Accident Analysis and Prevention 132 (2019) 105269

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

A multivariate-based variable selection framework for clustering traffic conflicts in a brazilian freeway

Miriam Rocha[a,b,*], Michel Anzanello[a], Felipe Caleffi[c], Helena Cybis[c], Gabrielli Yamashita[a]

[a] Department of Industrial Engineering, Federal University of Rio Grande do Sul, Porto Alegre, RS 90035-180, Brazil
[b] Center of Engineering, Federal Rural University of Semi-Arid, Mossoró, RN 59.625-900, Brazil
[c] Laboratory of transport systems, Federal University of Rio Grande do Sul, Porto Alegre, RS, 90035-180, Brazil

ARTICLE INFO

Keywords:
Collision risk
Traffic conflicts
variable selection
Clustering
NLPCA
SOM

ABSTRACT

More than one million people die or suffer non-fatal injuries annually due to road accidents around the world. Understanding the causes that give rise to different types of conflict events, as well as their characteristics, can help researchers and traffic authorities to draw up strategies aimed at mitigating collision risks. This paper proposes a framework for grouping traffic conflicts relying on similar profiles and factors that contribute to conflict occurrence using self-organizing maps (SOM). In order to improve the quality of the formed groups, we developed a novel variable importance index relying on the outputs of the nonlinear principal component analysis (NLPCA) that intends to identify the most informative variables for grouping collision events. Such index guides a backward variable selection procedure in which less relevant variables are removed one-by-one; after each removal, the clustering quality is assessed via the Davies-Bouldin (*DB*) index. The proposed framework was applied to a real-time dataset collected from a Brazilian highway aimed at allocating traffic conflicts into groups presenting similar profiles. The selected variables suggest that lower average speeds, which are typically verified during congestion events, contribute to conflict occurrence. Higher variability on speed (denoted by high standard deviation, and speed's coefficient of variation levels on that variable), which are also perceived in the assessed freeway near to congestion periods, also contribute to conflicts.

1. Introduction

According to the World Health Organization (WHO, 2018), around 1.35 million people die annually due to road accidents, while 20–50 million people suffer non-fatal injuries, some resulting in disabilities and long-term adverse health consequences. Although low and middle income countries concentrate approximately 54% of world's vehicles, such countries answer for more than 90% of road fatalities (WHO, 2013). When it comes to Latin America, The World Bank (TWB, 2013) estimates that the lack of road safety and careful driving in the region results in nearly 130,000 deaths per year, around six million injuries and some sort of disability for hundreds of thousands of people. In addition, traffic accidents are the major cause of death for people aged between 15 and 44 in Latin America. The WHO (2013) suggested that there were 46,935 deaths in Brazilian roads in 2013; in that year, the country ranked third in the world, behind China and India with 261,367 and 207,551 deaths, respectively. In addition, a mortality rate of 23.4 per 100 thousand population in road traffic is expected in Brazil, which is significantly higher than the world's rate of 17.4 per 100 thousand population (WHO, 2015).

Collision risk is deemed a relevant piece of information to estimate traffic conditions (e.g., level of congestion, prevailing speeds, and road volume, among others) most propitious to conflict occurrence (Caleffi et al., 2017). A traffic conflict occurs when two or more road users approach each other in space and time to such an extent that there is a risk of collision whether their movements remain unchanged (Davis et al., 2011). Understanding the causes that give rise to different types of conflict events, as well as their characteristics, can help researchers and traffic authorities to draw up strategies tailored to reducing the occurrence of collisions. In this sense, conflict clustering may be a useful resource as it allows inserting traffic conflict events into groups presenting similar characteristics and highlighting factors that contribute to conflict occurrence. Such groups can then be used to develop joint strategies aimed at reducing the probability of accident occurrence.

Among the various clustering techniques available in the literature, we highlight self-organizing maps (SOM). SOM is an unsupervised-learning neural-network method that provides a visual separation of the

[*] Corresponding author.
E-mail address: miriam.rocha@ufersa.edu.br (M. Rocha).

https://doi.org/10.1016/j.aap.2019.105269
Received 4 April 2019; Received in revised form 15 July 2019; Accepted 12 August 2019
Available online 21 August 2019

different clusters towards a similarity graph (Kohonen, 1995). Although SOM has been widely used in many areas for group formation, we found only five studies applying the SOM in the traffic safety area. Liu and Bucknall (2018) developed an algorithm based on SOM to avoid collisions between unmanned surface vehicles (USVs) that support complex ocean operations. Stoica et al. (2015) introduced a vehicular channel estimator structure based on a low complexity adaption of SOM complemented by a filtered decision feedback layer, while Wang et al. (2014) proposed a model for estimating changing lane probability which combines SOM and Back Propagation (BP). Prieto and Allen (2009) proposed a framework for real-time detection and recognition of traffic signs using SOM. Finally, Prato and Kaplan's (2012) study is the one that most closely resembles our propositions: they identified bus crash clusters by means of a two-stage approach. In that study, SOM was first applied to unveil natural crash groups, followed by Bayesian classification and unified distance matrix edge analysis. Although the aforementioned studies offered remarkable contributions towards reducing accident occurrence, none of them grouped and analyzed traffic conflicts with the SOM method, or proposed variable selection frameworks to enhance the analysis. In this way, it can be noticed that there is still room for research in such an area.

It is noteworthy that SOM, like other multivariate tools, may experience performance reduction when applied to datasets comprised of an elevated number of correlated and noisy variables. To avoid this, the use of variable selection techniques has been hailed an efficient way of increasing the efficiency of statistical techniques, while giving rise to simpler and easier to interpret models (Anzanello et al., 2014). It is expected that the clustering analysis carried out on a reduced subset of truly informative variables can improve SOM performance, making the interpretation of different groups of traffic conflicts easier. This paper proposes a novel approach for selecting the most informative variables for grouping collision events using SOM. For that matter, we first generate a novel importance variable index based on the nonlinear principal component analysis (NLPCA) outputs. Such index guides an one-by-one variable removal process (i.e. variables are iteratively removed from the less to the more important one); after each variable removal, the clustering quality is assessed through the Davies-Bouldin (*DB*) index. The proposed framework was used to identify the most informative variables for allocating traffic conflicts into groups presenting similar profiles in a real-time dataset collected from a Brazilian highway. The formed groups, which were subjectively assessed by traffic specialists, provided relevant information about characteristics that contribute to conflict occurrence.

We see at least two contributions arising from the framework here proposed. First, it offers a new variable selection index based on NLPCA which can assess the variability of datasets presenting nonlinear behavior. Nonlinear techniques can enhance the handling and description of more complex relations between variables, which typically tend to happen in real-world applications (Claveria and Poluzzi, 2017). Although the literature offers several indices for identifying the most informative variables, they are typically grounded in linear modeling and prone to information loss (Anzanello et al., 2011, 2016). A second contribution emerges from the application of a clustering technique unusually applied to the traffic safety area (i.e., SOM). Differently from other traditional clustering techniques, the SOM allows the user to visualize the formed groups by means of color maps, providing an efficient tool for subjective analysis. From a practical perspective, the stratification of traffic conflict events into groups presenting similar characteristics (and the identification of factors that contribute to conflict occurrence) contributes to the development of joint strategies aimed at reducing the probability and severity of conflicting events.

2. Background on NLPCA and SOM

Nonlinear Principal Components Analysis (NLPCA) is a multivariate technique for dimensional reduction inspired in the traditional PCA. In PCA, the original variables describing a conflict event are linearly combined to generate a new latent variable that reproduces as much variance from the original variables as possible (Linting et al., 2007). As for NLPCA, it generates nonlinear combinations of the original variables that allow the principal components to be represented by curves instead of lines (Scholz et al., 2008). Such property enhances the ability of the technique to describe variability in data collected from real-world applications where complex relations among the variables are verified (Li and Yu, 2002; Xie et al., 2003).

Mathematically, consider a matrix **A** consisting of M ($m = 1, \ldots, M$) conflict events described by J ($j = 1, \ldots, J$) variables. The NLPCA reduces the initial set of J variables to a smaller subset of C uncorrelated variables (i.e., principal components). The variance explained by each retained component c ($1, \ldots, C$) is λ_c; ideally, the first x components should explain as much variance as possible in a x dimensional subspace of the data (Scholz and Vigário, 2002). The NLPCA components' weights are associated to observation m, w_{cm}, and the factor loadings associated to variable j are represented by f_{cj} (Mori et al., 2016). The factor loadings are the result of the multiplication between the matrices describing components' weights and the original data **A**. Such parameters give rise to the variable importance index detailed in Section 4.

