Transportation Research Part C 122 (2021) 102917

Contents lists available at ScienceDirect

Transportation Research Part C

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

Classifying travelers' driving style using basic safety messages generated by connected vehicles: Application of unsupervised machine learning

Amin Mohammadnazar, Ramin Arvin, Asad J. Khattak *

Department of Civil & Environmental Engineering, The University of Tennessee, Knoxville, United States

ARTICLE INFO

Keywords:
Driving style classification
Unsupervised machine learning
Big data
Location-based services
Connected vehicles
Basic safety message

ABSTRACT

Driving style can substantially impact mobility, safety, energy consumption, and vehicle emissions. While a range of methods has been used in the past for driving style classification, the emergence of connected vehicles equipped with communication devices provides a new opportunity to classify driving style using high-resolution (10 Hz) microscopic real-world data. In this study, location-based big data and machine learning are used to classify driving styles ranging from aggressive to calm. This classification can be used to customize driver assistance systems, assess mobility, crash risk, fuel consumption, and emissions. This study's main objective is to develop a framework that harnesses Basic Safety Messages (BSMs) generated by connected vehicles to quantify instantaneous driving behavior and classify driving styles in different spatial contexts using unsupervised machine learning methods. To this end, a subset of the Safety Pilot Model Deployment (SPMD) with more than 27 million BSM observations generated by more than 1300 individuals making trips on diverse roadways and through several neighborhoods in Ann Arbor, Michigan, were processed and analyzed. To quantify driving style, the concept of temporal driving volatility, as a surrogate safety measure of unsafe driving behavior, was utilized and applied to vehicle kinematics, i.e., observed speeds and longitudinal/lateral accelerations. Specifically, six volatility measures are extracted and used for classifying drivers. K-means and K-medoids methods are applied for grouping drivers in aggressive, normal, and calm clusters. Clustering results indicate that not only does driving style vary among drivers, but the thresholds for aggressive and calm driving vary across different roadway types due to variations in environment and road conditions. The proportion of aggressive driving styles was also higher on commercial streets than on highways and residential streets. Notably, we propose a Driving Score to measure driving performance consistently across drivers.

1. Introduction

Driving behavior is a broad concept that can be a function of a considerable number of variables and factors including driving performance, environmental awareness, willingness to take risks, and reasoning abilities. Due to the large number of variables that influence driving behavior, it can be difficult to numerically analyze it. Therefore, although drivers develop their distinct driving habits (Elander et al., 1993), previous studies found similarities in driving styles of drivers which enabled the authors to classify drivers into

* Corresponding author.
E-mail address: akhattak@utk.edu (A.J. Khattak).

https://doi.org/10.1016/j.trc.2020.102917
Received 13 June 2020; Received in revised form 30 October 2020; Accepted 1 December 2020
Available online 16 December 2020

distinct groups (Aljaafreh et al., 2012; Kalsoom & Halim, 2013; Dörr et al., 2014; Brombacher et al., 2017; Feng et al., 2018). Driving style is a personal way of driving which can be either habitual or an act of instantaneous preference. The data used for driving style recognition can be obtained either from questionnaires (Taubman-Ben-Ari et al., 2004; Farah et al., 2007; van Huysduynen et al., 2015; Useche et al., 2019) or traceable driving information including position, speed, and acceleration of vehicles (Kalsoom & Halim, 2013; Dörr et al., 2014; Deng et al., 2017; Feng et al., 2018; Mousavi et al., 2019). Driving style recognition could be a very useful application in several different fields. For example, Driver Assistance Systems (DAS) can be programmed to detect different driving behavior in real-time and provide feedback to the drivers, even warning them of risks on the roadway (Meiring & Myburgh, 2015). Furthermore, insurance companies are interested in adopting DAS as a way to profile drivers and offer attractive insurance premiums to their clientele. Driving style classification can also be adapted for fuel consumption (Alessandrini et al., 2012), energy efficiency (Mensing et al., 2014; Ranacher et al., 2016), and emission reduction (Van Mierlo et al., 2004).

The development of connected vehicles (CVs) and location-based services (LBS) provide unprecedented access to information about a driver's location, maneuver, speed, and travel time in real-world driving conditions (Arvin et al., 2020; Hoseinzadeh et al., 2020a,b; Parsa et al., 2020; Khattak et al., 2019). LBS data are categorized into two major groups based on their location acquisition mechanism: 1) data collection via user-end hardware, e.g. smartphones and GPS receivers; 2) data collection via on-board sensors and vehicle-to-infrastructure (V2I) communication, e.g. connected vehicles (Lu & Liu, 2012). Such rich data that are available from advanced sensors capable of sending and receiving Basic Safety Messages (BSMs) is ideal for monitoring traffic conditions and driver behavior.

This study explores variations in driving styles under different roadway contexts using connected vehicle data. Driving styles can vary across different road types, as aggressive driving styles on highways could be perceived as normal behavior on commercial streets and vice versa. To this end, three road types (highways, commercial streets, and residential streets) are considered to be representatives of road types where driving style can be inherently different based on roadway context, and individual driving styles are evaluated separately in these types. The rest of the paper is structured as follows: Section 2 provides previous research on driving style classification and the contribution of this study to the field. Next, Section 3 describes the data used for the analysis. In Section 4, the methodological approach used for driving style classification is presented in detail. Then, Section 5 discusses the results of the study. Section 6 is where the limitations of the study are discussed. Finally, in the last section, the paper concludes with a summary of the key findings and suggestions for future studies.

2. Literature review

Driving style refers to the way drivers choose to drive as an act of instantaneous preference or habitual act. Driving style is reflected in drivers' actions, decisions, and maneuvers (Elander et al., 1993; Ishibashi et al., 2007). Generally, studies about driving style can be categorized into two major groups: survey studies and studies based on vehicle motion information. In the former group, driving style is determined based on self-reported driving behavior which can be obtained through questionnaires. These studies investigate the relation between driving style and driving behavior and personality of drivers (Taubman-Ben-Ari et al., 2004; Farah et al., 2007; van Huysduynen et al., 2015; Useche et al., 2019). The latter group of studies, including this study, uses the motion information of vehicles to classify driving styles of drivers. These studies usually classify driving style into three categories: Aggressive style, Normal style, and Calm (conservative) style (Murphey et al., 2009; Wang & Lukic, 2011; Kalsoom & Halim, 2013; Dörr et al., 2014; Qi et al., 2015; Deng et al., 2017; Feng et al., 2018). Among them, particular attention has been devoted to aggressive driving styles as unsafe driving behavior that increases collision risk. For example, a study carried out by the American Automobile Association Foundation for Traffic Safety revealed that aggressive driving was associated with 56 percent of fatal crashes in the United States (Aggressive driving: Research update, 2009). An aggressive driving style is generally associated with higher speeds, sudden acceleration and deceleration, abrupt changes in vehicle steering wheel angle, and harsh lateral and longitudinal maneuvers (Aljaafreh et al., 2012; Castignani et al., 2015; Deng et al., 2017; Wang et al., 2017), whereas a calm driving style is associated with relatively lower speeds, gradual acceleration, and deceleration, smooth lateral and longitudinal maneuvers, and mild changes in a vehicle steering wheel (Castignani et al., 2015; Deng et al., 2017). A moderate or normal driving style is positioned between these two styles (Aljaafreh et al., 2012; Deng et al., 2017; Wang et al., 2017).

A diverse range of methods and algorithms have been used to classify different driving styles. Generally, these methods can be categorized into two groups: machine learning-based methods and rule-based methods. In rule-based methods, a set of rules are defined for different driving features to detect different driving styles. Several studies used fuzzy logic as a rule-based method for driving style detection (Aljaafreh et al., 2012; Choudhary & Ingole, 2014; Dörr et al., 2014; Castignani et al., 2015; Sun et al., 2015; Alpar & Stojic, 2016), namely in (Dörr et al., 2014), where an online system using fuzzy logic detected driver styles in real-time. Machine learning methods used in the literature include both supervised and unsupervised methods. The Artificial Neural Network (ANN), the Support Vector Machine (SVM), and the Random Forest Decision method are among the most commonly applied methods in driving style classification. MacAdam et al. (1998) used ANN to analyze the aggressiveness levels of 36 drivers by measuring their preference for passing, chasing, or being passed by other vehicles. They found that younger drivers usually drive more aggressively than middle-aged and older drivers. In another study, Brombacher et al. (2017) used ANN to classify and score driving styles based on lateral and longitudinal maneuvers. They categorized driving styles into "very sporty", "sporty", "normal", "defensive", and "very defensive" with an accuracy of 81 percent. Wang et al. (2017) used a semi-supervised approach for driving style classification integrating the K-means method with the SVM method. After labeling the driving data of 20 drivers using the K-means method, they applied the SVM method to classify driving styles into aggressive and normal styles. Their results showed that this method could improve the accuracy of classification by 10 percent (Wang et al., 2017). Li et al. (2017) considered the transition between different

maneuvers as a measure of driving style. They used a Random Forest Decision technique to classify driving behavior into low, moderate, and high-risk styles. The authors found that maneuver transition probabilities are more accurate measures than maneuver frequencies for driving style classification. Combining different machine learning techniques is another approach adopted in (Bejani & Ghatee, 2018) where a Fusion technique was applied to aggregate SVM, K-nearest Neighbors, and Multi-Layer Perception for driving style recognition using smartphone data. One drawback of applying supervised learning methods for driving style classification is that these methods require labeled data as ground truth to train the algorithm, which is mostly not the case for data used in driving style detection. Therefore, unsupervised learning techniques such as clustering methods can be used for labeling different driving styles. The K-means clustering, the hierarchical clustering, the Support Vector Clustering, and the Mixture-based Clustering are among the Unsupervised methods applied in previous studies. For instance, Mantouka et al. (2019) applied a two-stage K-means clustering approach to detect unsafe trips based on acceleration profiles, speeding, and mobile usage obtained from smartphones. They found that the driving styles of drivers mostly change over time. Higgs & Abbas (2013) used the K-means method to evaluate the variation of driving styles within car-following periods. The authors found that drivers showed different driving styles within a car-following period. In another study, Yao et al. (2020) applied dynamic time wrapping and hierarchical clustering to investigate driver behavior patterns of taxi drivers based on the lateral and longitudinal position of the vehicles on curves. Then, driving behavior of the drivers were evaluated form safety and ecological point of view.

