UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

REVISTA UD Y LA GEOMÁTICA
https://revistas.udistrital.edu.co/index.php/UDGeo/index

UD y la GEOMÁTICA

INVESTIGACIÓN

Data mining techniques for road accidents: clustering versus complex

Técnicas de mineração de dados para acidentes rodoviários: clusterização versus redes complexas

María Lígia Chuerubim[1], *Alan D. B. Valejo*[2], *George Miguel Farha Diban*[3],
Bruno de Oliveira Lázaro[4], *Barbara Stolte Bezerra*[5], *Irineu da Silva*[6]

Para citar este artículo: Chuerubim, ML. Valejo, A.D.B, Farha-Diban, G.M., de Oliveira, B. Stolte-Bezerra, B. & da Silva, I. (2020) Data mining techniques for road accidents: clustering versus complex. UD y Geomática, (15), 69-79. DOI: https://doi.org/10.14483/23448407.16720

Fecha de envio: 29 de julio de 2020

Fecha de aceptación: 01 de noviembre de 2020

ABSTRACT:

This work analyses the performance of grouping methods based on complex networks and clusters, in order to identify main road accident groups and risk factors. The research included a balancing step of data classes, used in the classification and extraction process of decision rules applied in each grouping. Then, it was possible to assess and visualize the critical areas of traffic accidents involving victims with material damage, non-fatal and fatal victims. The results pointed out that complex networks present better possibility of generalization for different subsets of data, and higher accuracy in group formation when compared to traditional clustering methods. The use of complex networks aided in the process of acquiring decision rules with higher level of confidence, and higher probability of occurrence.

Key words: Road Safety. Clustering. Complex Networks. Decision Rules.

RESUMO

Este trabalho analisa o desempenho de métodos de agrupamento baseados em redes complexas e clusters, a fim de identificar os principais grupos de acidentes rodoviários e os fatores de risco. A pesquisa incluiu a etapa de balanceamento das classes de dados, usada no processo de classificação e extração das regras de decisão, aplicadas em cada agrupamento. Em seguida, foi possível a avaliação e visualização de áreas críticas de acidentes de trânsito, envolvendo vítimas com danos materiais, vítimas não fatais e fatais. Os resultados apontaram que as redes complexas apresentam melhor possibilidade de generalização para diferentes subconjuntos de dados e maior precisão na formação de grupos, quando comparadas aos métodos tradicionais de agrupamento. O uso das redes complexas auxiliou no processo de aquisição de regras de decisão com maior nível de confiança e maior probabilidade de ocorrência.

Palavras-Chaves: Segurança Viária. Clustering. Redes Complexas. Regras de Decisão.

[1] Correo eletrônico: marialigia@ufu.br. Vinculación institucional: Universidade Federal de Uberlândia, UFU, Faculdade de Engenharia Civil, Uberlândia, Minas Gerais, Brasil.Escola de Engenharia de São Carlos, USP, São Carlos, São Paulo, Brasil.

[2] Correo eletrônico: alanvalejo@gmail.com. Vinculación institucional: Departamento de Computação e Matemática - FFCLRP, USP, Ribeirão Preto, São Paulo, Brasil.

[3] Correo electronico: diban.george80@gmail.com. Vinculación institucional: Universidade Estadual Paulista, UNESP, Faculdade de Engenharia Civil , Bauru, São Paulo, Brasil

[4] Correo Electronico: brunodeoliveira38@gmail.com Vinculación institucional: Universidade Federal de Uberlândia, UFU, Faculdade de Engenharia Civil, Uberlândia, Minas Gerais, Brasil

[5] Correo electronico: irineu@sc.usp.br. Vinculación institucional: Escola de Engenharia de São Carlos, Departamento de Engenharia de Transportes, USP, São Carlos, São Paulo, Brasil

[6] Correo electronico: irineu@sc.usp.br. Vinculación institucional: Escola de Engenharia de São Carlos, Departamento de Engenharia de Transportes, USP, São Carlos, São Paulo, Brasil

Introduction

The development of new and modern technologies in traffic engineering such as speed sensors, real time video monitoring systems and Global Navigation Satellite System, has been making it possible to acquire large amount of data in a traffic environment. Nevertheless, this massive acquisition of information leads to the necessity for using techniques which allows the extraction or detection of patterns of interest from large data sets.

The large amount of information generated has been a challenge to researchers in the traffic engineering area, as well as to decision makers. In addition, the roads are dynamic, subject to spatial and time alterations, what makes it impracticable to apply traditional statistic methods in order to analyze the relations among traffic, road infrastructure, and weather conditions variables. Under this perspective, data mining techniques have been widely used, because it is a solid tool in the decision-making process, through the execution of grouping, data balancing and classifying tasks, as well as the extraction of decision rules.

Nevertheless, the main gap found in the literature is the absence of discussion regarding the procedures about detection and classification of road accident database. In addition, those researches did not discuss the limitations of these procedures.

In this context, this article proposes an approach that explores this subject through the combination of different techniques of data mining based on clustering and complex network clustering procedures. Its main objective is to compare the performance of these different approaches in detecting clusters or communities in traffic accident databases, with a pattern of enabling the detection of patterns in the occurrences of accidents with victims with different levels of severity.

Background

As it was previously mentioned, data mining techniques make possible to handle and extract information from large databases, through relations and interactions between a high number of variables, providing the necessary information to several processes that involve decision making (Fan, Han & Liu, 2014; Facelli et al., 2016; Halim et al., 2016). The large amount of data, which is inherent to the modern technologies, leads to the need for new alternatives for acquisition, manipulation and storage of this massive information. In order to overcome these challenges, several data mining approaches have been proposed in the literature (Aqib et al. 2019; Mehdizadeh et. al, 2020). From these data mining approaches it can be highlighted: clustering, data classes balancing, classification and extraction of decision rules.

Chang and Wong (2006) explored data mining techniques based on decision trees, aiming to assess the relation among several variables and the road accident severity level. The authors concluded that the severity of the accident for these data was directly related to the type of vehicle, and that the pedestrians were those that presented higher level of exposure to the traffic, and thus, those who had higher risk of injury.

