Unsupervised Approach to Investigate Urban Traffic Crashes Based on Crash Unit, Crash Severity, and Manner of Collision

Farzin Maniei, Ph.D.[1]; and Stephen P. Mattingly, Ph.D., A.M.ASCE[2]

Abstract: Both crash frequency analysis (CFA) and real-time crash prediction models (RTCPMs) divide a highway into small segments with a constant length [typically 0.161 km (0.10 mi)] for data aggregation. Many previous studies refer to this constant length as the segment length for data aggregation, but this paper adopts fragment size to avoid confusion with aggregation based on highway geometric features. Several studies have shown that segmentation length impacts the studies' results and recommend not using a length smaller than 0.161 km (0.10 mi) or greater than 0.402 km (0.25 mi) to segment and aggregate traffic data for urban/suburban highways and freeways. Despite the significant impact of the segmentation length on traffic crash aggregation, no specific recommendation for selecting or determining the segmentation length for crash data aggregation exists. This research investigates the impact of segmentation length on traffic crash data aggregation. It establishes a methodology for determining a recommended fragment size (RFS) using hidden heterogeneity in traffic crash data. The study defines featured traffic crash rates using three major traffic crash characteristics: number of vehicles in crash, manner of collision, and crash severity. The analysis uses the Laplacian score with distance-based entropy measure and K-means to cluster highway segments based on the featured crash rates (FCRs) and total crash rates (TCRs) for fragment sizes ranging from 0.161 to 0.402 km (0.10 to 0.25 mi) with an increment of 0.016 km (0.01 mi). The clustering results are compared using their silhouette coefficients. The sample results shows that FCR-based clustering outperforms TCR-based clustering by providing important traffic crash groups within a highway and the RFS to segment and aggregate traffic crash data. The proposed method provides a data-driven comparison of different fragment sizes, revealing the pattern of traffic crashes and a standardized approach for RFS, which reduces the likelihood of fragment misclassification and benefits traffic studies depending on segmentation length. **DOI: 10.1061/JTEPBS.TEENG-7852.**

Author keywords: Segment length; Number of vehicles involved in crash; Manner of collision; Crash severity; Unsupervised learning.

Introduction

Traffic crashes represent one type of incident, defined as an "unplanned randomly occurring traffic event that adversely affects normal traffic operation" (Wang and Feng 2019). Previous studies arbitrarily select the segment length as a constant value between 0.161 km (0.1 mi) and 1.6 km (1.0 mi) (or, in some studies, 100 m to 1.6 km) based on the study's objectives [TxDOT (Texas DOT) Traffic Safety Division 2020]. Choosing different segment lengths for aggregation may result in some variables becoming either statistically significant or insignificant (Ahmed and Abdel-Aty 2012). It is recommended not to use a segmentation length smaller than 0.161 km (0.1 mi) (AASHTO 2010) or a spacing interval greater than 0.402 km (0.25 mi) to segment and aggregate traffic data for urban/suburban highways and freeways (Alabama DOT 2015); however, no specific method currently exists to select segment length. This paper adopts the term fragment size to avoid confusion because the tern *segment length* is used to refer to not only explanatory variable representing the length of roadway section in some studies but also the length selected to divide a roadway to smaller units for data aggregation in some other studies. This study proposes an innovative method to provide a recommended fragment size for data aggregation based on historical crash risk.

Since selecting of fragment size (segment length) for aggregation may cause variables to become statistically significant or insignificant, creating a standard methodology for selecting a suitable fragment size (segment length) for aggregation appears essential for future research. Previous studies argue that the selection of arbitrary fixed-size fragments (segments) for aggregating crash data generates fundamental problems in crash frequency analysis (Pedregosa et al. 2011). Previous research fails to provide any standardized guidance or methodology to select the fragment size (segment length) to aggregate crash data. Since the selection of fragment size (segment length) impacts traffic safety research, this study seeks to investigate and propose a method to find a recommended segment length.

Generally, safety studies can investigate traffic crashes based on different crash characteristic dimensions such as number of vehicles involved (Xu et al. 2018), manner of collision (Cheng et al. 2017; Bhowmik et al. 2018; Mahmud and Gayah 2021), and crash severity (Yu and Abdel-Aty 2013; Afghari et al. 2020). This study also seeks to capture the crash patterns and transitions between crash combinations across highways based on three major traffic crash characteristics: number of vehicles involved in crashes (crash units), manner of collision, and crash severity, simultaneously.

The number of vehicles involved in a crash represents an important crash characteristic dimension that will affect the results

[1] Dept. of Civil Engineering, Univ. of Texas at Arlington, Box 19308, Arlington, TX 76019 (corresponding author). ORCID: https://orcid.org/0000-0002-2071-2043. Email: farzin.maniei@mavs.uta.edu

[2] Professor, Dept. of Civil Engineering, Univ. of Texas at Arlington, Box 19308, Arlington, TX 76019. ORCID: https://orcid.org/0000-0001-6515-6813. Email: mattingly@uta.edu

Note. This manuscript was submitted on December 12, 2022; approved on February 9, 2024; published online on May 20, 2024. Discussion period open until October 20, 2024; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Transportation Engineering, Part A: Systems***, © ASCE, ISSN 2473-2907.

of aggregate traffic crash analyses. Previous studies investigate traffic crashes based on number of vehicles involved by grouping the crashes into two categories: single-vehicle (SV) and multivehicle (MV) crashes because the crash contributing factors may differ or demonstrate different impacts for SV and MV crashes (Abdel-Aty et al. 2006; Ivan et al. 1999; Islam and Pande 2020). Yu and Abdel-Aty (2013) show that the selected crash contributing factors have different impacts on SV and MV crashes and recommend that future safety analyses need to consider the number of vehicles involved as a traffic characteristic for both aggregate and disaggregate approaches. This study includes the number of vehicles involved in crashes by creating SV and MV categories for crash features. The manner of collision, which refers to the first event in a crash, represents another important traffic crash characteristic. Some previous studies refer to the manner of collision as crash type and show that including the manner of collision (crash type) reveals facts about traffic crashes that traffic studies conducted based on total crashes would fail to recognize (Golob et al. 2008). This and other studies (Islam et al. 2017; Cheng et al. 2017) support the importance of including the manner of collision in safety analyses; therefore, the authors integrate the manner of collision as another traffic crash feature dimension. Several studies also investigate the impact of traffic crash contributing factors on crash severity (Abdel-Aty 2003; Islam and Pande 2020). This study combines crash severity as a crash feature with the number of vehicles involved and the manner of collision to create a more refined crash combination than TCR.

This study investigates the effect of segment length for aggregating data and clustering roadway segments using the number of vehicles involved in the crash, manner of collision, and crash severity simultaneously. The clustering approach is selected for roadway segmentation because it can mitigate crash heterogeneity for within-group elements by grouping roadway segments with similar crash distributions into homogeneous groups, according to (Lu et al. 2013). The focus on the crash characteristics makes grouping the data based on the crash characteristics critical for understanding patterns in the crash data. However, some temporal instability (Islam and Mannering 2020) and unobserved heterogeneity associated with environmental characteristics and driver behaviors (Islam et al. 2020) may affect the study result. To reduce computation complexities and ease implementation, the study excludes the temporal instability and unobserved heterogeneity associated with environmental characteristics and driver behaviors. The authors also propose a standard method to provide a recommended fragment size (RFS) for aggregating crash data that can be used as a foundation for all future traffic crash analyses requiring data aggregation, which may reduce the impact of arbitrary selection of fragment size (segment length) on crash frequency analysis (CFA).

Literature Review

Fragment Size (Segment Length)

As aforementioned, selecting the segment length to aggregate traffic and crash data impacts both CFA and RTCPM since it may affect the variables' statistical significance; therefore, the impact of segment length on safety analyses requires further investigation. Thomas (1996) studies the effect of segment length on crash count and density. Thomas (1996) argues that the arbitrary selection of segment length to aggregate data creates an unaddressed problem called a size problem. According to AASHTO Highway Safety Manual (AASHTO 2010), creating segments with consistent geometry and Annual Average Daily Traffic (AADT) may address this concern. However, it introduces new issues due to the inconsistent and small segment lengths and the need for universal data availability for all segments (Ghadi and Torok 2019). Segment length selection to aggregate crash data impacts the identification of crash hotspots (Cook et al. 2011) and affects the consistency of hotspot identification (Geyer et al. 2008). Also, the safety analysis outcomes can be affected for both extremely long and short roadway segments (Lu et al. 2013). Despite the importance of segmentation length, there is minimal guidance on segmentation.

Segmentation Approaches

Various approaches to segmentize a roadway using a subset of sources, including traffic data, roadway characteristics, and traffic crash data exist but a typical approach segments a roadway based on its characteristics to account for unobserved heterogeneity. However, roadway segmentation by roadway characteristics may lead to long segments since many roadways may have little to no variation in roadway attributes over a long stretch (Green 2018). For example, a very long segment length may occur because a long stretch of a highway has constant shoulder width, the number of lanes, cross slope, and median width on a straight section (Green 2018). While a homogeneous long segment can be divided to smaller segments to redistribute traffic crashes into resulting smaller segments, dividing the homogeneous long segments into small segments may lead to an arbitrary selection of break points or selection of a (segment) length with no specific guidelines (Green 2018). Besides, quality roadway characteristics data may not be available, requiring costly data collection. Other than roadway characteristics, traffic data can be used to develop a homogeneous segment when variation in roadway attributes is negligible (Borsos et al. 2014). Even though traffic data may help to divide long segments into smaller segments, it may not be helpful for roadways with limited access over a long distance due to minor changes in traffic volume (Green 2018).

Other alternatives to roadway segmentation by roadway attributes exist. These alternatives include continuous risk profile (Kwon et al. 2013), sliding moving window (Qin and Wellner 2012; Kwon et al. 2013), peak searching (Kwon et al. 2013), fixed length and variable length segmentation (Koorey 2009), clustering methods (Valent et al. 2002; Depaire et al. 2008; Lu et al. 2013). Among these alternatives, the clustering techniques are beneficial for roadway segmentation using traffic crash data, especially when quality data on traffic and roadway attributes are unavailable because they may reveal undiscovered relationships in traffic crash data (De Luca et al. 2012; Depaire et al. 2008; Golob et al. 2004; Lu et al. 2013). Valent et al. (2002) applied a clustering method using a specific crash type to analyze traffic crashes. The clustering method can mask the underlying contributing factors for the specific crash type (Valent et al. 2002). Depaire et al. (2008) utilized latent class clustering by using the heterogeneity of traffic crash data to segment a roadway. Lu et al. (2013) used Fisher's clustering to create a segmentation based on sections with similar crash distributions. The segmentation produced by Fisher's clustering improved the predictive model performance. Due to the lack of quality data on roadway attributes, this study performs the clustering method using the heterogeneity of crash data.

An essential aspect of traffic safety studies is unobserved heterogeneity. Studies can only include some information to capture data for all potentially contributing causes of traffic crashes (Chang et al. 2021; Mannering et al. 2016). A popular approach to address unobserved heterogeneity is to group the traffic crash data into homogeneous groups by different attributes (Mannering and Bhat 2014).