Data used for studying driving style can be collected in several ways such as driving simulators (Murphey et al., 2009; Dörr et al., 2014; Wang et al., 2017; Ahangari et al., 2018), test vehicles (MacAdam et al., 1998; Li et al., 2017; Feng et al., 2018; Suzdaleva & Nagy, 2018), and smartphones (Bejani and Ghatee, 2018; Castignani et al., 2015; Mantouka et al., 2019; Sadeghinasr et al., 2019). Although driving simulators are safe, low-cost, and easy to set up, they are not fully representative of real-world conditions (Blana & Golias, 2002; Godley et al., 2002; Groeger & Murphy, 2020) since it is hard to simulate real-world traffic conditions with all their complexity and variety. Also, drivers might lose their spontaneity when they know their driving is being monitored. Data that comes from test vehicles more closely resemble real-world driving conditions. However, due to the high cost of field tests and the difficulty of data collection, the scope of these studies is limited in terms of the number of distinguished drivers, the number of trips, different types of roads driven on, different types of vehicles driven by the drivers, and the area covered by the tests. Data collected from user-end hardware, e.g. smartphones, are more comprehensive than test vehicle data as they contain information from a diverse population of drivers covering various types of roadways and neighborhoods. However, depending on the service a location-based application is providing to users, data might be biased (e.g. drive-safe apps offered by the insurance companies). Furthermore, different factors such as weather conditions, satellite signal blockage, and receiver quality can decrease a GPS's speed measurement accuracy (GPS.gov, 2020). This study, however, benefits from large-scale high-resolution BSM data generated by connected vehicles equipped with vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication devices. The sampled data used in the study contains trajectories from more than 1300 unique drivers over two months. This is a high-resolution comprehensive database that represents real-world naturalistic driving conditions. Therefore, many limitations in previous studies are addressed to a large extent. Note that here the connectivity between vehicles does not affect driving style but the instrumented vehicles provide useful data about vehicle motion information. Furthermore, a notable gap in previous studies was a lack of driving style comparisons in roadways with different contexts. This paper classifies driving style in different road types to show how the perception of an aggressive or calm driving style can vary across different roadways. Therefore, this study contributes to the field in three ways:

1. Developed a framework for harnessing BSMs generated by CVs to comprehensively study driving style under a naturalistic driving condition in different road types considering that perception of an aggressive or calm driving style can vary across different road types.
2. Proposed an approach to quantify variations in instantaneous driving behavior and style over space by developing several temporal volatility measures.
3. Proposed an approach for calculating driving score as a measure of driver performance.

3. Data

This paper used BSMs collected from connected vehicles participating in the Safety Pilot Model Deployment (SPMD) program in Ann Arbor, Michigan. The main goal of this program was to deploy connected vehicle technology in the real-world and evaluate the feasibility of dedicated short-range communication for safety benefits (Bezzina & Sayer, 2014). This one-year program, run by the US Department of Transportation, started in August 2012 with vehicles equipped with V2V and V2I communication devices to transmit BSMs with a frequency of 10 Hz. Containing information from 2836 vehicles equipped with V2V technology and 30 roadside equipment (RSE) covering more than 73 lane-miles on public streets, the SPMD is one of the largest real-world data collection programs ever undertaken in the field (Bezzina & Sayer, 2014). Considering the so-called four Vs of big data - Volume, Velocity, Variety, and Veracity (Yang et al., 2017)- the SPMD data can be considered as big data because they satisfy at least two of the Vs, namely Volume and Veracity. Overall, the SPMD data collection effort consists of light vehicle drivers with comprehensive data collection on the surrounding environment (63 vehicles), light vehicle drivers (2836 drivers), heavy vehicle fleet (3 vehicles), and a transit vehicle fleet (3 vehicles). This study focuses on a sample of 2836 drivers from Ann Arbor community drivers and employees of the University of Michigan. The participants had a valid driver license and drove on average more than 32 miles/day mostly in the model deployment geographic area. After recruiting more than 2500 drivers and vehicles, the vehicles were equipped with four main sensors including a DSRC antenna, an onboard unit (OBU), a GPS antenna (installed on the roof), and a power supply. The data collection was performed for 6 months, and 2 months of the data collected in October 2012 and April 2013 is publicly available through the Research Data

Exchange (RDE) website (https://www.its.dot.gov/data/). This dataset contains high-resolution records (N ~ 225 million BSMs) of vehicles' position (latitudinal, longitudinal, altitudinal) and motion (speed and acceleration). Fig. 1 plots the BSMs generated by the CVs in the study area. This figure demonstrates the spread of trajectories over the study area with high resolution.

4. Methodology

The key objectives of this study are to develop a framework to harness big data generated by CVs, quantify instantaneous drivers' behavior in different roadway contexts, and classify individuals driving style using unsupervised machine learning (Fig. 2). The proposed framework consists of several steps: 1- Data acquisition, 2- Data pre-processing and integration, 3- Quantifying driving volatility, and 4- Driving style classification. The initial hypothesis is that driving styles on highways is different from local streets due to the differences in complexity and driving environmental variation we expect on these roadways since the context of these roadways is different such as land use, built environment, development patterns, access points, and road users. In addition to the comparison of driving styles on highways and local streets, this study further tests the hypothesis that driving styles also vary on different local streets. To do so, the Urban Street Design Guide classification was used, which classifies local streets into two major groups: commercial streets (commercial strip corridors and downtown streets) and residential streets (Urban Street Design Guide, 2013). In order to quantify driving behavior of drivers the concept of driving volatility, as a surrogate safety measure of unsafe driving behavior was used. The concept of driving volatility has been explored in several studies (Wang et al., 2015; Kamrani et al., 2018; Arvin et al., 2019a; Mantouka et al., 2019). Driving volatility shows the degree of deviation from the norm. Therefore, higher driving volatilities indicate higher fluctuation in speed, acceleration, deceleration in both the lateral and longitudinal directions, and as a result, higher driver instability. It has been found that aggressive driving has a positive correlation with driving volatility (Arvin et al., 2019b). Also, previous studies show that driving volatility has a strong positive association with crash frequency (Kamrani et al., 2018) and severity (Arvin et al., 2019b). Therefore, it can be concluded that volatile driving is aggressive behavior which can significantly increase crash risk. Although SPMD data provides us with trajectory and motion information, the threshold values for different driving styles have not been determined. In other words, although variations in drivers' acceleration in a period can be captured, the amount of variation in a driver's acceleration or speed which represents an aggressive, normal, or calm driving style cannot be determined. Therefore, in such cases of dealing with unlabeled data, it is best to use unsupervised learning methods. In the literature, several approaches have been

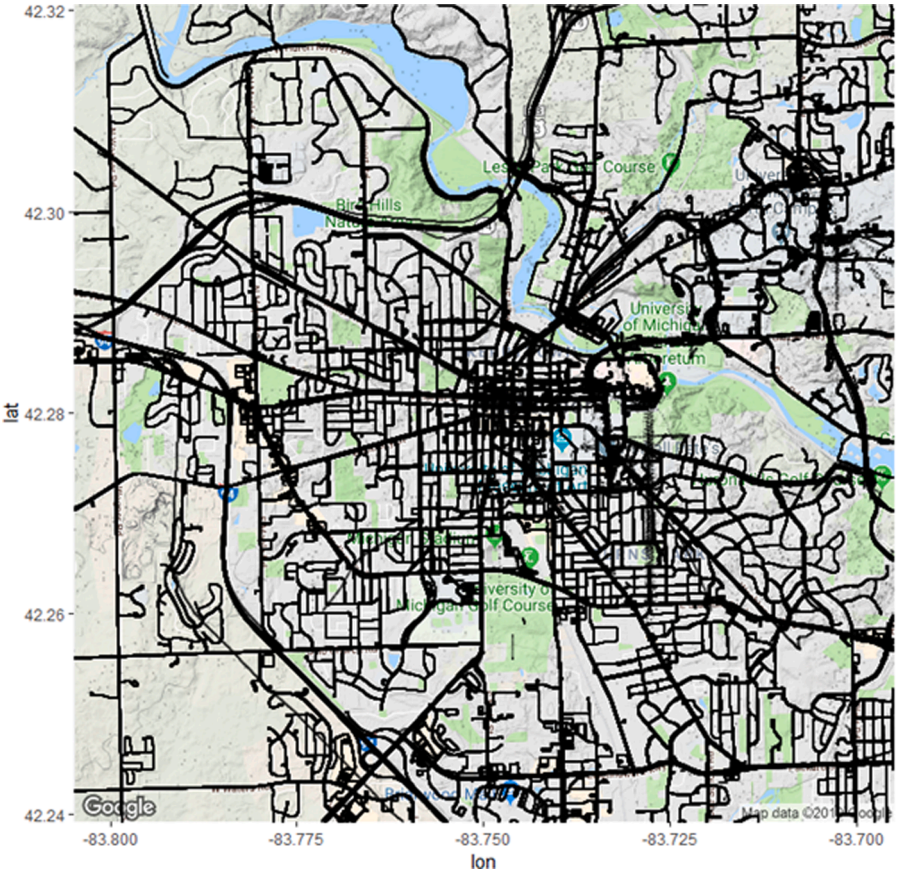

Fig. 1. Study area and generated map by connected vehicles.

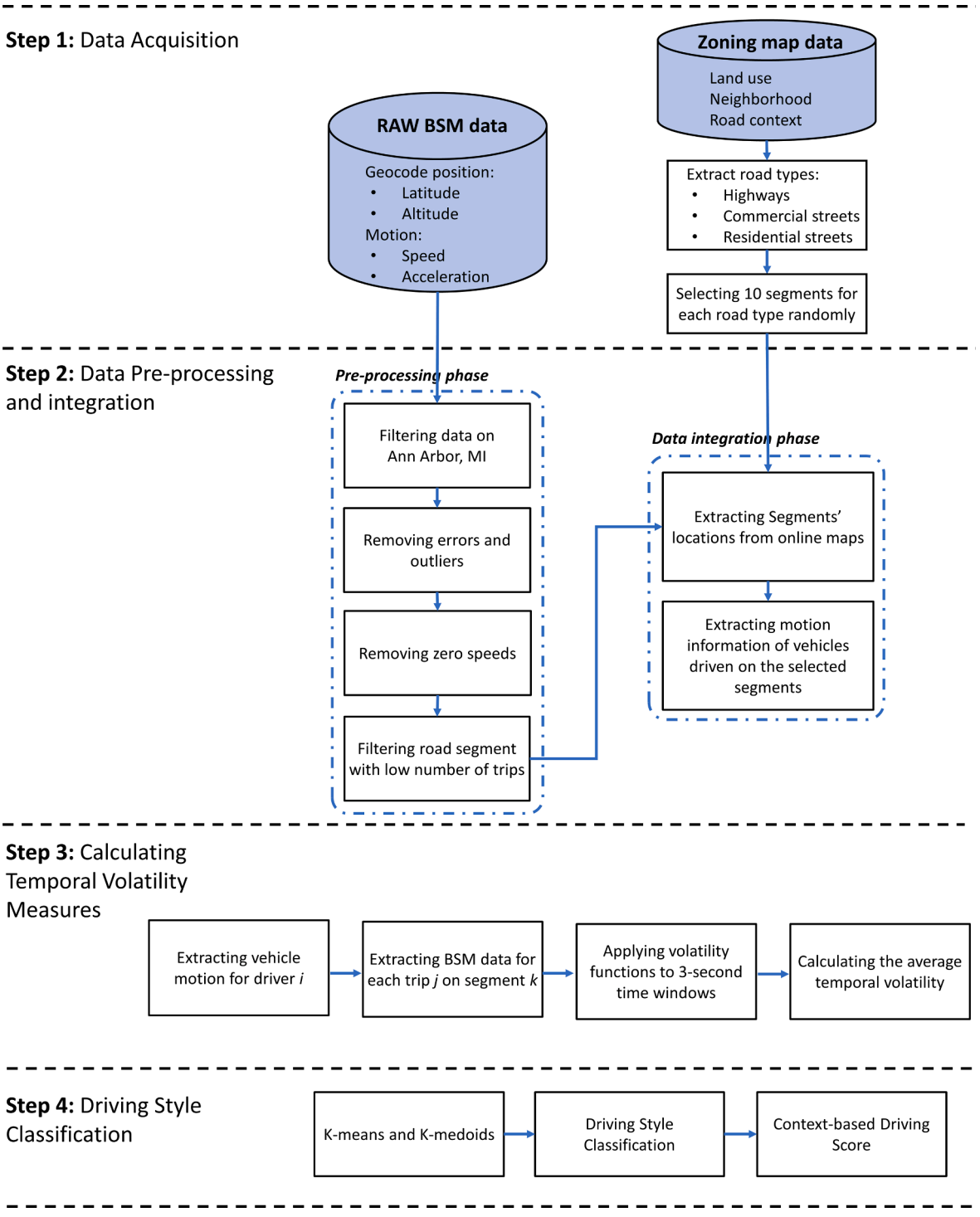

Fig. 2. The methodological approach of the study.