Abugessaisa (2008) approached the use of data grouping techniques based on clustering together with the classification of the data through decision trees. Lin et al. (2014) have explored the potential of complex networks and association rules in data mining traffic. These authors highlighted the effectiveness of the network grouping method while detecting communities and the prevalence variables in accident frequency.

Data grouping with clustering

Data clustering methods are divided in hierarchical and no hierarchical. In the first one, the elements of a group become elements of the superior group from a series of grouping decisions, which combine observation in a hierarchical structure. Whereas in non-hierarchical methods, in general, k initial groups are defined and then the observable n is allocated to the k groups. In these methods the initial allocation is arbitrary and iteratively seeks the optimal allocation (Hair et al., 2009).

In this work, the no hierarchical method will be considered, using *k-means* algorithm. The *k-means* allocates each of the *n* elements in one *k* group previously established. At the same time, it minimizes the sum of the squares of the residual inside each group, in order to increase its homogeneity, while maximize this sum between groups, which increases the difference between them.

Performance measures in clusters detection

In order to evaluate the performance of a clustering algorithm, some measures can be used, such as Silhouette coefficient (Equation 1), which quantifies the level of dissimilarity between variables that compose each cluster and selects the best number of clusters (Rousseeuw, 1987):

$$S(i) = \frac{b_i - a_i}{\max(a_i, b_i)}, \qquad (1)$$

where, S is the Silhouette coefficient calculated for a variable i; a_i is the average dissimilarity existent between i in relation with the other variables belonging to the same cluster, and b_i is the average dissimilarity of the variable i, in relation with the other variables belonging to the closest neighbour cluster. The average Silhouette coefficient to all variables i in relation with the data analysed is obtained by Equation 2:

$$S_{WC} = \frac{1}{N} \Sigma_{i=1}^{N} S(i), \qquad (2)$$

where, S_{wc} corresponds to the average coefficient obtained after the Silhouette calculation for each variable analysed, and N the total number of variables analysed. Coefficients $S(i)$ and S_{wc} vary from -1 to 1.

Data grouping by homogeneous complex networks

A R network can be described by $R = (V, E)$, where V corresponds to the set of vertices type $V = \{v_1, v_2, ..., v_n\}$, and E to the set of edges or links type $E = \{e_1, e_2,..., e_n\}$, (Newman, 2004). Each corner is connected to at most two vertices, v and u, being one corner defined by $e_{v,u} = \{(v, u) = (u, v) \mid v, u \in V\}$.

Measures of performance for detecting communities

Complex networks approach is based on building homogeneous complex networks, by using $k - NN$ (*k–Nearest Neighbors*), (Berton, 2016). For each network generated, it was applied the algorithm of detection of communities *Fast-Greedy* (Clauset *et al.*, 2004). The goal here is to find the division of the network that maximizes the modularity, which is obtained according to Equation 3, (Clauset *et al.*, 2004):

$$Q = \frac{1}{2m} \sum_{i,j} (A_{i,j} - \frac{k_i k_j}{2m}) \delta\left(c_i, c_j\right), \qquad (3)$$

Where m corresponds to the total number of edges of the network, and $A_{i,j}$ to the adjacent matrix between vertices i and j (in case there is a connection between vertices i and j, having $A_{i,j} = 1$, otherwise $A_{i,j} = 0$), k_i and k_j correspond, respectively, to the number of vertices connected to vertices i and j (that means, k_i is equivalent to the degree of vertices i, and k_j to the degree of vertices j). Function δ returns 1 in case vertices i and j belong to the same community, and 0 in other cases.

Data classification, data balancing and decision trees

The data modeling is made from sets of test and training, which trains the classifier used and estimate the new values to the classes' attributes of the variables in analysis, based on input values or attributes. Within the existent classification methods, those ones based on hierarchic tree structures stands out.

The unbalance of a database occurs when a number of instances of a given class are much higher than the others. In the literature, the unbalancing of classes is minimized by the application of techniques of data classes balancing such as undersampling, oversampling, and SMOTE (Synthetic Minority Over-sampling Technique), (Li *et al.*, 2017). In the process of balancing of classes by SMOTE, the minor class is balanced considering each of its instances, where samples are introduced based on the neighborhood criteria, using the closest neighbor algorithm.

Decision rules are disjointed, that means, one rule is only triggered when a new example is classified. A decision rule is an implication if A then B, in which A represents a set of conditions, and each condition is defined by a relation type attribute and value, in which value appertain to the domain of attribute in analysis. So, metrics are used to maximize the purity score in each node within the possible input variables (De Oña *et al.*, 2013).

Materials and Methods

Database

The database used in this work is composed by registers of road accidents observed during the period between 2009 and 2016, within km 125 to km 145+500 of Dom Pedro I Highway, in the urban area of Campinas city (Brazil), provided by "Rota das Bandeiras" concessionaire and road infrastructure variables georeferenced collected by the authors in the area of study. The database originally provided by the concessionaire, together with variables collected on the field, is composed by 34 original variables: type of accident, probable cause, period of the day, kilometer milestone, direction of the road, location of the accident, fatal victims (FV), victims with material damage (VMD), non-fatal victims (NFV), visibility condition, weather condition, road condition, road pattern, shoulder, shoulder width, illumination, speed radar, interchange, ramp, maximum speed, road profile, road inclination, construction signaling, horizontal signaling, vertical signaling, latitude, longitude, number of lanes, critical lane's width, central reservation, width central reservation, barriers, average daily volume (AADT) of personal cars and average daily volume (AADT) of commercial vehicles.

The database was divided in three sub data set in order to apply and assess the performance of grouping techniques based on clusters and complex networks, while analyzing registers on accidents with VMD, NFV, and FV for before period (2009-2012) and after period of interventions (2012-2016). So, each of the three data subsets is composed by 31 variables.

The accidents occurring between 2009 and 2012 have resulted in 3,087 victims, in which 79% correspond to VMD, 20% NFV, and 1% of FV. On the other hand, for the period between 2012 and 2016, 2,891 victims were registered in traffic, in which 77% correspond to VMD, 22% NFV, and 1% FV. These data were used in the exploratory phase of the research, with the goal to aid in the descriptive interpretation of the databases, reducing the heterogeneity of the data, and in the detection of possible subjective relations between variables.