As for the SOM, it consists of a powerful visualization tool that has been applied in several segments, including industrial, telecommunications, biomedical and financial (Kohonen, 2014). The SOM maps a high-dimensional distribution onto a regular low-dimensional grid, thus converting complex, nonlinear statistical relationships between high-dimensional data items into simple geometric relationships in a low-dimensional display. The clustering carried out in SOM can be shown graphically based on the magnitudes of vectorial distances between neighboring models in the map (Kohonen, 2014). In order to make visual inspection of the data easier, Ultsch (1993) and Kraaijveld et al. (1995) developed the U-matrix, which consists of a graphic display for illustrating the degree of clustering stratification reached by the SOM. In U-matrix, the average (smoothed) distances between the nearest SOM models are represented by colors (Kohonen, 2014). Typically, darker colors denote small mean differences among conflict events (i.e., denoting conflict events presenting similar characteristics), while light colors indicate larger mean differences among events.

3. Study site

Data used in this study were collected from a segment of the Brazilian freeway BR-290/RS, kilometer 94, northbound direction, deemed the main access to the city of Porto Alegre, RS. Such a location was selected among other BR-290/RS sections given its congestion extent, reliable traffic surveillance cameras, and existence of a high flow access ramp that tends to disturb traffic stream. The section under study is comprised of a 3-lane freeway with a single lane access ramp (see Fig. 1). Since this freeway segment does not present loop detectors in the access ramp area where conflicts are likely to occur, data was collected by means of surveillance cameras. For the purpose of this study, a traffic conflict is defined as an event involving the intersection of two or more road users where one or both drivers take evasive maneuvers to avoid a collision (Williams, 1981).

The assessed data were collected during morning peak hours on weekdays and under fine weather conditions, and followed the steps proposed by Huang et al. (2013). The camera was positioned above the freeway to achieve adequate viewing height to cover the bottleneck area 50 m downstream the access ramp. 150 h of traffic data were

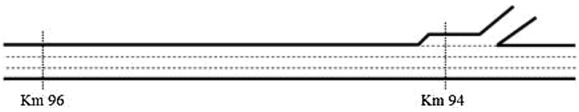


Fig. 1. Freeway section under study.

recorded; from those, 120 h presented fine visibility. The recorded videos were later assessed in laboratory to obtain traffic flow, speeds and conflicts data. All the recorded videos were reviewed by a trained graduate student to ensure that consistent and reliable criteria were used to identify traffic conflicts. The video analysis relied on a computer algorithm to extract traffic volumes and speeds, and on another algorithm for tracking vehicles trajectories. Vehicles trajectories were tracked using an implementation of the Kanade-Lucas-Tomasi Feature Tracker algorithm. Conflicts between vehicles could then be determined by evaluating whether any future position of a specific vehicle would coincide spatially and temporally with other vehicles. Such technique for analysis of surveillance camera videos is a recognized course of action that has been widely used in transport studies (see Autey et al., 2012; Essa and Sayed, 2018, 2015).

Traffic conflicts were identified by assessing vehicles' evasive actions, which included breaking, swerving and deceleration. In order to identify a traffic conflict, the observer monitored vehicles' speed and brake lights. Swerving was considered an indicator of traffic conflict as drivers tend to change direction or lane they are on instead of applying brakes to avoid a collision. As for deceleration, it consisted of a subjective indicator, and was only pointed by the observer when vehicles' brake lights had a mechanical failure, were obstructed, or could not be seen due to inadequate position of video cameras. All conflict events classified in the 50 m area covered by the camera were computed.

Information related to average speed, occupancy and total flow for each lane were collected once a conflict was identified in the intervals of 0–5, and 5–10 minutes. These time intervals were selected due to their consistent results in previous studies; see Islam et al. (2013), Zheng et al. (2010), and Abdel-Aty et al. (2004). In addition, we also collected traffic data for corresponding non-conflict events aimed at performing a matched case-control technique. For instance, if a particular conflict occurred on Monday 09:00 am, data referring to a non-conflict event data were collected in the closest available previous or successive day for the exact hour, provided that no conflict happened at that span of collected time (i.e. 08:45 to 09:00 am). As stated by Roshandel et al. (2015), previous studies have collected multiple non-collision/conflict event data corresponding to each collision event ranging from 1:1 to 1:5.

In order to enhance the description of the assessed events, other variables were created based on the 3 collected variables (i.e., speed, occupancy and flow data for each lane), totalizing 26 variables refereed to traffic conflict observations. The initial pool of variables was comprised of (i) Average speed for the whole section [Av.Speed]; (ii) Standard deviation of speeds [Std.Dev.Speed]; (iii) Speed's coefficient of variation [Coeff.Var.Speed] – ratio between standard deviation and mean; (iv) Difference between standard deviation of lanes' speeds [Diff.Std.Dev.Speed]; (v) Average occupancy for the whole section [Av.Occ]; (iv) Standard deviation of occupancy [Std.Dev.Occ]; (vii) Occupancy's coefficient of variation [Coeff.Var.Occ]; (viii) Difference between standard deviation of lanes' occupancy [Diff.Std.Dev.Occ]; (ix) Total flow for the whole section [Total.Flow]; (x) Standard deviation of flow [Std.Dev.Flow]; (xi) Average flow [Av.Flow]; (xii) Flow's coefficient of variation [Coeff.Var.Flow]; and (xiii) Difference between standard deviation of lanes' flow [Diff.Std.Dev.Flow]. The aforementioned variables were created for 0–5, and 5–10 minutes, accounting for a set of 26 variables. Several previous studies have used this variable unfolding strategy to increase the volume of data and enrich the analysis of conflict events (Lee et al., 2003; Oh et al., 2005; Abdel-Aty et al., 2006; Park and Oh, 2009; Zheng et al., 2010; Islam et al., 2013; Xu et al., 2013).

4. Framework for selecting the most informative clustering variables

The proposed framework for grouping traffic conflicts into clusters presenting similar profiles and characteristics is comprised of four operational steps: (i) collect and prepare the data describing conflicting events; (ii) define the number of clusters NC to be formed; (iii) derive a variable importance index based on NLPCA to guide the iterative removal of less relevant variables; and (iv) cluster observations and eliminate less relevant variables. The proposed steps are now detailed.

Step 1 - Collect and prepare the data describing conflicting events

The dataset bringing the assessed conflict events was collected in a segment of the Brazilian freeway BR-290/RS, as presented in section 3, and it is comprised of 99 traffic conflict observations described by 26 variables. In the following steps, we refer to the dataset as matrix **A** comprised of M observations (i.e. conflicting events) described by J variables. Matrix **A** is then assessed in terms of outliers using the Max-Eigen difference (MED) technique (Gao et al., 2005); outliers can negatively affect the calculation of distances between observations and lead to biased conclusions. In case outliers are identified, analysts are required to decide between their removal from the dataset or treatment.

Step 2 - Define the number of clusters to be formed

A hierarchical tree (i.e. dendrogram) is applied to the dataset resulting from Step 1 aimed at defining a proper number of clusters, NC, to be formed; the distances between clusters are calculated based on Ward's algorithm (Rencher, 2002). Although the dendrogram may not indicate the optimal number of clusters to be considered, it typically provides the decision maker with a reasonable range of clusters capable of describing the underlying structure of the assessed data. In addition to the dendrogram, we also rely on the qualitative assessment of traffic experts to define a number of clusters suitable to further interpretation of formed groups.