used for clustering such as k-means (Mantouka et al., 2019), hierarchical clustering (Yao et al., 2020), and block clustering (Rahimi et al., 2019). In this study, K-means and K-medoids are performed to aid in driving style classification, which is discussed in detail in the clustering approach section. Driving volatility measures defined by Kamrani et al. (2018) and temporal driving volatility were calculated and used as classification features. After driving style classification, driving style events belonging to each driver were aggregated to score his or her driving. Therefore, the final product of the study is the driving score of drivers in different road types

both separately and overall. As mentioned previously, this study uses connected vehicle data only to access vehicle motion information. Therefore, evaluating the effect of connectivity (e.g., for collision avoidance) between vehicles on the driving style is not within the scope of this analysis. The next sections first explain the pre-processing procedure applied to the SPMD data. Next, the calculations of temporal driving volatility measures as classification features are described in detail. Finally, the clustering approach used for driving style classification is described.

4.1. Data acquisition, pre-processing, and integration

The raw BSM data were obtained from the online SPMD database. Then, to prepare the data for further analysis, pre-processing and data integration steps were applied. First, the data was filtered on the study area, Ann Arbor city, MI. Then, outliers and errors were removed from the dataset based on the recorded speed, longitudinal and lateral acceleration of CVs. The main issue with the dataset is a coding error in acceleration values, which were coded as $\pm g$, made by developers when transferring data from DSRC to CSV files. These errors are removed in the pre-processing step. Finally, zero speeds values were omitted from the data set because they had the potential to affect driving volatility measures substantially.

Online maps were used to distinguish highway segments (e.g., arterials) from local streets using geometric characteristics of the segments (e.g., shoulder width, median width, and level of accessibility). Also, to identify commercial and residential streets, we first need to identify neighborhoods with commercial and residential land use. For this purpose, zoning codes provided on the Ann Arbor zoning map were used (Ann Arbor Zoning Map, 2019). Then, after identifying segments in each area, those segments with a low number of trips were removed from the database. Finally, 10 different segments for each road type were selected randomly from the remaining segments, and their coordinates were extracted from online maps. Fig. 3 illustrates these segments' locations on a map of Ann Arbor. Finally, vehicle kinematic information on the selected segments was extracted from the SPMD data. Therefore, after following these steps, three subsets (highways, commercial streets, and residential streets) were prepared for the classification.

4.2. Calculating temporal driving volatility

In this study, in order to classify travelers' driving style, the concept of temporal driving volatility is utilized. In the following, the volatility functions used for the calculation of driving volatility measures and the concept of temporal driving volatility are presented.

4.2.1. Measures of driving volatility

Previous studies have developed several volatility functions which have been applied to speed (Kamrani et al., 2018; Arvin et al., 2019b, 2019a; Kamrani et al., 2019), lateral acceleration(Arvin et al., 2019a), longitudinal acceleration (Kamrani et al., 2018; Arvin et al., 2019b, 2019a; Kamrani et al., 2019), and vehicular jerk (Kamrani et al., 2018) to evaluate alteration in driving movement. This

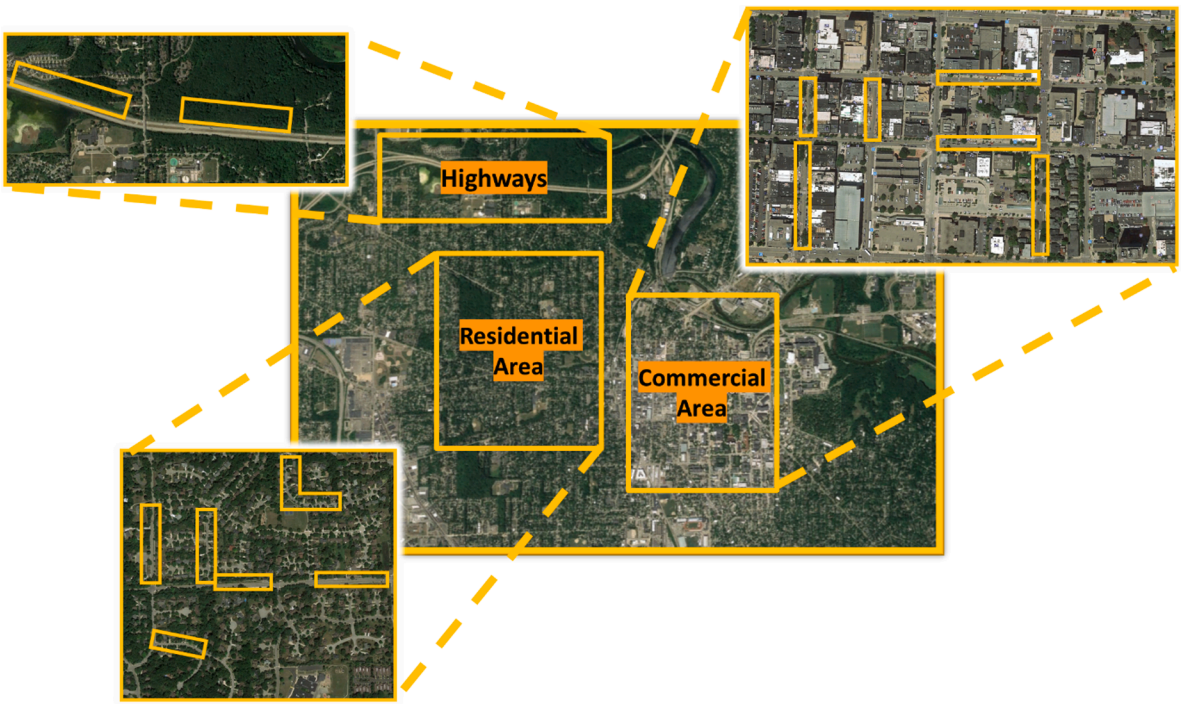

Fig. 3. Location of segments selected in highways, commercial, and residential areas of Ann Arbor, MI.

paper applies several mathematical functions on CV data to develop three groups of driving volatilities:

1. Speed-based volatility
2. Longitudinal acceleration-based volatility
3. Lateral acceleration-based volatility

Volatility functions applied to speed, lateral, and longitudinal acceleration are discussed as follows (Kamrani et al., 2018).

Coefficient of Variation (C_v): This measure takes into account the standard deviation and the mean value to capture the dispersion using the following equation:

$$C_v = \frac{S_{dev}}{|x|} * 100 \tag{1}$$

where S_{dev} is the standard deviation, and $\bar{x}$ is the mean. It can be applied to vehicular speed, acceleration, and deceleration separately.

Mean Absolute Deviation (D_{mean}): This measure calculates the mean distance between the observation of a variable and the central tendency of the data:

$$D_{mean} = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}| \tag{2}$$

where D_{mean} is the mean absolute deviation, n the number of observations, and $\bar{x}$ is the mean. It can be applied to the vehicular speed, acceleration, and deceleration. Note that this measure captures acceleration and deceleration simultaneously.

Quartile Coefficient of Variation (Q_{cv}): This measure considers the dispersion of a dataset as follows:

$$Q_{cv} = \frac{Q_3 - Q_1}{Q_3 + Q_1} * 100 \tag{3}$$

where Q_3 is the third quartile of a variable, and Q_1 is the first quartile.

Time-Varying Stochastic Volatility (V_f): This measure requires observations with a positive time series. Therefore, it can only be applied to vehicular speed which only contains positive values. This measure can be calculated as below:

$$V_f = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (r_i - \bar{r})^2} \tag{4}$$

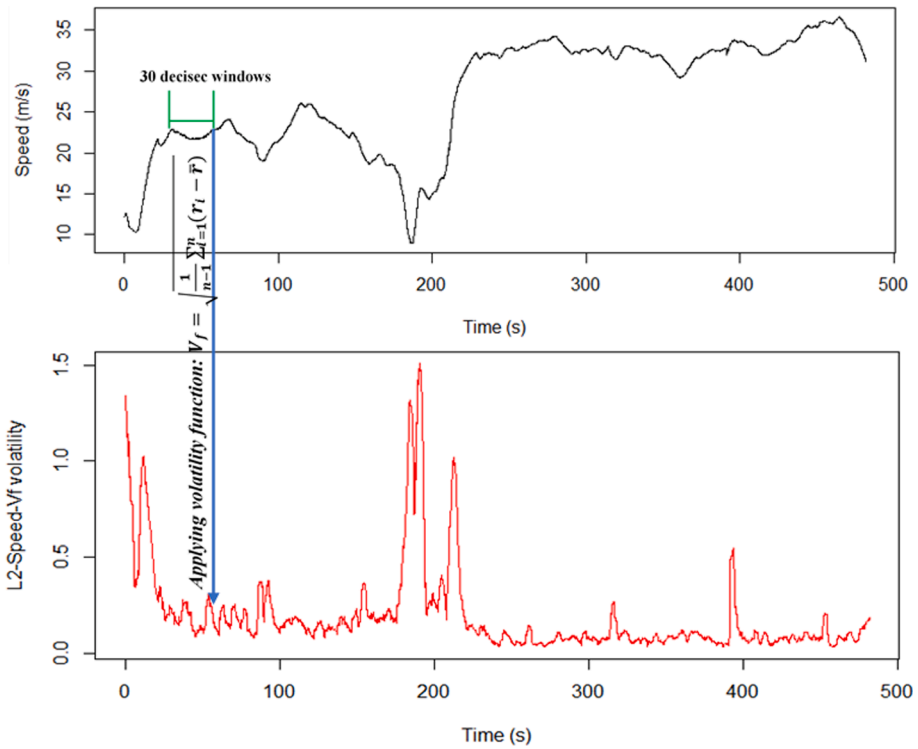

Fig. 4. Calculation of temporal driving volatility.

$$r_i = \ln\left(\frac{x_i}{x_i - 1}\right) * 100 \tag{5}$$

where x_{i-1} is the previous observation regarding the observation x_i, and $\bar{r}$ is the mean value of parameter "r".

