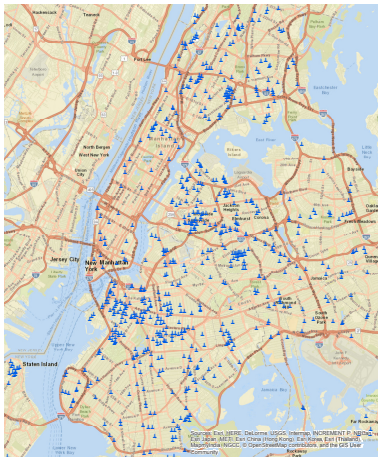

Figure 7: Speed Radar Guns Locations

hard acceleration as well as to generate speed profiles.

3.3. Validation of the Sample Telematics Data with Spot-Speed Radar Data

The research team recognizes the challenge of demonstrating that the fleet under study is a representative sample of the actual driving population in New York City. For that reason, two measures were taken. First, as mentioned earlier, the team has only included the sedans that are classified as "light duty" vehicles by the NYC DCAS. Also, spot-speeds data collected using radar speed guns that measure the speed of moving vehicles, were compared with the telematics maximum speed profiles by time buckets. Two-sample T-test for mean comparison, and F-Test for equality of two variances were conducted to compare city-vehicles maximum speed histograms and spot-speeds histograms at a significance level of $\alpha = 0.001$. Spot-speed data are collected in 1 to 2-hour random time intervals at 770 locations across New York City, as shown in Figure 7 over the 2005-2018 period.

Figure 8 reports some examples of matching speed histograms at different locations along with the T-test, F-test p-values and time period. Table 1 summarizes the validation results. Overall, 39% of the city fleet speed histograms match the spot-speed histograms. At 34% of the locations, T-test and F-test have rejected the hypothesis that both speed profiles have the same mean but the difference in speed is less than 5 mph.

Speed Difference	Percent of locations (%)
Same mean and variance	39
Difference in mean $<$ 5mph	34
Difference in mean $>$ 5mph	27

Table 1: Validation of the Telematics data with Spot-Speeds Radar Data

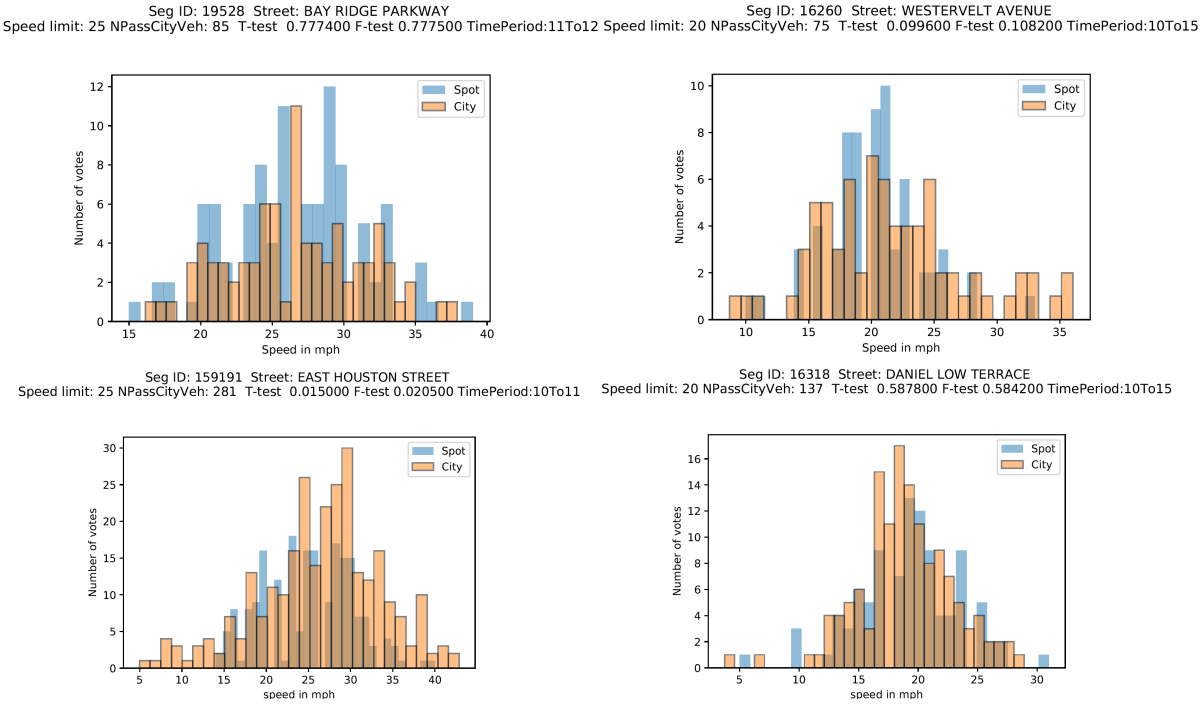

Figure 8: Validation of city maximum speed profile with spot-speed profile

As shown in Figure 9, the majority of these locations have city drivers' average maximum speed lower than the average spot-speeds. This could be due to the fact that spot-speed data are collected before some street improvements were done as revealed in Figure 10, compared to the telematics city data collected in 2015 and 2016. Some street improvements involve lane reduction, adding parking lanes, which may result in lower speed values. We also noticed that some of the locations, where telematics vehicles speeds are higher than spot-speeds (Figure 11), belong to the parallel service road locations, where the map matching problem is challenging. In other words, the map matching engine assigns these large speed values to the parallel service road instead of the main priority road.

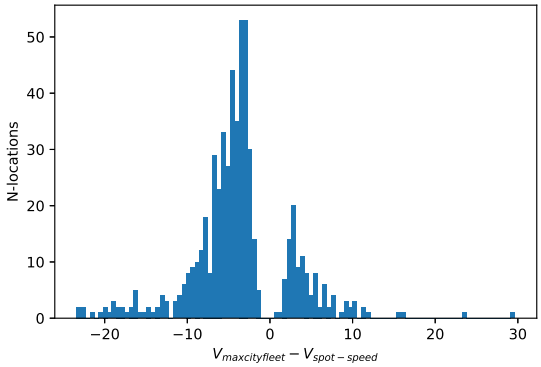

Figure 9: Difference in speed between city fleet and spot-speed radar data

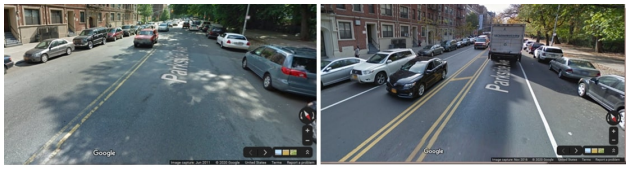

Figure 10: Street improvement on Parkside Ave in Brooklyn using Google Maps view

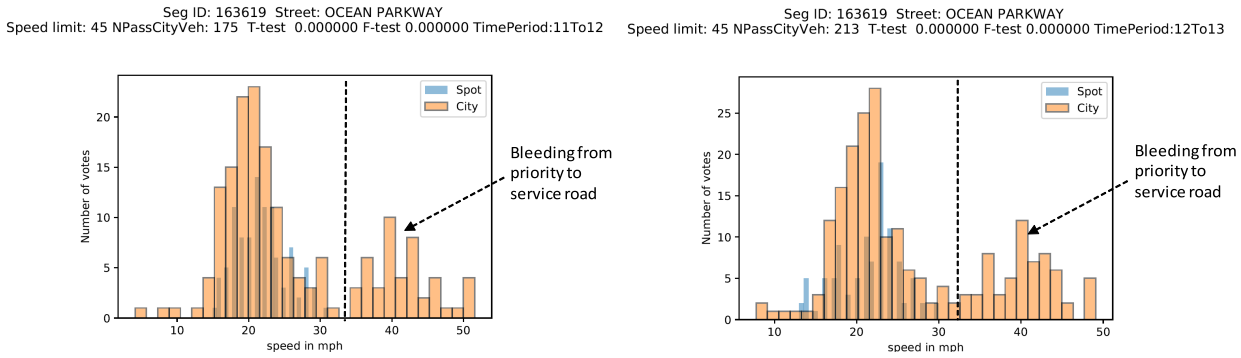

Figure 11: Data from the priority road falsely mapped to the service road on Ocean Parkaway in Brooklyn

3.4. Extraction of The Driver Behavior Indices

In this paper, we aggregate the proposed driver behavior indices for each road segment based on the 1.5-year large GPS vehicle fleet telematics data. The driver behavior indices are estimated per road segment. It is noteworthy that some of the important driver behavior indices considered in this study are normalized by the appropriate quantities (e.g., hard braking data are normalized by the total count of vehicles that

traveled over the corresponding road segment.)