Step 3 - Generate a variable importance index based on NLPCA to guide the removal of less relevant variables

This step proposes a new variable importance index aimed at guiding the removal of less important and informative clustering variables. Indices similar to the one here proposed have been widely employed in variable selection frameworks in order to reduce the computational efforts arising from the combinatorial nature of the selection problem (Anzanello et al., 2016). First, NLPCA is applied to matrix **A**. The outputs of interest include the percentage of variance explained, λ_c, and the factor loadings, f_{cj}, described in section 2. As component's weights are not unique, we intend to maximize the variance with a minimum number of principal components. The NLPCA outputs are then combined under a new variable importance index, v_j, to guide the variable removal process; see Eq. (1). Based on the assumption that variables presenting high factor loadings and large variance are preferred, variables with higher v_j are deemed more relevant for grouping conflict events.

$$v_j = \sum_{c=1}^{C} |f_{cj}|\ \lambda_c,\ \forall\ j = 1,\ \ldots, J \tag{1}$$

Step 4 - Cluster observations and eliminate less relevant variables

The assessed conflict events are initially clustered into NC groups using all the J variables through the SOM method; clustering quality is evaluated using the Davies-Bouldin (DB) index (Liu et al., 2013) in Eq. (2). In that equation, C_a refers to the a-th cluster; n_a is the number of objects in C_a; c_a represents the a-th cluster center; $d(m;\ c_a)$ is the distance between observation m and c_a; and, $d(c_a;\ c_b)$ is the distance between c_a and c_b. The objective is to minimize the DB, suggesting proper clustering results (Maulik and Bandyopadhyay, 2002). Thus, the smaller the DB, the better the allocation of conflict events to their final clusters.

$$DB = \frac{1}{NC}\sum_{a} max_{b,b\neq a}\left\{\left[\frac{1}{n_a}\sum_{m\in C_a} d(m,\,c_a) + \frac{1}{n_b}\sum_{m\in C_b} d(m,\,c_b)\right] /d(c_a,\,c_b)\right\} \quad (2)$$

Next, remove the variable with the smallest v_j, perform a new clustering using the J-1 remaining variables with SOM, and recalculate the clustering quality based on DB index. Repeat this iterative procedure (i.e., remove the next attribute with the smallest v_j and cluster conflict events using the remaining variables) until a single variable is left. This removal process yields a clustering quality profile consisting of DB in the vertical axis, and number of retained variables in the horizontal axis. The subset of variables yielding the smallest DB is chosen for future clustering of conflict events; in case multiple subsets lead to the same DB, choose the one retaining the smallest number of variables.

5. Selected variables and clustering results

We now describe the quantitative results yielded by the proposed framework. All experimental procedures were performed using Matlab toolboxes (i.e. Nonlinear PCA and SOM).

Given that the 26 assessed variables were derived from speed, occupancy and flow data for each lane (as explained in Section 3), a large correlation and eventual redundancy between such variables is expected. High collinearity is observed when vectors describing two or more variables lie on the same or on very close lines (Belsley et al., 2005). Given their statistical properties, variables presenting elevated collinearities can jeopardize the performance of several statistical techniques. In order to reduce the impact of highly correlated variables on the proposed framework, we employed the BKW collinearity diagnostics (Belsley, 1991) to identify and remove such problematic variables (e.g. whether three variables are highly correlated and bring very similar information, the BKW removes two of them and retains the one that provides most information).

Condition indexes and variance-decomposition proportions are essential parameters to run BKW. The first quantifies the presence of collinear relations within different subsets of variables, while the second points out the variables candidate to be removed from each subset comprised of collinear variables. In our propositions, variables that exceeded the threshold of 30 for the condition index and presented variance-decomposition proportion superior to 0.5 were deemed highly correlated and removed; see Fig. 2. Eight variables were removed by the BKW: Av.Occ$_5$, Std.Dev.Occ$_5$, Coeff.Var.Occ$_5$, Av.Flow$_5$, Av.Speed$_{10}$, Std.Dev.Speed$_{10}$, Std.Dev.Occ$_{10}$, and Coeff.Var.Occ$_{10}$. Thus, 18 out of the 26 original variables were assessed in further steps of the method.

A dendrogram generated on the 18 remaining variables suggested three (NC = 3) as a proper number of clusters to be formed; see Fig. 3. The assessment of traffic experts corroborated three groups as a promising number of clusters towards providing relevant information about conflict events.

For the SOM clustering, parameters were defined according to three phases: initialization, coarse and fine training. A linear algorithm was chosen for the initial phase, while a sequential algorithm was set for both coarse and fine training phases. The following parameters were also trained: map grid size, lattice (rectangular or hexagonal), neighborhood function (Gaussian, cutgauss, bubble or ep), neighborhood radius in coarse and fine training, cycles in coarse and fine training. Final parameters are presented in Table 1; details on such parameters can be found in Kohonen (2014).

NLPCA was then applied to the dataset comprised of 99 observations and 18 variables; 2 principal components responding for 80.81% of the total variance were retained. The outputs of the NLPCA gave rise to the variable importance index in Eq. (1), which ranked the 18 remaining variables according to their importance for the clustering procedure. Next, we carried out the iterative procedure described in step 4 of section 4 (i.e. cluster observations and eliminate less relevant variables) using SOM to generate 3 groups. That process yielded a graph relating clustering quality assessed by the DB (vertical axis) and number of retained variables used in the clustering procedure (horizontal axis); see Fig. 4. For comparative reasons, a profile depicting DB when variables are removed according to the order defined by the traditional PCA index is also presented in Fig. 4 and later discussed.

In Fig. 4, the DB presented a substantial reduction as variables were removed according to the order suggested by the NLPCA index (note that such reduction suggests an improvement on the quality of formed groups as less informative variables are discarded from the dataset used for clustering). The two smallest DB values are 0.3404 and 0.5262, obtained when one and three variables are used for clustering, respectively. As both scenarios present satisfactory DB values, we decided to retain the three-variable result as it yields more information to interpret the formed clusters. The three retained variables are Av.Speed$_5$,

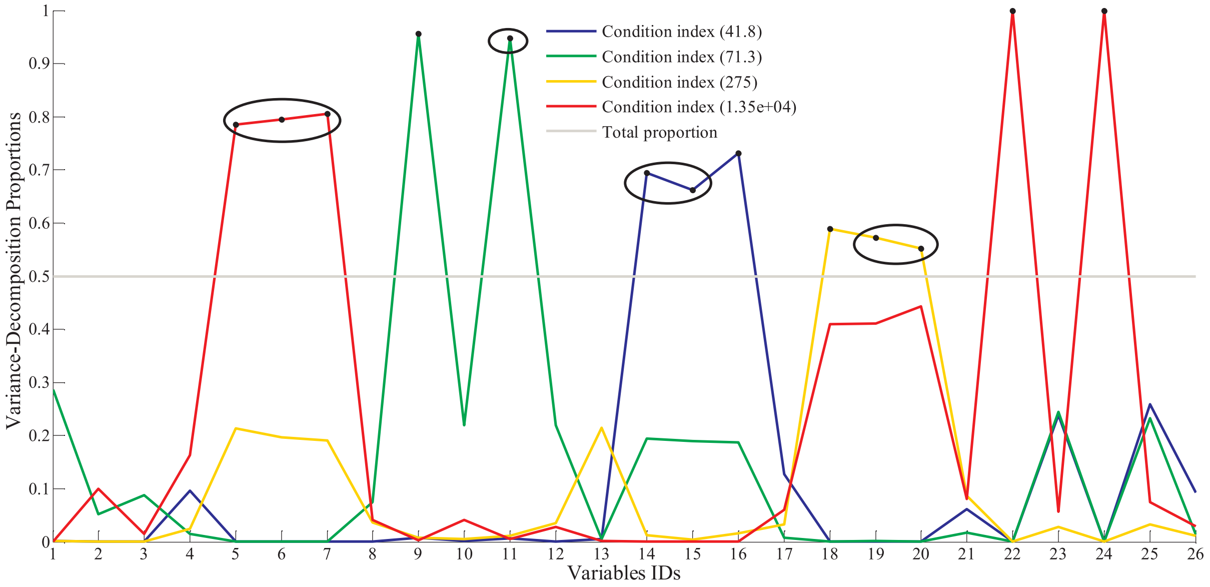


Fig. 2. BWK diagnostics (variables removed are indicated by an ellipse in the graph).