Methodology

Based on the three data subsets with VMD, NFV, and FV, the grouping algorithms were applied using clusters and homogeneous complex networks with the purpose of reduce the heterogeneity of the databases. In the clustering process it was applied the *k-means* algorithm with *k* varying from 3 to 10 clusters. Previous experiences run with the data set have shown that for $k<10$ there were significant variations in the group's formation in relation with WSS (*Within-Cluster Sum of Square*).

The lower the WSS value the higher the homogeneity inside each grouping. It must be adopted a *k* value from which there is no significant increase in the total sum of the data variance, while adding new clusters. The clusters obtained from *k-means* method were evaluated in terms of WSS, and by using the performance metric of Silhouette partitioning for different values of *k*, with the goal to find the best value for this parameter in SPSS software.

In the approach using homogeneous complex networks, the algorithm *k-NN* was applied for community detection that selects, in previously iteration, the closest neighbor values of *k* based on the highest number of attributes correlated. In this work, networks with *NN* value (Number of Nearest Neighbors) were generated by experimentation, having values equals to 3, 5, 10, 15, 20, and 30, in order to validate the best parameter of *k-NN*, considering that for $NN>30$, there were no significant alterations observed in the network.

In order to detect communities using complex networks, the algorithm Fastgreedy was used, which optimizes the measurement of modularity proposed by Newman (2004). The partitioning obtained with Fastgreedy and that presents the highest values of modularity suggests a more robust network structure, and consequently at a better division of the network in communities.

Aiming to compare groupings generated by *k-means* in relation with Fastgreedy, communities were generated with the same grouping rate in the interval from 2 to 10, seeking for detecting the best value of *k*, to construct the network, and define the optimal number of communities (C). The communities will be discriminated by C_i^n, in which *i* indicates the index of the community, and *n* the total number of communities obtained.

Finally, by obtaining the optimal number of groupings in before and after period data subsets (VMD, NFV, and FV), the classification process was applied by decision tree algorithm CART using Gini metric, which seeks for maximizing the homogeneity of nodes in relation with the dependent variable, through computational functions implemented in SPSS software.

Analysis and Discussion of Results

Data clustering

The Figures 1 to 12 illustrate until the point at which data partitioning in clusters result in significant variations in the groupings obtained. For the database with VMD, for the before period (2009-2012), 4 communities were obtained (Figures 1 and 2) with 2,433 registers of accidents with respectively total occurrences in the clusters 1, 2, 3 and 4: 517 (21%); 753 (31%); 1,049 (47%) and 114 (5%).

For the after period (2012-2016), 2,235 accidents with VMD were recorded and 5 communities were obtained (Figures 3 and 4), with respectively the total occurrences: 366 (16%), 1; 289 (13%); 794 (36%); 420 (19%); and 366 (16%). In the after period 2009 and 2012, 616 registers of accidents with NFV have been detected.

By analyzing the WSS variation (Figure 5), the optimal structure of groupings is observed for $k = 4$, with average Silhouette value equals to 0.94 (Figure 6). According to this, cluster 1, 2, 3 and 4 concentrates, respectively, observations of: 163 (26%), 124 (20%); 95 (15%) and 234 (38%).

For the after period 2012 to 2016, 627 registers were observed with NFV. Thus, the best value for this parameter is $k = 4$(Figure 7 and 8). In this setting, it is noted the total occurrences, respectively: 108 (17%); 191 (30%); 204 (33%) and 124 (20%).

The analysis of groupings obtained for the after period considering FV accidents indicate that the algorithm converges with $k = 3$ (Figure 9 and Figure 10). Cluster 1, 2 and 3, respectively, presents 14 observations (37%), 16 observations (42%) and 8 observations (21%).

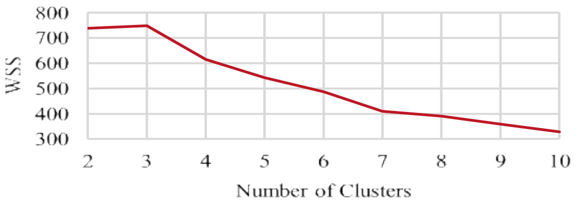

Figure 1 - WSS for different values of *k* with VMD for 2009-2012.

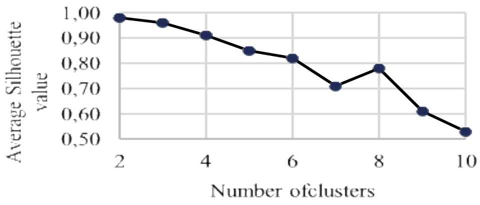

Figure 2 - Average value of the Silhouette for different values of *k* with VMD for 2009-2012.

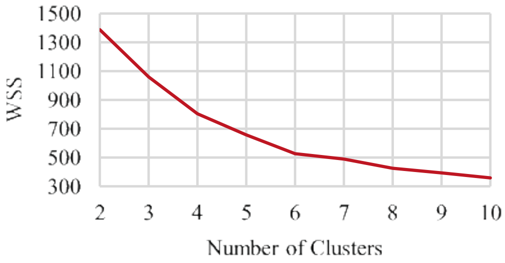

Figure 3 - WSS for different values of *k* with VMD for 2012-2016.

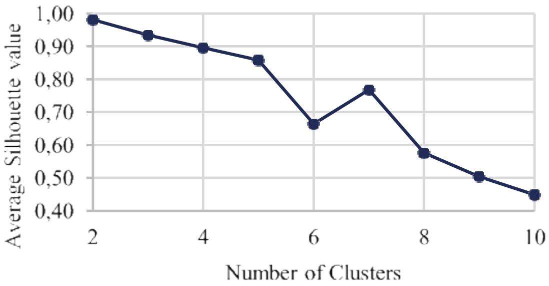

Figure 4 - Average Silhouette value for different values of *k* with VMD for 2012-2016.

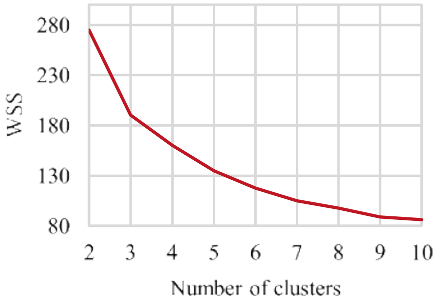

Figure 5 - WSS for different values of *k* with NFV for 2009-2012.