3.4.1. Maximum Speed

Speed has widely been believed to be a safety surrogate of crash risk, especially on highways. The maximum speed $V_{\max,i}$ for each road segment e_i is calculated as the average of all sampled vehicle OBD-II maximum speeds $\sum_{j=1}^{N_c} V_{\max,ij}$, matched by the map matching engine to that specific segment, divided by the total counts N_c:

$$V_{\max,i} = \frac{\sum_{j=1}^{N_c} V_{\max,ij}}{N_c}. \tag{3}$$

The above calculation results in the maximum speed profile shown in Figure 15 for the entire NYC road network.

3.4.2. Mean Speed

We define the mean speed value as the average of all travel speeds mapped to a given segment. As previously mentioned, the travel speed between each pair of GPS points (p_{i-1}, p_i) is determined as the distance traveled $P(c_{i-1}, c_i)$, divided by the sampling time interval Δt, as shown in Equation 2. The mean speed $V_{\text{mean},i}$ is indicative of travel time as it includes the stop and go activity on a traffic signal. It is calculated for each road segment e_i as the average of all travel speeds $\sum_{j=1}^{N_c} V_{\text{travel},ij}$, matched by the map matching engine to that specific segment divided by the total counts N_c:

$$V_{\text{mean},i} = \frac{\sum_{j=1}^{N_c} V_{\text{travel},ij}}{N_c}. \tag{4}$$

The above calculation results in the mean speed profile shown in Figure 16 for the entire NYC road network.

3.4.3. Speed Variation

The literature summarized in Section 2, in general, focuses on understanding the direction of the relationship between speed variations and collision risks. Speed variations could be a better indicative candidate for safety than the speed magnitude itself, although this statement is biased to studies on freeways. We chose the standard deviation of the sampled maximum speeds on a given road segment e_i, as a measure of speed variation:

$$SV_i = \sqrt{\frac{\sum_{j=1}^{N_c} (V_{\max,ij} - V_{\max,i})}{N_c - 1}}, \tag{5}$$

where $V_{\max,ij}$ is the sampled vehicle OBD-II maximum speed and $V_{\max,i}$ is the average maximum speed determined in Equation 3. Figure 17 clearly shows the speed variations as high on arterials and major roads where vehicles are more likely to change speeds.

3.4.4. Fleet Traffic Flow

Traffic volume has been considered as an indicator of crash risk since a higher traffic volume means a higher exposure and therefore a higher chance for a crash. We estimate the fleet traffic flow $\mathrm{TV}_{\mathrm{fleet},i}$ across the entire NYC road network based on the sum of speed counts N_c that have fallen on each road segment e_i as follows:

$$\mathrm{TV}_{\mathrm{fleet},i} = N_c \tag{6}$$

Figure 18 shows how the telematics sample data are distributed by the map matching engine across the entire road network.

3.4.5. Free Flow Condition

We define the free flow condition FFC_i parameter as the ratio of mean speed $V_{\mathrm{mean},i}$ defined in Equation 4 and the speed limit SPL_i for each segment e_i. This represents a measure of the level of congestion on the streets. But since the speed limit is 25 mph on non-highway New York City local streets, the free flow condition profile is similar to the mean speed profile.

$$FFC_i = \frac{V_{\mathrm{mean,i}}}{SPL_i} \tag{7}$$

3.4.6. Hard Braking and Hard Acceleration Rate

The last driver behavior measure that were examined in this study are the rate of hard braking HBR_i and hard acceleration HAR_i on a given road segment e_i. 10,121,892 hard braking events with a g-value density distribution shown in Figure 12, were collected over the 1.5-year 2015-2016 period. Also, 9,369,525 hard acceleration events were registered and distributed as shown in Figure 13. We count the number of sampled hard braking and acceleration events HB_i and HA_i, respectively, map the event counts HB_i and HA_i to the corresponding road segment e_i. Then, the event counts HB_i and HA_i are divided by the fleet traffic flow $\mathrm{TV}_{\mathrm{fleet},i}$, in an effort to normalize them with the road segment e_i's traffic volume. Figures 19 and 20 plot, respectively, the hard braking and hard acceleration spatial distributions. Figures 21 and 22 plot, respectively, the spatial profiles of the hard braking and hard acceleration rates, in order to remove the confounding variables effect defined in Frank (2000), when correlating with collision data.

The safety surrogate measures are evaluated at 115,289 road segments. Correlation strength between the derived indices are provided in Figure 14.

$$HBR_i = \frac{HB_i}{\mathrm{TV}_{\mathrm{fleet},i}} \tag{8}$$

$$HAR_i = \frac{HA_i}{\mathrm{TV}_{\mathrm{fleet},i}} \tag{9}$$

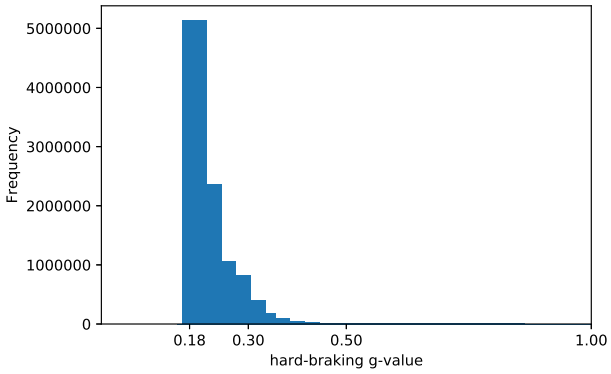

Figure 12: g-value density distribution of hard braking events over the 1.5-year 2015-2016 period

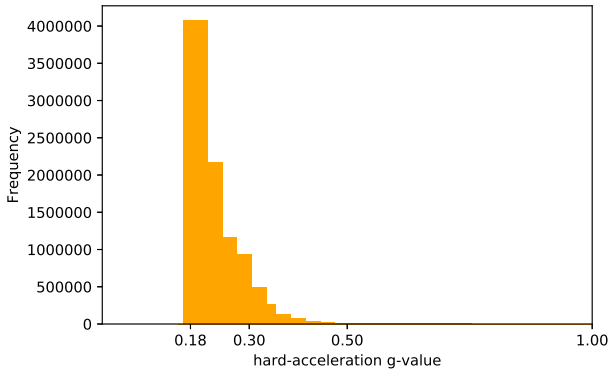

Figure 13: g-value density distribution of hard acceleration events over the 1.5-year 2015-2016 period