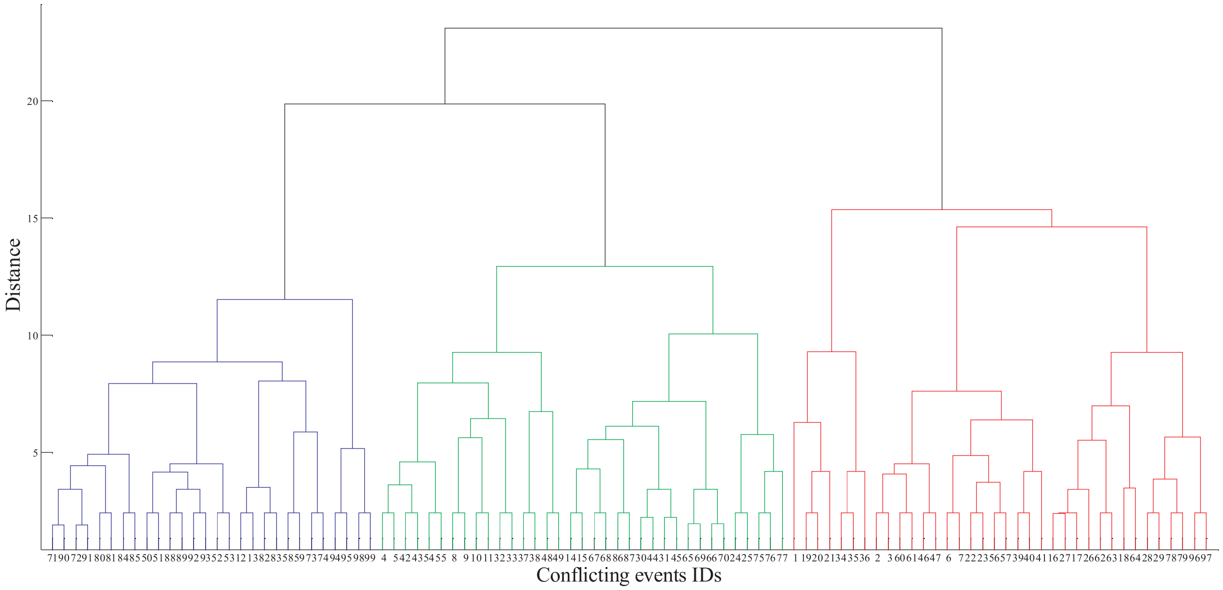


Fig. 3. Dendrogram suggesting the formation of 3 clusters.

Table 1
Estimated SOM parameters.

Parameter	Specification
Map grid size	[9 9]
Lattice	Hexagonal
Neighborhood function	Bubble
Neighborhood radius, coarse	[5 5]
Cycles in coarse training	500
Neighborhood radius, fine	[5 5]
cycles in fine training	250

Coeff.Var.Speed$_5$, and Std.Dev.Speed$_5$; such variables are qualitatively discussed in section 6.

For comparative purposes, we also run the selection process using a variable importance index based on the traditional linear PCA (Anzanello et al., 2014, 2016), and grouped the conflict events using SOM (Rencher, 2002); see Fig. 4. *DB* values are typically smaller for the NLPCA-based index when a reduced number of retained variables is considered for clustering. When the range from 1 to 8 variables is used as input for the SOM, the quality of the grouping procedure using the NLPCA-based index performs better in all cases, except for the scenario when two variables pointed by the PCA-based index are used for the clustering (DB = 0.5414). That *DB*, however, does not outperform the three-variable scenario suggested by the NLPCA-based index, which yields DB = 0.5262. The superior performance of the NLPCA-based index can be justified by the ability of the NLPCA to unveil complex information present in the conflict events dataset, which typically tends to rely on nonlinear relations among the variables. The robustness of the proposed variable importance index is also remarkable when the SOM is replaced by the *k*-means, a traditional and widely used clustering technique (such results are not presented due to space restrictions).

Next, we built the U-matrices aimed at visually assessing the clusters yielded by the subset of selected variables. As aforementioned, the U-matrix consists of a graphic display based on the magnitudes of

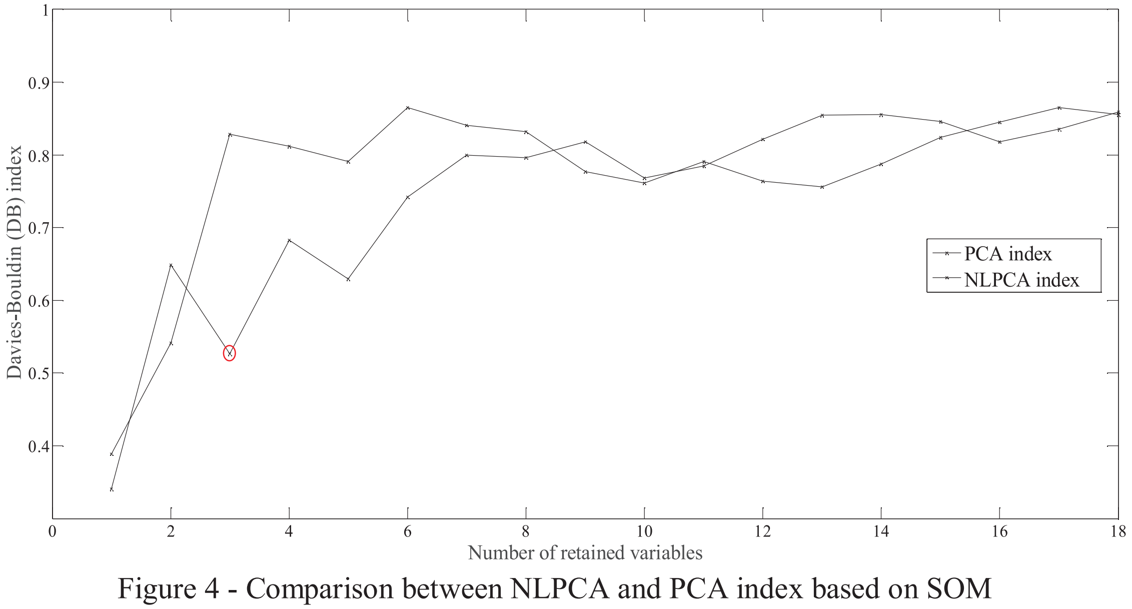
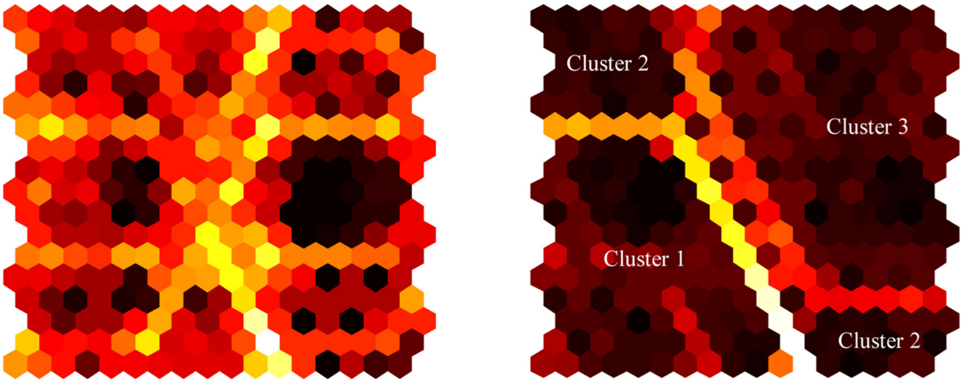


Fig. 4. Comparison between NLPCA and PCA index based on SOM.



(a) U-matrix of the 26 original variables

(b) U-matrix of the 3 retained variables

Fig. 5. U-matrices for entire and selected datasets.

vectorial distances between neighboring models in a map (Kohonen, 2014). The U-matrix in Fig. 5(a) refers to groups generated when the clustering was carried out using the 26 original variables, while the U-matrix in Fig. 5(b) illustrates the groups formed when only the 3 selected variables were used by the SOM. The U-matrix based on the entire set of variables suggests a higher distance between the neurons, and no clear separation of clusters is observed. As for the U-matrix based on the 3 selected variables, the dark, compact regions denote a significant proximity between the neurons and a better group formation. In addition, Fig. 5(b) displays a well-defined stratification between clusters represented by the lighter lines that serve as borders for clustering separation. The comparison between these two graphs corroborates the importance of selecting an informative subset of relevant variables in order to improve the stratification of the conflict events.

6. Qualitative assessment of retained variables and formed clusters

In the quantitative analysis (section 5), the proposed framework retained 3 out of the 26 original variables capable of improving the grouping of conflict events: Av.Speed$_5$, Coeff.Var.Speed$_5$, and Std.Dev.Speed$_5$. As mentioned in section 3, such variables represent, respectively, average speed for the whole section, speed coefficient of variation, and speed standard deviation in five minutes.