Some traffic crash attributes are crash units (number of vehicles involved in crashes), crash type (manner of collision), and level of crash severity. Generally, previous research classifies crashes based on crash units by grouping crashes into two major classic groups: single-vehicle (SV) and multivehicle (MV). Previous traffic crash studies based on total crashes have failed to identify some contributing factors and hotspots with a false positive tendency (Cheng et al. 2017). Regardless of applying an aggregate or disaggregate approach, a crash analysis should be performed based on the crash units (number of vehicles involved in crashes) (Yu et al. 2013). Another typical dimension in traffic safety studies is the manner of collision (crash type), which refers to the first event in a crash; other studies refer to it as crash type. Previous studies document the importance of including the manner of collision (crash type) in traffic crash analysis (Pande et al. 2010). The traffic crash type can be considered a dimension of group traffic crashes since it helps mask the underlying contributing factors associated with a manner of collision (Valent et al. 2002). It is also highlighted that the traffic crashes need to be separately investigated by manner of collision since the crash mechanism may potentially vary for different manner of collision (Bhowmik et al. 2018). The previous studies confirm that the contributing factors and their statistical significance are different for various manner of collisions (Mahmud and Gayah 2021). Crash severity represents another dimension to consider in capturing the heterogeneity of traffic crashes. According to Yang et al. (2009), crash severity is determined by the most seriously injured individuals in the crash, ranging from low-cost property damage to extremely costly severe injuries or fatalities. The analysis of all crashes together may conceal the injury level of crashes (Valent et al. 2002). For unobserved heterogeneity, this study considers crash severity alongside the crash unit (number of vehicles involved in the crash) and the crash type (manner of collision).

This study proposes a method to identify a RFS using an unsupervised clustering method on traffic crash data. The study addresses the heterogeneity of traffic crash data by grouping traffic crashes based on crash unit, crash type, and crash severity. A feature for each group of crashes is defined, and its corresponding crash rate is calculated, known as the featured crash rate (FCR). To discover the most critical features for clustering, the Laplacian score with distance-based entropy measure (LSDBEM) is used for K-means clustering feature selection identifies the features providing the most information to capture the similarities between segments. The LSDBEM-selected features significantly improve K-mean clustering results by forming homogeneous clusters (Liu et al. 2009). Additional dimensions, such as roadway geometry, can be included and investigated in future studies to address unobserved spatial heterogeneity. While roadway geometry attributes may represent a better approach to form homogeneous segments. In the absence of quality geometry attributes, the proposed K-means clustering using crash units, crash type, and crash severity provides another strategy for crash data aggregation.

Data Description

This study uses crash data from the network of urban freeways within Dallas County in Texas. The study area includes mainlane segments for both directions of Texas Loop 12, IH-20, IH-30, IH-35E, IH-45, IH-635, and US-75 (see Fig. 1). The data includes crash data, roadway geometric characteristics, and traffic characteristics for the 5-year period of 2015–2019. A statistics summary of crash units with the manner of collision and crash severity is provided in Tables 1 and 2, respectively.

Crash Data Features

The crash data from the Texas Department of Transportation (TxDOT) C.R.I.S. (Crash Record Information System) includes features from three groups: crash fields, unit fields, and person fields. The crash fields provide information about crashes. These include geospatial data such as latitude, longitude, reference marker, offset distance, highway system, roadway part, highway name, and the roadway geometry at the crash location. The crash fields also include crash characteristics like manner of collision and crash severity. This study only uses the information in the crash fields. Also, traffic count data for the study area is obtained from TxDOT for the 5-year period of 2015–2019.

Data Preparation

The crash data provides a separate entry for every individual involved in a traffic crash sharing the same crash ID as other individuals but with a different case number. The analysis aggregates the traffic crash entries for each day by crash ID and the total number of vehicles involved in the crashes to form a new crash data set. The new crash IDs include the crash date and time to avoid loss during when fusing five years of data together. To standardize crash location, the analysis calculates the milepost values from the crash location reference marker and offset values provided in the crash data. The analysis only uses crash data for the main segment of each roadway and excludes the crashes involving active work zones, construction areas, pedestrians, or wrong-way driving. The researchers geovalidated the crash data points by importing crash data points as KMZ files to Google Earth® to ensure the feature values for roadway segments, and vehicle travel directions are consistent with the location of crash data points.

The Instruction to Police for Reporting Crashes [TxDOT (Texas DOT) Traffic Safety Division 2020] categorizes crash severity levels as A—suspected serious injury, B—suspected minor injury, C—possible injury, K—fatal injury, N—not injured, and 99—unknown (see Table 3 for the definitions). The study area traffic crash data shows that crash severity at levels A, B, C, K, and N are 2.05%, 10.55%, 21.15%, 0.48%, and 64.57% of total crashes for 2015–2019 in Dallas County, respectively. Since fatal crash percentages remain very small, a separate fatal crash characteristic may not be necessary. Therefore, the analysis groups fatal and suspected serious injury crashes together since they are close in terms of severity level and represent a low portion of total crashes. Similarly, the analysis groups suspected minor and possible injury crashes together because they do not necessarily represent distinct crash severities and likely experience a significant overlap, which would make distinctive clustering more difficult. Noninjury remains a separate crash characteristic and the authors exclude crashes with unknown severity from the study.

Methodology

Introduction

K-means clustering is an unsupervised learning method to group unlabeled objects by similarities (Pedregosa et al. 2011). Previous traffic crash studies use this technique to cluster traffic data based on similarities. Using clustering approach, recent research captures congestion-sensitive spots (Bhatia et al. 2020), groups traffic flow data (Azizi and Hadi 2021; Xu et al. 2012, 2013), classifies the crash risk for urban expressways (Cheng et al. 2022) or other objectives. In this study, K-means clustering segmentizes urban freeway highways with features defined as crash rates calculated based

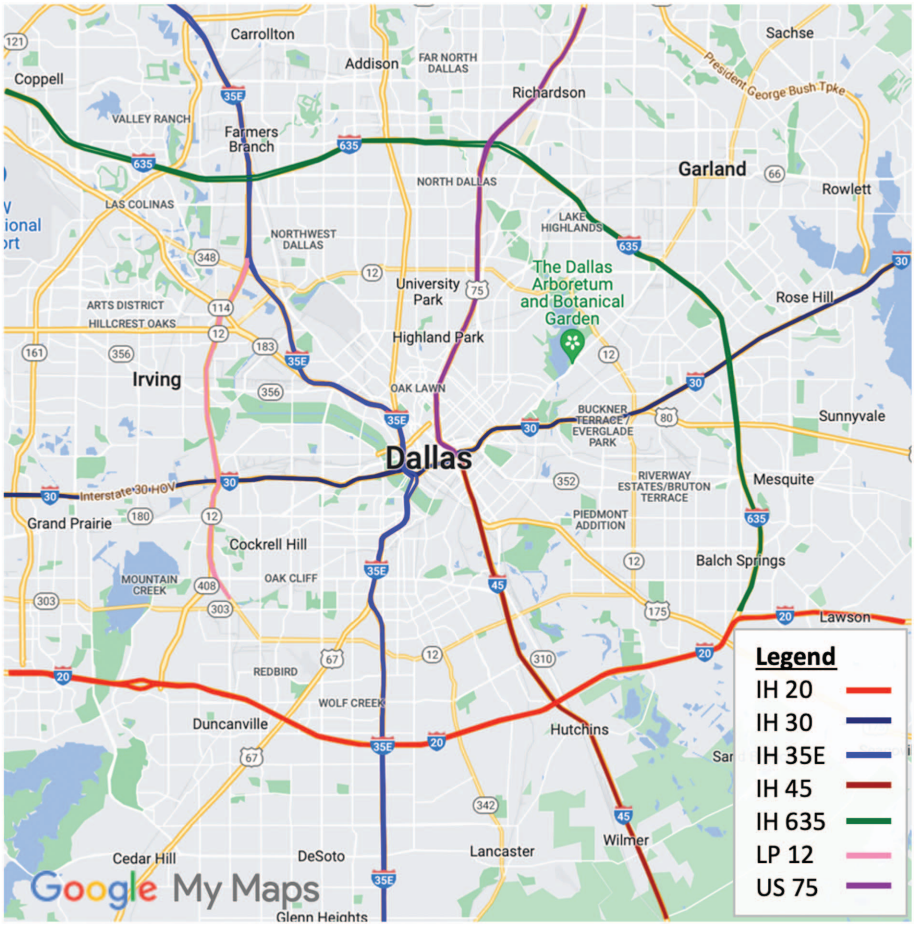


Fig. 1. (Color) Study area map. (Map data © 2024 Google.)

Table 1. Crash units and manner of collision summary (2015–2019)

	Single-vehicle (SV)			Multivehicle (MV)				
Highway	Object related (OBJ)	Overturned (OVT)	Other (OTH)	Angled (ANGL)	Rear-end (RRND)	Sideswipe (SDSW)	Stopped (STPD)	Total
IH-20 EB	464	60	41	2	839	838	244	2,488
IH-20 WB	475	31	43	5	854	754	215	2,377
IH-30 EB	585	22	19	3	804	936	406	2,775
IH-30 WB	552	27	23	2	979	842	451	2,876
IH-35E NB	1,011	71	52	10	2,109	1,853	1,121	6,227
IH-35E SB	825	50	34	9	1,673	1,750	945	5,286
IH-45 NB	166	10	6	2	150	156	47	537
IH-45 SB	231	10	7	4	174	174	27	627
IH-635 NB	846	45	7	8	1,819	1,604	473	4,802
IH-635 SB	802	62	4	5	1,924	1,442	562	4,801
LP-12 NB	218	16	6	2	357	340	131	1,070
LP-12 SB	235	16	3	4	236	313	62	869
US-75 NB	352	19	1	2	1,138	779	373	2,664
US-75 SB	370	13	3	1	1,321	791	492	2,991
Dallas County	7,132	452	249	59	14,377	12,572	5,549	40,390

on jointly considering the number of vehicles involved in the crash, manner of collision, and crash severity (crash combination).

Feature Selection

The K-means clustering results heavily depend on the features selected for grouping the objects into the clusters. The main goal is to compare and group highway segments by crash combination crash rates, which creates 21 features. Before applying the K-means clustering, the methodology implements feature reduction approaches to avoid redundancies and improve clustering results. This study deals with a multivariate problem in which feature values form a sparse matrix for each highway and freeway

Table 2. Crash severity summary

Highway and travel direction	Suspected serious injuries (A)	Suspected minor injuries (B)	Possible injuries (C)	Fatal (K)	Not injured (N)	Total
	Crash severity (2015–2019)					
IH-20 EB	54	313	479	17	1,625	2,488
IH-20 WB	65	288	455	12	1,557	2,377
IH-30 EB	37	191	463	18	2,066	2,775
IH-30 WB	45	264	513	7	2,047	2,876
IH-35E NB	80	509	1,116	25	4,497	6,227
IH-35E SB	88	409	899	23	3,867	5,286
IH-45 NB	23	54	121	4	335	537
IH-45 SB	14	79	144	8	382	627
IH-635 NB	135	651	1,203	21	2,792	4,802
IH-635 SB	124	659	1,108	23	2,887	4,801
LP-12 NB	21	117	332	3	597	1,070
LP-12 SB	21	112	234	5	497	869
US-75 NB	63	322	769	10	1,500	2,664
US-75 SB	58	337	837	9	1,750	2,991
Dallas County	828	4,305	8,673	185	26,399	40,390

direction of travel. The methodology requires an appropriate unsupervised feature selection method to address the multivariate nature of the problem, potential redundancy, and existing sparsity in the features. The recent review by Solorio-Fernández et al. (2020) categorizes feature selection candidates for this study under multivariate spectral/sparse learning methods. This study adopts the Laplacian score combined with distance-based entropy measure (LSDBEM) (Liu et al. 2009) because it finds the best subset of features capturing underlying clustering structures before performing clustering methods. Unlike the supervised and semisupervised feature selection approaches, the unsupervised feature selection methods have no privilege of relying on labeled data to alleviate irrelevant and redundant features. As an unsupervised feature selection, the LSDBEM employs evaluation metrics to eliminate redundant features (He et al. 2017). Several studies utilized the LSDBEM as unsupervised feature selection to capture the relevancy, eliminate the redundancy, and identify the most important features for unsupervised clustering, such as K-means clustering (Barile et al. 2022; Karim et al. 2020; Wang et al. 2022). Karim et al. (2020) extensively implemented the LSDBEM for feature selection. They compared it with two other unsupervised feature selections, principal component analysis (PCA) and multicluster-based feature selection (MCFS). The feature selection results show that 75% of the features selected by LSDBEM are in common with features selected by PCA and MCFS (Karim et al. 2020). Also, Karim et al. (2020) utilized various clustering methodologies, including balanced iterative reducing and clustering using hierarchies (BIRCH), hierarchical distance-based spatial clustering of applications with noise (DBSCAN), ordering points to identify cluster structure (OPTICS), K-modes, spectral, and K-means. They evaluated the clustering results using the Davies-Bouldin index, Calinski-Harabasz, and silhouette coefficient score. The K-means clustering results showed a significant purity with a very negligible difference (0.1%) compared to the outperforming clustering method OPTICS. As the method name implies, LSDBEM is a combination of the Laplacian score and an entropy measure that are separately explained in separate subsections. Prior to LSDBEM, all-zero and single nonzero features are discussed in the following subsection.