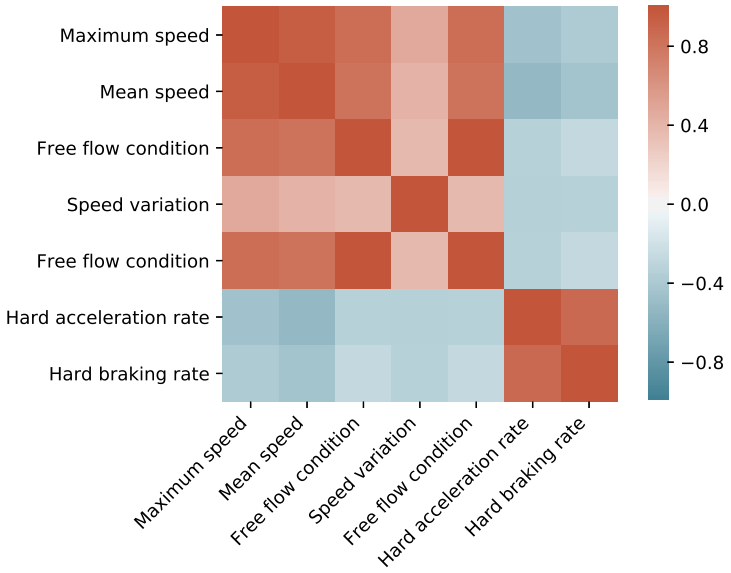

Figure 14: Correlation strength between the derived driver behavior indices

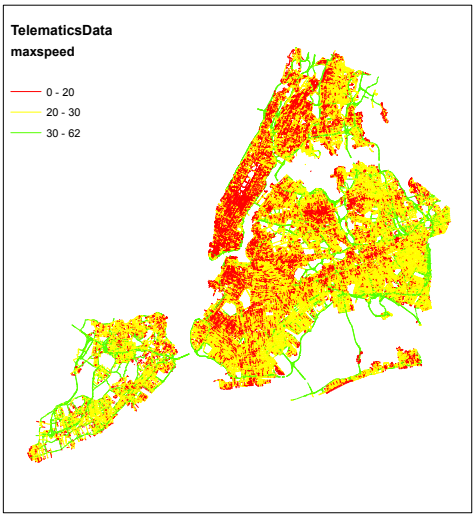

Figure 15: OBD-II Max speed map $V_{\max,i}$ in mph

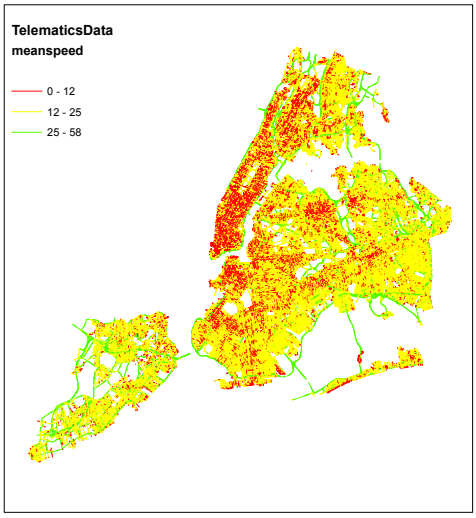

Figure 16: Mean speed map $V_{\text{mean},i}$ in mph

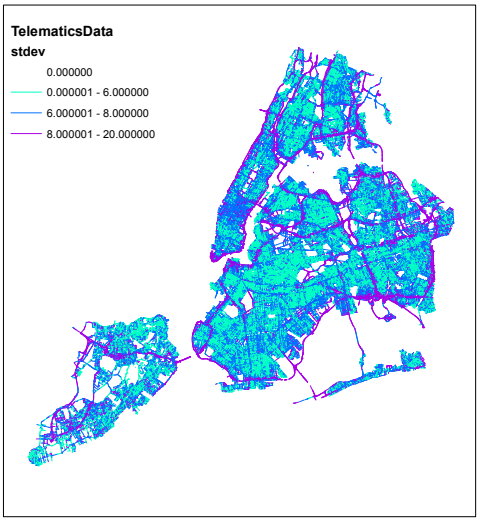

Figure 17: Speed variation SV_i in mph

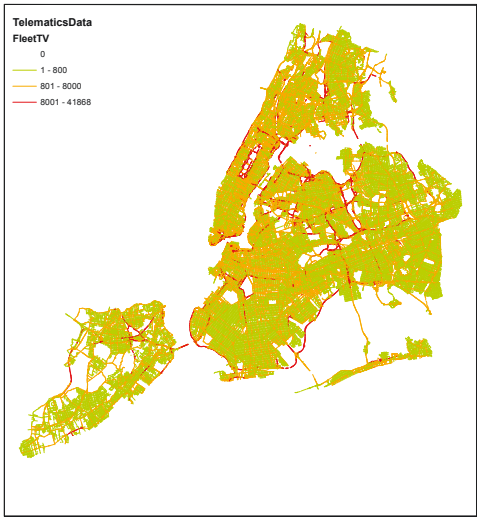

Figure 18: Fleet traffic volume $\text{TV}_{\text{fleet},i}$

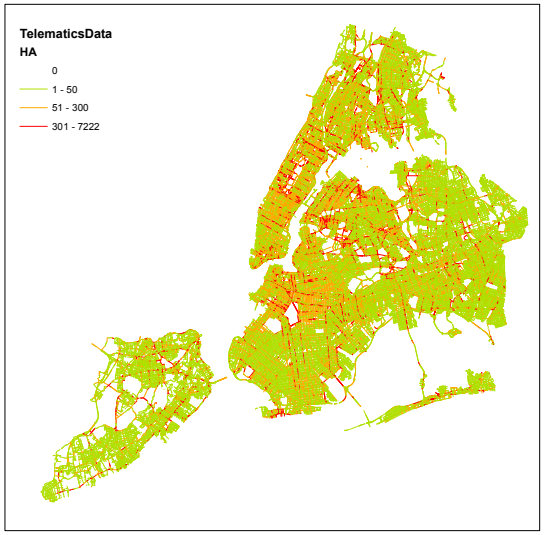

Figure 19: Hard acceleration counts HA_i

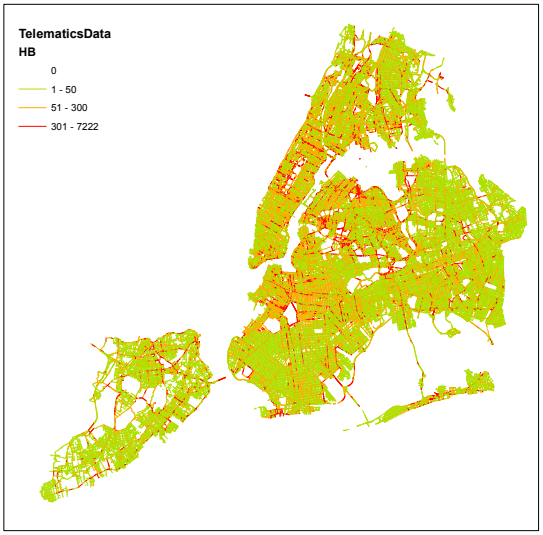

Figure 20: Hard braking counts HB_i

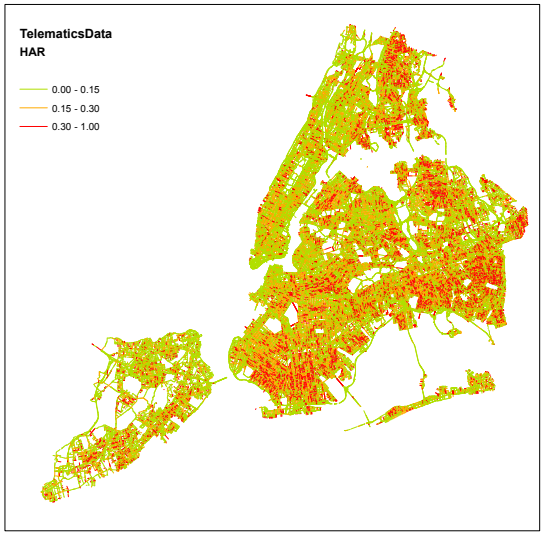

Figure 21: Hard acceleration rate HAR_i

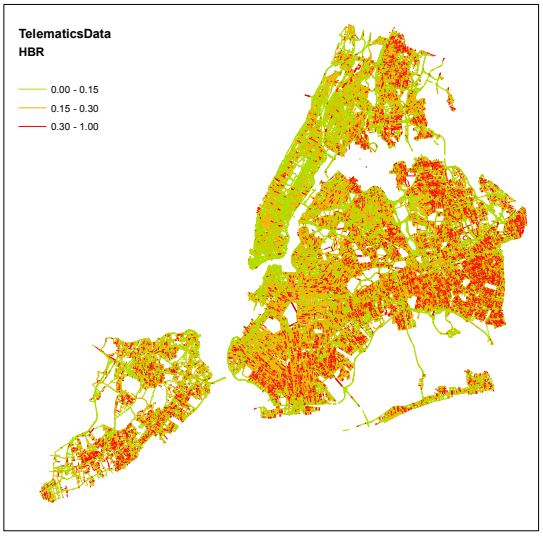

Figure 22: Hard braking rate HBR_i

3.5. Crash Data

In this study, we used the New York City Police Department Motor vehicle collisions data, in order to determine its correlation with the proposed safety surrogate measures. The data set is publicly available on the NYC Open Data portal in New York Police Department (2019). A collision is recorded when at

least one person is injured or killed, or when there is at least $1,000 worth of damage. The collision data considered in this study are collected over a longer period (2012-2019), compared to telematics data collected from 2015 to 2016 since collisions do not occur frequently and need a wider time range in order to have a granular classification of streets. The authors are aware of changes in driving conditions and driving behavior throughout the time of day and throughout the years with road design changes, however, preliminary correlation analysis between the crash data of the two periods (2012-2019 and 2015-2016) shows a strong correlation with a correlation coefficient of 0.78.

3.5.1. Crash Frequency

In total, 1,360,911 collisions with known latitude and longitude coordinates were considered, of which 276,316 were with injuries. One observation to note is that the spatial distribution of the injury collision map (Figure 25) looks similar to the spatial distribution of the total crash map (injuries and non-injuries), shown in (Figure 24). The total crash data was used in this study for correlating with safety surrogate measures since the injury collisions data are sparse. Crash events that happened on road intersections are jointed to the inbound road segments, proportionally to their respective traffic volume. The crash data are mapped to the nearest road segment using the spatial join feature in the ArcMap ESRI software (CA: Environmental Systems Research Institute). In this study, the correlation between the telematics data and crashes is carried at the segment level to maintain accuracy of the telematics data. The illustration of the crash data structure is shown in Figure 23. Each crash record has the following information: 1) timestamp, 2) the number of fatalities, 3) the number of injuries, 4) the number of injured pedestrians, 5) the number of injured cyclists, 6) contributing factor, 7) latitude, and 8) longitude.