4.2.2. Temporal driving volatility

This study employs the concept of temporal driving volatility developed by Arvin et al. (2021), which captures alterations in instantaneous driving behavior by creating a time-series stream of data at the driver level. This provides us with the opportunity to capture driving behavior as a time-dependent variable. To calculate temporal driving volatility, a 3-second time window is considered, and the measures are allocated to the subject time. It was found that using a 3-second time-frame window for the calculation of volatility measures results in the highest correlation of volatility measures with crash risk compared to 1, 2, and 5 s time windows (Arvin & Khattak, 2020). Fig. 4 depicts the calculation of temporal driving volatility using the moving window. Finally, the average temporal volatility is used as a measure of driving style classification.

4.3. Clustering approach

4.3.1. K-means

K-means is an unsupervised method for partitioning objects into k clusters with k mean values called centroids so that each object clusters around the nearest centroid (Jain, 2010). The final output of the K-means method is a set of clusters with their centroids, which minimizes the error function defined below (Gan et al., 2007):

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} d(x, \mu(C_i)) \tag{6}$$

where $C_1, C_2, \cdots, C_k$ are the k clusters, $\mu(C_i)$ is the centroid of cluster C_i, and $d(x, \mu(C_i))$ represents the distance between the observation $\times$ and $\mu(C_i)$.

There are several formulas for the distance calculation. However, in this paper, Euclidean distance is used. If $x = \{x_1, x_2, \cdots, x_n\}$ and $\mu = \{\mu_1, \mu_2, \cdots, \mu_n\}$ are a point and a centroid of a cluster respectively, then the Euclidean Distance from x to μ is calculated using the equation below:

$$d = \sqrt{\sum_{k=1}^{n} (x_k - \mu_k)^2} \tag{7}$$

Assuming D as the data set, and k is the number of clusters, the steps of the K-means algorithm are:

Step 1: Randomly choose sets of centroids $(C_1, C_2, \cdots, C_k)$ from D
Step 2: Assign the remaining data points to the clusters with the closest centroids (min d)
Step 3: Recompute the cluster centroids of the changed clusters above based on the mean value of the data points in each cluster
Step 4: Reassign the data points to the closest clusters based on the new centroids calculated in the last step
Step 5: Repeat steps 3 and 4 until no changes happen in the clusters' membership (no change in the error function)
Step 6: Report the results

4.3.2. K-medoids

In the K-medoids method, instead of centroids, objects cluster around representative objects called medoids (Park & Jun 2009). The error function of the K-medoids algorithm calculates the sum of the dissimilarities between each data point and its corresponding medoid:

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} (x - m(C_i)) \tag{8}$$

where $C_1, C_2, \cdots, C_k$ are the k clusters, and $m(C_i)$ is the medoid of the cluster C_i.

Many algorithms have been proposed for K-medoids clustering. However, this paper applies Partitioning Around Medoids (PAM) as it is one of the most widely used K-medoid algorithms, proposed by Kaufman and Rousseeuw (2009). Assuming D as the data set, and k as the number of clusters, the steps of the PAM algorithm are:

Step 1: Select k data points from the data set randomly as cluster medoids
Step 2: Assign the remaining data points to the clusters with the closest medoids
Step 3: Calculate the error function in equation (8)
Step 4: Select a non-medoid data point from D randomly and each time compute the error function assuming that one of the current medoids are swapped for the selected non-medoid data

Step 5: If the error function calculated in the last step is lower than the one calculated in step 4, then swap the old medoid with the new one

Step 6: Repeat steps 2 to 5 until no changes happen in the clusters' membership (no change in the error function)

4.3.3. Cluster number and clustering method selection criterion

From the perspective of driving style classification, each cluster represents a driving style. A cluster's number is usually selected based on either the subjective judgment of the researcher or clustering quality measures. In this study, using the former approach, 3-cluster classification is considered because the authors believe it is more consistent with previous studies conducted on driving style classification (Canale & Malan, 2002; Murphey et al., 2009; Wang & Lukic, 2011; Kalsoom & Halim, 2013; Dörr et al., 2014; Qi et al., 2015; Deng et al., 2017; Li et al., 2017; Feng et al., 2018).

In this paper, the average silhouette width criterion (*ASWC*) was used for the selection of the clustering method. Generally, ranging from −1 to +1, (*ASWC*) demonstrates how well the objects are classified in the clusters. Higher ASWC values indicate a higher quality of clusters in terms of within-cluster homogeneity and between-cluster separation (Kaufman & Rousseeuw, 2009). Assuming that the data has been clustered in k clusters, for each data point x_i in cluster C_i, the silhouette coefficient of datapoint x_i is calculated as follows (Kaufman & Rousseeuw, 2009):

$$S_{x_i} = \frac{b_{x_i} - a_{x_i}}{\max(a_{x_i}, b_{x_i})} \tag{9}$$

where S_{x_i} is the silhouette coefficient of the datapoint x_i, a_{x_i} is the average distance between x_i and other data points in the same cluster, and b_{x_i} is the minimum average distance between datapoint x_i and data points in any other clusters. Then, the silhouette average of each cluster and the silhouette average of all clusters can be calculated using equations (10) and (11):

$$SWC_j = \frac{1}{n}\sum_{i=1}^{n} S_{x_i} \tag{10}$$

$$ASWC = \frac{1}{k}\sum_{j=1}^{k} S_j \tag{11}$$

where n is the number of data points in the same cluster, and k is the number of all clusters.

5. Results and discussion

5.1. Descriptive statistics

Table 1 shows the descriptive statistics of the driving volatility measures at the driver-level in different road types. Based on the results, driving volatility measures have higher values in commercial streets compared to residential streets and highways, indicating that driving in commercial streets is more volatile than in other roadways. For example, the mean values of time-varying stochastic volatility (*Speed-*V_f) for the drivers in highways, commercial streets, and residential streets are 0.044, 0.155, and 0.053, respectively. The higher values of driving volatility in commercial streets are expected because of the higher variation in environment and road conditions in commercial districts, e.g. traffic signals, access points, and commercial establishments, and more complex interaction with other road users (pedestrian and bicyclists). Similarly, driving in residential streets is more volatile than in highways given the fact that driving environments in residential streets are more complex and variable. This evidence indicates that the idea of classifying the driving style separately for different road classes is logical.

Fig. 5 illustrates the distribution of BSMs generated by CVs at different times of the day over two months on different road types. As expected, a higher number of BSMs have been generated on highways compared to residential and commercial streets which is

Table 1
Descriptive statistics of the temporal volatility measures at driver-level.

Volatility Measures	Highway Drivers N = 1302				Commercial Street Drivers N = 975				Residential Street Drivers N = 840			
	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD
***Speed-*V_f**	0.028	35.983	0.185	0.597	0.000	138.11	3.740	7.560	0.082	27.695	0.673	1.157
***Speed-*D_{mean}**	0.026	5.935	0.304	0.276	0.000	9.967	0.870	0.550	0.065	15.915	0.715	0.515
***Speed-*C_v**	0.000	0.729	0.006	0.013	0.000	1.110	0.120	0.103	0.003	0.561	0.033	0.031
***Speed-*Q_{CV}**	0.000	0.693	0.005	0.012	0.000	0.857	0.101	0.085	0.002	0.459	0.028	0.027
***Accl*$_x$*-*D_{mean}**	0.016	0.717	0.130	0.042	0.000	1.510	0.354	0.109	0.058	0.883	0.236	0.082
***Accl*$_y$*-*D_{mean}**	0.000	0.758	0.076	0.044	0.000	1.941	0.095	0.084	0.000	2.701	0.094	0.088

Note: V_f: Time-Varying Stochastic Volatility C_v: Coefficient of variation; Q_{cv}: Quartile coefficient of variation; D_{mean}: Mean absolute deviation; $Accl_y$: Lateral acceleration; $Accl_x$: Longitudinal acceleration.

consistent with the higher traffic volumes on these roads. Generally, the figure indicates that the BSM data used in this study covers a wide range of traffic conditions at different times of the day.

5.2. Clustering method selection

Each observation in the dataset represents a driving event which will be classified in the nearest cluster. The data was standardized to make the classification features comparable. Then, the "*cluster*" and "*factoextra*" packages in *R* software were utilized to perform K-means and K-medoids clustering (Maechler et al., 2013; Kassambara & Mundt, 2016). Fig. 6 shows the chart of *ASWC* values for different numbers of clusters on highways, commercial streets, and residential streets using K-means and K-medoids methods. From the figure, the K-means method has higher *ASWC* values meaning that it provides more quality clusters (in terms of within-cluster homogeneity and between-cluster separation) compared to the K-medoids method. Particularly, for 3-cluster classification, *ASWC* values for the K-means method provide better results, and therefore, the K-means method was selected over the K-medoids method.

5.3. Classification results

When performing K-means clustering on the dataset, each event was assigned a number (1, 2, or 3) to represent the number of the cluster in which the driving event has been placed. Deciding which driving style is represented by each cluster was based on the researcher's judgment. The mean values of classification features for each cluster were used to determine the driving style that was being represented. Table 2 shows the scaled mean values of classification features for highways, commercial streets, and residential streets. This table shows that for highways and residential streets, clusters 1, 2, and 3 have the highest driving volatility measures, respectively. Also, for commercial streets, except for Mean Absolute Value of Speed ($Speed\text{-}D_{mean}$) and Mean Absolute Value of Lateral acceleration ($Accl_y\text{-}D_{mean}$), the clusters follow the same order. This has been depicted more clearly in Fig. 7 using radar charts of cluster centers. In these charts, classification features are illustrated on axels starting from the same point in which the length of each spoke shows the magnitude of the corresponding feature in the cluster. Drawing a line to connect these values creates a blue region in which a larger size indicates the higher a features' mean values are in the cluster. Given the fact that high values of driving volatilities represent aggressive and risky driving (Arvin et al., 2019b), it is logical to allocate the large, medium, and small regions (clusters 1, 2, and 3) to