Dropping All-Zero Features and Features with Single Nonzero Value

A feature (crash group) that has a zero value (zero crash count) for all the objects (subsegments) has no impact on the clustering result. Therefore, a zero-value feature can be excluded from the set of selected features for clustering. The single nonzero feature (a crash group with nonzero crash count for only one subsegment) may be excluded because it will either not affect clustering or form a trivial single object cluster with a single object.

Feature Selection Using Laplacian Score (fsulaplacian)

He et al. (2005) introduce an unsupervised method to rank features based on a Laplacian score calculated using the nearest neighbor similarity graph as a feature selection method called Laplacian score. This method has a proven record of capturing significant features. A detailed Laplacian score algorithm may be found in a study by Pande and Abdel-Aty (2006). The algorithm favors features

Table 3. Traffic crash categories

Categories		Description
Number of vehicle Involved in crashes	Single-vehicle (SV)	Crashes that only involves one motor vehicle.
	Multi-vehicle (MV)	Crashes that involve two or more motor vehicles.
Manner of collision	Fixed object (OBJ)	Crashes that involve hiding fixed objects as the first harmful event.
	Over-turned (OVT)	Crashes that the first harmful event is identified as vehicle overturn.
	Angled (ANG)	Crashes that two motor vehicles are collided at an angle caused by at-least one vehicle deviating, turning left/right, or backing
	Rear-end (RRND)	Crashes that a motor vehicle is rear-ended by another motor vehicle.
	Sideswipe (SDSW)	Crashes that a motor vehicle is sideswiped by another motor vehicle.
	Stopped (STPD)	Crashes that a motor vehicle that is stopped on travel way is collided by a motor vehicle in motion.
	Other (OTH)	Crashes that the manner of collision is none of the items above.
Crash severity	A—suspected serious injury	Severe injury that prevents continuation of normal activities leading to temporarily or permanent incapacitation.
	B—suspected minor injury	Evident injury such as bruises, abrasions, or minor lacerations which do not incapacitate.
	C—possible injury	Injury claimed, reported, or indicated by behavior but without visible wounds, includes limping or complaint of pain
	K—fatal	If death resulted due to injuries sustained from the crash, at the scene or within 30 days of crash.
	N—not injured	The person involved in the crash did not sustain as A, B, C, or K injury.
	99—unknown	Unable to determine whether injuries exist. Some examples may include hit and run, fled scene, fail to stop or render aid.

with large variance because "the algorithm assumes that two data points of an important feature are close if and only if the similarity graph has an edge between the two data points" (Pande and Abdel-Aty 2006). A feature with a large score s_r represents an important feature. This can be used with the distance-based entropy measure to determine important features.

Distance-Based Entropy Measure

Liu et al. (2009) showed that the best subset for clustering can be identified by combining the Laplacian score method with the distance-based entropy measure. The process starts sorting features by their corresponding Laplacian score in ascending order, i.e., from the most important feature to the least important feature [note that the lowest the Laplacian score, the highest the importance of the feature (Liu et al. 2009)]. Then, the top two important features are selected as the current subset of features and the distance-based entropy measure is calculated. In the next step, the next subset is formed by adding the next important feature to the current subset and the corresponding distance-based entropy measure is calculated. This process is iterated until all features are in the current subset. Among all the subsets that are investigated in the process, the subset with the highest distance-based entropy measure is the best subset of the features for clustering purposes (Liu et al. 2009).

Feature Selection Steps

The feature selection procedure for this study is as follows:

1. The features are generated for all the possible combinations of traffic crash groups. Fig. 2 shows traffic crash groups, their abbreviations, and the generated features. The crash rates calculated for each of the generated features are called featured crash rates (FCRs). The naming convention of features is in the format of A-B-C in which A, B, and C are the traffic crash abbreviations for the number of vehicles involved in crashes, manner of collision, and crash severity. For instance, SV-OBJ-N is the feature for single-vehicle object-related crashes with no injuries. Also, MV-RRND-B+C is the feature for multivehicle rear-end crashes with suspected minor or possible injuries.
2. All the unknown severity, all-zero, and single nonzero features are dropped.
3. The function *fsulaplacian* is applied to the current set to find all the feature scores.
4. The distance-based entropy measure is applied to the features with their corresponding Laplacian scores. The subset with the highest distance-based entropy measure is selected as the best subset of features for clustering.

K-Means Clustering Algorithm

As aforementioned, K-means is an unsupervised learning technique that clusters unlabeled objects by similar features. The K-means algorithm starts with k centroids to group the objects in k clusters (a centroid for each cluster). After assigning all objects to their nearest cluster, the algorithm calculates a new set of centroids by finding the mean values of the objects in each cluster. This process iterates until the associated cost function, the sum of squared error (SSE) within each cluster (also known as cluster inertia), reaches its minimum value and determines the final clusters and their corresponding centroids (Bhatia et al. 2020). Raschka and Mirjalili (2017) provide the formal definition of a K-means clustering algorithm as follows:

Step 1: Randomly pick k centroids from the sample points as initial cluster centers.

Step 2: Assign each sample to the nearest centroid $\mu^j, j \in \{1, \ldots, k\}$.

Step 3: Move each centroid to the center of the samples that were assigned to it.

Step 4: Repeat steps 2 and 3 until the cluster assignments do not change or a user-defined tolerance or maximum number of iterations is reached.

In "Step 2", the term nearest implies the distance comparison requiring a measure. The distance refers to the differences between values of features for each sample (object) and values of features for the centroids. The shorter the distance to a centroid, the closer the sample (object) to a centroid. For "Step 4", the K-means function (***KMeans***) from Python libraries ***"sklearn.cluster"*** has input variables for a user-define tolerance and maximum number of iteration as stop conditions to terminate the iterative process and report the clustering results. The parameter of the K-means function (***KMeans***) are discussed in the result section under algorithm implementation.

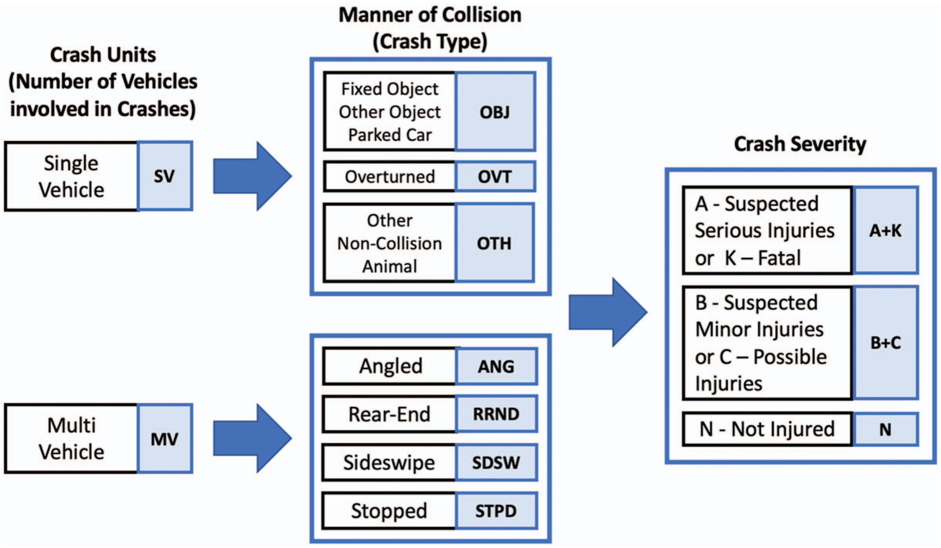

Manner of Collision (Crash Type)
Crash Units (Number of Vehicles involved in Crashes)
Single Vehicle
SV
Fixed Object Other Object Parked Car
OBJ
Overturned
OVT
Other Non-Collision Animal
OTH
Crash Severity
A - Suspected Serious Injuries or K – Fatal
A+K
B - Suspected Minor Injuries or C – Possible Injuries
B+C
N - Not Injured
N
Multi Vehicle
MV
Angled
ANG
Rear-End
RRND
Sideswipe
SDSW
Stopped
STPD

Fig. 2. (Color) Three dimensions of traffic crashes and the generated features.

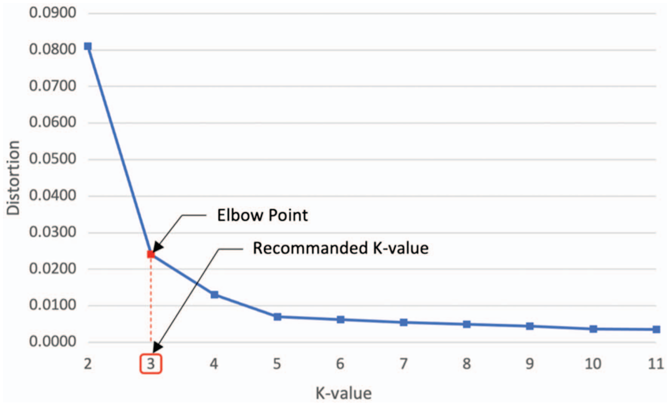


Fig. 3. (Color) Elbow curve and elbow point.

Elbow Curve and Silhouette Coefficient

As described in the previous section, the K-means clustering algorithm starts with randomly selected k centroids to group the objects in k clusters but selecting a preferred k value represents a challenge (Bhatia et al. 2020). Running K-means clustering for a range of k-values and monitoring the cost function value associated with each k-value can overcome this obstacle (Bhatia et al. 2020). The elbow method can assist in finding a preferred k based on the marginal improvement associated with adding another cluster (Bhatia et al. 2020). An elbow curve, which is a plot of the cost function against the number of clusters, visualizes this process. In an elbow curve, a point where the marginal gain drops such that it generates an angular point called an elbow point should occur. The number of clusters corresponding to the elbow point is the optimal number of clusters, k^* (Bhatia et al. 2020). Mathematically, the maximum absolute value of the second derivative of the elbow curve is the elbow point (Bhatia et al. 2020). Fig. 3 shows an elbow curve and its elbow point. Silhouette analysis evaluates the tightness of objects within the clusters and assesses the clustering quality using the silhouette coefficient (Anon 2011). In fact, the silhouette coefficient measures cluster cohesion and separation simultaneously. Cluster cohesion refers to how objects within a cluster are similar to each other. Cluster separation represents how cluster objects are different from the objects in other clusters. The greater the silhouette coefficient, the stronger the cohesion and the greater the separation. The silhouette coefficient ranges from -1 to 1, and it equals zero when the cluster cohesion and separation are the same; a value that approaches one indicates that separation greatly exceeds the within-cluster distance (Anon 2011).