Timestamp,	Num_of_fat;	Num_of_inj,	Number of Ped.inj	, Number of cycl. inj	, Contributing Factor,	Latitude,Longitude
2015/01/01 13:15	, 0	, 2	, 0	, 0	, Rear end	, 42.7156, -73.1
2016/01/13 16:12	, 0	, 2	, 2	, 0	, Overtaking	, 43.8538, -73.2
••••• ••••• ••••						
2017/12/12 11:17	, 0	, 1	, 1	, 1	, Left Turn	, 41.9215 ,-73.05

Figure 23: Crash data structure

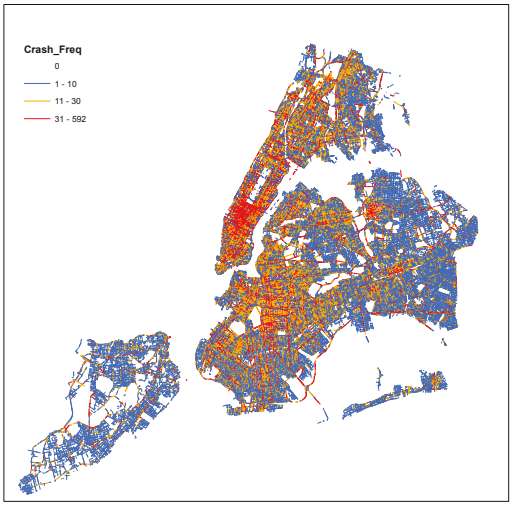

Figure 24: Total Crash frequency map in New York City

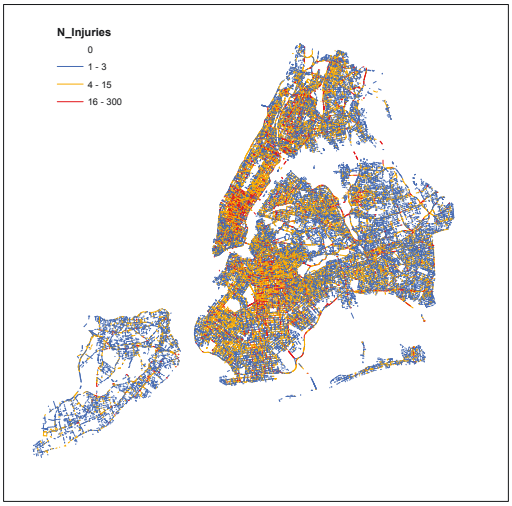

Figure 25: Injury Crash frequency map in New York City

3.5.2. Crash Rate

In order to remove the confounding effect of traffic volume when studying correlations, it is necessary to normalize the crash frequency map with traffic exposure. Exposure on highways is determined using the fleet traffic volume, whereas exposure on local streets is determined using a traffic regression model. The traffic regression model was developed by Datakind DataKind (2017). Figure 26 displays the result of the model. The reason why the fleet traffic volume is not used to normalize crash frequencies is that some of the locations might be biased to city facilities and may fail to represent the public driving population. The model estimates the average annual daily vehicular volume. Assuming proximate streets behave the same, the model uses traffic count data on selected streets and propagates these count values to the neighboring streets. When there are no count data available, a Random Forest regression model, explained in Liaw et al. (2002), predicts the traffic volume of a street given a set of predefined features. Figure 27 shows the calculated crash rate in percentage across New York City. The correlation coefficient between crash rate and crash frequencies is significantly high ($\rho = 0.94$), indicating that when normalizing crash frequencies with respect to traffic volume, crash hotspots stay the same as shown in Figure 27.

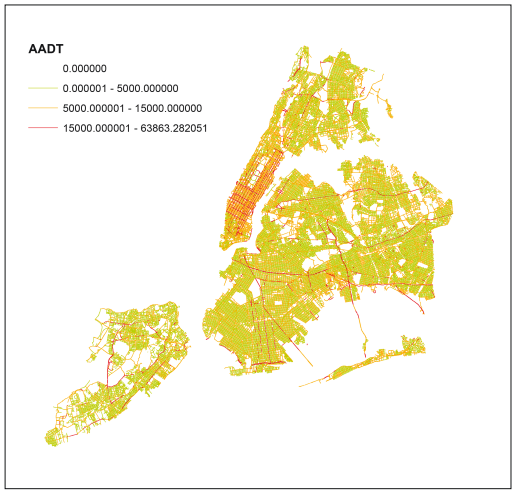

Figure 26: Average Annual Daily Traffic Map in New York City developed by DataKind

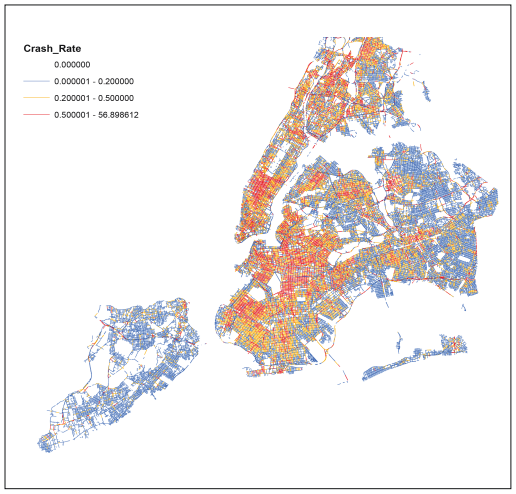

Figure 27: Calculated crash rate map (unit: count per AADT)

3.6. Statistical Dependence

In order to examine if high-collision locations have high harsh driving indices, the rank correlations between the predefined driver behavior indices and crash data are estimated using the Spearman's rank correlation coefficient (Myers et al., 2013). The coefficient ranges from -1 to 1. The negative value means a negative relationship; and the positive value means a positive relationship, whereas a zero value is an indication of randomness. The Spearman's rank correlation coefficient for two variables $\vec{x}_1$ and $\vec{x}_2$, each of size n is defined as follows:

$$\rho_s = 1 - \frac{6 \times \sum_{i=1}^{N} (r_{1,i} - r_{2,i})^2}{n(n^2 - 1)} \tag{10}$$

Where $r_{1,i}$ is the rank of the segment e_i based on $\vec{x}_1$'s value, $r_{2,i}$ is the rank of segment e_i based on $\vec{x}_2$'s value, and n is the size of vectors $\vec{x}_1$ and $\vec{x}_2$ (i.e, n is the total number of road segments). In this study, $\vec{x}_1$ represents one of the predefined driver behavior indices; and $\vec{x}_2$ is either the collision data or collision frequency variable.

3.7. Safety Metrics Generation Using Safety Surrogate Measures

The presence of safety surrogate measures derived from the telematics data, as large-scale objective performance measures, provides better and broader insights into roads safety. Safety surrogate measures can help develop corridors with similar characteristics and performance. The following subsection discusses the methodology used in this study to build a safety corridor map for New York City, based on the maximum

speed value, as well as hard braking and hard acceleration hot-spots generation, which may indicate the presence of an unsafe design defects.

3.7.1. Spatial Partitioning Based on Snake Similarities

We adopt the existing methodology of snake algorithms explained in Saeedmanesh & Geroliminis (2016) to partition a road network into homogeneous corridors that have the same characteristics and behavior. We define a corridor as a cluster of segments that have: 1) geometric similarity between the streets, 2) connectivity within the corridor, i.e., the vehicle can travel from one segment to the other, and 3) low variance of maximum speed, derived from the OBD-II port of the vehicles. In order to achieve the three objectives, we start with grouping road segments that have similar geometric features as well as having spatial connectivity. We used the street name (*st-name*) encoded in the digital map as an indication of geometric similarity and the directionality information (*node-from*,*node-to*) to test for spatial connectivity. Figure 28 shows the resulting process in dividing the large road network into smaller arterials in Brooklyn. Each arterial is displayed with different color.

Figure 28: The resulting process in dividing the large road network into smaller arterials

Then, an iterative process is applied. A snake starts at each arterial and expands by adding to it the segments that belong to the same arterial and have a similar maximum speed value. We define a similar

segment to the existing segments in a snake, every segment with a difference between its maximum speed and the average maximum speed value of the current snake less than 5 mph (a typical buffer used in Forbes (2012) to round speed data in order to determine speed limits on roads). For instance, a road segment e_j is added to snake $S_{i,k}$ of arterial i if :

$$|V_{max}(e_j) - \bar{V}_{max}(S_{i,k_{-e_j}})| < 5 \text{ mph}, \tag{11}$$

where $\bar{V}_{max}(S_{i,k})$ is the average maximum speed value of snake $S_{i,k}$ including the speed value of road segment e_j and $\bar{V}_{max}(S_{i,k_{-e_j}})$ is the average maximum speed value of snake $S_{i,k}$ without including road segment e_j. In the event where road segment e_j is not added to the snake $S_{i,k}$, a new snake $S_{i,k+1}$ is initiated. The process is repeated until all segments within an arterial belongs to one of the snakes.