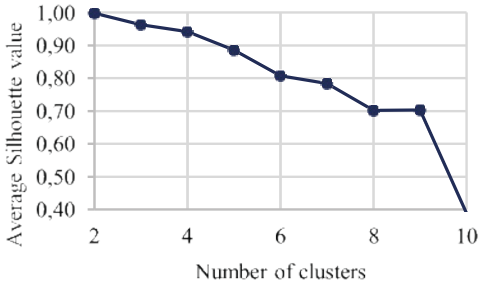

Figure 6 - Average Silhouette value for different values of *k* with NFV for 2009-2012.

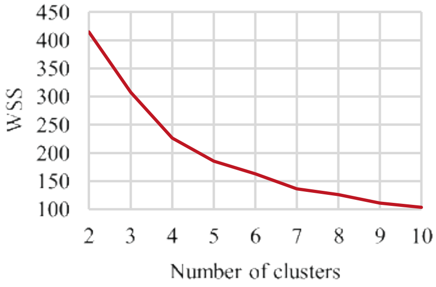

Figure 7 - WSS for different values of *k* with NFV for 2012-2016.

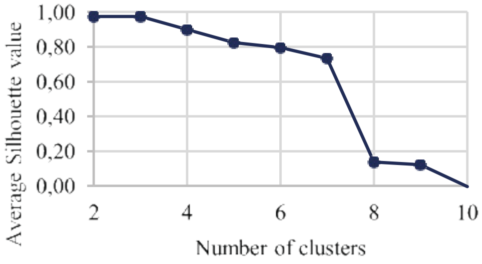

Figure 8 - Average Silhouette value for different values of *k* with NFV for 2012-2016.

For the after setting (2012 to 2016), the algorithm converges with $k = 2$ (Figures 11 and 12). The formation of groupings in this setting indicates that cluster 1 concentrates 10 observations (34%), while cluster 2 presents 19 observations (66%).

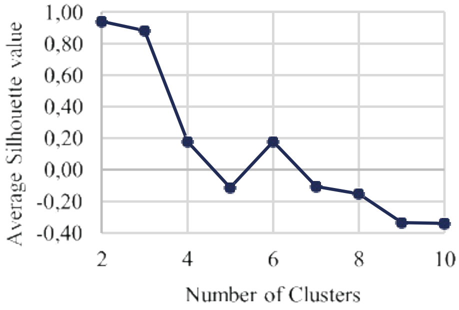

Figure 9 - WSS for different values of *k* with FV for 2009-2012.

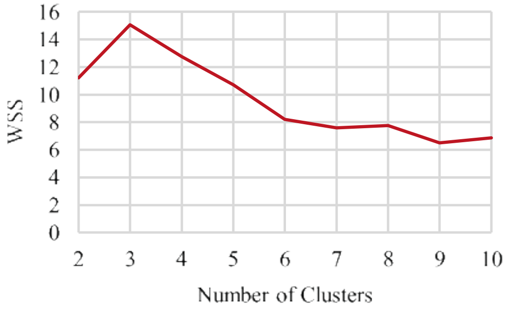

Figure 10 - Average Silhouette value for different values of *k* with FV for 2009-2012.

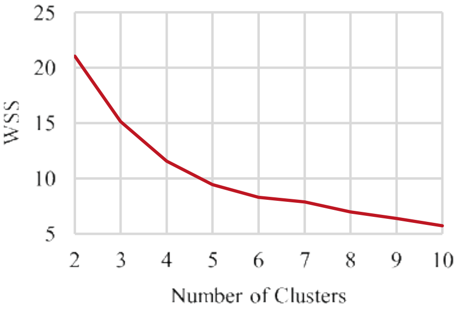

Figure 11 - WSS for different values of k with FV for 2012-2016.

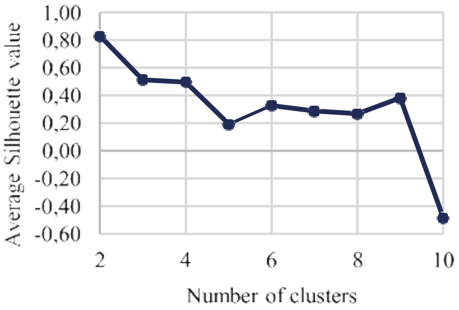

Figure 12 - Average Silhouette value for different values of k with FV for 2012-2016.

Groupings by complex networks

The results obtained for the data set for VMD analyzed for the before period (2009-2012) made it possible to identify that the most robust structure is identified for 3-NN (Figure 13) and that with the highest average Modularity value equals to 0.76 (Figure 14). The network $3-NN$ presents more homogeneous communities for $C = 3$, (Figure 16), with respectively, 36% , 18% and 47% of the observations.

As for the after period (2012-216), together with the database including VMD, it is noted that $k-NN$ algorithm converged with $k = 25$ (Figure 16).

Network $25-NN$ presents average Modularity value of 0.79 (Figure 17), and more homogeneous communities for $C = 5$, as it is presented in Figure 19, with, respectively, 15%, 15%, 29%, 25% and 16% of the observations.

The analysis of the after period (2012-216), including accidents with NFV, has presented a more robust network structure with $k = 10$ (Figure 19). Network $10-NN$ presents average modularity value of 0.78 (Figure 20) and converges with $C = 3$ (Figure 21), with, respectively,. 36%, 28% and 37% of the observations.

The after period (2012-2016) for accidents with NFV presents the best network structure with $k = 10$, as it can be seen in Figure 22.

Network $10-NN$ converges at an optimal rate with $C = 4$ (Figures 23 and 24) with, respectively, 26%, 30%, 22% and 21% of the observations. The after-period database (2012-216), with FV registers, has presented the best network structure for $k = 3$, as it can be seen in Figure 25.

Network $3-NN$ converges at an optimal rate for $C = 3$ (Figure 26 and 27), with, respectively, 39%, 37% and 24% of the observations. Regarding the number of communities, it can be noted that WSS variation (Figure 27) is insignificant starting from $C = 3$, what indicates a strong structure of groups.

The network structure with FV for the after setting has converged with $k = 5$, and then this is the best parameter for defining groups, with average Modularity of 0.57 (Figure 28).