Search Algorithm

This section describes the algorithm to search for the RFS. This algorithm utilizes the K-means clustering algorithm to cluster the highway segments as the objects with the FCR based on the dimensions described in the data preparation section. The study calculates the featured crash rates for highway segments with fragment size (segment length) ranging from 0.161 km (0.10 mi) to 0.402 km (0.25 mi) in the study. The search process starts with the initial value of 0.161 km (0.10 mi) to perform K-means clustering to cluster highway segments and continues through the remaining segment lengths using increments of 0.016 km (0.01 mi). The algorithm normalizes the FCRs by dividing each feature value by its corresponding maximum FCR. As described in the feature selection section, the method investigates all the features to select the final features for K-means clustering. The K-means clustering algorithm uses the final features to cluster highway segments. Applying the K-means clustering provides a path to group highway segments by comparing feature similarities of the segments at an aggregate level. After completing the clustering for all fragment sizes (segment lengths), the recommended clustering corresponds to the result with the greatest silhouette coefficient since it provides clusters with higher cohesion and better separation. Also, this will provide a sufficient range of segment length scenarios to investigate the effect of segment length on data aggregation and find the significant crash combinations.

Results

This section discusses the search algorithm results. The search algorithm is applied to crash data for mainlane segments in both directions of Texas Loop 12, IH-20, IH-30, IH-35E, IH-45, IH-635, and US-75 within Dallas County limits. By applying the search algorithm, the results provide the best set of features for clustering, preferred number of cluster k^*, and silhouette coefficient for each fragment sizes (segment lengths) ranging from 0.161 to 0.402 km (0.10 to 0.25 mi). This section also compares the FCR k-means clustering results with the findings from TCR k-means clustering results to evaluate the benefits of using FCR over TCR.

Algorithm Implementation

This study methodology develops a library of functions in Python 3 to perform the entire process, from data cleaning and preparation to feature selection and K-means clustering. The K-means clustering and elbow point detection use the ***KMeans*** and ***Kneelocator*** functions from Python libraries ***sklearn.cluster*** and ***kneed***, respectively. The ***KMeans*** function requires values for the attributes ***n_init=50*** and ***max_iter=1000***. ***n_init*** is the number of times that the k-means algorithm will be applied with different centroid seeds. The final k-means clustering result is the best output of ***n_init*** successive runs in terms of inertia. ***max_iter*** sets the maximum number of iterations that the k-means algorithm will be applied in a single run (Raschka and Mirjalili 2017; Solorio-Fernández et al. 2020). The ***KMeans*** function is applied with large enough values for the attributes ***n_init=50*** and ***max_iter=1000*** to minimize the impact of random centroids on the final result. For each run, the average computing time is 155 s and 58 s for FCR and TCR (6-Core Intel Core i7, 2.6 GHz CPU, 16 GB memory), respectively.

Clustering Results

The study forms traffic crash clusters by applying K-means to FCRs and TCRs data for each highway mainlane travel direction. As a sample, the clustering results for IH-20 EB (all 16 values) are shown in Table 4. Compared with TCR, the FCR-based clustering results consistently provide clusters with greater cohesion within the cluster and better separation between clusters based on their silhouette scores. For each highway travel direction, the recommended FCR-based cluster reaches silhouette scores between 0.7415 and 0.9699, which is significantly greater than the recommended TCR-based clustering results with silhouette scores between 0.6056 and 0.7255. To evaluate the significance of FCR over TCR, paired T-test is performed on $d = SC_{FCR} - SC_{TCR}$, in which SC_{FCR} and SC_{TCR} are the silhouette scores of FCR and TCR-based clustering across all highways. By calculating d for all highways, it is obtained that $\mu_d = 0.2177$ and $S_d = 0.0054$. The hypothesis test is defined as $H_0{:}\mu \le 0$ and $H_a{:}\mu > 0$. Considering the level of significance $\alpha = 0.01$ and $n = 14$, the value of t for the

Table 4. Clustering results comparison (IH-20 EB)

	Featured crash rate			Total crash rate	
Frag. size km (mi.)	Recom′d K-value	Silhouette coefficient	Set of features	Recom′d K-value	Silhouette coefficient
0.161 (0.10)	3	0.9699	[′SV-OBJ-A+K′, ′SV-OVT-A+K′]	4	0.6276
0.177 (0.11)	3	0.9647	[′SV-OBJ-A+K′, ′SV-OVT-A+K′]	4	0.6294
0.193 (0.12)	5	0.3910	[′SV-OBJ-A+K′, ′SV-OBJ-B+C′, ′SV-OTH-N′, ′SV-OVT-A+K′, ′SV-OVT-N′, ′MV-RRND-B+C′, ′MV-RRND-N′, ′MV-SDSW-B+C′, ′MV-STPD-B+C′, ′MV-STPD-N′]	4	0.5958
0.209 (0.13)	3	0.9573	[′SV-OBJ-A+K′, ′SV-OVT-A+K′]	4	0.6115
0.225 (0.14)	3	0.9540	[′SV-OBJ-A+K′, ′SV-OVT-A+K′]	4	0.6228
0.241 (0.15)	3	0.9041	[′SV-OBJ-A+K′, ′MV-RRND-A+K′]	4	0.6115
0.257 (0.16)	3	0.9451	[′SV-OBJ-A+K′, ′SV-OVT-A+K′]	3	0.6448
0.274 (0.17)	4	0.6872	[′SV-OVT-A+K′, ′MV-SDSW-B+C′]	3	0.6385
0.290 (0.18)	4	0.8219	[′SV-OBJ-A+K′, ′SV-OTH-N′, ′MV-STPD-A+K′]	5	0.5840
0.306 (0.19)	3	0.9346	[′SV-OBJ-A+K′, ′SV-OVT-A+K′]	4	0.5881
0.322 (0.20)	4	0.6661	[′SV-OBJ-B+C′, ′SV-OVT-A+K′]	3	0.6103
0.338 (0.21)	3	0.8605	[′SV-OBJ-A+K′, ′MV-RRND-A+K′]	3	0.6396
0.354 (0.22)	2	0.9123	[′SV-OVT-A+K′, ′MV-STPD-A+K′]	5	0.5614
0.370 (0.23)	3	0.9190	[′SV-OBJ-A+K′, ′SV-OVT-A+K′]	3	0.6530
0.386 (0.24)	5	0.4367	[′SV-OTH-N′, ′SV-OVT-N′, ′MV-SDSW-B+C′, ′MV-STPD-B+C′]	4	0.6632
0.402 (0.25)	3	0.6906	[′SV-OTH-N′, ′SV-OVT-N′, ′MV-STPD-A+K′]	4	0.6564

right-tailed test is $t(13, 0.01) = 2.6503$. the value of critical t, t_c is $(\mu_d - \mu)/(S_d/\sqrt{n})$. Then $t_c = 151.16$. Thus, $t_c = 151.16 \gg 2.6503$. It yields to reject H_0 and accept H_a, i.e. $SC_{FCR} - SC_{TCR,} > 0$. Therefore, $SC_{FCR} > SC_{TCR}$, with significance level of $\alpha = 0.01$ and $C.I. = (0.2139, 0.2251)$. This shows that FCR-based clusters outperformed TCR-based clusters. For each highway travel direction, the recommended FCR-based cluster reaches silhouette scores between 0.7415 and 0.9699, which is significantly (p-value < 0.0000) greater than the recommended TCR-based clustering results with silhouette scores between 0.6056 and 0.7255.

Feature Selection

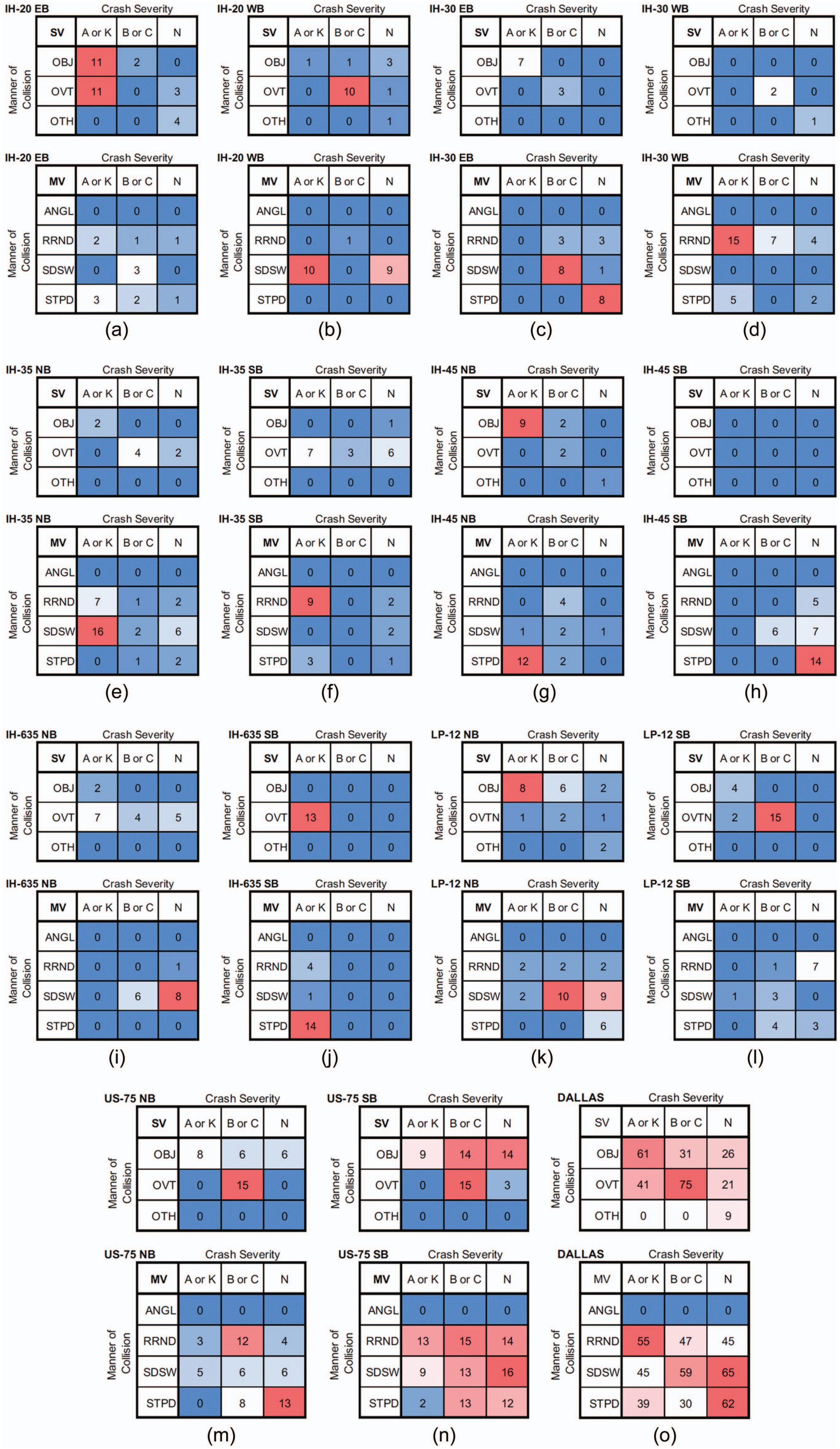
The FCR-based clustering results provide the sets of significant features associated with the clustering. Also, Fig. 4 shows heatmap representations of feature significance for the urban highway travel directions for the sixteen segment length values ranging from 0.161 to 0.402 km (0.10 to 0.25 mi). Due to the sixteen values, the frequency of features appearing significant varies between 0 and 16. The results demonstrate that the significant features differ depending on the urban highway and travel direction; however, some features appear frequently in most trials generated by different segment lengths. For IH-20 EB, the methodology selects ‘SV-OBJ-A+K’ and ‘SV-OVT-A+K’ as the significant features for more trials (segment length), including the RFS, than other features. For IH-20 EB, severe single-vehicle crashes with a clear crash class create the best crash data clusters (see Fig. 2 for abbreviations). The feature significance appears relatively insensitive to the segment length selected to aggregate the crash data. In most cases, the most frequently significant features (during the sixteen trials) for each highway appear in the cluster with the highest silhouette score. However, a few less frequently selected features also appear in the clusters with the highest silhouette scores, such as features SV-OBJ-A+K and SV-OVT-B+C for IH-30 EB and feature SV-OVT-N for IH-35E NB. Other less frequently significant features include SV-OBJ-N, SV-OVT-N, and SV-OTH-N, which makes sense because these crashes may be uniformly distributed along a highway since no injuries occur and they only involve a single vehicle. For most freeways, one to three features frequently appear for clustering with the first and second-ranked highest silhouette scores; however, US-75 SB has ten frequently appearing features. Fig. 4(o) shows the Dallas County heatmap that summarizes the total frequency of the significant features for the studied highways. The potential range of values in this figure is [0, 224]. Based on Fig. 4(o), the most frequently significant features are SV-OVT-B+C, MV-SDSW-N, MV-STPD-N, SV-OBJ-A+K, MV-SDSW-B+C, and MV-RRND-A+K, in descending order.