3.7.2. Dynamic Clustering of Hard Braking and Hard Acceleration Events

Hard braking and hard acceleration events are aggregated to a specific road segment on the digital map. Two road segments might have the exact counts of hard baking and hard acceleration events recorded, but these events might be: 1) clustered in a specific location on the segment (indicating the possibility of the presence of a pothole or a street design defect) or 2) uniformly distributed along the length of the segment. Therefore, it is of great benefit for city planners to locate hard braking and acceleration hot-spots. The geo-location and the direction of travel for the sampled hard braking and hard acceleration data on a given segment are identified by the map matching engine. We present below a dynamic clustering algorithm that relies on the Gaussian mixture models explained in Reynolds (2009). This algorithm detects the number and locations of hard braking and acceleration clusters on a segment.

Assume we have the hard braking events $B = (b_i | i = 1, \ \ldots, \ N_b)$ distributed on a given road segment as shown in Figure 29. Given that each road segment has two nodes: V_{From} and V_{To}, we define the scalar parameter d_i as the distance from V_{From} to hard braking event b_i. Then, we determine if the set $d = (d_i | i = 1, \ \ldots, \ N_b)$ fits a uniform distribution using the Kolmogorov–Smirnov goodness of fit test (K-S test) (Massey Jr, 1951). If this is the case (i.e., K-S test *p-value* > 0.05), then the hard braking events are considered randomly dispersed along a segment. If not, the events are clustered in specific locations along the segment. In order to find the number and center of the clusters, we model the distribution of d_i as a weighted sum of K Gaussian components $p(d_i) = \sum_{j=1}^{K} \pi_j \mathcal{N}(\mu_k, \sigma_k)$. We fit three models with $K \in [1, 2, 3]$. Each component is a Gaussian density defined in Equation 12 with $\theta_k = \{\mu_k, \sigma_k\}$. μ_k is the cluster's center and σ_k indicates the precision. The number of clusters K on a segment is chosen from the model with the

lowest Bayesian information criteria (BIC) defined in Burnham & Anderson (2004).

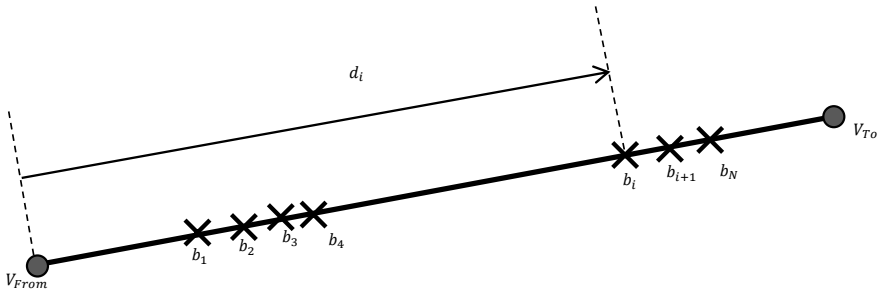

Figure 29: Hard braking events on a road segment

$$p(d_i|\theta_k) = \frac{1}{(2\pi)^{1/2}|\sigma_k|^{1/2}} e^{-\frac{1}{2}(d_i-\mu_k)^T \sigma_k^{-1}(d_i-\mu_k)} \tag{12}$$

4. Results

4.1. Statistical Dependence

Previous studies have separated roads according to their functional types. Jun. et al. (2007) found large behavioral differences between driver-groups from their activities on highways, and on local roadways. Thus, the statistical analysis was conducted by classifying roads into two categories: highway and non-highway local roads. We define road segments that are not accessible to pedestrians as highway. The information about pedestrians' accessibility to roads is encoded in the digital map. Figure 30 categorises the road segments: roads classified as highway (14,240 road segments) are colored in red, whereas the roads classified as non-highway are colored in green (101,049 road segments).

New York City				
Driver Behavior Indices	Crash rate		Crash frequency	
	Highway	Non-Highway	Highway	Non-Highway
Maximum speed ($V_{\mathbf{max}}$)	0.19	-0.18	**0.21**	-0.04
Mean speed ($V_{\mathbf{mean}}$)	**0.21**	-0.18	**0.23**	-0.02
Free flow condition (FFC)	-0.07	-0.17	-0.09	-0.02
Speed variation(SV)	0.11	**-0.23**	0.12	-0.14
Hard acceleration rate (HAR)	0.09	**0.38**	0.06	**0.25**
Hard braking rate (HBR)	0.14	**0.33**	0.12	**0.2**

Table 2: Spearman's rho correlation of safety surrogate measures with crash rate and crash frequency in New York City

Manhattan driver behavior indices	Crash rate		Crash frequency	
	Highway	Non-Highway	Highway	Non-Highway
Maximum speed ($V_{\mathbf{max}}$)	**0.2**	-0.13	**0.24**	0.02
Mean speed ($V_{\mathbf{mean}}$)	**0.22**	-0.14	**0.25**	0.02
Free flow condition (FFC)	0.04	-0.15	0.06	-0.1
Speed variation (SV)	0.12	-0.15	0.13	-0.05
Hard acceleration rate (HAR)	0.17	**0.56**	0.14	**0.5**
Hard braking rate (HBR)	0.19	**0.51**	0.17	**0.45**

Table 3: Spearman's rho correlation of safety surrogate measures with crash rate and crash frequency in Manhattan

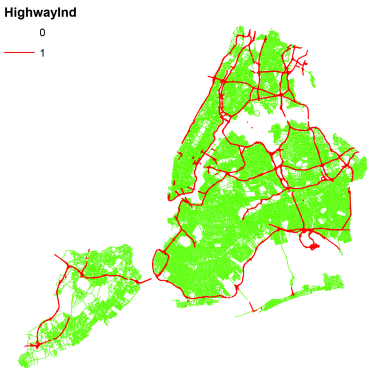

Figure 30: Red roads are classified as *highways* whereas green roads are classified as *non-highway*

The Spearman's rank correlations between the predefined driver behavior indices and crash data are estimated for both highway and non-highway roads for the entire New York City road network and summarized in Table 2. Tables 3 - 7 display the correlations bucketed by borough. All correlation values reported were statistically significant ($p\text{-}value < 0.05$) except when p-value is reported explicitly.

Overall, as shown in Table 2, speeding on highways is found to be correlated with crash rate and crash

Brooklyn driver behavior indices	Crash rate		Crash frequency	
	Highway	Non-Highway	Highway	Non-Highway
Maximum speed ($V_{\mathbf{max}}$)	0.04(*p-value=0.028*)	**-0.21**	0.07	0.01(*p-value=0.017*)
Mean speed ($V_{\mathbf{mean}}$)	0.06	**-0.2**	0.09	0.04
Free flow condition (FFC)	-0.11	-0.18	-0.14	0.05
Speed variation (SV)	0.11	**-0.28**	0.14	-0.15
Hard acceleration rate (HAR)	**0.26**	**0.48**	**0.23**	**0.3**
Hard braking rate (HBR)	**0.26**	**0.45**	**0.24**	**0.27**

Table 4: Spearman's rho correlation of safety surrogate measures with crash rate and crash frequency in Brooklyn

Queens driver behavior indices	Crash rate		Crash frequency	
	Highway	Non-Highway	Highway	Non-Highway
Maximum speed ($V_{\mathbf{max}}$)	**0.23**	-0.13	**0.24**	0.04
Mean speed ($V_{\mathbf{mean}}$)	**0.24**	-0.12	**0.26**	0.05
Free flow condition (FFC)	-0.16	-0.11	-0.17	0.06
Speed variation(SV)	0.15	**-0.26**	0.16	-0.17
Hard acceleration rate (HAR)	0.02(*p-value=0.22*)	**0.29**	-0.005	0.16
Hard braking rate (HBR)	0.09	**0.27**	0.07	0.13

Table 5: Spearman's rho correlation of safety surrogate measures with crash rate and crash frequency in Queens

The Bronx driver behavior indices	Crash rate		Crash frequency	
	Highway	Non-Highway	Highway	Non-Highway
Maximum speed ($V_{\mathbf{max}}$)	0.16	**-0.2**	0.18	-0.02
Mean speed ($V_{\mathbf{mean}}$)	0.17	-0.19	0.19	-0.03
Free flow condition (FFC)	-0.05	-0.19	-0.08	-0.03
Speed variation(SV)	0.09	**-0.27**	0.09	-0.16
Hard acceleration rate (HAR)	0.11	**0.45**	0.08	**0.35**
Hard braking rate (HBR)	0.16	**0.44**	0.13	**0.33**