The retained variables are aligned with results obtained by previous similar studies (Lee et al., 2003; Oh et al., 2005; Abdel-Aty et al., 2006; Park and Oh, 2009; Zheng et al., 2010; Islam et al., 2013; Xu et al., 2013). The variables that account for the speed standard deviation (Std.Dev.Speed$_5$) and the ratio between standard deviation and mean (Coeff.Var.Speed$_5$) were created to account for speed pattern heterogeneity. These variables were proven to be effective in identifying the observed speed differences between vehicles travelling along a lane. This traffic pattern, noticeable of Brazilian motorways, leads to aggressive maneuvers, resulting in high rates of lane changes and rear-end conflicts, thus increasing the probability of conflict occurrence.

Table 2
Conflict *versus* Non-conflict (comparison on the three retained variables).

	Av.Speed$_5$		Coeff.Var.Speed$_5$		Std.Dev.Speed$_5$	
	Conflict	Non-conflict	Conflict	Non-conflict	Conflict	Non-conflict
Mean	69.01	76.20	0.38	0.31	23.91	21.19
Minimum value	52.65	57.67	0.24	0.21	19.41	17.05
Maximum value	81.13	83.92	0.60	0.51	32.96	28.04

Table 2 depicts the differences in the 3 retained variables for conflict and non-conflict events. It is noteworthy that average speeds for non-conflict events are significantly higher, suggesting that lower average speeds, which are typically verified during congestion events, contribute to conflict occurrence. Also, the retained variable Std.Dev.Speed$_5$ was considered effective in identifying the noticeable Brazilian's speed differences between vehicles. For conflict events, this variable presents higher values, indicating more disturbance in the traffic stream. From a practical perspective, such variables can provide relevant information tailored to the improvement of dynamic control algorithms. Dynamic traffic control measures (e.g. variable speed limits or ramp metering) can help to reduce speed variability, lane changing and rear-end collision risk.

We now discuss the three clusters generated by the SOM using the 99 conflict events and the three selected variables (Av.Speed$_5$, Coeff.Var.Speed$_5$, and Std.Dev.Speed$_5$). Clusters 1, 2 and 3 consist of 41, 19, and 39 conflict events, respectively.

Figs. 6–8 depict boxplots for the three retained variables with regards to each cluster. Cluster 1 presents higher values of Av.Speed$_5$ and lower values of Coeff.Var.Speed$_5$ and Std.Dev.Speed$_5$. The opposite happens in cluster 3: it presents smaller values of Av.Speed$_5$ and larger values of Coeff.Var.Speed$_5$ and Std.Dev.Speed$_5$. Cluster 2 assumes an intermediate position. In order to corroborate differences on such variables within clusters, we applied ANOVA and multiple mean comparisons (Rencher, 2002) with post hoc of Tukey. Results suggest that the 3 clusters are significantly different from each other for the 3 variables for a confidence level of 0.05.

Figs. 9 and 10 depict the visual separation of events based on the retained variables. Observations belonging to clusters 1 and 3 are clearly stratified when variables Av.Speed$_5$ versus Coeff.Var.Speed$_5$ (Fig. 9), and Av.Speed$_5$ versus Std.Dev.Speed$_5$ (Fig. 10) are considered for group formation. It is important to note that all the conflict events in this study occurred at morning peak hour, when the freeway operates close to maximum capacity.

In Cluster 3 presented in Fig. 9, the events are related to traffic conditions close to maximum road capacity (i.e. during breakdown occurrence). Moments before and during this period, speed difference between vehicles are at higher rate. This condition leads to an increase in lane changing and evasive maneuvers, which tends to cause conflicts with greater severity. In these events, average speeds are around 60% of the freeway speed limit, and the standard deviation of speeds are higher.

On the other hand, cluster 1 presents the group of conflicts relying on higher average speeds. Average speeds are around 75% of the freeway speed limit, and the standard deviation of speeds has the smallest differences. In these events, even during peak hours, the vehicles have more freedom than vehicles in Cluster 3. This freedom

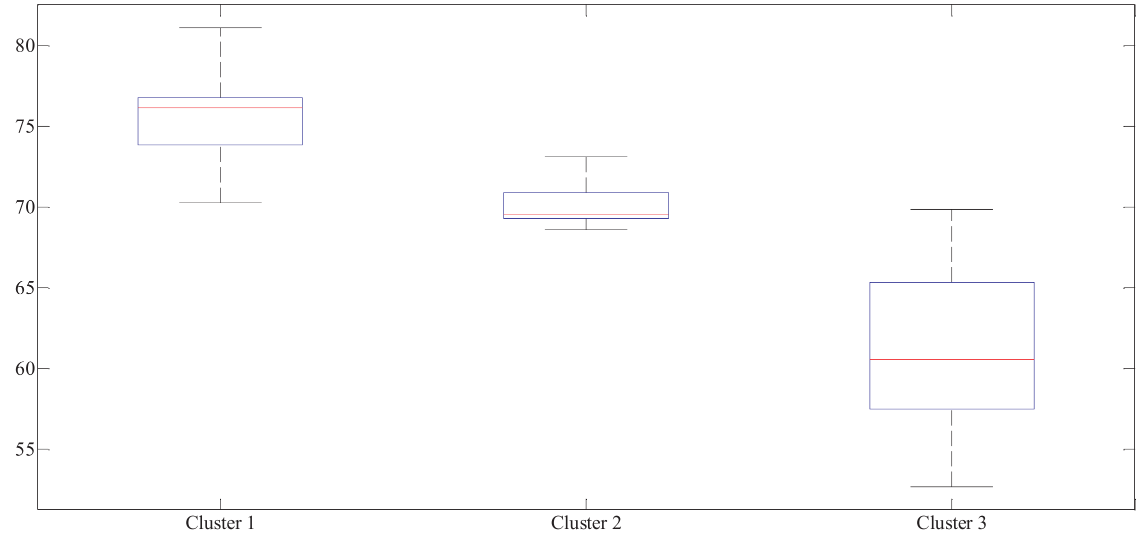


Fig. 6. Boxplot for variable Av.Speed$_5$.

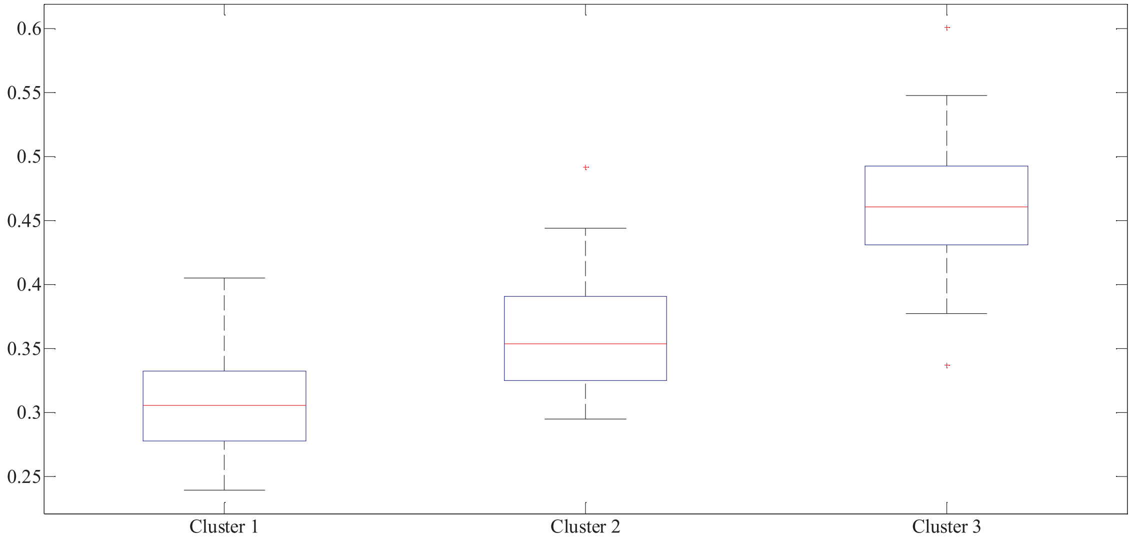


Fig. 7. Boxplot for variable Coeff.Var.Speed$_5$.