With $C = 4$, Figures 29 and 30) it is noted that $C^{4}_{i=1}$, $C^{4}_{i=2}$, $C^{3}_{i=3}$ and $C^{4}_{i=4}$ concentrates, respectively, the total observations: 4 (14%), 9 (31%), 11 (38%) and with 5 observations (17%).

Extraction of decision rules from groupings of accidents with victims

The database with VMD registers for the before and the after settings has presented the best and the highest assertiveness rate with Fastgreedy, being it respectively 83% and 86%. In terms of accuracy, the before setting was 96%, and after 93%. Through these decision rules applied to each grouping obtained by Fastgreedy in the before setting of the intervention with VMD, it was identified that the most probable accidents in Grouping 1 are rear-end impact type, which occur in the direction South of the road, with level profile between km 125 and km 141+900. Grouping 2 and Grouping 3 own accidents presenting the highest probability of occurrence in direction North, in parts with barriers, or mixed, with critical lane's width superior to 2.96 m and maximum speed of 100 km/h.

As for the after setting with VMD with groupings obtained with Fastgreedy, it was noted that in Grouping 1, accidents present the highest probability of occurrence starting from km 132+900 in segments without barriers, direction North. In Grouping 2, accidents in km 132+700 are more probable, and finally Grouping 3 between km 125 and km 132+700. In Grouping 4 the most probable accidents start in km 132+900, in segments with barriers, or mixed (with and without barriers), direction North of the road, and associated to the probable cause "road/environment" and "others".

For the database with registers of NFV, the before setting of the intervention presents the best performance with Fastgreedy, with assertiveness rate of decision rules 100%, and accuracy of 94%. As for the after setting with NFV, it is noted that both partitioning methods present assertiveness rate of 100%, and they converge with the same number of groupings ($k = 4$). Nonetheless, the accuracy of classification was superior for method k-means, with magnitude of 98%.

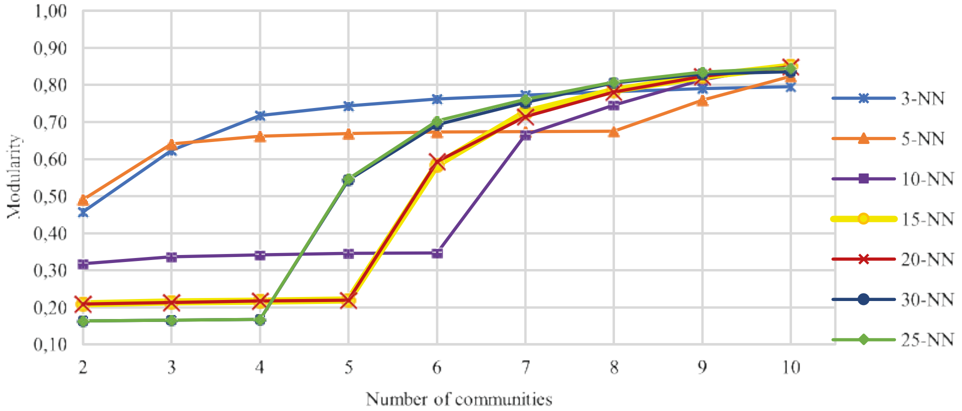

Figure 13 - Construction of complex networks for VMD before period 2009-2012 with variation of parameter $k-NN$, and number of communities (C).

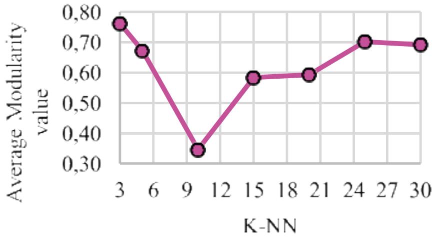

Figure 14 - Average Modularity value for different values of $k-NN$ for VMD between 2009-2012.

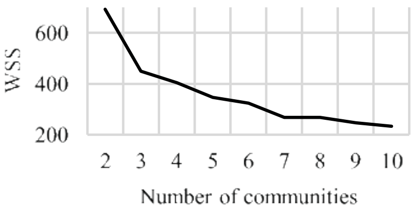

Figure 15 - WSS for different values of C with VMD between 2009-2012.

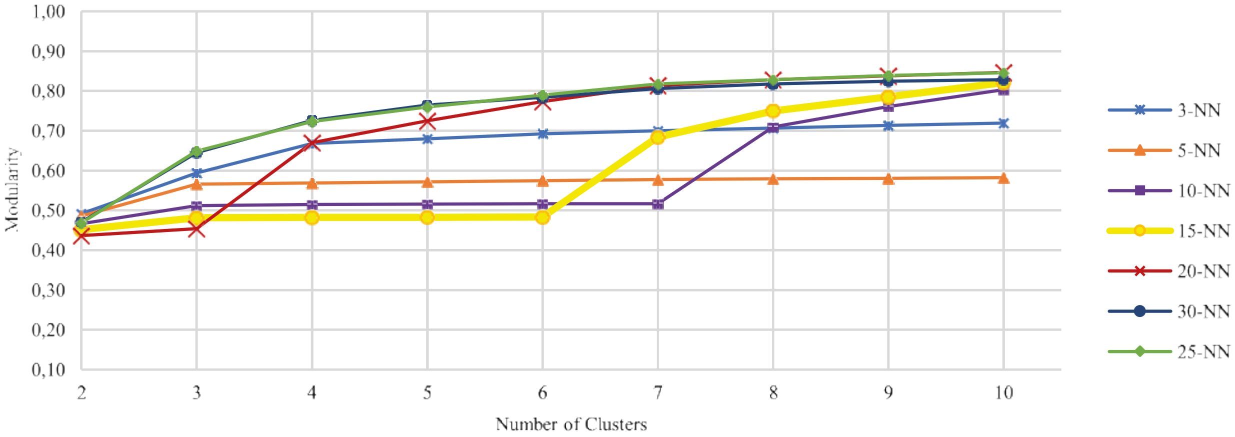

Figure 16 - Construction of complex networks for VMD from 2012-2016 with variation of $k-NN$, and the number of communities (C).