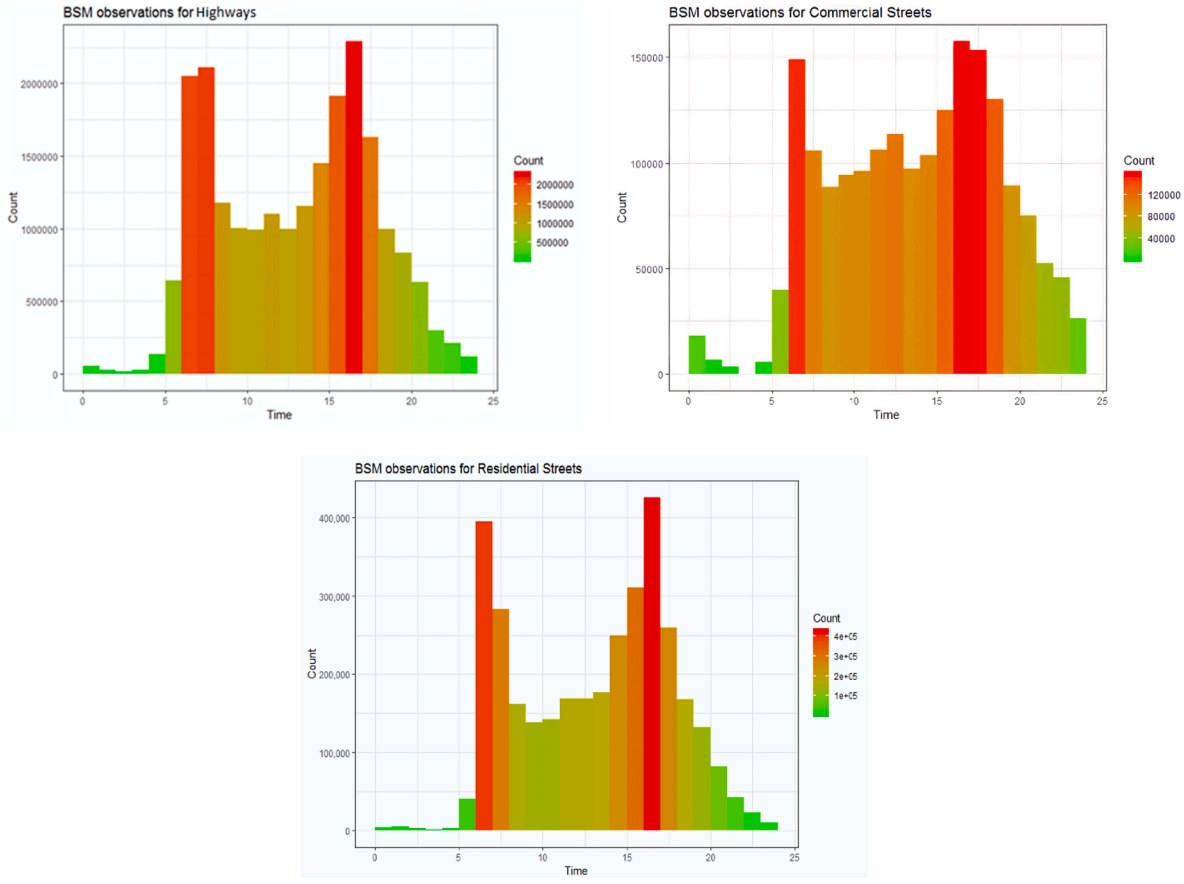

Fig. 5. Histograms of BSM observations over time for highways, commercial streets, and residential streets.

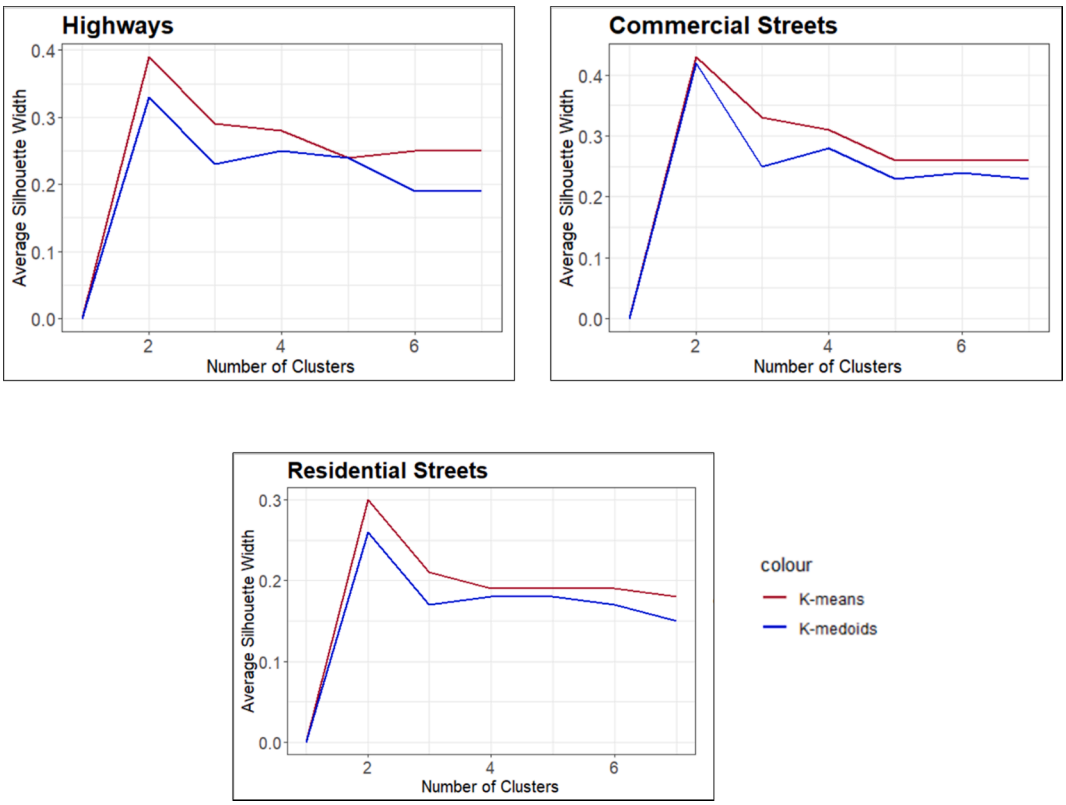

Fig. 6. Comparison of K-means and K-medoids clustering performance using *ASWC*.

Table 2
The scaled cluster centers for highways, commercial streets, and residential streets.

Cluster	*Speed-V_f*	*Speed-D_{mean}*	*Speed-C_v*	*Speed-Q_{CV}*	*$Accl_x$-D_{mean}*	*$Accl_y$-D_{mean}*
Highways						
Cluster 1 (Aggressive)	1.572	1.705	1.971	1.940	1.096	0.654
Cluster 2 (Normal)	0.147	0.316	0.186	0.187	0.334	0.337
Cluster 3 (Calm)	−0.570	−0.775	−0.717	−0.710	−0.630	−0.514
Commercial streets						
Cluster 1 (Aggressive)	1.486	0.669	1.500	1.500	0.505	−0.245
Cluster 2 (Normal)	−0.182	0.712	−0.079	−0.073	0.761	0.818
Cluster 3 (Calm)	−0.569	−0.584	−0.617	−0.620	−0.533	−0.231
Residential streets						
Cluster 1 (Aggressive)	2.209	1.803	2.356	2.338	1.789	1.297
Cluster 2 (Normal)	0.399	0.607	0.494	0.493	0.332	0.151
Cluster 3 (Calm)	−0.558	−0.636	−0.639	−0.636	−0.458	−0.274

aggressive, normal, and calm driving styles respectively.

Table 3 shows the results of classifying driving style events for each road type including the number and percentage of aggressive, normal, and calm driving. Results show that of the 7537 events recorded from 1302 drivers in highways, 11.85, 44.13, and 44.02 percent were classified as aggressive, normal, and calm driving, respectively. This indicates that aggressive driving is not a recurring behavior in highways. Instead, the majority of drivers have normal and calm driving styles on highways. In residential streets, the proportion of aggressive events is smaller than highways (7.96 percent), and calm driving accounts for the majority of events (56 percent). These values indicate that overall driving behavior on residential streets is consistent with the nature of this neighborhood, which requires cautious and slow driving. Although the proportion of aggressive driving is higher in commercial streets than in highways and residential streets (23.47 percent), calm driving events are still the most frequent style in this road type (54.18 percent). A higher number of aggressive events in commercial streets could be because of the more complex and various environments in these segments, which makes drivers have sharp acceleration and deceleration more frequently.

By performing Principal Component Analysis (PCA), the number of features was reduced from a 6-dimensional space to a 2-dimensional space to visualize the clusters. Table 4 shows the contribution percentage of the classification features in the first two principal

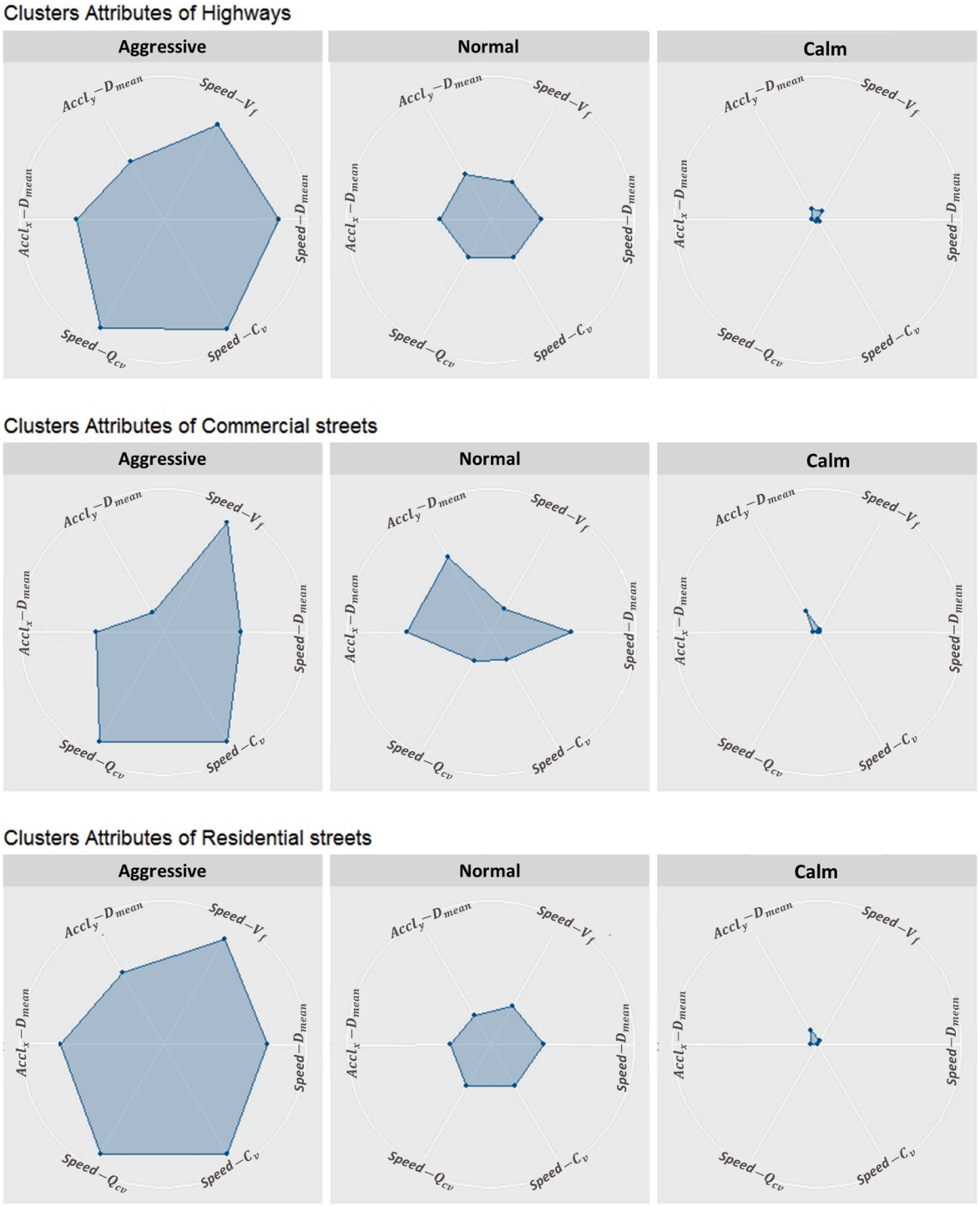

Fig. 7. Radar charts of classification features mean values in the clusters for highways, commercial, and residential streets.

components and the percentage of variance explained by each principal component. From the table, it can be found that the first two principal components explain 78.61, 82.29, and 81.93 percent of the variance in highways, commercial, and residential streets, respectively. Also, the contribution percent values indicate that the first principal component (PC_1)in all three road types is largely influenced by the variables which capture speed variation (*Speed-*C_v, *Speed-*Q_{CV}, *Speed-*D_{mean} and *Speed-*V_f*)*, while the second principal component (PC_2) is mostly influenced by the variables which capture acceleration variation (*Accleration*$_x$*-*D_{mean} and *Accleration*$_y$*-*D_{mean}).