Fragment Sizes (Segment Lengths)

The results show that the fragment sizes (segment lengths) impacts the clustering results and their corresponding silhouette scores. Table 5 provides a comparison between the top two RFS values for FCR and TCR. The FCR clustering tends to recommend much shorter segment lengths than the TCR because they also capture trends in specific crash combinations more effectively than the TCR. For almost all the highways, the FCR clustering methodology selects two features, which generate clusters with silhouette scores at least 0.1 larger than the best corresponding TCR result. The additional information provided by the FCR strengthens the clustering and segregates the freeway into segments with different crash risks for the selected features.

Z-Score Analysis of FCR-Based Clusters

The features’ Z-scores for the clusters with highest silhouette score is provided in Table 6. For each highway travel direction, the features F_1, F_2, and F_3 correspond to the set of features in Table 5 for clustering with the highest silhouette scores. In most two-cluster and two-feature cases, the clustering results for $k^* = 2$ (two clusters) show that one feature appears with a large positive Z-score in one cluster while the other feature shows a small value (somewhat close to zero) and the feature values reverse in the other cluster. For instance, the clustering result for IH-20 WB shows that single-vehicle overturned crashes with minor or possible injuries has a Z-score of 4.32 for cluster #2, meaning, cluster #2 represents single vehicle overturned crashes with minor or possible injuries but not multivehicle sideswipe fatal and serious crashes; cluster #1 represents risky locations for multivehicle sideswipe fatal and serious crashes but not single-vehicle object crashes with minor or possible injuries. The large Z-score also indicates the intensity of the risk for cluster #2 is much higher than cluster #1. The same pattern



IH-20 EB Crash Severity (Manner of Collision)

SV	A or K	B or C	N
OBJ	11	2	0
OVT	11	0	3
OTH	0	0	4

IH-20 EB Crash Severity (Manner of Collision)

MV	A or K	B or C	N
ANGL	0	0	0
RRND	2	1	1
SDSW	0	3	0
STPD	3	2	1

(a)

IH-20 WB Crash Severity (Manner of Collision)

SV	A or K	B or C	N
OBJ	1	1	3
OVT	0	10	1
OTH	0	0	1

IH-20 WB Crash Severity (Manner of Collision)

MV	A or K	B or C	N
ANGL	0	0	0
RRND	0	1	0
SDSW	10	0	9
STPD	0	0	0

(b)

IH-30 EB Crash Severity (Manner of Collision)

SV	A or K	B or C	N
OBJ	7	0	0
OVT	0	3	0
OTH	0	0	0

IH-30 EB Crash Severity (Manner of Collision)

MV	A or K	B or C	N
ANGL	0	0	0
RRND	0	3	3
SDSW	0	8	1
STPD	0	0	8

(c)

IH-30 WB Crash Severity (Manner of Collision)

SV	A or K	B or C	N
OBJ	0	0	0
OVT	0	2	0
OTH	0	0	1

IH-30 WB Crash Severity (Manner of Collision)

MV	A or K	B or C	N
ANGL	0	0	0
RRND	15	7	4
SDSW	0	0	0
STPD	5	0	2

(d)

IH-35 NB Crash Severity (Manner of Collision)

SV	A or K	B or C	N
OBJ	2	0	0
OVT	0	4	2
OTH	0	0	0

IH-35 NB Crash Severity (Manner of Collision)

MV	A or K	B or C	N
ANGL	0	0	0
RRND	7	1	2
SDSW	16	2	6
STPD	0	1	2

(e)

IH-35 SB Crash Severity (Manner of Collision)

SV	A or K	B or C	N
OBJ	0	0	1
OVT	7	3	6
OTH	0	0	0

IH-35 SB Crash Severity (Manner of Collision)

MV	A or K	B or C	N
ANGL	0	0	0
RRND	9	0	2
SDSW	0	0	2
STPD	3	0	1

(f)

IH-45 NB Crash Severity (Manner of Collision)

SV	A or K	B or C	N
OBJ	9	2	0
OVT	0	2	0
OTH	0	0	1

IH-45 NB Crash Severity (Manner of Collision)

MV	A or K	B or C	N
ANGL	0	0	0
RRND	0	4	0
SDSW	1	2	1
STPD	12	2	0

(g)

IH-45 SB Crash Severity (Manner of Collision)

SV	A or K	B or C	N
OBJ	0	0	0
OVT	0	0	0
OTH	0	0	0

IH-45 SB Crash Severity (Manner of Collision)

MV	A or K	B or C	N
ANGL	0	0	0
RRND	0	0	5
SDSW	0	6	7
STPD	0	0	14

(h)

IH-635 NB Crash Severity (Manner of Collision)

SV	A or K	B or C	N
OBJ	2	0	0
OVT	7	4	5
OTH	0	0	0

IH-635 NB Crash Severity (Manner of Collision)

MV	A or K	B or C	N
ANGL	0	0	0
RRND	0	0	1
SDSW	0	6	8
STPD	0	0	0

(i)

IH-635 SB Crash Severity (Manner of Collision)

SV	A or K	B or C	N
OBJ	0	0	0
OVT	13	0	0
OTH	0	0	0

IH-635 SB Crash Severity (Manner of Collision)

MV	A or K	B or C	N
ANGL	0	0	0
RRND	4	0	0
SDSW	1	0	0
STPD	14	0	0

(j)

LP-12 NB Crash Severity (Manner of Collision)

SV	A or K	B or C	N
OBJ	8	6	2
OVTN	1	2	1
OTH	0	0	2

LP-12 NB Crash Severity (Manner of Collision)

MV	A or K	B or C	N
ANGL	0	0	0
RRND	2	2	2
SDSW	2	10	9
STPD	0	0	6

(k)

LP-12 SB Crash Severity (Manner of Collision)

SV	A or K	B or C	N
OBJ	4	0	0
OVTN	2	15	0
OTH	0	0	0

LP-12 SB Crash Severity (Manner of Collision)

MV	A or K	B or C	N
ANGL	0	0	0
RRND	0	1	7
SDSW	1	3	0
STPD	0	4	3

(l)

US-75 NB Crash Severity (Manner of Collision)

SV	A or K	B or C	N
OBJ	8	6	6
OVT	0	15	0
OTH	0	0	0

US-75 NB Crash Severity (Manner of Collision)

MV	A or K	B or C	N
ANGL	0	0	0
RRND	3	12	4
SDSW	5	6	6
STPD	0	8	13

(m)

US-75 SB Crash Severity (Manner of Collision)

SV	A or K	B or C	N
OBJ	9	14	14
OVT	0	15	3
OTH	0	0	0

US-75 SB Crash Severity (Manner of Collision)

MV	A or K	B or C	N
ANGL	0	0	0
RRND	13	15	14
SDSW	9	13	16
STPD	2	13	12

(n)

DALLAS Crash Severity (Manner of Collision)

SV	A or K	B or C	N
OBJ	61	31	26
OVT	41	75	21
OTH	0	0	9

DALLAS Crash Severity (Manner of Collision)

MV	A or K	B or C	N
ANGL	0	0	0
RRND	55	47	45
SDSW	45	59	65
STPD	39	30	62

(o)

Fig. 4. (Color) Heatmap of significant features for the highways and Dallas County.

Table 5. RFS values (FCR vs TCR)