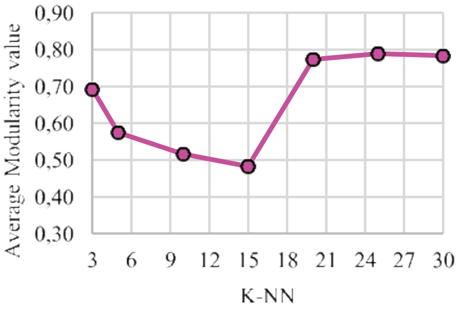

Figure 17 - Average Modularity value for different values of $k-NN$- for VMD between 2012-2016.

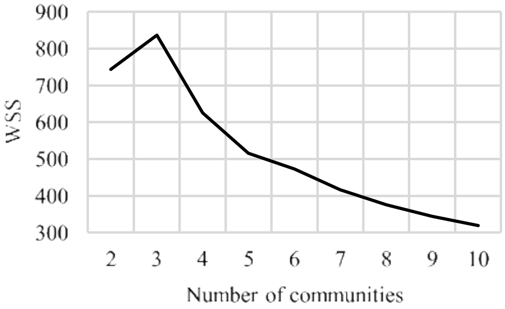

Figure 18 - WSS for different values of C with VMD between 2012-2016.

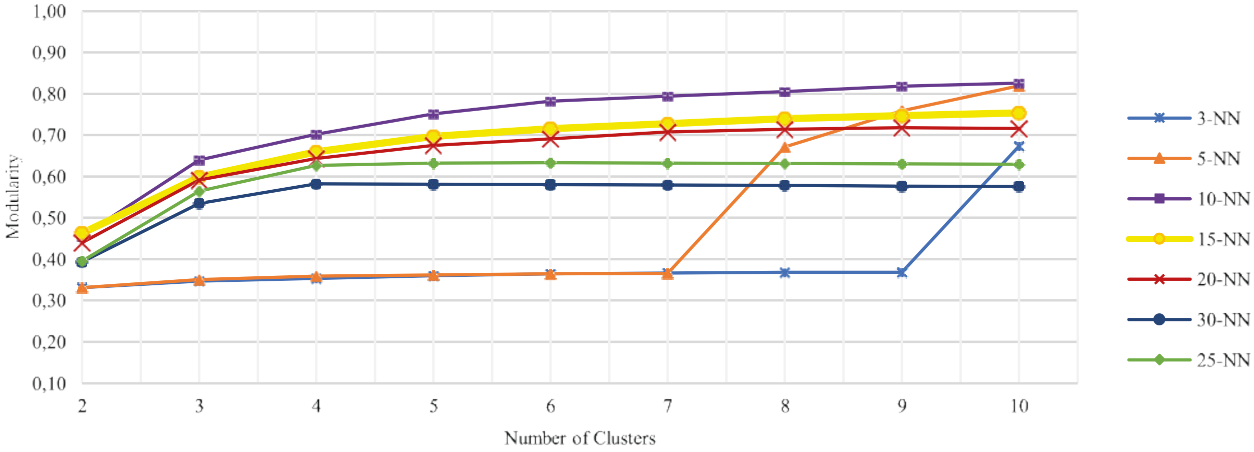

Figure 19 - Construction of complex networks for NFV between 2009-2012 with variation of $k-NN$ and number of communities (C).

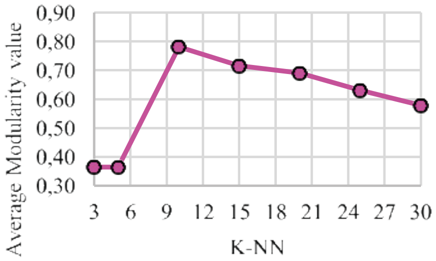

Figure 20. Average Modularity value for different values of $k-NN$ for NFV between 2009-2012.

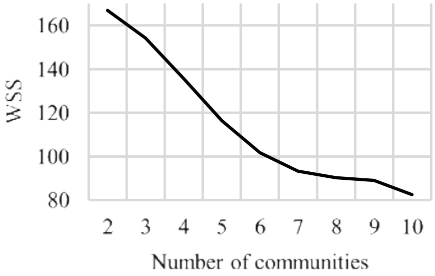

Figure 21. WSS for different values of C with NFV between 2009-2012.

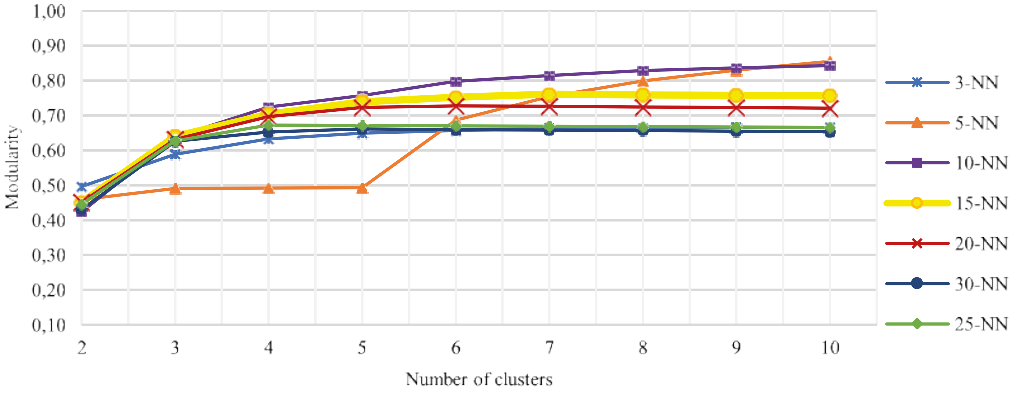

Figure 22 - Construction of complex networks for NFV between 2012-2016 with variation of $k-NN$ and number of communities (C)

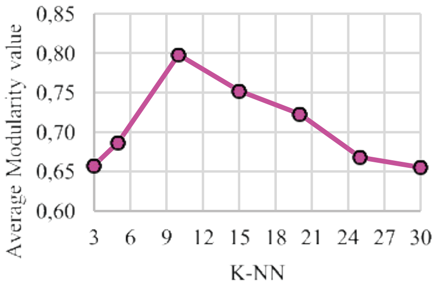

Figure 23. Average Modularity value for different values of $k-NN$ for NFV between 2012-2016.