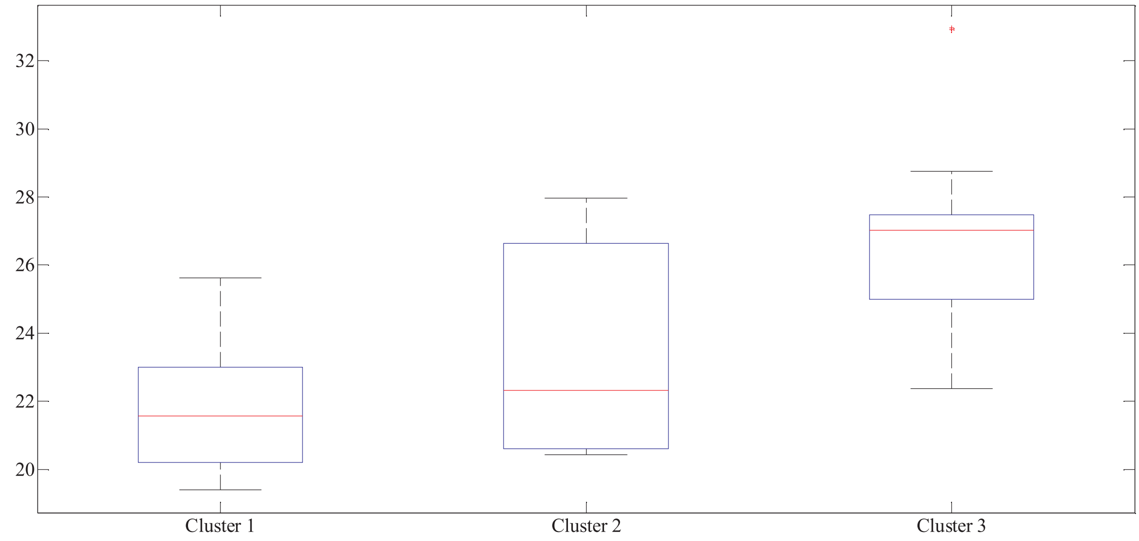


Fig. 8. Boxplot for variable Std.Dev.Speed$_5$.

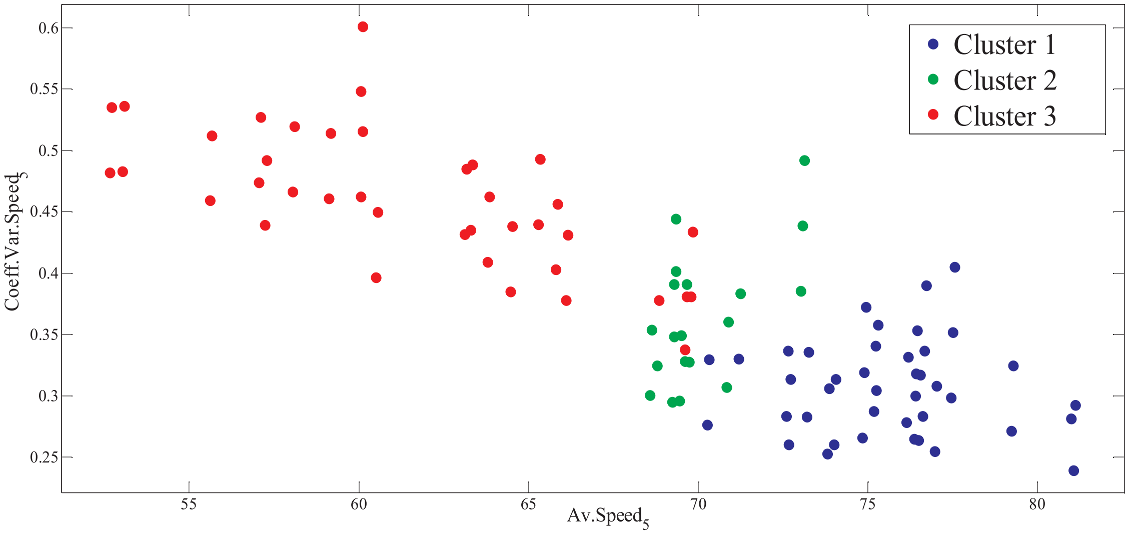


Fig. 9. Cluster analysis (Coeff.Var.Speed$_5$ *versus* Av.Speed$_5$).

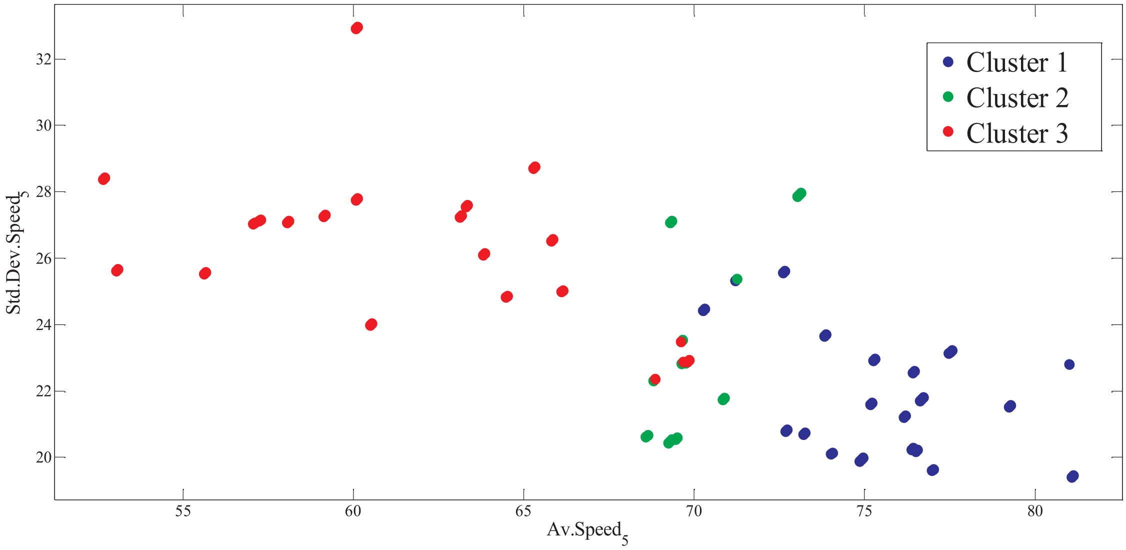


Fig. 10. Cluster analysis (Std.Dev.Speed$_5$ *versus* Av.Speed$_5$).

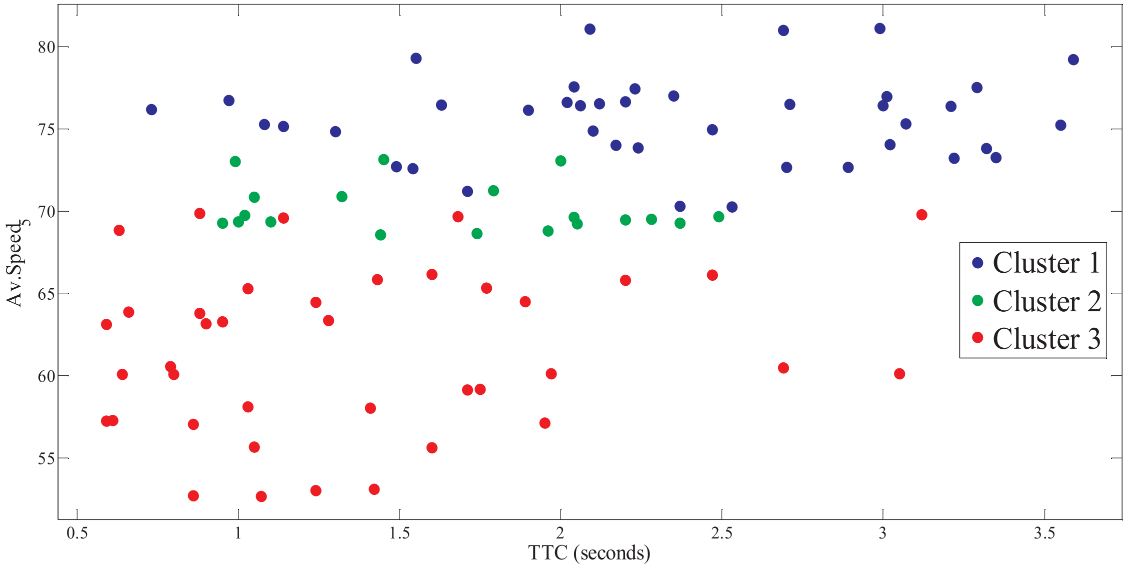


Fig. 11. Av.Speed$_5$ *versus* TTC.