	Featured crash rate					Total crash rate			
Roadway	RLS rank	Frag. size km (mi.)	K-value	Silh. score	Set of features	RLS rank	Frag. size km (mi.)	K-value	Silh. score
IH 20 EB	1st	0.161 (0.10)	3	0.9699	[′SV-OBJ-A+K′, ′SV-OVT-A+K′]	1st	0.386 (0.24)	4	0.6632
IH 20 EB	2nd	0.177 (0.11)	3	0.9647	[′SV-OBJ-A+K′, ′SV-OVT-A+K′]	2nd	0.402 (0.25)	4	0.6564
IH 20 WB	1st	0.161 (0.10)	2	0.9223	[′SV-OVT-B+C′, ′MV-SDSW-A+K′]	1st	0.322 (0.20)	3	0.6575
IH 20 WB	2nd	0.177 (0.11)	2	0.9153	[′SV-OVT-B+C′, ′MV-SDSW-A+K′]	2nd	0.209 (0.13)	3	0.6413
IH 30 EB	1st	0.225 (0.14)	3	0.8880	[′SV-OBJ-A+K′, ′SV-OVT-B+C′]	1st	0.386 (0.24)	3	0.6704
IH 30 EB	2nd	0.241 (0.15)	3	0.8716	[′SV-OBJ-A+K′, ′SV-OVT-B+C′]	2nd	0.402 (0.25)	4	0.6470
IH 30 WB	1st	0.161 (0.10)	2	0.9128	[′MV-RRND-A+K′, ′MV-STPD-A+K′]	1st	0.402 (0.25)	3	0.6852
IH 30 WB	2nd	0.193 (0.12)	2	0.8994	[′MV-RRND-A+K′, ′MV-STPD-A+K′]	2nd	0.257 (0.16)	4	0.6680
IH 35 NB	1st	0.161 (0.10)	2	0.8726	[′SV-OVT-N′, ′MV-SDSW-A+K′]	1st	0.386 (0.24)	3	0.6478
IH 35 NB	2nd	0.177 (0.11)	2	0.8601	[′MV-RRND-A+K′, ′MV-SDSW-A+K′]	2nd	0.193 (0.12)	3	0.6435
IH 35 SB	1st	0.177 (0.11)	2	0.9366	[′SV-OVT-A+K′, ′MV-RRND-A+K′]	1st	0.209 (0.13)	2	0.7255
IH 35 SB	2nd	0.193 (0.12)	2	0.9363	[′SV-OVT-A+K′, ′SV-OVT-N′]	2nd	0.225 (0.14)	3	0.6754
IH 45 NB	1st	0.209 (0.13)	2	0.9240	[′SV-OBJ-A+K′, ′MV-STPD-A+K′]	1st	0.257 (0.16)	3	0.6532
IH 45 NB	2nd	0.193 (0.12)	2	0.9216	[′SV-OBJ-A+K′, ′MV-STPD-A+K′]	2nd	0.225 (0.14)	3	0.6473
IH 45 SB	1st	0.257 (0.16)	4	0.8114	[′MV-SDSW-B+C′, ′MV-STPD-N′]	1st	0.161 (0.10)	2	0.6817
IH 45 SB	2nd	0.338 (0.21)	3	0.7530	[′MV-SDSW-B+C′, ′MV-STPD-N′]	2nd	0.290 (0.18)	2	0.6800
IH 635 NB	1st	0.193 (0.12)	3	0.9042	[′SV-OVT-A+K′, ′SV-OVT-B+C′, ′SV-OVT-N′]	1st	0.274 (0.17)	2	0.6886
IH 635 NB	2nd	0.290 (0.18)	2	0.8915	[′SV-OVT-A+K′, ′SV-OVT-N′]	2nd	0.193 (0.12)	4	0.6579
IH 635 SB	1st	0.177 (0.11)	2	0.9358	[′SV-OVT-A+K′, ′MV-STPD-A+K′]	1st	0.177 (0.11)	3	0.6406
IH 635 SB	2nd	0.193 (0.12)	2	0.9341	[′SV-OVT-A+K′, ′MV-STPD-A+K′]	2nd	0.274 (0.17)	3	0.6215
LP 12 NB	1st	0.177 (0.11)	4	0.8260	[′SV-OBJ-A+K′, ′MV-STPD-N′]	1st	0.306 (0.19)	3	0.6514
LP 12 NB	2nd	0.241 (0.15)	4	0.7981	[′SV-OBJ-A+K′, ′MV-STPD-N′]	2nd	0.225 (0.14)	3	0.6374
LP 12 SB	1st	0.177 (0.11)	2	0.9167	[′SV-OVT-A+K′, ′SV-OVT-B+C′]	1st	0.322 (0.20)	4	0.6056
LP 12 SB	2nd	0.290 (0.18)	2	0.8664	[′SV-OVT-A+K′, ′SV-OVT-B+C′]	2nd	0.274 (0.17)	4	0.6035
US 75 NB	1st	0.354 (0.22)	3	0.7627	[′SV-OVT-B+C′, ′MV-RRND-B+C′]	1st	0.370 (0.23)	4	0.6147
US 75 NB	2nd	0.402 (0.25)	3	0.7512	[′SV-OVT-B+C′, ′MV-RRND-B+C′]	2nd	0.338 (0.21)	4	0.5945
US 75 SB	1st	0.354 (0.22)	5	0.7415	[′SV-OVT-B+C′, ′MV-SDSW-N′]	1st	0.177 (0.11)	4	0.6905
US 75 SB	2nd	0.161 (0.10)	3	0.6037	[′SV-OBJ-A+K′, ′SV-OBJ-B+C′, ′SV-OBJ-N′, ′SV-OVT-B+C′, ′MV-RRND-A+K′, ′MV-RRND-B+C′, ′MV-RRND-N′, ′MV-SDSW-A+K′, ′MV-SDSW-B+C′, ′MV-SDSW-N′, ′MV-STPD-B+C′, ′MV-STPD-N′]	2nd	0.161 (0.10)	4	0.6107

for cluster #1 and #2 applies to other highway travel directions with $k^* = 2$ (two clusters) IH-35 NB, IH-35 SB, IH-45 NB, IH-635 SB, and LP-12 SB for their corresponding features. For IH-35 NB, cluster #2 identifies high-risk multivehicle sideswipe fatal and serious crash locations. For IH-35 SB, cluster #2 identifies high-risk multivehicle rear-end fatal and serious injury crash locations. For IH-45 NB and IH 635 SB, cluster #2 identifies high-risk multivehicle stopped fatal and serious injury crash locations. For LP-12 SB, cluster #2 identifies high-risk single-vehicle overturned minor and possible injury crash locations. Another two-cluster case, IH-30 WB, follows a different pattern where cluster #1 represents a low crash risk for both features and cluster #2 represents a high crash risk for fatal and serious multivehicle rear-end and stopped crashes. For $k^* = 3$ (three clusters), one cluster indicates a high-risk location for one crash type and another cluster indicates a high-risk location for the other selected crash type; the third cluster indicates low-risk crash locations for both selected crash features. IH-20 EB identifies high-risk locations for single-vehicle object and overturn crashes with fatal and serious injury, IH-30 EB identifies high-risk locations for single-vehicle object fatal and serious injury crashes and single-vehicle overturn crashes with minor and possible injury, and US-75 NB identifies high-risk locations for single-vehicle overturned minor and possible injury crashes and multivehicle rear-end minor and possible injury crashes. Another three-cluster case, IH-635 NB, adds a third feature to the clustering results; this case creates a low-risk crash cluster for single-vehicle overturn crashes. The other clusters separate high-risk single-vehicle overturned fatal and serious crash locations from high-risk single-vehicle overturned minor and possible injury crash locations. Only two freeway corridors (IH-45 SB and LP-12 NB) showed $k^* = 4$ (four clusters). For the IH-45 SB case, one cluster identifies low-risk locations for multivehicle sideswipe crashes with minor and possible injuries and multivehicle stopped crashes with property damage only. Another cluster identifies locations with high-risk for multivehicle sideswipe minor and possible injury crashes and low-risk for multivehicle stopped crashes with property damage only. The final two clusters contain moderate risk for multivehicle sideswipe minor and possible injury crash locations and high and moderate risk for multivehicle stopped crashes with property damage only. The LP-12 NB case identifies clusters with low risk for both features (single-vehicle object fatal and serious injuries and multivehicle stopped property damage), high risk for both features, and high risk for one feature/low risk for the other feature. Finally, US-75 SB demonstrated $k^* = 5$ (five clusters), as with all clusters with $k^* > 2$, one cluster represents low crash risk locations for the selected features. Similar to other cluster amounts, one cluster characterizes locations with high risk for single-vehicle overturned minor and probable injury crashes and low risk for multivehicle sideswipe property damage only crashes. Two other clusters identify locations with high and moderate risk for multivehicle sideswipe property damage only crashes and low risk for single-vehicle overturned minor and probable injury crashes. The final

Table 6. Z-score values of selected features used in LSDBEM/K-means clustering for FCR

Highway travel direction	Cluster ID	Feature mean per cluster			Feature total mean			Feature variance			Feature Z-score		
		F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3
IH-20 EB	1	0.00	0.00	—	0.07	0.02	—	0.04	0.02	—	−0.35	−0.17	—
	2	0.63	0.00	—	0.07	0.02	—	0.04	0.02	—	2.69	−0.17	—
	3	0.00	0.78	—	0.07	0.02	—	0.04	0.02	—	−0.35	5.74	—
IH-20 WB	1	0.00	0.03	—	0.03	0.03	—	0.02	0.02	—	−0.23	0.01	—
	2	0.68	0.00	—	0.03	0.03	—	0.02	0.02	—	4.32	−0.21	—
IH-30 EB	1	0.00	0.00	—	0.11	0.07	—	0.06	0.04	—	−0.45	−0.32	—
	2	0.60	0.00	—	0.11	0.07	—	0.06	0.04	—	2.05	−0.32	—
	3	0.10	0.66	—	0.11	0.07	—	0.06	0.04	—	−0.01	2.89	—
IH-30 WB	1	0.00	0.04	—	0.04	0.04	—	0.03	0.03	—	−0.22	−0.02	—
	2	0.84	0.09	—	0.04	0.04	—	0.03	0.03	—	4.31	0.31	—
IH-35E NB	1	0.04	0.00	—	0.04	0.04	—	0.02	0.03	—	0.00	−0.27	—
	2	0.03	0.60	—	0.04	0.04	—	0.02	0.03	—	−0.04	3.35	—
IH-35E SB	1	0.02	0.00	—	0.02	0.05	—	0.02	0.04	—	0.01	−0.26	—
	2	0.00	0.77	—	0.02	0.05	—	0.02	0.04	—	−0.16	3.71	—
IH-45 NB	1	0.04	0.00	—	0.04	0.05	—	0.03	0.04	—	0.01	−0.24	—
	2	0.00	0.91	—	0.04	0.05	—	0.03	0.04	—	−0.23	4.14	—
IH-45 SB	1	0.01	0.00	—	0.14	0.12	—	0.05	0.08	—	−0.61	−0.42	—
	2	0.21	0.87	—	0.14	0.12	—	0.05	0.08	—	0.32	2.71	—
	3	0.47	0.00	—	0.14	0.12	—	0.05	0.08	—	1.51	−0.42	—
	4	0.20	0.44	—	0.14	0.12	—	0.05	0.08	—	0.29	1.16	—
IH-635 NB	1	0.00	0.00	0.03	0.03	0.08	0.03	0.03	0.06	0.03	−0.20	−0.32	−0.04
	2	0.00	0.81	0.08	0.03	0.08	0.03	0.03	0.06	0.03	−0.20	3.08	0.26
	3	0.83	0.16	0.07	0.03	0.08	0.03	0.03	0.06	0.03	4.89	0.35	0.22
IH-635 SB	1	0.02	0.00	—	0.02	0.04	—	0.02	0.02	—	0.01	−0.26	—
	2	0.00	0.60	—	0.02	0.04	—	0.02	0.02	—	−0.15	3.69	—
LP-12 NB	1	0.00	0.03	—	0.07	0.15	—	0.04	0.08	—	−0.32	−0.42	—
	2	0.72	0.65	—	0.07	0.15	—	0.04	0.08	—	3.22	1.80	—
	3	0.00	0.73	—	0.07	0.15	—	0.04	0.08	—	−0.32	2.11	—
	4	0.63	0.10	—	0.07	0.15	—	0.04	0.08	—	2.79	−0.18	—
LP-12 SB	1	0.04	0.00	—	0.04	0.07	—	0.04	0.05	—	0.02	−0.28	—
	2	0.00	0.87	—	0.04	0.07	—	0.04	0.05	—	−0.21	3.45	—
US-75 NB	1	0.00	0.94	—	0.14	0.19	—	0.07	0.03	—	−0.52	4.05	—
	2	0.00	0.15	—	0.14	0.19	—	0.07	0.03	—	−0.52	−0.23	—
	3	0.62	0.22	—	0.14	0.19	—	0.07	0.03	—	1.79	0.13	—
US-75 SB	1	0.00	0.07	—	0.08	0.10	—	0.05	0.02	—	−0.37	−0.25	—
	2	0.94	0.08	—	0.08	0.10	—	0.05	0.02	—	3.97	−0.18	—
	3	0.46	0.13	—	0.08	0.10	—	0.05	0.02	—	1.75	0.14	—
	4	0.00	1.00	—	0.08	0.10	—	0.05	0.02	—	−0.37	6.05	—
	5	0.00	0.35	—	0.08	0.10	—	0.05	0.02	—	−0.37	1.63	—

cluster includes locations with moderate risk for single-vehicle overturned minor and possible injury crashes and slightly above average risk for multivehicle sideswipe property damage only crashes. Overall, the clustering represents an effective strategy for identifying data patterns for the selected crash features, which can directly identify high and low risk locations for these crash combinations.

Silhouette Scores and Fragment Sizes

A stairs-type stacked plot of silhouette scores for FCR and TCR clusters versus various fragment sizes for all highway travel directions is shown in Fig. 5. The silhouette scores for the FCR and TCR clustering results for the selected features are illustrated in blue and orange color, respectively. Overall, the silhouette scores of the TCR-based clustering results show greater stability across the various fragment sizes than the silhouette scores of the FCR-based clustering results. While the TCR-based clustering is more resistant to changes in the fragment sizes used for data aggregation, its silhouette scores remain under 0.80 while FCR-based clustering shows silhouette scores greater than 0.80 for some fragment sizes. However, the TCR-based clustering result supersedes the FCR-based clustering for US-75 SB for all fragment sizes but 0.370 km (0.23 mi) where FCR-based clustering result reaches the highest silhouette score. For IH-635 SB, the FCR-based clustering show highest silhouette scores for all fragment sizes comparing to TCR-based. These trends can be related to the traffic crash data distribution along US-75 SB and IH-635 SB.