Table 6: Spearman's rho correlation of safety surrogate measures with crash rate and crash frequency in The Bronx

Staten Island driver behavior indices	Crash rate		Crash frequency	
	Highway	Non-Highway	Highway	Non-Highway
Maximum speed ($V_{\mathbf{max}}$)	**0.35**	0.13	**0.38**	**0.31**
Mean speed ($V_{\mathbf{mean}}$)	**0.35**	0.13	**0.38**	**0.31**
Free flow condition (FFC)	0.07(*p-value=0.06*)	0.14	0.04 (*p-value=0.256*)	**0.29**
Speed variation (SV)	0.009 (*p-value=0.8*)	-0.12	0.07 (*p-value=0.08*)	-0.15
Hard acceleration rate (HAR)	-0.02 (*p-value=0.62*)	0.12	-0.05 (*p-value=0.166*)	-0.025
Hard braking rate (HBR)	-0.01 (*p-value=0.71*)	0.11	-0.03 (*p-value=0.41*)	-0.05

Table 7: Spearman's rho correlation of safety surrogate measures with crash rate and crash frequency in Staten Island

frequency ($\rho_{sV_{\mathbf{mean}},\text{crash rate}} > 0.2$, $\rho_{sV_{\mathbf{max}},\text{crash frequency}} > 0.2$, $\rho_{sV_{\mathbf{mean}},\text{crash frequency}} > 0.2$). Whereas lower maximum speed values and lower mean speed values (i.e. higher travel times), are found to have low to moderate negative correlations with crash rates ($\rho_{sV_{\mathbf{max}},\text{crash rate}} = -0.18$, $\rho_{sV_{\mathbf{mean}},\text{crash rate}} = -0.18$ and $\rho_{s\text{FFC},\text{crash rate}} = -0.17$) on non-highway roads. Speed variations are found to be negatively correlated with crash rates on urban roads ($\rho_{sSV,\text{crash rate}} = -0.23$). This is expected as higher speed variations on urban roads are located on those with higher mean speeds.

Hard braking and hard accelerations had moderate positive correlations with crash rates and crash frequencies on non-highway roads ($\rho_{s\text{HAR},\text{crash rate}} = 0.38$ and $\rho_{s\text{HBR},\text{crash rate}} = 0.33$, $\rho_{s\text{HAR},\text{crash frequency}} = 0.25$, and $\rho_{s\text{HBR},\text{crash frequency}} = 0.2$). It is noteworthy that overall, on highways, hard braking is more indicative of crashes than hard acceleration. On the other hand, on non-highway local roads, the correlation strength of hard acceleration rates with crash rates is stronger than for deceleration rates.

When bucketing road segments by borough, correlation trends remain consistent in general. However, as shown in Table 3, Manhattan borough exhibits stronger correlations between hard acceleration, hard deceleration and crash rates and crash frequencies ($\rho_{s\text{HAR},\text{crash rate}} = 0.56$ and $\rho_{s\text{HBR},\text{crash rate}} = 0.51$, $\rho_{s\text{HAR},\text{crash frequency}} = 0.5$, and $\rho_{s\text{HBR},\text{crash frequency}} = 0.45$).

Also, as shown in Table 7, Staten island non-highway road segments do not have the same speed-crash direction of relationship as the other boroughs. Staten island non-highway road segments with higher mean speed and maximum speed values have higher crash rates and crash frequencies ($\rho_{sV_{\mathbf{max}},\text{crash frequency}} = 0.31$, $\rho_{sV_{\mathbf{mean}},\text{crash frequency}} = 0.31$ on non-highway roads). This finding reveals that Staten island roads labeled as non-highway have indeed different characteristics than the rest of New York City.

4.2. Safety Metrics Generation

Now that the derived indices are proved to be effective safety surrogate measures, given their correlation with collision data, safety metrics are generated. Speed is an important design input for city planners and policy makers' projects. These projects are often carried out at the corridor level since it would be unrealistically granular to carry them out at the segment level. Figures 31 and 32 show, respectively, the maximum speed profile by segment and the maximum speed safety corridors for the off-peak hours 10:00 A.M–16:00 P.M in New York City, using the methodology described in Section 3.7.1. Moreover, Figure 33 shows a map of hard braking and acceleration hot-spots generated using the steps described in Section 3.7.2.

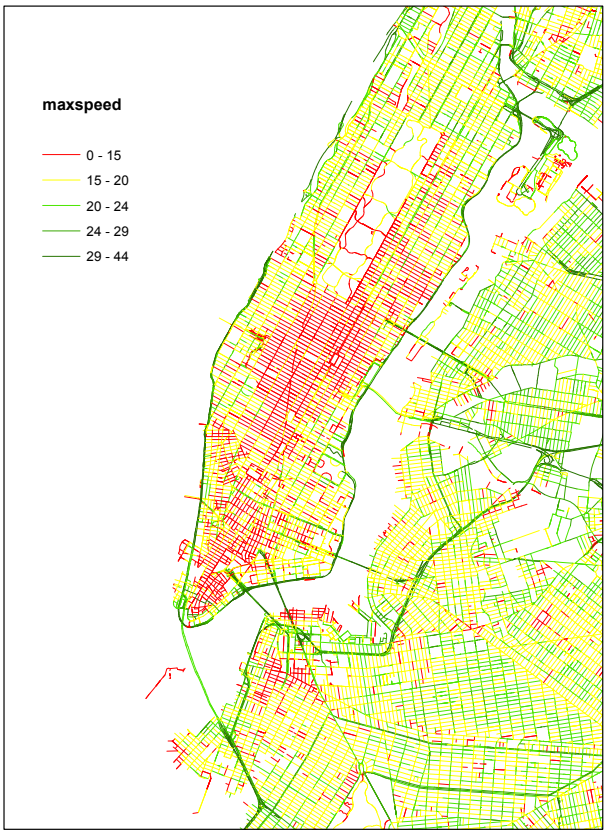

Figure 31: Maximum speed profile by segment for the off-peak hours 10:00 A.M–16:00 P.M in New York City in mph

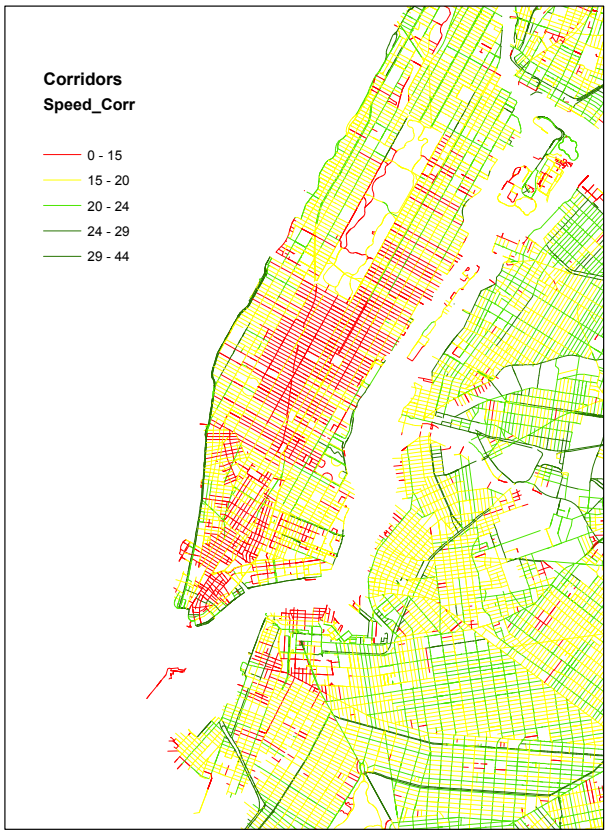

Figure 32: Maximum speed corridors map for the off-peak hours 10:00 A.M–16:00 P.M in New York City in mph

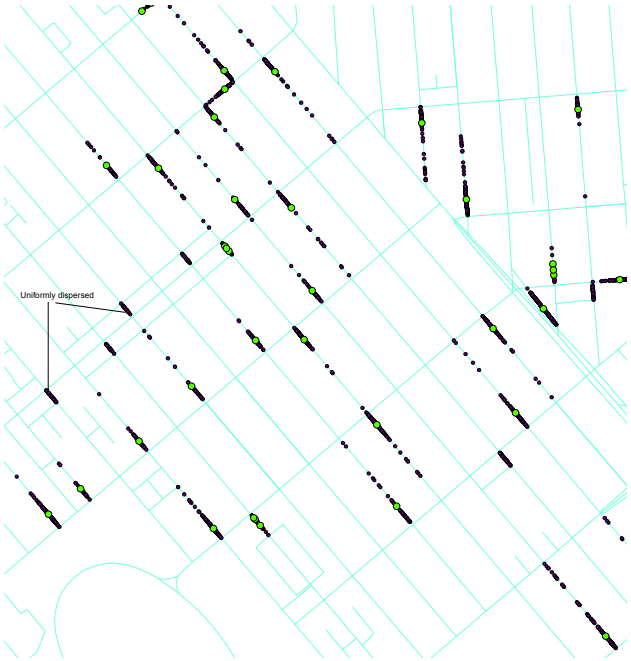

Figure 33: Hard braking clusters

5. Discussion

The results in Section 4 indicate that urban streets that have longer travel times generally present higher safety risk. This result supports the hypotheses found in Quddus (2013); Stipancic et al. (2017). On the other hand, the finding that speeding on highways is more safety critical is also consistent with the conclusions in Taylor et al. (2000); Martin (2002); Wang et al. (2018); Gargoum & El-Basyouny (2016). This study confirms that separating roads according to their functional types is essential and a pre-requisite, before conducting a statistical analysis between SSMs and crashes.