Table 3
Number of trips classified in each driving style cluster.

Type of road	Total num. of events	Total num. of drivers	Driving style					
			Aggressive		Normal		Calm	
			Number of events	Percent %	Number of events	Percent %	Number of events	Percent %
Highways	7537	1302	893	11.85	3326	44.13	3318	44.02
Commercial streets	3387	975	795	23.47	757	22.35	1835	54.18
Residentialstreets	1912	840	147	7.96	690	36.09	1075	56.22

Table 4
PCA loadings and percentage of variance explained with each component for the two first principal components.

Classification Features	Highways		Commercial Streets		Residential Streets	
	PC1* (63.92%) **	PC2 (14.69%)	PC1 (59.60%)	PC2 (22.69%)	PC1 (66.98%)	PC2 (14.95%)
Speed-V_f	14.39***	3.06	22.71	5.60	17.43	0.33
Speed-D_{mean}	21.67	3.00	15.25	12.40	19.24	6.52
Speed-C_v	23.61	5.37	25.51	4.34	22.77	6.32
Speed-Q_{CV}	23.10	6.33	25.56	4.08	22.57	6.62
$Accl_x$-D_{mean}	11.73	1.15	10.66	18.91	11.96	14.37
$Accl_y$-D_{mean}	5.50	81.09	0.30	54.66	6.02	65.85

* Principal component.
** Percentages of variance explained by the component.
*** Percentage of the feature contribution in the principal component.

Fig. 8 illustrates the labeled clusters in a 2-dimensional area. As depicted in the figure, the clusters for calm and normal driving styles on highways are more compact, containing a higher number of events. However, for commercial and residential streets, the majority of the events are clustered in the blue cluster, showing that most of the driving events in these roadways are classified as a calm driving style. This figure is consistent with the results presented in table 3. The significant lower proportion of aggressive driving events could be the result of several factors.

5.4. Driver tracking and driving score

Now that driving events are classified in different roads, we are able to track drivers in different roadways by aggregating the driving events corresponding to each individual. Table 5 shows the proportion of drivers who had a constant driving style. For instance, on highways, 23 drivers always possessed an aggressive driving style, 1.76 percent of the whole population. Similarly, in commercial streets, 4.31, 8.72, and 28.00 percent of the drivers always exhibited aggressive, normal, and calm driving, respectively. Remarkably, in residential streets, 40.71 percent of the drivers always drove calmly on the streets. Overall, considering all the segments, 1.73, 6.35, and 3.46 percent of the1385 drivers in the dataset always had an aggressive, normal, and calm driving style, while 88.59 percent of the individuals had various behavior at different locations.

In the next step, individual drivers were tracked and profiled to evaluate their driving performance in different road types. In order to quantify driving styles, values of 1, 2, or 3 were assigned to aggressive, normal, and calm driving, respectively. Therefore, driving style is a discrete variable that takes on values from 1 to 3. As discussed, there can be different values of driving style for a driver as he/she can have different driving behavior on different road segments. Therefore, to have a single value as a measure of driving performance, we define "Driving Score" which is the average of these values for each road type. To this end, driving events were aggregated based on drivers' ID in highways, commercial streets, and residential streets. Therefore, for each driver, there are three Driving Scores ranging between 1 and 3 which reflect the driver's performance on different roads. Unlike driving style, Driving Score is a continuous variable that can take on any values between 1 and 3 which gives us more flexibility in assessing drivers' performance. Fig. 8 illustrates the schematics of the Driving Score as a measure of an excellent, good, or bad driver, and the histogram charts of the Driving Score in all road types. A Driving Score of 3 represents calm and low-risk driving whereas a driving score of 1 represents aggressive driving which increases collision risk. The Driving Score is a variable measure that changes by drivers' performance, meaning that a driver with a low score can drive more calmly in order to improve his or her score and vice versa. In Fig. 9, the right-skewed histograms of Driving Scores on all three road types indicate that there is a higher quantity of drivers with good and excellent performance than the quantity of drivers of poor and very poor performance. This is confirmed by the shape of the 'overall' histogram.

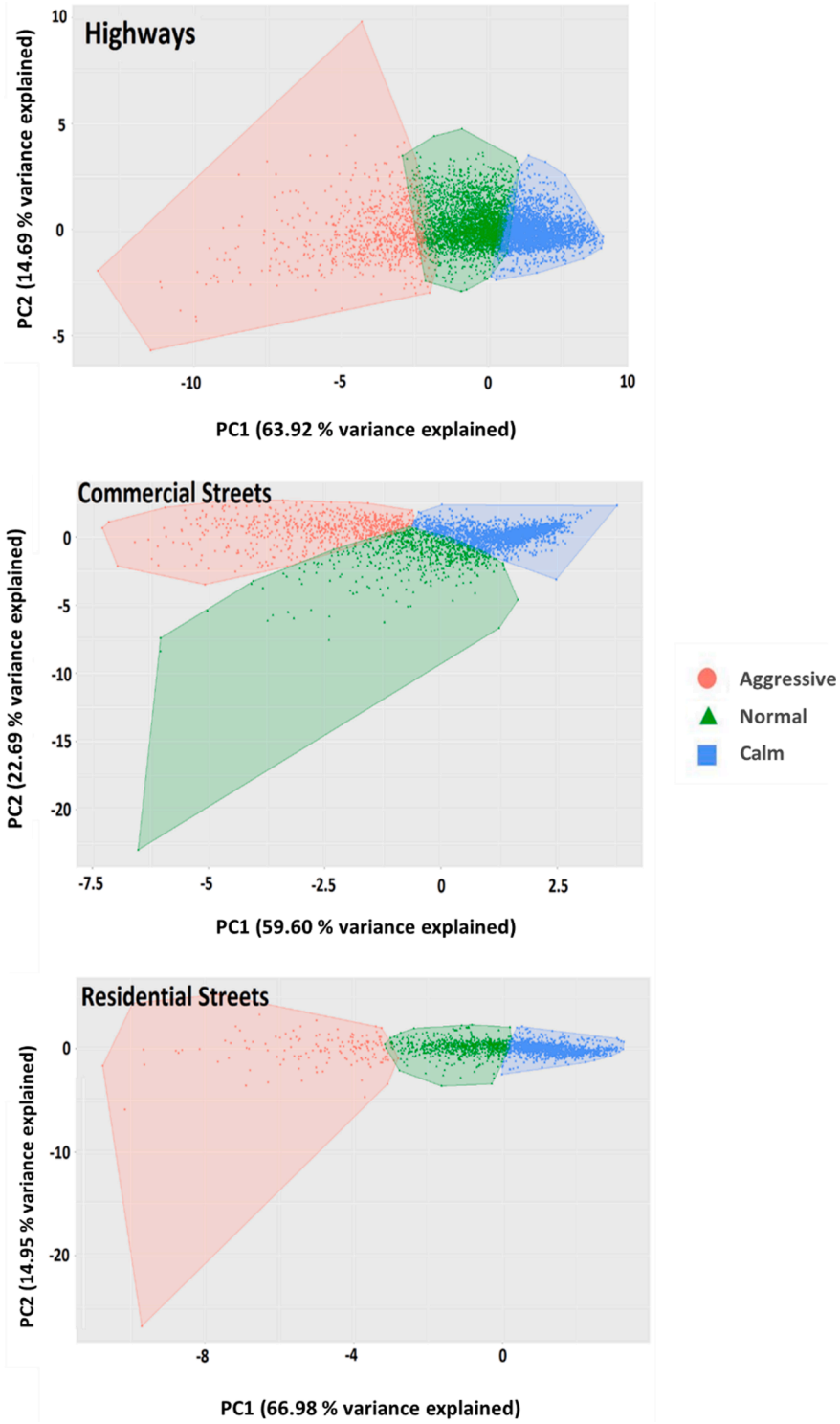

Fig. 8. Visualization of clusters in 2-dimensional space.

5.5. Discussion

Previous studies proved that not only does driving style vary among drivers (Aljaafreh et al., 2012; Castignani et al., 2015; Deng et al., 2017; W. Wang et al., 2017), but that a driver can change his or her driving style over time (Higgs & Abbas, 2013; Dörr et al.,

Table 5
Numbers and proportion of drivers with a constant driving style in different roads.

Type of road	Number of drivers	Aggressive drivers		Normal drivers		Calm drivers		Drivers with various behavior	
		Count	Percent %	Count	Percent %	Count	Percent %	Count	Percent %
Highways	1302	23	1.76	126	9.67	91	6.98	1062	81.56
Commercial streets	975	42	4.31	85	8.72	273	28.00	575	58.97
Residential streets	840	15	1.79	107	12.74	342	40.71	376	44.76
Overall	1385	11	0.79	80	5.78	67	4.84	1227	88.59

2014; Hong et al., 2014; Suzdaleva & Nagy, 2018). Harnessing BSMs generated by connected vehicles, this study extends the understanding of driving style by showing that driving styles vary among different roadway types. Different thresholds obtained for aggressive, normal, and calm driving styles on freeways, commercial, and residential streets indicate that the perception of an aggressive or calm driving style varies across different road types. This is because of the differences in complexity, environmental variation, number of access points, traffic conditions, and interactions with other road users (pedestrian and bicyclists) among these roadways. Given the fact that traffic flow on uninterrupted flow facilities such as highway arterials is more uniform than commercial and residential streets (Liu & Khattak, 2016), sudden acceleration and deceleration, abrupt changes in vehicle steering wheel angle, and harsh lateral and longitudinal maneuvers are more likely to be conceived as an aggressive driving style on these roads. This study also provides new insight into driving styles at both microscopic and macroscopic scales. At the macroscopic scale, this study shows the proportions of different driving styles on different roadways. For example, it was found that the majority of the driving style events on highways, commercial streets, and residential streets were classified as normal and calm driving styles. In other words, an aggressive driving event is not always a recurrent behavior on roadways segments analyzed in this study. This finding is in line with the results of the 2008 AAA Foundation's Traffic Safety Culture Index survey where most of the drivers reported that they had engaged in risky driving behaviors only a few times in the past 30 days (Traffic safety culture index, 2008). At a microscopic scale, however, this study tracks individual drivers to evaluate differences in their driving styles on different roadways. One advantage of this approach is detecting risky drivers who continuously drive aggressively during the observation period or drivers who always drive calmly. The results show that a significant proportion of the drivers had various driving styles while driving on different roadways, e.g. normal drivers who sometimes drive aggressively. It is consistent with the result of previous studies that found that the driving style of individual drivers changes over time (Dörr et al., 2014; Hong et al., 2014; Suzdaleva & Nagy, 2018). Furthermore, this study proposes a driving score that assesses drivers' performance on different roadways. Ranging from 1 to 3, the driving score converts the driving style as a discrete variable to a continuous variable that can change by drivers' performance. Another advantage of using driving scores is to find the distribution of driving performance on different roadways. Right-skewed histograms of driving scores on a roadway indicate that on average drivers had good performances on the roadway.