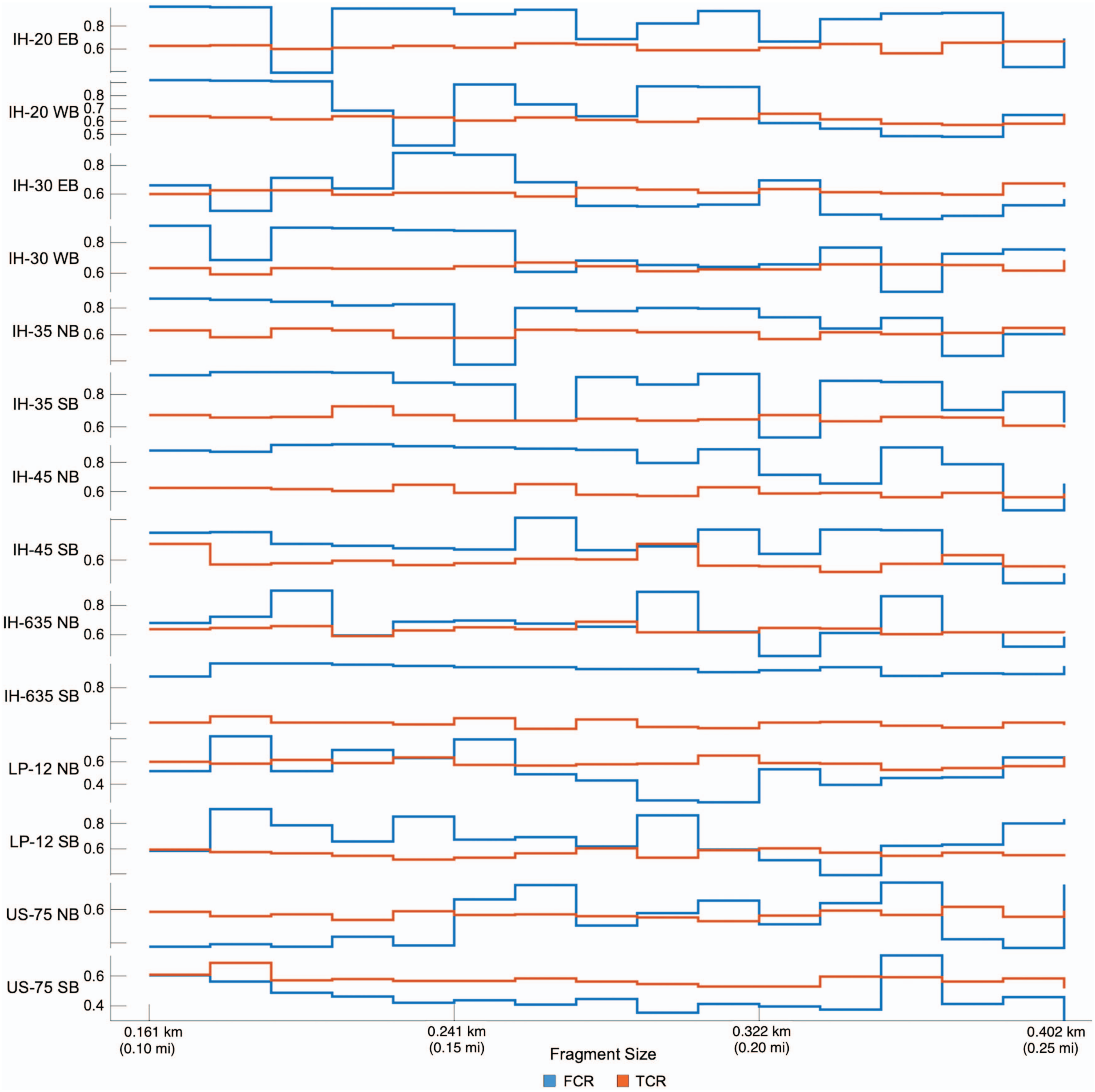


Fig. 5. (Color) Stairs-type stacked plot of silhouette scores for FCR and TCR clusters versus fragment size.

Conclusions and Recommendations

This paper develops a recommended fragment size (segment length) using three dimensions of traffic crashes (i.e., number of vehicles involved in the crash, manner of collision, and crash severity) and clustering methods as an innovative data-driven method to aggregate crash data. This strategy provides a standard approach for future studies to aggregate crashes and resolves the previously identified concern associated with the arbitrary selection of segment length in previous research. The proposed method harnesses the advantages of LSDBEM and K-means clustering algorithm as unsupervised learning applied to highway segments as the objects. The study defines featured crash rates (FCRs) using three dimensions of traffic crash characteristics: number of vehicles involved in the crash, manner of collision, and crash severity. The FCR-based clustering results show that RFS varies for each highway travel direction. The typical segment length of 0.161 km (0.10 mi) that has been used in several studies matches the RFS only for IH-20 EB, IH-20 WB, IH-30 WB, IH-35E NB, and US-75 SB that, which is less than forty percent of highway travel directions. The RFS based on FCR clustering varies between 0.161 and 0.354 km (0.10 and 0.22 mi), while the RFS based on TCR clustering cover the entire range from 0.161 and 0.402 km (0.10 to 0.25 mi). The variation in RFS across the different highways and travel directions indicates that a single best segment length does not exist, and the segment length should be selected based on observed crash data. However, the RFS based on FCR clustering and TCR clustering is the same for US 75 SB [0.161 km (0.10 mi)] and IH 635 SB

[0.177 km (0.11 mi)] (see Table 5). The FCR-based clustering results not only provide a RFS using three dimensions of traffic crashes characteristics but also identify the significant features for each highway travel direction which is impractical using TCR-based clustering. This paper proposed a data-driven methodology that overcomes the arbitrary selection of segment length using three dimensions of traffic crash characteristics.

The significant improvement in silhouette score between the FCR and TCR clustering methods indicates more cohesive and distinct clusters. This improvement will make the aggregated crash data more valuable and guarantee that the within-cluster segments experience similar crash risk for the selected features. The highest FCR-based silhouette scores range between 0.7415 and 0.9699. The methodology typically chooses two features for the best silhouette scores. However, the methodology evaluated several sets of features before selecting the set of features to represent the data clusters best. While the selected features vary significantly between freeways and travel directions, the features used to select the clusters associated with the RFS typically reflect the most commonly selected features for a particular freeway and travel direction. This study provides a foundation for highway segmentation that benefits future traffic and crash studies and RTCPMs using aggregated data.

Because this study establishes a standardized method for selecting a segment length to aggregate crash data for future safety analyses and RTCPMs, many opportunities for future research exist. The total assessment of this method's impact requires investigating the improvement in crash modeling that results. In addition, this method may eliminate the need for disaggregating locationally specific static crash modification factors for RTCPMs if the clustering can effectively capture aggregate static crash contributing factors. Future research should also examine the RFS's temporal stability and cluster structure's temporal stability. An extension of this study is to consider the temporal instability and unobserved heterogeneity associated with the environmental characteristics and driver behaviors by introducing featured crash rates (FCRs) for each year, including the environmental characteristics, and applying the LSDBEM/K-means. The study only investigates the clustering and recommended fragment length using all three traffic crash characteristics combined. The LSDBEM/K-means clustering can be applied to crash groups for scenarios including crash units only, and crash units and manner of collision combined to compare with FCR and TCR clustering results. The future study should investigate the value or importance of including additional crash characteristics in predicting crash risk and identifying contributing factors. Future studies need to extend this study by investigating each traffic crash characteristic separately and comparing the results with all three traffic crash characteristics considered. This study considered each highway and travel direction separately and created distinctive clusters for each. The future research can also consider the network wide clustering for a comparison. Future studies should apply this method on other freeway networks and explore applying it (or a variation) for two-lane highways and arterials. While this study includes three crash dimensions in its features, future studies may consider fewer (e.g., number of vehicles and manner of collision) and more crash dimensions (e.g., roadway geometry or AADT). The clustering may also involve other noncrash features and incorporate spatial correlation. A future study may expand the proposed RFS method to segmentize highways with a variable segment length rather than a constant length of the segment. The fragment size (segment length) selected for data aggregation may impact the statistical significance of explanatory variables in crash prediction models; a future study investigates these impacts and investigates the potential advantages of the recommended fragment size (RFS) for crash prediction models.

Data Availability Statement

Some or all data, models, or code that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments

This work was depended on the traffic count data which was provided by TxDOT Traffic Analysis and System Support. The authors would like to thank Laura Dabalain and Eric Oeding to support this research by providing the traffic count data. The authors would like to acknowledge partial support from the National Institute for Transportation and Communities (NITC; Grant No. #1578).

Author contributions: The authors confirm contribution to the paper as follows: conceptualization: Farzin Maniei and Stephen P. Mattingly; formal analysis: Farzin Maniei; Resources: Farzin Maniei; data curation: Farzin Maniei; methodology: Farzin Maniei and Stephen P. Mattingly; programming and implementation supporting algorithms: Farzin Maniei; interpretation of results: Farzin Maniei; writing–original draft: Farzin Maniei; writing review and editing: Stephen P Mattingly; draft manuscript preparation: Farzin Maniei; project administration: Stephen P. Mattingly; and funding acquisition: Stephen P. Mattingly. All authors reviewed the results and approved the final version of the manuscript.

References

AASHTO. 2010. *Highway safety manual*. Washington, DC: AASHTO.

Abdel-Aty, M., R. Pemmanaboina, and L. Hsia. 2006. "Assessing crash occurrence on urban freeways by applying a system of interrelated equations." *J. Transp. Res. Board* 1953 (1): 1–9. https://doi.org/10.1177/0361198106195300101.

Abdel-Aty, M. A. 2003. "Analysis of driver injury severity levels at multiple locations using ordered Probit models." *J. Saf. Res.* 34 (5): 597–603. https://doi.org/10.1016/j.jsr.2003.05.009.

Afghari, A. P., M. M. Haque, and S. Washington. 2020. "Applying a joint model of crash count and crash severity to identify road segments with high risk of fatal and serious injury crashes." *Accid. Anal. Prev.* 144 (Sep): 1–11. https://doi.org/10.1016/j.aap.2020.105615.

Ahmed, M. M., and M. A. Abdel-Aty. 2012. "The viability of using automatic vehicle identification data for real-time crash prediction." *IEEE Trans. Intell. Transp. Syst.* 13 (2): 459–468. https://doi.org/10.1109/TITS.2011.2171052.

Alabama DOT. 2015. "Alabama speed management manual." Accessed July 16, 2022. https://www.dot.state.al.us/publications.

Anon. 2011. "sklearn.cluster.KMeans." Accessed July 11, 2021. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.

Azizi, L., and M. Hadi. 2021. "Using traffic disturbance metrics to estimate and predict freeway traffic breakdown and safety events." *Transp. Res. Rec.* 2675 (10): 723–733. https://doi.org/10.1177/03611981211012422.

Barile, C., C. Casavola, G. Pappalettera, and V. Paramsamy Kannan. 2022. "Laplacian score and K-means data clustering for damage characterization of adhesively bonded CFRP composites by means of acoustic emission technique." *Appl. Acoust.* 185 (Jan): 108425. https://doi.org/10.1016/j.apacoust.2021.108425.