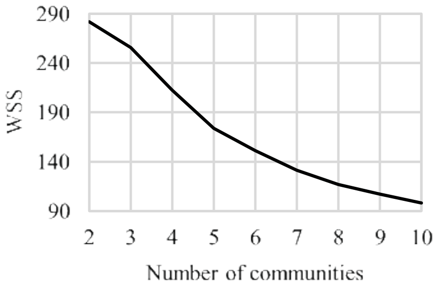

Figure 24. WSS for different values of C with NFV between 2012-2016.

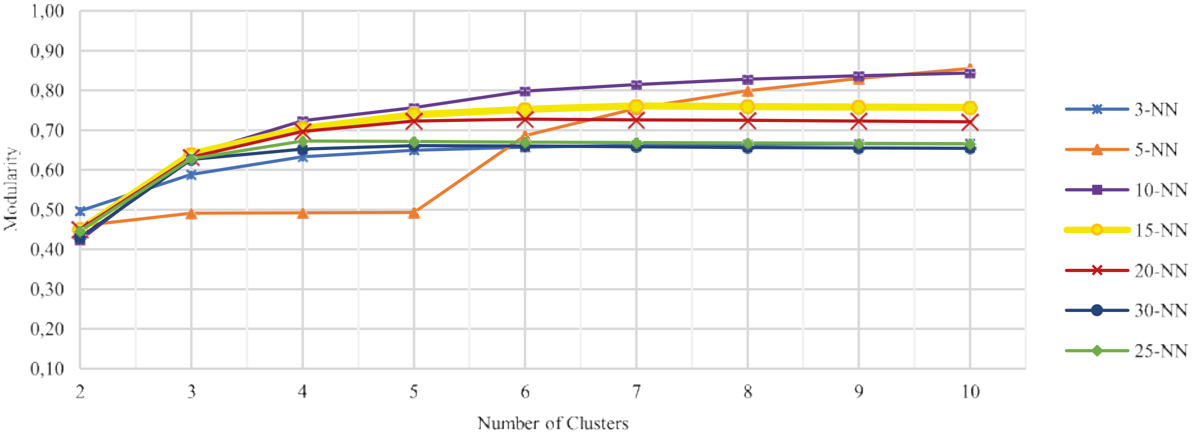

Figure 25 - Construction of complex networks for FV between 2009-2012 with variation of $k-NN$ and number of communities (C).

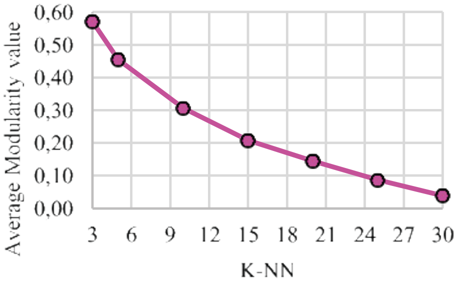

Figure 26 - Average modularity value for different $k-NN$ with FV between 2009-2012.

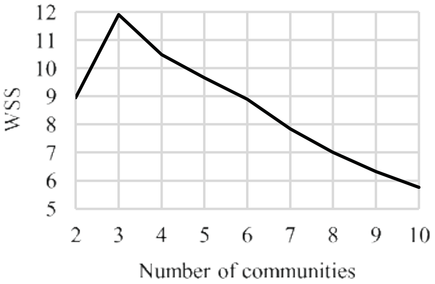

Figure 27 - WSS for different values of C with FV between 2009-2012.

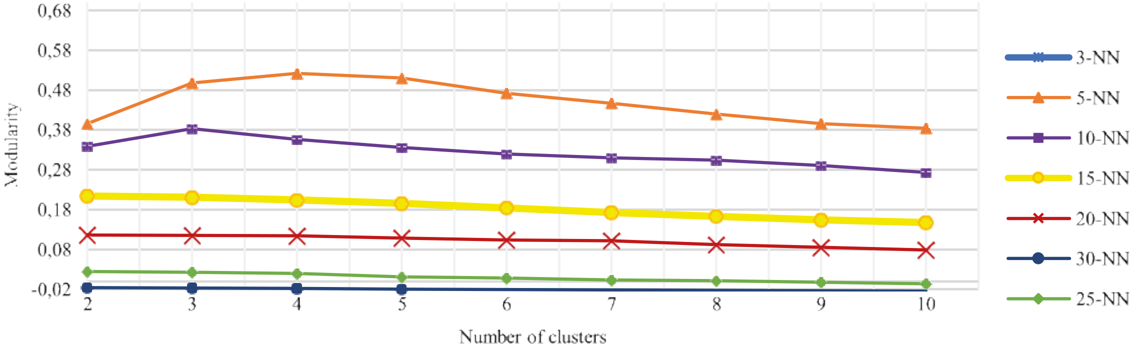

Figure 28 - Construction of complex networks for FV after period 2012-2016 with variation of $k-NN$ and number of communities (C).

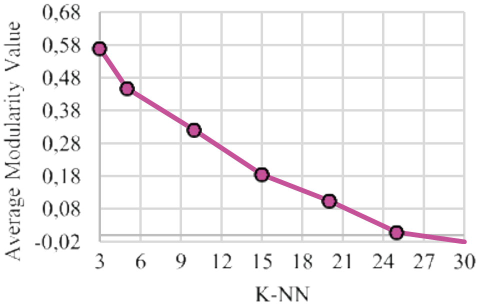

Figure 29 - Average Modularity value for different values of $k-NN$ para FV between 2012-2016.

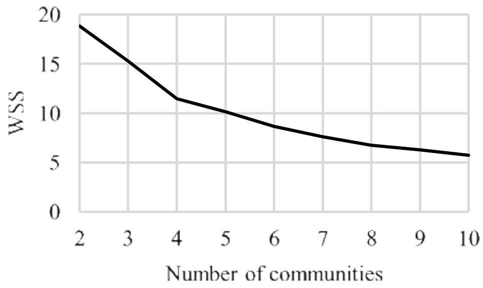

Figure 30 - WSS for different values of C with FV for 2012-2016.