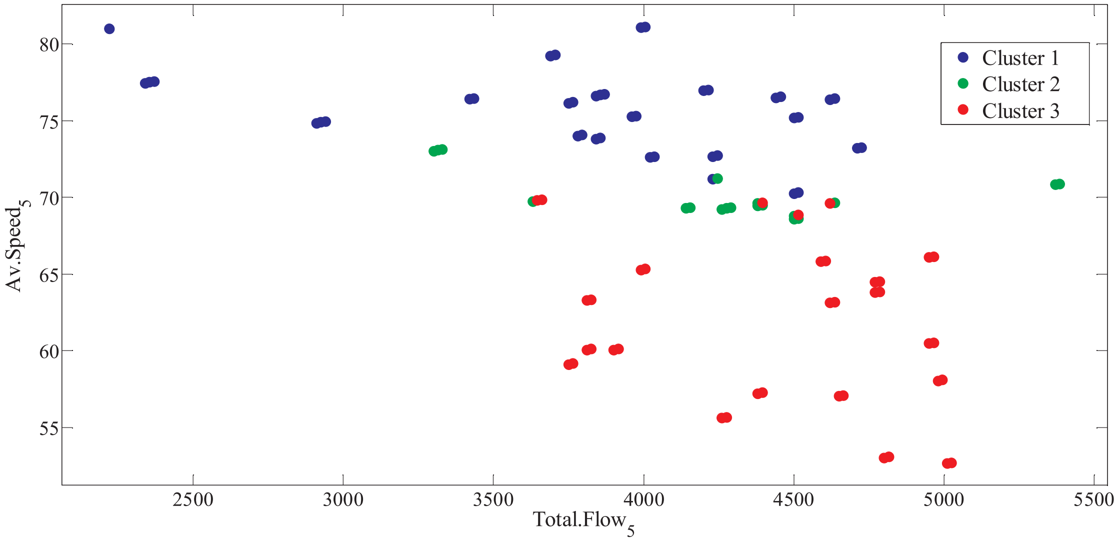


Fig. 12. Speed-flow relationship between $Av.Speed_5$ and $Total.Flow_5$.

condition in the road (i.e. road occupancy level is smaller) leads to events with larger time-to-collisions (TTC), even under a conflict state; such events are represented by the blue points in Fig. 11. Relative positions and velocities between road users or between a user and obstacles can be characterized by a collision course and the corresponding TTC. As point out by Davis et al. (2011), the severity of a conflict decreases as TTC increases. TTC is an example of an indicator that indicates collision probability primarily: the smaller the TTC, the less likely drivers have time to perceive and react before a collision, and thus the higher the probability of a collision outcome. Thus, Cluster 3 could be classified as the group containing conflicts with lower severity potential. This statement is also corroborated by Fig. 11, which depicts the TTC for the 3 clusters.

It is also noteworthy that Cluster 2 is comprised of events belonging to a transition area between Clusters 1 and 3. Cluster 2 presents smallest number of events (19 of 99 events), and does not represent a specific traffic state. However, it offers a clear visual boundary between the other clusters. TTC values in Cluster 2 also make clear the difference between Cluster 1 and 3: while Cluster 1 presents high TTCs, Cluster 3 presents TTC falling in a range that could be considered as critical regarding conflict severity, since lower TTC values correspond to higher conflict severities, as pointed out by Gettman and Head (2003) and Kruysse (1991).

Since the data were collected during peak hours – when the majority of conflicts occur –all conflicts are related to levels-of-service D, E and F (for more details about levels-of-service, see Lu et al., 2008). While Cluster 2 only presents a visual boundary between the remaining clusters, Cluster 1 can clearly be classified as level-of-service D, and Cluster 3 as levels-of-service E and F. Conflicts in Cluster 3 occur under congestion conditions, which are verified when the freeway reaches its capacity, and lead to higher conflict severities (Fig. 12). Zhou and Sisiopiku (2007), who previously studied the correlation between volume/capacity (v/c) and accident rates, corroborate this statement. In their study, it was found that high ranges of v/c ratio increase traffic conflicts and may contribute to accident occurrence.

7. Conclusions

This paper proposed a framework to group traffic conflicts presenting similar profiles from a real-time database collected from a Brazilian highway. In order to identify the most informative variables for such grouping, we generated a novel variable importance index based on NLPCA to guide a backwards variable selection procedure. Three variables were retained ($Av.Speed_5$, $Coeff.Var.Speed_5$, and $Std.Dev.Speed_5$) and used by the SOM to cluster the traffic conflicts. The new proposed index improved the SOM results. Differently from the U-matrix using the 26 variables, the U-matrix based on the 3 selected variables presented well-defined borders between the clusters, suggesting an improvement on the clustering quality when a selected subset of variables is used for grouping the conflicts.

These 3 variables also suggest that lower average speeds, which are typically verified during congestion events, contribute to conflict occurrence. Higher variability on speed (denoted by high standard deviation, and speed's coefficient of variation levels on that variable), which are also perceived in the assessed freeway near to congestion periods, will also contribute to conflicts. This type of analysis, based on data obtained by traffic detectors, can contribute to the improvement of dynamic control algorithms. Dynamic traffic control measures (e.g. variable speed limits or ramp metering) have the potential to reduce speed variability, which shown to be directed linked to conflict occurrence. By harmonizing the speeds between vehicles, these techniques can reduce lane changing and rear-end collision risk.

Future research includes the application of supervised multivariate techniques (e.g., k-Nearest Neighbor or Support Vector Machine) to insert events into categories of conflict severity. The use of the parameters derived from Partial Least Squares regression to build a new variable importance index is also promising.

References

Abdel-Aty, M., Dilmore, J., Dhindsa, A., 2006. Evaluation of variable speed limits for real-time freeway safety improvement. Accid. Anal. Prev. 38 (2), 335–345. https://doi.org/10.1016/j.aap.2005.10.010.

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, F., Hsia, L., 2004. Predicting freeway crashes from loop detector data by matched case-control logistic regression. Transportation Research Record: Journal of the Transportation Research Board 1897, 88–95. https://doi.org/10.3141/1897-12.

Anzanello, M.J., Fogliatto, F.S., Rossini, K., 2011. Data mining-based method for identifying discriminant attributes in sensory profiling. Food Qual. Prefer. 22 (1), 139–148. https://doi.org/10.1016/j.foodqual.2010.08.010.

Anzanello, M.J., Fogliatto, F.S., Ortiz, R.S., Limberger, R., Mariotti, K., 2014. Selecting relevant Fourier transform infrared spectroscopy wavenumbers for clustering authentic and counterfeit drug samples. Sci. Justice 54 (5), 363–368. https://doi.org/10.1016/j.scijus.2014.04.005.

Anzanello, M., Fogliatto, F., Marcelo, M.C.A., Pozebon, D., Ferrão, M.F., 2016. Wavelength selection framework for classifying food and pharmaceutical samples into multiple classes. J. Chemom. 30 (6), 346–353. https://doi.org/10.1002/cem.2799.

Autey, J., Sayed, T., Zaki, M.H., 2012. Safety evaluation of right-turn smart channels using automated traffic conflict analysis. Accid. Anal. Prev. 45, 120–130. https://doi.org/10.1016/j.aap.2011.11.015.

Belsley, D.A., 1991. A guide to using the collinearity diagnostics. Comput. Sci. Econ. Manag. 4 (n. 1), 33–50. https://doi.org/10.1007/BF00426854.

Belsley, D.A., Kuh, E., Welsch, R.E., 2005. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity Vol. 571 John Wiley & Sons.

Caleffi, F., Anzanello, M.J., Cybis, H.B.B., 2017. A multivariate-based conflict prediction model for a Brazilian freeway. Accid. Anal. Prev. 98, 295–302. https://doi.org/10.1016/j.aap.2016.10.025.

Claveria, O., Poluzzi, A., 2017. Positioning and clustering of the world's top tourist destinations by means of dimensionality reduction techniques for categorical data. J. Destin. Mark. Manag. 6 (1), 22–32. https://doi.org/10.1016/j.jdmm.2016.01.008.