In regard to the relationships between HBRs, HARs and collisions, the findings in this study represent a notable contribution to the existing literature, that is deficient on understanding the spatial relationship between harsh driving and collisions, but rather focuses on the causality between the two variables. On highways, hard braking correlates stronger with crashes than hard acceleration. This is consistent with the fact that hard braking is more indicative of an aggressive driver than hard acceleration, since accelerating on highways is normal as drivers tend to pick up speed. On the other hand, part of the hard braking and hard acceleration on non-highway roads and residential streets are due to the "stop and go" activity that results from congestion, traffic lights and stop signs, as well as driver maneuvers (parking entrance/exit). Looking at the network scale in Figures 21 - 22, one can clearly see that higher rates of hard braking and

hard accelerations occur also in low traffic residential areas. The driver gains speed on these empty streets, causing the accelerometer to capture a hard braking/acceleration event when approaching an intersection, but that intersection will have low collision rates given the small chance of a driver being exposed to another car.

Associating non-spatial driver behavior data (speed, hard braking and hard acceleration) with the correct streets of a road network, and then computing safety metrics on top, such as speed safety corridors or harsh driving clusters, represents a substantial improvement to the strategies of road planners, who often make design interventions based on standards many of which have not been evaluated for their impact on safety (Administration, 2011). To the best of the authors' knowledge, there has been no similar comprehensive study of driver behavior in an urban metropolitan city with a complex road network. The results of this paper present a substantial contribution to the existing literature that is deficient on this topic, when considering metropolitan cities and dense urban environments. Future work will focus on developing a network screening model that incorporates the SSMs to be used for safety classification of streets and intersections.

6. Conclusion

This study examines driver behavior indices derived directly from the vehicle's on-board diagnostic port as promising safety surrogate measures for the entire road network. Methods for assigning the telematics data to the road network, sample validation and extracting the SSMs are presented. The statistical relationship between speed indices, hard braking and hard acceleration rates and collision frequency and rate (normalized with traffic volume) was determined using Spearman's rank correlation method. Highway and non-highway roads are considered separately. Two road safety metrics were then derived, as an example usage of the driver behavior indices for safety assessment: *speed corridor maps* based on spatial partitioning and snake similarities, and *hard braking and hard acceleration hotspots* based on Gaussian mixture dynamic clustering.

This study considers the city-scale, whereas many previous studies considered individual corridors or links. It also presents a substantial addition to the literature as the driver behavior correlations with crashes are studied both on highways and on local roads. It also concludes that the hypotheses that hold on highways does not hold in dense and complex metropolitan cities. Hard braking is more indicative of collision rates on highways than hard acceleration, whereas hard acceleration is found to be a stronger safety indicator than hard braking on dense urban roads. The correlation direction of speed varies also by road type. Longer travel times are linked to crashes in dense urban roads. However, speeding on highways is

more indicative of collision risks. Future work will focus on building a crash prediction model that will incorporate the driver behavior indices in conjunction with other factors such as road geometry, pedestrian exposure and the left-turn and right-turn activity on an intersection.

Integrating vehicular data with map data make generating data-enabled safety metrics and planning possible. It provides an invaluable insight into road safety and driver behavior, compared to sparse and manually collected crash data. Moreover, the specific designs of the interventions, while obviously well intentioned, are typically not supported by large objective performance data that accounts for all variables specific to the location. The potential now exists to radically alter this paradigm by using driver behavior data collected over the entire road network, to both gain insights to locations of anomalous behaviors as well as to quantitatively evaluate the before and after behavior performance of street improvement projects (SIPs).

Acknowledgement

We would like to thank the New York City Department of Transportation for their financial support and guidance in this study. This work is also partially supported by the U.S. National Science Foundation (OAC-1948066).

References

Administration, U.-D. F. H. (2011). Highway safety improvement program manual foundations. URL: https://safety.fhwa.dot.gov/hsip/resources/fhwasa09029/sec1.cfm.

Agerholm, N., & Lahrmann, H. (2012). Identification of Hazardous Road Locations on the basis of Floating Car Data: Method and first results.

Alrassy, P., Jang, J., & Smyth, A. W. (2019). A Novel Vehicle Fleet Data-Assisted Map Matching Algorithm for Safety Ranking and Road Classification in Metropolitan Areas using Low-Sampled GPS Trajectories. URL: https://trid.trb.org/view/1572532.

Amin, S., Andrews, S., Arnold, J., Bayen, A., Chiou, B., Claudel, C., Claudel, C., Dodson, T., Flens-Batina, C., Gruteser, M., Herrera, J., & Herring, R. (2019). Mobile Century Using GPS Mobile Phones as Traffic Sensors: A Field Experiment, .

Bagdadi, O. (2013). Assessing safety critical braking events in naturalistic driving studies. *Transportation Research Part F: Traffic Psychology and Behaviour*, *16*, 117–126. URL: http://www.sciencedirect.com/science/article/pii/S1369847812000770. doi:10.1016/j.trf.2012.08.006.

Boonsiripant, S., Rodgers, M. O., & Hunter, M. P. (2011). Speed profile variation as a road network screening tool. *Transportation research record*, *2236*, 83–91.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, *33*, 261–304.

CA: Environmental Systems Research Institute (). Arcmap—arcgis desktop. URL: https://desktop.arcgis.com/en/arcmap/.

Cai, X., Lei, C., Peng, B., Tang, X., & Gao, Z. (2020). Road traffic safety risk estimation method based on vehicle onboard diagnostic data. *Journal of advanced transportation*, *2020*.

DataKind (2017). DataKind | Creating Safer Streets Through Data Science. URL: https://www.datakind.org/projects/creating-safer-streets-through-data-science.

Davis, G. A., Hourdos, J., Xiong, H., & Chatterjee, I. (2011). Outline for a causal model of traffic conflicts and crashes. *Accident Analysis & Prevention*, *43*, 1907–1919. URL: http://www.sciencedirect.com/science/article/pii/S0001457511001205. doi:10.1016/j.aap.2011.05.001.

Dreyfus, S. E. (1969). An appraisal of some shortest-path algorithms. *Operations research*, *17*, 395–412.

El-Basyouny, K., & Sayed, T. (2013). Safety performance functions using traffic conflicts. *Safety Science*, *51*, 160–164. URL: http://www.sciencedirect.com/science/article/pii/S0925753512001671. doi:10.1016/j.ssci.2012.04.015.

Ellison, A. B., Greaves, S. P., & Bliemer, M. C. (2015). Driver behaviour profiles for road safety analysis. *Accident Analysis & Prevention*, *76*, 118–132.

Eren, H., Makinist, S., Akin, E., & Yilmaz, A. (2012). Estimating driving behavior by a smartphone. In *2012 IEEE Intelligent Vehicles Symposium* (pp. 234–239). doi:10.1109/IVS.2012.6232298.

Forbes, G. (2012). *Methods and practices for setting speed limits: An informational report*. IR-133.

Frank, K. A. (2000). Impact of a confounding variable on a regression coefficient. *Sociological Methods & Research*, *29*, 147–194.

Gargoum, S. A., & El-Basyouny, K. (2016). Exploring the association between speed and safety: A path analysis approach. *Accident Analysis & Prevention*, *93*, 32–40. URL: http://www.sciencedirect.com/science/article/pii/S0001457516301361. doi:10.1016/j.aap.2016.04.029.

Hayward, J. C. (1972). Near-Miss Determination Through Use Of A Scale Of Danger.

Johnson, D. A., & Trivedi, M. M. (2011). Driving style recognition using a smartphone as a sensor platform. In *2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)* (pp. 1609–1615). doi:10.1109/ITSC.2011.6083078.