6. Limitations

While the connected vehicles in the SPMD study collect information of subject vehicle using OBU, the data does not include information on the surrounding environment and traffic conditions of roadways for the trips undertaken by drivers. We acknowledge that variation in traffic conditions of roadways may affect driving behaviors. However, this limitation is mitigated to some extent by evaluating the driving style on highways, commercial streets, and residential streets separately. Also, this study takes into account the overall performance of the drivers on roadways over two months. Therefore, the driving performance of drivers is the aggregate under different traffic conditions that can happen during this period. Furthermore, in this study, 10 segments on highways, commercial and residential streets were selected as representative segments for the driving classification of each road type. Although a significant proportion of the drivers who participated in the SPMD program drove on these segments, they may not be representative of the whole population of segments for a type of road. If a larger sample size was used, the classification would likely be more accurate. Also, this paper only takes into account variations in longitudinal and lateral vehicle movements. However, other measures can represent driving styles such as lane change behavior, tailgating behavior, and time to collision. However, the calculation of these measures requires information regarding the surrounding environment which is not collected in this dataset. With higher market penetration of CVs in future, this information will be available to the researchers for future studies.

7. Conclusion

With the emergence of connected vehicles, researchers have unprecedented access to location-based data, which can be coupled with big data analytics and machine learning methods, to extract valuable information and get new insights about micro and macro level transportation performance. This research benefits from a sample of the SPMD data as a large-scale real-world database containing the BSMs generated by connected vehicles. This dataset includes position and motion information of more than 1300 vehicles making trips on diverse roadways and through several neighborhoods in Ann Arbor. Focusing on driver behavior, the main objective of this study is to develop a framework to harness BSM data to study and quantify instantaneous driving behavior in order to classify driving styles. The main contribution of this study is to quantify volatility in driving behavior using high-volume real-world CV data and classify driving styles in different road types using unsupervised machine learning methods. Finally, an approach was proposed for scoring drivers' aggressiveness which can be used as a measure of driving performance on different road types.

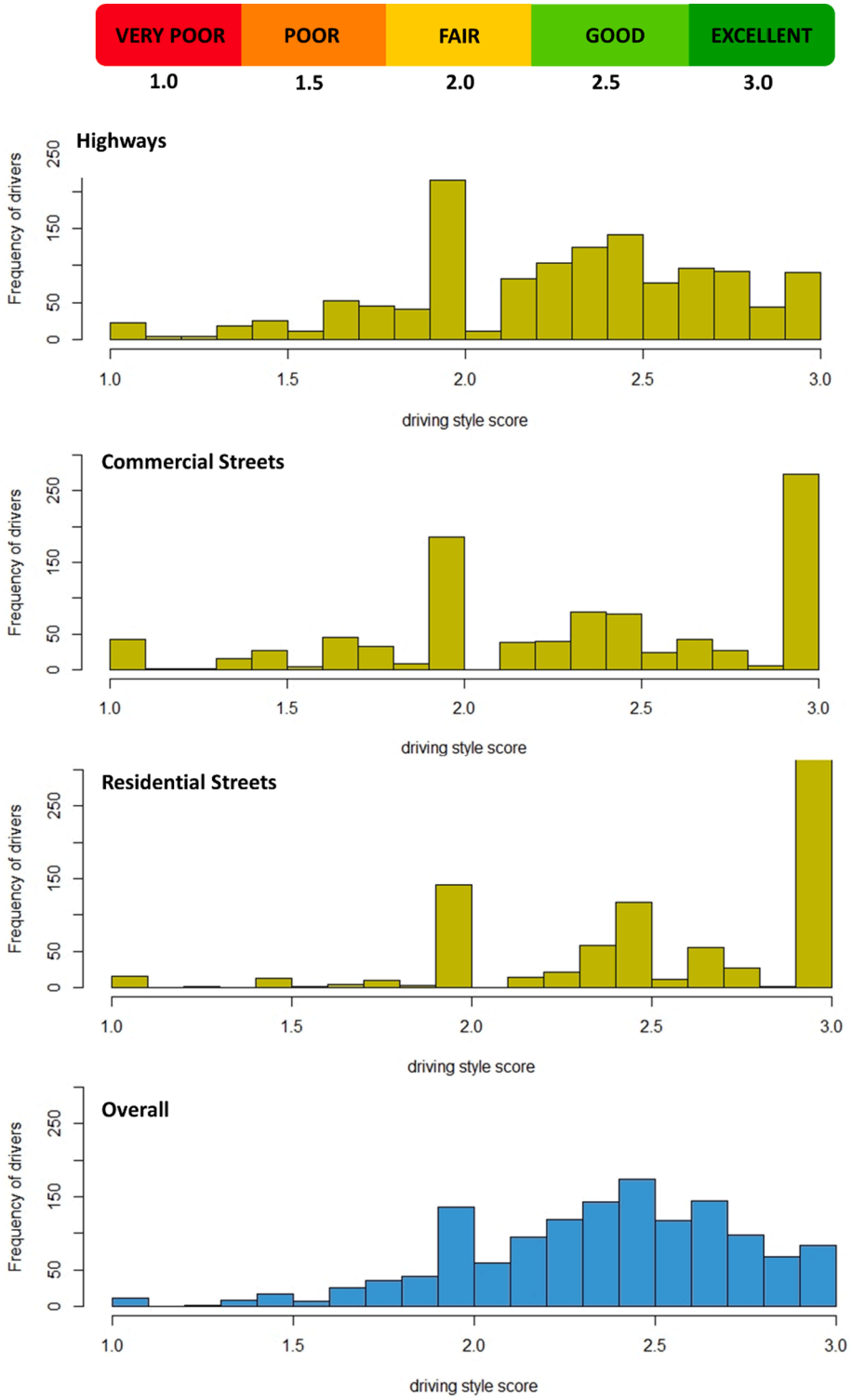

Fig. 9. Range of Driving Score criterion and histogram of driving scores both separately and overall.

The results show that the proposed method in this study could successfully extract the driving volatility features from raw BSM data to quantify driving behavior at different road types. The clustering results indicate that the K-means method provides more accurate classification results compared to the K-medoids method. It was found that there was a significant difference between the three groups of drivers in terms of temporal driving volatilities which helped the authors classify drivers based on these measures. Clustering results

show that perception and thresholds of an aggressive, normal, and calm driving vary in different road types. In other words, an aggressive style in highways can be perceived as normal driving in commercial streets and vice versa. Moreover, the analysis indicates that aggressive driving, as one of the most important contributing factors in motor vehicle crashes, is not a recurrent behavior on roads. The results show that the proportion of aggressive driving events is highest in commercial streets (4.31 percent) and lowest in highways (1.76 percent). Also, the histograms of the driving scores in different roadways demonstrate that drivers showed good performance overall. This study uses one of the most high-resolution datasets for driving style classification, containing driving information from more than 1300 unique drivers. Therefore, the methodology and results of this study can be generalized to a high extent in other areas.

The findings of the study can be used in location-based services to improve the safety of vehicles on different types of roadways. For example, it can have an application in Driving Assistance Systems (DAS) for monitoring driving behavior and giving feedback about their driving performance. The driving score proposed in this study can be a good measure of driving performance of the drivers. If the driving score of a driver indicates that he or she is driving aggressively, the driver could be alerted/advised to drive more calmly. This methodology can also be applied to individual CVs which could then be submitted by the car owners to the insurance companies as a record of driving style history. Another application of the study is in fuel consumption and emission studies. Because volatile and aggressive driving styles increase fuel consumption and emissions (Alessandrini et al., 2012; Mahdinia et al., 2020), air pollution hot spots could be identified with high accuracy.

Future studies could integrate crash data with CV data to evaluate the association of crash frequency/severity with driving styles on different road types. Also, temporal evaluation of driving style can be considered in future studies to assess the variation of driving styles over time (e.g. variation of driving style at different hours of a day). Furthermore, although this study uses criteria such as the number of trips and number of connected vehicles passing through the segments for selecting the representative segments, considering additional selection criteria such as safety conditions (i.e., segments with high crash rates) and traffic conditions (e.g. segments with high traffic volumes) can be studied in future research.

CRediT authorship contribution statement

Amin Mohammadnazar: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft, Writing - review & editing. **Ramin Arvin:** Conceptualization, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft, Writing - review & editing. **Asad J. Khattak:** Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing.

References

Aggressive driving: Research update. (2009). American Automobile Association Foundation for Traffic Safety, 5–6.

Ahangari, S., Jeihani, M., & Jarju, A. (2018). Drivers' behavior analysis under rainy weather conditions using a driving simulator. Paper presented at the Transportation Research Board 97th Annual Meeting, Washington DC.

Alessandrini, A., Cattivera, A., Filippi, F., Ortenzi, F., 2012. Driving style influence on car CO2 emissions. Paper presented at the 2012 international emission inventory conference.

Aljaafreh, A., Alshabatat, N., Al-Din, M.S.N., 2012. Driving style recognition using fuzzy logic. In: Paper presented at the 2012 IEEE International Conference on Vehicular Electronics and Safety (ICVES 2012).

Alpar, O., Stojic, R., 2016. Intelligent collision warning using license plate segmentation. J. Intell. Transp. Syst. 20 (6), 487–499.

Ann Arbor Zoning Map, 2019. Retrieved from https://gisappsecure.ewashtenaw.org/Html5Viewer_2_12/index.html?viewer=A2Zoning.City_of_Ann_Arbor_Zoning.

Arvin, R., Kamrani, M., Khattak, A.J., 2019a. How instantaneous driving behavior contributes to crashes at intersections: extracting useful information from connected vehicle message data. Accid. Anal. Prev. 127, 118–133.

Arvin, R., Kamrani, M., Khattak, A.J., 2019b. The role of pre-crash driving instability in contributing to crash intensity using naturalistic driving data. J. Accident Anal. Prevent. 132, 105226.

Arvin, R., Khattak, A.J., 2020. Harnessing big data generated by connected vehicles to monitor safety performance: Application of geographically weighted negative binomial regression. In: Paper presented at the Transportation Research Board 99th Annual Meeting, Washington DC.

Arvin, R., Khattak, A., Kamrani, M., Rios-Torres, J., 2020. Safety evaluation of connected and automated vehicles in mixed traffic with conventional vehicles at intersections. Journal of Intelligent Transportation Systems 1–18.

Arvin, R., Khattak, A., Qi, H., 2021. Safety critical event prediction through unified analysis of driver and vehicle volatilities: Application of deep learning methods. Accident Analysis and Prevention. In press.

Bejani, M.M., Ghatee, M., 2018. A context aware system for driving style evaluation by an ensemble learning on smartphone sensors data. Transp. Res. Part C: Emerg. Technol. 89, 303–320.