Davis, G.A., Hourdos, J., Xiong, H., Chatterjee, I., 2011. Outline for a causal model of traffic conflicts and crashes. Accid. Anal. Prev. 43 (6), 1907–1919. https://doi.org/10.1016/j.aap.2011.05.001. issn: 0001-4575.

Essa, M., Sayed, T., 2015. Transferability of calibrated microsimulation model parameters for safety assessment using simulated conflicts. Accid. Anal. Prev. 84, 41–53. https://doi.org/10.1016/j.aap.2015.08.005.

Essa, M., Sayed, T., 2018. Traffic conflict models to evaluate the safety of signalized intersections at the cycle level. Transp. Res. Part C Emerg. Technol. 89 (February), 289–302. https://doi.org/10.1016/j.trc.2018.02.014.

Gao, S., Li, G., Wang, D., 2005. A new approach for detecting multivariate outliers. Communications in Statistics—Theory and Methods 34 (8), 1857–1865. https://doi.org/10.1081/STA-200066315.

Gettman, D., Head, L., 2003. Surrogate safety measures from traffic simulation models. Transp. Res. Rec. J. Transp. Res. Board 1840 (1), 104–115. https://doi.org/10.3141/1840-12.

Huang, F., Liu, P., Yu, H., Wang, W., 2013. Identifying if VISSIM simulation model and SSAM provide reasonable estimates for field measured traffic conflicts at signalized intersections. Accid. Anal. Prev. 50, 1014–1024. https://doi.org/10.1016/j.aap.2012.08.018.

Islam, M., Hadiuzzaman, M., Fang, J., Qiu, T., El-Basyouny, K., 2013. Assessing mobility and safety impacts of a variable speed limit control strategy. Transp. Res. Rec. 2364, 1–11. https://doi.org/10.3141/2364-01.

Kohonen, T., 1995. Self-organizing Maps. Springer, Berlin.

Kohonen, T., 2014. MATLAB Implementations and Applications of the Self-organizing Map. Unigrafia *Oy*, Helsinki, Finland, pp. 11–23.

Kraaijveld, M.A., Mao, J., Jain, A.K., 1995. A nonlinear projection method based on Kohonen's topology preserving maps. IEEE Trans. Neural Netw. 6 (3), 548–559. https://doi.org/10.1109/72.377962.

Kruysse, H.W., 1991. The subjective evaluation of traffic conflicts based on an internal concept of dangerousness. Accid. Anal. Prev. 23 (1), 53–65. https://doi.org/10.1016/0001-4575(91)90035-4.

Lee, C., Hellinga, B., Saccomanno, F., 2003. Real-time crash prediction model for application to crash prevention in freeway traffic. Transp. Res. Rec. 1840, 67–77. https://doi.org/10.3141/1840-08.

Li, E. and Yu, J.(2002). "An input-training neural network-based nonlinear principal component analysis ap-proach for fault diagnosis". 4, pp. 2755–2759.

Linting, M., Meulman, J.J., Groenen, P.J.F., van der Kooij, A.J., 2007. Nonlinear principal components analysis: introduction and application. Psychol. Methods 12 (3), 336–358. https://doi.org/10.1037/1082-989X.12.3.336.

Liu, Y., Bucknall, R., 2018. Efficient multi-task allocation and path planning for unmanned surface vehicle in support of ocean operations. Neurocomputing 275, 1550–1566. https://doi.org/10.1016/j.neucom.2017.09.088. issn: 0925-2312.

Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J., Wu, S., 2013. Understanding and enhancement of internal clustering validation measures. IEEE Trans. Cybern. 43 (3), 982–994. https://doi.org/10.1109/TSMCB.2012.2220543. issn: 2168-2267.

Lu, J., Pan, F., Xiang, Q., 2008. Level-of-Safety service for safety performance evaluation of highway intersections. Transp. Res. Rec. 2075 (1), 24–33. https://doi.org/10.3141/2075-04.

Maulik, U., Bandyopadhyay, S., 2002. Performance evaluation of some clustering algorithms and validity indices. IEEE Trans. Pattern Anal. Mach. Intell. 24 (12), 1650–1654. https://doi.org/10.1109/TPAMI.2002.1114856.

Mori, Y., Kuroda, M., Makino, N., 2016. Nonlinear Principal Component Analysis and Its Applications. Springer.

Oh, C., Oh, J.-S., Ritchie, S.G., 2005. Real-time hazardous traffic condition warning system: framework and evaluation. Ieee Trans. Intell. Transp. Syst. 6 (3), 265–272. https://doi.org/10.1109/TITS.2005.853693.

Park, J., Oh, C., 2009. Relating freeway traffic accidents to inductive loop detector data using logistic regression. 4th IRTAD Conference 223–231.

Prato, C.G., Kaplan, S., 2012. Promoting safe transit: Analyzing bus accident patterns 1–18.

Prieto, M.S., Allen, A.R., 2009. Using self-organising maps in the detection and recognition of road signs. Image Vis. Comput. 27 (6), 673–683. https://doi.org/10.1016/j.imavis.2008.07.006. issn: 0262-8856.

Rencher, A.C., 2002. Methods of Multivariate Analysis, 2nd ed. John Wiley & Sons.

Roshandel, S., Zheng, Z., Washington, S., 2015. Impact of real-time traffic characteristics on freeway crash occurrence: systematic review and meta-analysis. Accid. Anal. Prev. 79, 198–211. https://doi.org/10.1016/j.aap.2015.03.013.

Scholz, M., Fraunholz, M., Selbig, J., 2008. Nonlinear principal component analysis: neural network models and applications. Lecture Notes in Computational Science and Engineering 58, 44–67. https://doi.org/10.1007/978-3-540-73750-6_2.

Scholz, M.S., Vigário, R., 2002. Nonlinear PCA: a new hierarchical approach. Esann 439–444.

Stoica, R.-A., Severi, S., De Abreu, G.T.F., 2015. Learning the vehicular channel through the self-organization of frequencies. 2015 IEEE Vehicular Networking Conference (VNC) 68–75. https://doi.org/10.1109/VNC.2015.7385549. vol. 2016-January.

Ultsch, A., 1993. Self-organizing neural networks for visualisation and classification. Information and Classification. Springer, pp. 307–313. https://doi.org/10.1007/978-3-642-50974-2_31.

Wang, J., Chai, R., Wu, Q. (2014). "Changing lane probability estimating model based on neural network". pp. 3915–3920. DOI: 10.1109/CCDC.2014.6852864.

WHO, World Health Organization (2015). Global status report on road safety 2015. World Health Organization.

Williams, M.J., 1981. Validity of the traffic conflicts technique. Accid. Anal. Prev. 13 (2), 133–145. https://doi.org/10.1016/0001-4575(81)90025-7.

Xie, L., Zhang, Q.-L., Guo, M., Wang, S.-Q. (2003). "Linear pruning techniques for neural networks - Based on projection latent structure". 2, 1304–1309.

Xu, C., Tarko, A.P., Wang, W., Liu, P., 2013. Predicting crash likelihood and severity on freeways with real-time loop detector data. Accid. Anal. Prev. 57, 30–39. https://doi.org/10.1016/j.aap.2013.03.035. issn: 0001-4575.

Zheng, Z., Ahn, S., Monsere, C.M., 2010. Impact of traffic oscillations on freeway crash occurrences. Accid. Anal. Prev. 42 (2), 626–636. https://doi.org/10.1016/j.aap.2009.10.009.

Zhou, M., Sisiopiku, V.P., 2007. Relationship between volume-to-Capacity ratios and accident rates. Transp. Res. Rec. J. Transp. Res. Board 1581, 47–52. https://doi.org/10.3141/1581-06.

Web References

null

TWB (2013). *Latin America: Time to Put a Stop to Road Deaths*. Ed. by The World Bank (TWB). url: http://www.worldbank.org/en/news/feature/2013/05/10/accidentes-trafico-carreteras-america-latina (visited on 06/29/2018).

null

WHO (2013). *Road traffic deaths: Data by country*. Ed. by World Health Organization (WHO). url: http://apps.who.int/gho/athena/data/GHO/RS_196,RS_198?filter=COUNTRY:*&format=xml&profile=excel (visited on 10/10/2017).

null

WHO (2018). *Road traffic injuries*. Ed. by World Health Organization (WHO). url: http://www.who.int/en/ news-room/fact-sheets/detail/road-traffic-injuries (visited on 08/14/2019).