Johnsson, C., Laureshyn, A., & Ceunynck, T. D. (2018). In search of surrogate safety indicators for vulnerable road users: a review of surrogate safety indicators. *Transport Reviews*, *38*, 765–785. URL: https://doi.org/10.1080/01441647.2018.1442888. doi:10.1080/01441647.2018.1442888.

Jun, J., Ogle, J., & Guensler, R. (2007). Relationships between Crash Involvement and Temporal-Spatial Driving Behavior Activity Patterns Using GPS Instrumented Vehicle Data.

Kockelman, K. M., & Kweon, Y.-J. (2002). Driver injury severity: an application of ordered probit models. *Accident Analysis & Prevention*, *34*, 313–321. URL: http://www.sciencedirect.com/science/article/pii/S0001457501000288. doi:10.1016/S0001-4575(01)00028-8.

Laureshyn, A., Åström, K., & Brundell-freij, K. (2009). From Speed Profile Data To Analysis Of Behaviour: Classification by Pattern Recognition Techniques. *IATSS Research*, *33*, 88–98. URL: http://www.sciencedirect.com/science/article/pii/S0386111214602478. doi:10.1016/S0386-1112(14)60247-8.

Li, P., Abdel-Aty, M., & Yuan, J. (2021). Using bus critical driving events as surrogate safety measures for pedestrian and bicycle crashes based on gps trajectory data. *Accident Analysis & Prevention*, *150*, 105924.

Liaw, A., Wiener, M. et al. (2002). Classification and regression by randomforest. *R news*, *2*, 18–22.

Lord, D. (2006). Modeling motor vehicle crashes using poisson-gamma models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. *Accident Analysis & Prevention*, *38*, 751–766.

Martin, J.-L. (2002). Relationship between crash rate and hourly traffic flow on interurban motorways. *Accid Anal Prev*, *34*, 619–629.

Massey Jr, F. J. (1951). The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, *46*, 68–78.

Mi, X., Shao, C., Dong, C., Zhuge, C., & Zheng, Y. (2020). A framework for intersection traffic safety screening with the implementation of complex network theory. *Journal of advanced transportation*, *2020*.

Moosavi, S., Samavatian, M. H., Parthasarathy, S., Teodorescu, R., & Ramnath, R. (2019). Accident risk prediction based on heterogeneous sparse data: New dataset and insights. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 33–42).

Myers, J. L., Well, A. D., & Lorch Jr, R. F. (2013). *Research design and statistical analysis*. Routledge.

N. Algerholm, H. L. (2012). Identification of Hazardous Road Locations on the basis of Floating Car Data — Aalborg University's Research Portal, . URL: `https://vbn.aau.dk/en/projects/identifikation-af-sorte-pletter-ved-hj%C3%A6lp-af-gps-data-fra-k%C3%B8rende`.

New York Police Department (2019). Motor Vehicle Collisions, NYC Open Data. URL: `https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95`.

NYC Department of City Planning (2020). Lion single line street base map. URL: `https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-lion.page`.

Oh, C., Oh, J.-S., Ritchie, S., & Chang, M. (2001). Real-time estimation of freeway accident likelihood. In *80th Annual Meeting of the Transportation Research Board, Washington, DC*.

Pei, X., Wong, S. C., & Sze, N. N. (2012). The roles of exposure and speed in road safety analysis. *Accident Analysis & Prevention*, *48*, 464–471. URL: `http://www.sciencedirect.com/science/article/pii/S0001457512000942`. doi:`10.1016/j.aap.2012.03.005`.

Quddus, M. (2013). Exploring the Relationship Between Average Speed, Speed Variation, and Accident Rates Using Spatial Statistical Models and GIS. *Journal of Transportation Safety & Security*, *5*, 27–45. URL: `https://doi.org/10.1080/19439962.2012.705232`. doi:`10.1080/19439962.2012.705232`.

Ramamoorthy, K., Ramanouudjam, V., Subramaniam, R. G., & Samuel, J. E. (2014). Remotely monitoring vehicle information using wi-fi direct. US Patent App. 13/901,493.

Rémy, G., Senouci, S.-M., Jan, F., & Gourhant, Y. (2012). Lte4v2x—collection, dissemination and multi-hop forwarding. In *2012 IEEE international conference on communications (ICC)* (pp. 120–125). IEEE.

Reynolds, D. A. (2009). Gaussian mixture models. *Encyclopedia of biometrics*, *741*.

Rodrigue, J.-P. (2017). Urban Transport Challenges. URL: `https://transportgeography.org/?page_id=4621`.

Rolison, J. J., Regev, S., Moutari, S., & Feeney, A. (2018). What are the factors that contribute to road accidents? an assessment of law enforcement views, ordinary drivers' opinions, and road accident records. *Accident Analysis & Prevention*, *115*, 11–24.

Saeedmanesh, M., & Geroliminis, N. (2016). Clustering of heterogeneous networks with directional flows based on "snake" similarities. *Transportation Research Part B: Methodological*, *91*, 250–269.

Smith, D. L., Najm, W. G., & Lam, A. H. (2003). Analysis of braking and steering performance in car-following scenarios.

SAE transactions, (pp. 248–255).

Stipancic, J., Miranda-Moreno, L., & Saunier, N. (2017). Impact of Congestion and Traffic Flow on Crash Frequency and Severity: Application of Smartphone-Collected GPS Travel Data. *Transportation Research Record*, *2659*, 43–54. URL: https://doi.org/10.3141/2659-05. doi:10.3141/2659-05.

Stipancic, J., Miranda-Moreno, L., & Saunier, N. (2018a). Vehicle manoeuvers as surrogate safety measures: Extracting data from the gps-enabled smartphones of regular drivers. *Accident Analysis & Prmoevention*, *115*, 160–169. URL: http://www.sciencedirect.com/science/article/pii/S000145751830109X. doi:10.1016/j.aap.2018.03.005.

Stipancic, J., Miranda-Moreno, L., Saunier, N., & Labbe, A. (2018b). Surrogate safety and network screening: Modelling crash frequency using GPS travel data and latent Gaussian Spatial Models. *Accident Analysis & Prevention*, *120*, 174–187. URL: http://www.sciencedirect.com/science/article/pii/S0001457518303117. doi:10.1016/j.aap.2018.07.013.

Tageldin, A., & Sayed, T. (2016). Developing evasive action-based indicators for identifying pedestrian conflicts in less organized traffic environments. *Journal of Advanced Transportation*, *50*, 1193–1208. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/atr.1397. doi:10.1002/atr.1397.

Tageldin, A., Sayed, T., & Wang, X. (2015). Can Time Proximity Measures Be Used as Safety Indicators in All Driving Cultures? *Transportation Research Record: Journal of the Transportation Research Board*, *2520*, 165–174. doi:10.3141/2520-19.

Tarko, A., A. Davis, G., Saunier, N., & Sayed, T. (2009). Surrogate Measures of Safety, .

Taylor, M. C., Lynam, D. C., & Baruya, A. (2000). The effects Of Drivers' Speed On The Frequency Of Road Accidents. *TRL REPORT 421*, . URL: https://trid.trb.org/view/651648.

U.S. Department of Transportation Federal Highway Administration (2016). *Evaluation of Four Network Screening Performance Measures*. Technical Report. URL: https://safety.fhwa.dot.gov/rsdp/downloads/fhwasa16103.pdf.

Wang, C., Quddus, M. A., & Ison, S. G. (2009). Impact of traffic congestion on road accidents: a spatial analysis of the M25 motorway in England. *Accid Anal Prev*, *41*, 798–808. doi:10.1016/j.aap.2009.04.002.

Wang, X., Zhou, Q., Quddus, M., Fan, T., & Fang, S. (2018). Speed, speed variation and crash relationships for urban arterials. *Accident Analysis & Prevention*, *113*, 236–243. URL: http://www.sciencedirect.com/science/article/pii/S0001457518300381. doi:10.1016/j.aap.2018.01.032.

Yannis, G., Tselentis, D., Paradimitrior, E., Mavromatis, S. et al. (2016). Star rating driver traffic and safety behavior through obd and smartphone data collection, .

Yuan, Z., Zhou, X., & Yang, T. (2018). Hetero-ConvLSTM: A Deep Learning Approach to Traffic Accident Prediction on Heterogeneous Spatio-Temporal Data. (pp. 984–992). doi:10.1145/3219819.3219922.

Zheng, L., Ismail, K., & Meng, X. (2014). Freeway safety estimation using extreme value theory approaches: A comparative study. *Accident Analysis & Prevention*, *62*, 32–41. URL: http://www.sciencedirect.com/science/article/pii/S000145751300359X. doi:10.1016/j.aap.2013.09.006.