The decision rules applied to each grouping obtained with Fastgreedy in the before setting of the intervention with NFV made it possible to check that the most probable accidents in grouping 1 present the highest probability to occur in parts with barriers or mixed, direction South, and personal use VDM ≤ 14,200. In grouping 2, accidents taking place where there is no barrier and shoulder's width superior to 2.67 m are more probable. In grouping 3, the most probable accidents occur in parts with and without barriers (mixed), direction North, and critical lane's width inferior or equals to 3.45 m.

As for the after setting with NFV, the decision rules obtained from k-means groupings have resulted in accidents more probable in grouping 1 in parts of the road superior to km 133, whose probable cause is the road/environment, or others, direction South. Grouping 2 concentrates those ones that had occurred in parts of the road inferior or equals to km 133. Grouping 3 concentrates accidents that occurred in parts of the road starting from km 133, with probable cause being the driver or the vehicle. Grouping 4 concentrates those accidents that had taken place from km 133, however exclusively direction North of the road, and associated to the probable cause "road/environment", and "others".

Considering the databases of accidents with FV in the before setting, it is verified that Fastgreedy has presented the highest assertiveness rate in extracting rules (100%), and the best accuracy (83%). As for the after setting with FV, k-means has resulted in an assertiveness rate of 100%, and accuracy of 86%.

The decision rules obtained with FV from Fastgreedy for the before period have shown that for the before period Grouping 1 concentrates the accidents that occur between km 125 and km 138, and Groupings 2 and 3 accidents starting from km 138. In the after period, those groupings obtained with FV from k-means suggest that the most probable accidents of Grouping 1 are associated to the probable cause "vehicle" or "road/environment", and that they had occurred between km 125 and km 133+900. Grouping 2 and Grouping 3 are associated to the probable cause "vehicle" or "road/environment", and they had occurred starting from km 133+900. Grouping 4 hosts accidents whose probable cause is the driver. For the after setting, those groupings obtained with *k-means* with FV data revealed that accidents from Grouping 1 had presented higher probability of occurrence between km 125, and km 132+600. Accidents from Grouping 2 had presented higher probability of occurrence starting from km 132+600.

Conclusions

In this paper it was found a significant differences between clustering and complex networks grouping techniques, considering different levels of accident severity, in before and after period of interventions in the road. Also, it was notice that, in general, Fastgreedy method, based on complex networks modeling, presents high level of reliability in extracting decision rules, as well as the greater general accuracy in the process of classification of the observations attributed to each grouping.

From the number of optimal groups obtained from the algorithm with the best performance in detection of groupings (cluster or complex networks), it is possible to extract a reduced number of decision rules, which makes it easier for a general understanding of the database. Even if it is a binary hierarchical method, the decision trees obtained with the groupings, derived from the complex networks approach enabled the association of a greater number of parameters in decision rules when compared to *k-means*, which contributes to simplify or reduce the dimensionality of databases by selecting variables of higher contribution to the groups formation.

Although the algorithm of complex networks presents the best generalization of data to the model, for future works it is recommended to test the approach proposed in this paper for different databases, and with different dimensionalities. Furthermore, it is relevant to explore other networks construction and detection of communities' algorithms, as well as testing different metrics which allow the assessment of performance of these models.

Acknowledgments

This research was partly funded by the State of São Paulo Research Foundation (FAPESP) grants 19/14429-5, 15/50122-0 and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

References

Abugessaisa, I. Knowledge discovery in road accidents database. *International Journal of Public Information Systems*, v. 2008, n. 1, p. 59–85, 2008.

Aqib, M., Mehmood, R., Alzahrani, A., Katib, I., Albeshri, A., & Altowaijri, S. M. Smarter Traffic Prediction Using Big Data, In-Memory Computing, Deep Learning and GPUs. *Sensors* (Basel), Switzerland, 19 (9), 2019.

Berton, L.; Faleiros, T.; Valejo, A.; Valverde-Rebaza, J.; And Lopes, A. A. Rgcli: Robust Graph that Considers Labeled Instances for Semi-Supervised Learning. *Neurocomputing*, (2016).

Chang, L.; Wang, H. Analysis of traffic injury severity: An application of non-parametric classification tree techniques. *Accident Analysis and Prevention*, v. 38, p. 1019–1027, 2006.

Clauset, A., Newman, M. E., Moore, C. Finding community structure in very large networks, *Physical Review E.*, Vol 70, Nº. 6, 2004.

De Oña, J.; López, G.; Mujalli, R.; Calvo, F. J. Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks. *Accident Analysis and Prevention*, v. 51, p. 1–10, 2013. LI, J.;

Facelli, K., Lorena, A. C., Gama, J., Carvalho, A.c.p.l.f. *Inteligência Artificial: Uma abordagem de aprendizado de máquina*. Rio de Janeiro: LTC, 2011. 378p.

Fan, J.; Han, F.; Liu, H. Challenges of Big Data Analysis. *National science review*, 1 (2), 293–314, 2014.

Fong, S.; Wong, R. K.; Mohammed, S.; Fiaidhi, J.; Sung, Y. A suite of swarm dynamic multi-objective algorithms for rebalancing extremely imbalanced datasets. *Applied Soft Computing Journal*, p. 1–22, 2017.

Hair, J. F.; Black, W. C.; Babin, B. J.; Anderson, R. E. *Multivariate data analysis*. New York: Prentice Hall. 2009. 816p.

Halim, Z.; Kalsoom, R.; Bashir, S.; Abbas, G. Artificial intelligence techniques for driving safety and vehicle crash prediction. *Artificial Intelligence Review*, v. 46, n. 3, p. 351–387, 2016.

Lin, L.; Wang, Q.; Sadek, A. W. Data Mining and Complex Networks Algorithms for Traffic Accident. *Transportation Research Record Journal of the Transportation Research Board*, January 2014.

Mehdizadeh, A., Cai, M., Hu, Q., Alamdar Yazdi, M. A., Mohabbati-Kalejahi, N., Vinel, A., Rigdon, S. E.; Davis, K. C.; Megahed, F. M. A Review of Data Analytic Applications in Road Traffic Safety. *Sensores (Basel)*, v. 20 (4), 2020.

Newman, M. E. J. Coauthorship networks and patterns of scientific collaboration. In: *Proceedings of the National Academy of Science of the Uninted States*, v. 101, n. PNAS'04, p. 5200–5205, 2004. ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, v. 20, 53–65, 1987.