Bezzina, D., Sayer, J., 2014. Safety pilot model deployment: Test conductor team report (DOT HS 812 171). Retrieved from https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/812171-safetypilotmodeldeploydeltestcondrtmrep.pdf.

Blana, E., Golias, J., 2002. Differences between vehicle lateral displacement on the road and in a fixed-base simulator. Hum. Factors 44 (2), 303–313.

Brombacher, P., Masino, J., Frey, M., Gauterin, F., 2017. Driving event detection and driving style classification using artificial neural networks. Paper presented at the 2017 IEEE International Conference on Industrial Technology (ICIT).

Canale, M., Malan, S., 2002. Analysis and classification of human driving behaviour in an urban environment. Cogn. Technol. Work 4 (3), 197–206.

Castignani, G., Derrmann, T., Frank, R., Engel, T., 2015. Driver behavior profiling using smartphones: a low-cost platform for driver monitoring. IEEE Intell. Transp. Syst. Mag. 7 (1), 91–102.

Choudhary, A.K., Ingole, P.K., 2014. Smart phone based approach to monitor driving behavior and sharing of statistic. Paper presented at the 2014 Fourth International Conference on Communication Systems and Network Technologies.

Deng, C., Wu, C., Lyu, N., Huang, Z., 2017. Driving style recognition method using braking characteristics based on hidden Markov model. PLoS ONE 12 (8), e0182419.

Dörr, D., Grabengiesser, D., Gauterin, F., 2014. Online driving style recognition using fuzzy logic. Paper presented at the 17th International IEEE Conference on Intelligent Transportation Systems (ITSC).

Elander, J., West, R., French, D., 1993. Behavioral correlates of individual differences in road-traffic crash risk: An examination of methods and findings. Psychol. Bull. 113 (2), 279.
Farah, H., Polus, A., Bekhor, S., Toledo, T., 2007. Study of passing gap acceptance behavior using a driving simulator. Adv. Transp. Stud. Int. J. 9–16.
Feng, Y., Pickering, S., Chappell, E., Iravani, P., Brace, C., 2018. Driving style analysis by classifying real-world data with support vector clustering. Paper presented at the 2018 3rd IEEE International Conference on Intelligent Transportation Engineering (ICITE).
Gan, G., Ma, C., Wu, J., 2007. Data Clustering: Theory, Algorithms, and Applications, Vol. 20. Siam.
Godley, S.T., Triggs, T.J., Fildes, B.N., 2002. Driving simulator validation for speed research. Accid. Anal. Prev. 34 (5), 589–600.
GPS.gov, 2020. Official U.S. government information about the Global Positioning System (GPS) and related topics, GPS Accuracy. Retrieved from https://www.gps.gov/systems/gps/performance/accuracy/#speed.
Groeger, J.A., Murphy, G., 2020. Driver performance under simulated and actual driving conditions: validity and orthogonality. Accid. Anal. Prev. 143, 105593.
Higgs, B., Abbas, M., 2013. A two-step segmentation algorithm for behavioral clustering of naturalistic driving styles. In: Paper presented at the 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013).
Hong, J.-H., Margines, B., Dey, A.K., 2014. A smartphone-based sensing platform to model aggressive driving behaviors. Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.
Hoseinzadeh, N., Arvin, R., Khattak, A.J., Han, L.D., 2020a. Integrating safety and mobility for pathfinding using big data generated by connected vehicles. J. Intell. Transp. Syst. 1–17.
Hoseinzadeh, N., Liu, Y., Han, L.D., Brakewood, C., Mohammadnazar, A., 2020b. Quality of location-based crowdsourced speed data on surface streets: a case study of Waze and Bluetooth speed data in Sevierville, TN. Comput. Environ. Urban Syst. 83, 101518.
Ishibashi, M., Okuwa, M., Doi, S.i., Akamatsu, M., 2007. Indices for characterizing driving style and their relevance to car following behavior. In: Paper presented at the SICE Annual Conference 2007.
Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern Recogn. Lett. 31 (8), 651–666.
Kalsoom, R., Halim, Z., 2013. Clustering the driving features based on data streams. Paper presented at the INMIC.
Kamrani, M., Arvin, R., Khattak, A.J., 2018. Extracting useful information from Basic Safety Message Data: an empirical study of driving volatility measures and crash frequency at intersections. Transp. Res. Rec. 2672 (38), 290–301.
Kamrani, M., Arvin, R., Khattak, A.J., 2019. The role of aggressive driving and speeding in road safety: Insights from SHRP2 naturalistic driving study data. In: Paper presented at the Transportation Research Board 98th Annual Meeting, Washington DC.
Kassambara, A., Mundt, F., 2016. Package 'factoextra'. Extract and Visualize the Results of Multivariate Data Analyses.
Kaufman, L., Rousseeuw, P.J., 2009. Finding Groups in Data: An Introduction to Cluster Analysis, Vol. 344. John Wiley & Sons.
Li, G., Li, S.E., Cheng, B., Green, P., 2017. Estimation of driving style in naturalistic highway traffic using maneuver transition probabilities. Transp. Res. Part C: Emerg. Technol. 74, 113–125.
Liu, J., Khattak, A.J., 2016. Delivering improved alerts, warnings, and control assistance using basic safety messages transmitted between connected vehicles. Transp. Res. Part C: Emerg. Technol. 68, 83–100.
Lu, Y., Liu, Y., 2012. Pervasive location acquisition technologies: opportunities and challenges for geospatial studies. Comput. Environ. Urban Syst. 36 (2), 105–108.
MacAdam, C., Bareket, Z., Fancher, P., Ervin, R., 1998. Using neural networks to identify driving style and headway control behavior of drivers. Veh. Syst. Dyn. 29 (S1), 143–160.
Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., 2013. Package 'cluster'. Dosegljivo na.
Mahdinia, I., Arvin, R., Khattak, A.J., Ghiasi, A., 2020. Safety, energy, and emissions impacts of adaptive cruise control and cooperative adaptive cruise control. Transp. Res. Rec. 0361198120918572.
Mantouka, E.G., Barmpounakis, E.N., Vlahogianni, E.I., 2019. Identifying driving safety profiles from smartphone data using unsupervised learning. Saf. Sci. 119, 84–90.
Meiring, G., Myburgh, H., 2015. A review of intelligent driving style analysis systems and related artificial intelligence algorithms. Sensors 15 (12), 30653–30682.
Mensing, F., Bideaux, E., Trigui, R., Ribet, J., Jeanneret, B., 2014. Eco-driving: an economic or ecologic driving style? Transp. Res. Part C: Emerg. Technol. 38, 110–121.
Mousavi, S.M., Zhang, Z., Parr, S.A., Pande, A., Wolshon, B., 2019. Identifying high crash risk highway segments using jerk-cluster analysis. Paper presented at the International Conference on Transportation and Development 2019: Smarter and Safer Mobility and Cities.
Murphey, Y.L., Milton, R., Kiliaris, L., 2009. Driver's style classification using jerk analysis. Paper presented at the 2009 IEEE Workshop on Computational Intelligence in Vehicles and Vehicular Systems.
Park, H.-S., Jun, C.-H., 2009. A simple and fast algorithm for K-medoids clustering. Expert Syst. Appl. 36 (2), 3336–3341.
Parsa, A.B., Shabanpour, R., Mohammadian, A., Auld, J., Stephens, T., 2020. A Data-Driven Approach to Characterize the Impact of Connected and Autonomous Vehicles on Traffic Flow.
Qi, G., Du, Y., Wu, J., Xu, M., 2015. Leveraging longitudinal driving behaviour data with data mining techniques for driving style analysis. IET Intel. Transport Syst. 9 (8), 792–801.
Rahimi, A., Azimi, G., Asgari, H., Jin, X., 2019. Clustering approach toward large truck crash analysis. Transp. Res. Rec. 2673 (8), 73–85.
Ranacher, P., Brunauer, R., Van der Spek, S.C., Reich, S., 2016. A model to estimate and interpret the energy-efficiency of movement patterns in urban road traffic. Comput. Environ. Urban Syst. 59, 152–163.
Sadeghinasr, B., Akhavan, A., Wang, Q., 2019. Estimating Commuting Patterns from High Resolution Phone GPS Data. Computing in Civil Engineering 2019: Data, Sensing, and Analytics. American Society of Civil Engineers Reston, VA, pp. 9–16.
Sun, R., Ochieng, W.Y., Feng, S., 2015. An integrated solution for lane level irregular driving detection on highways. Transp. Res. Part C: Emerg. Technol. 56, 61–79.
Suzdaleva, E., Nagy, I., 2018. An online estimation of driving style using data-dependent pointer model. Transp. Res. Part C: Emerg. Technol. 86, 23–36.
Taubman-Ben-Ari, O., Mikulincer, M., Gillath, O., Prevention, 2004. The multidimensional driving style inventory—scale construct and validation. Accident Anal. Prevent. 36 (3), 323–332.
Traffic safety culture index, 2008. Washington. AAA Foundation for Traffic Safety, DC.
Urban Street Design Guide, 2013. National Association of City Transportation Officials, Washington, DC.
Useche, S.A., Cendales, B., Alonso, F., Pastor, J.C., Montoro, L., behaviour, 2019. Validation of the Multidimensional Driving Style Inventory (MDSI) in professional drivers: how does it work in transportation workers? Transp. Res. Part F: Traffic Psychol. 67, 155–163.
van Huysduynen, H.H., Terken, J., Martens, J.-B., Eggen, B., 2015. Measuring driving styles: a validation of the multidimensional driving style inventory. Paper presented at the Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications.
Van Mierlo, J., Maggetto, G., Van de Burgwal, E., Gense, R., 2004. Driving style and traffic measures-influence on vehicle emissions and fuel consumption. Proc. Inst. Mech. Eng., Part D: J. Automobile Eng. 218 (1), 43–50.
Wang, R., Lukic, S.M., 2011. Review of driving conditions prediction and driving style recognition based control algorithms for hybrid electric vehicles. Paper presented at the 2011 IEEE Vehicle Power and Propulsion Conference.
Wang, W., Xi, J., Chong, A., Li, L., 2017. Driving style classification using a semisupervised support vector machine. IEEE Trans. Hum.-Mach. Syst. 47 (5), 650–660.
Wang, X., Khattak, A.J., Liu, J., Masghati-Amoli, G., Son, S., 2015. What is the level of volatility in instantaneous driving decisions? Transp. Res. Part C: Emerg. Technol. 58, 413–427.
Yang, C., Yu, M., Hu, F., Jiang, Y., Li, Y., 2017. Utilizing Cloud Computing to address big geospatial data challenges. Comput. Environ. Urban Syst. 61, 120–128.
Yao, Y., Zhao, X., Wu, Y., Zhang, Y., Rong, J., 2020. Clustering driver behavior using dynamic time warping and hidden Markov model. J. Intell. Transp. Syst. 1–14.
Khattak, A., Mahdinia, I., Mohammadi, S., Mohammadnazar, A., & Wali, B. (2019). *Big Data Generated by Connected and Automated Vehicles for Safety Monitoring, Assessment and Improvement, Final Report (Year 3)*. Retrieved from https://doi.org/10.13140/RG.2.2.22542.18246.