Bhatia, J., R. Dave, H. Bhayani, S. Tanwar, and A. Nayyar. 2020. "SDN-based real-time urban traffic analysis in VANET environment." *Comput. Commun.* 149 (Jan): 162–175. https://doi.org/10.1016/j.comcom.2019.10.011.

Bhowmik, T., S. Yasmin, and N. Eluru. 2018. "A joint econometric approach for modeling crash counts by collision type." *Anal. Methods Accid. Res.* 19 (Sep): 16–32. https://doi.org/10.1016/j.amar.2018.06.001.

Borsos, A., J. N. Ivan, and G. Orosz. 2014. "Development of safety performance functions for two-lane rural first-class main roads in Hungary." In *Proc., Transport Research Arena (TRA) 5th Conf.:*

Transport Solutions from Research to Deployment. New York: Wiley. https://doi.org/10.1002/9781119307853.ch6.

Chang, F., S. Yasmin, H. Huang, and A. H. Chan. 2021. "Injury severity analysis of motorcycle crashes: A comparison of latent class clustering and latent segmentation based models with unobserved heterogeneity." *Anal. Methods Accid. Res.* 32 (Dec): 1–28. https://doi.org/10.1016/j.amar.2021.100188.

Cheng, W., G. S. Gill, R. Dasu, M. Xie, X. Jia, and J. Zhou. 2017. "Comparison of Multivariate Poisson lognormal spatial and temporal crash models to identify hot spots of intersections based on crash types." *Accid. Anal. Prev.* 99 (Part A): 330–341. https://doi.org/10.1016/j.aap.2016.11.022.

Cheng, Z., J. Yuan, B. Yu, J. Lu, and Y. Zhao. 2022. "Crash risks evaluation of urban expressways: A case study in Shanghai." *IEEE Trans. Intell. Transp. Syst.* 23 (9): 15329–15339. https://doi.org/10.1109/TITS.2022.3140345.

Cook, D., R. Souleyrette, and J. Jackson. 2011. "Effect of road segmentation on highway safety analysis." In Vol. 11 of *Proc., Transportation Research Board 90th Annual Meeting*. Washington, DC: Transportation Research Board.

De Luca, M., R. Mauro, R. Lamberti, and G. Dell'Acqua. 2012. "Road safety management using Bayesian and cluster analysis." *Procedia Social Behav. Sci.* 54 (Oct): 1260–1269. https://doi.org/10.1016/j.sbspro.2012.09.840.

Depaire, B., G. Wets, and K. Vanhoof. 2008. "Traffic accident segmentation by means of latent class clustering." *Accid. Anal. Prev.* 40 (4): 1257–1266. https://doi.org/10.1016/j.aap.2008.01.007.

Geyer, J., E. Lankina, C. Y. Chan, D. R. Ragland, T. Pham, and A. Sharafsaleh. 2008. *Methods for identifying high collision concentration locations for potential safety improvements*. Berkeley, CA: Univ. of California at Berkeley.

Ghadi, M., and A. Torok. 2019. "A comparative analysis of black spot identification methods and road accident segmentation methods." *Accid. Anal. Prev.* 128 (Jul): 1–7. https://doi.org/10.1016/j.aap.2019.03.002.

Golob, T. F., W. Recker, and Y. Pavlis. 2008. "Probabilistic models of freeway safety performance using traffic flow data as predictors." *Saf. Sci.* 46 (9): 1306–1333. https://doi.org/10.1016/j.ssci.2007.08.007.

Golob, T. F., W. W. Recker, and V. M. Alvarez. 2004. "Freeway safety as a function of traffic flow." *Accid. Anal. Prev.* 36 (6): 933–946. https://doi.org/10.1016/j.aap.2003.09.006.

Green, E. R. 2018. *Segmentation strategies for road safety analysis*. Lexington, KY: UKnowledge. https://doi.org/10.1016/j.aap.2003.09.006.

He, W., X. Cheng, R. Hu, Y. Zhu, and G. Wen. 2017. "Feature self-representation based hypergraph unsupervised feature selection via low-rank representation." *Neurocomputing* 253 (Aug): 127–134. https://doi.org/10.1016/j.neucom.2016.10.087.

He, X., D. Cai, and P. Niyogi. 2005. "Laplacian score for feature selection." In Vol. 18 of *Advances in neural information processing systems*. Lexington, KY: MIT Press.

Islam, M., N. Alnawmasi, and F. Mannering. 2020. "Unobserved heterogeneity and temporal instability in the analysis of work-zone crash-injury severities." *Anal. Methods Accid. Res.* 28 (Dec): 100130. https://doi.org/10.1016/j.amar.2020.100130.

Islam, M., and F. Mannering. 2020. "A temporal analysis of driver-injury severities in crashes involving aggressive and non-aggressive driving." *Anal. Methods Accid. Res.* 27 (Sep): 100128. https://doi.org/10.1016/j.amar.2020.100128.

Islam, M., and A. Pande. 2020. "Analysis of single-vehicle roadway departure crashes on rural curved segments accounting for unobserved heterogeneity." *Transp. Res. Rec.* 2674 (10): 146–157. https://doi.org/10.1177/0361198120935877.

Islam, M., D. Perez-Bravo, and K. K. Silverman. 2017. *Performance-based assessment to transportation safety planning for metropolitan travel improvement study.* Washington, DC: Transportation Research Board.

Ivan, J. N., R. K. Pasupathy, and P. J. Ossenbruggen. 1999. "Differences in causality factors for single and multi-vehicle crashes on two-lane roads." *Accid. Anal. Prev.* 31 (6): 695–704. https://doi.org/10.1016/S0001-4575(99)00030-5.

Karim, A., S. Azam, B. Shanmugam, and K. Kannoorpatti. 2020. "Efficient clustering of emails into spam and ham: The foundational study of a comprehensive unsupervised framework." *IEEE Access* 8: 154759–154788. https://doi.org/10.1109/ACCESS.2020.3017082.

Koorey, G. 2009. "Road data aggregation and sectioning considerations for crash analysis." *Transp. Res. Rec.* 2103 (1): 61–68. https://doi.org/10.3141/2103-08.

Kwon, O. H., M. J. Park, H. Yeo, and K. Chung. 2013. "Evaluating the performance of network screening methods for detecting high collision concentration locations on highways." *Accid. Anal. Prev.* 51 (Mar): 141–149. https://doi.org/10.1016/j.aap.2012.10.019.

Liu, R., N. Yang, X. Ding, and L. Ma. 2009. "An unsupervised feature selection algorithm: Laplacian score combined with distance-based entropy measure." In *Proc., 2009 3rd Int. Symp. on Intelligent Information Technology Application*, 65–68. New York: IEEE.

Lu, J., A. Gan, K. Haleem, and W. Wu. 2013. "Clustering-based roadway segment division for the identification of high-crash locations." *J. Transp. Saf. Secur.* 5 (3): 224–239. https://doi.org/10.1080/19439962.2012.730118.

Mahmud, A., and V. V. Gayah. 2021. "Estimation of crash type frequencies on individual collector roadway segments." *Accid. Anal. Prev.* 161 (Oct): 106345. https://doi.org/10.1016/j.aap.2021.106345.

Mannering, F. L., and C. R. Bhat. 2014. "Analytic methods in accident research: Methodological frontier and future directions." *Anal. Methods Accid. Res.* 1 (Jan): 1–22. https://doi.org/10.1016/j.amar.2013.09.001.

Mannering, F. L., V. Shankar, and C. R. Bhat. 2016. "Unobserved heterogeneity and the statistical analysis of highway accident data." *Anal. Methods Accid. Res.* 11 (Sep): 1–16. https://doi.org/10.1016/j.amar.2016.04.001.

Pande, A., and M. Abdel-Aty. 2006. "Comprehensive analysis of the relationship between real-time traffic surveillance data and rear-end crashes on freeways." *Transp. Res. Rec.* 1953 (1): 31–40. https://doi.org/10.1177/0361198106195300104.

Pande, A., M. Abdel-Aty, and A. Das. 2010. "A classification tree based modeling approach for segment related crashes on multilane highways." *J. Saf. Res.* 41 (5): 391–397. https://doi.org/10.1016/j.jsr.2010.06.004.

Pedregosa, F., et al. 2011. "Scikit-learn: Machine Learning in Python." *J. Mach. Learn. Res.* 12 (Nov): 2825–2830.

Qin, X., and A. Wellner. 2012. "Segment length impact on highway safety screening analysis." *Transp. Res. Rec.* 12: 0644.

Raschka, S., and V. Mirjalili. 2017. *Python machine learning*. 2nd ed. Birmingham, UK: Packt Publishing.

Solorio-Fernández, S., J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad. 2020. "A review of unsupervised feature selection methods." *Artif. Intell. Rev.* 53 (2): 907–948. https://doi.org/10.1007/s10462-019-09682-y.

Thomas, I. 1996. "Spatial data aggregation: Exploratory analysis of road accidents." *Accid. Anal. Prev.* 28 (2): 251–264. https://doi.org/10.1016/0001-4575(95)00067-4.

TxDOT (Texas DOT) Traffic Safety Division. 2020. *Instruction to police for reporting crashes*. Austin, TX: TxDOT.

Valent, F., et al. 2002. "Risk factors for fatal road traffic accidents in Udine, Italy." *Accid. Anal. Prev.* 34 (1): 71–84. https://doi.org/10.1016/S0001-4575(00)00104-4.

Wang, D., et al. 2022. "Assessing dynamic metabolic heterogeneity in non-small cell lung cancer patients via ultra-high sensitivity total-body [18F]FDG PET/CT imaging: Quantitative analysis of [18F]FDG uptake in primary tumors and metastatic lymph nodes." *Eur. J. Nucl. Med. Mol. Imaging* 49 (13): 4692–4704. https://doi.org/10.1007/s00259-022-05904-8.

Wang, X., and M. Feng. 2019. "Freeway single and multi-vehicle crash safety analysis: Influencing factors and hotspots." *Accid. Anal. Prev.* 132 (Feb): 1–12. https://doi.org/10.1016/j.aap.2019.105268.

Xu, C., D. Li, Z. Li, W. Wang, and P. Liu. 2018. "Utilizing structural equation modeling and segmentation analysis in real-time crash risk assessment on freeways." *KSCE J. Civ. Eng.* 22 (7): 2569–2577. https://doi.org/10.1007/s12205-017-0629-3.

Xu, C., P. Liu, W. Wang, and Z. Li. 2012. “Evaluation of the impacts of traffic states on crash risks on freeways.” *Accid. Anal. Prev.* 47 (Jul): 162–171. https://doi.org/10.1016/j.aap.2012.01.020.

Xu, C., A. P. Tarko, W. Wang, and P. Liu. 2013. “Predicting crash likelihood and severity on freeways with real-time loop detector data.” *Accid. Anal. Prev.* 57 (Aug): 30–39. https://doi.org/10.1016/j.aap.2013.03.035.

Yang, Y., X. Ding, and L. Ma. 2009. “An unsupervised feature selection algorithm: Laplacian score combined with distance-based entropy measure.” In Vol. 3 of *Proc., 2009 3rd Int. Symp. on Intelligent Information Technology Application*, 65–68. New York: IEEE.

Yu, R., and M. Abdel-Aty. 2013. “Multi-level Bayesian analyses for single- and multi-vehicle freeway crashes.” *Accid. Anal. Prev.* 58 (Sep): 97–105. https://doi.org/10.1016/j.aap.2013.04.025.

Yu, R., M. Abdel-Aty, and M. Ahmed. 2013. “Bayesian random effect models incorporating real-time weather and traffic data to investigate mountainous freeway hazardous factors.” *Accid. Anal. Prev.* 50 (Jan): 371–376. https://doi.org/10.1016/j.aap.2012.05.011.