IDENTIFICAÇÃO DE *OUTLIERS* EM DADOS DE ACIDENTES DE TRÂNSITO NO BRASIL: ANÁLISE DE *CLUSTER VERSUS* MÉTODOS ESTATÍSTICOS

Philippe Barbosa Silva (Instituto Federal Goiano; Universidade de Brasília) E-mail: philippe.silva@ifgoiano.edu.br
Sara Ferreira (Universidade do Porto, Portugal) E-mail: sara@fe.up.pt
Zafeiris Kokkinogenis (Universidade do Porto) E-mail: zafeiris.kokkinogenis@gmail.com
Michelle Andrade (Universidade de Brasília) E-mail: michelleandrade@unb.br

Resumo: A avaliação e tratamento inicial dos dados é fundamental em qualquer análise de acidentes de trânsito e desenvolvimento de modelos. Um dos aspectos que pode induzir ao enviesamento dos resultados é a não identificação ou tratamento de valores discrepantes, os *outliers*. Neste trabalho investigou-se o potencial do uso de análise de cluster para detecção de *outliers* frente às técnicas estatísticas tradicionalmente utilizada para tal finalidade. Foram utilizados 28.034 registros de acidentes, acumulados de 2011 a 2016 num trecho de 170 km da BR-116/RJ. Dentre as técnicas estatísticas, o método do desvio-padrão apresentou razoabilidade nos resultados, mas dificuldade na aplicação conjunta dos limites univariados de valores discrepantes. Já os métodos Boxplot e MAD se revelaram inadequados para a detecção de *outliers* na base de dados analisada, uma vez que conduziram a resultados incoerentes e sem consistência prática. A análise de cluster (algoritmo *k-means*), doutro lado, demonstrou ter potencial para aplicação a este tipo de problema, tendo identificado conjuntos coerentes de *outliers* para a base de dados. O método não tem rígidas limitações a pressupostos estatísticos, é adequado a grandes bases de dados, permite a avaliação multivariada dos dados e ainda, a análise combinada de dados categóricos e numéricos. Ainda assim, o emprego do método deve ser feito de forma a tirar proveito dos pontos fortes da técnica e minimizar suas limitações.

Palavras-chave: Acidentes de trânsito, *Outliers*, Análise de *cluster*.

OUTLIERS IDENTIFICATION IN TRAFFIC ACCIDENTS DATA IN BRAZIL: CLUSTER ANALYSIS VERSUS STATISTICAL METHODS

Abstract: Initial evaluation and treatment of data is critical in any analysis of crashes and also in the development of models. One of the aspects that can induce the bias of results is the non-identification or treatment of discrepant values, the outliers. In this work, it was investigated the potential of using cluster analysis to detect outliers instead of statistical techniques traditionally used for this purpose. 28,034 accident records from 2011 to 2016 were used in a 170 km segment of BR-116/RJ. Among the statistical techniques, the standard deviation method has presented reasonableness in the results, but difficulty in applying the univariate limits of discrepant values together. The Boxplot and MAD methods have proven to be inadequate for the detection of outliers in the analyzed database, since they led to incoherent results with no practical consistency. The cluster analysis (k-means algorithm), however, have shown potential for application to this type of problem, being able to identify coherent sets of outliers for the database. The method does not have strict limitations to statistical assumptions, is suitable for large databases, allows the multivariate evaluation of data and also the combined analysis of categorical and numerical data. Still, this method should be employed in order to take advantage of the strengths of the technique and minimize its limitations.

Keywords: Traffic accidents, Outliers, Cluster analysis.

1. Introdução

Uma das consequências mais danosas da operação de sistemas de transporte, especialmente o rodoviário, é a ocorrência de acidentes. Em números absolutos, o Brasil figura em terceiro lugar no ranking dos países com maior número de mortes no trânsito. Só no ano de 2014 foram registradas mais de 44 mil vítimas fatais (WHO, 2015; DATASUS, 2018).

Conforme estudo da Organização Mundial da Saúde (OMS), no ano de 2010 ocorreram 1,24 milhão de mortes decorrentes de acidentes de trânsito, além de ter gerado lesão (de vários níveis) em mais de 50 milhões de pessoas. Já em 2013, o número de mortes por acidentes de trânsito foi de 1,25 milhão. Estes acidentes consomem, a nível mundial, 518 bilhões de dólares por ano (OMS, 2010; WHO, 2015). Tal cenário tem feito a segurança dos usuários se constituir por um dos principais objetivos do planejamento e operação do transporte rodoviário.

A segurança viária tem é um dos maiores desafios de gestores e técnicos, o que tem levado à necessidade de investigação e modelagem de acidentes. A existência dos registros dos acidentes é o passo inicial para a análise da acidentalidade. É comum a existência de elevado número de dados de acidentes e de tráfego sem, no entanto, a explicitação de padrões e relacionamento entre os dados da operação e os acidentes de trânsito. De acordo com estudos realizados por Guo *et al*. (2015); Xuensong *et al*. (2014) e Prasad *et al*. (2013), as técnicas de mineração de dados tem grande potencial para a descoberta de relações ocultas em grandes bases de dados, como as bases de acidentes supracitadas.

Anterior a qualquer análise e modelagem de acidentes, é necessário assegurar a qualidade dos dados. Para tanto, deve-se identificar e definir estratégia para tratamento de dados faltantes, dados inconsistentes, dados redundantes, dados duplicados e *outliers*. Esses últimos são especialmente importantes pois podem enviesar e comprometer a análise.

De forma clássica, *outlier* pode ser entendido como uma observação cujos desvios são acentuados em relação aos outros elementos da amostra em que ele ocorre. Tal discrepância dos demais dados leva a suspeitar que este *outlier* foi gerado por um mecanismo diferente, carecendo de investigação (HAWKINS, 1980; GRUBBS, 1969).

Para Barnett e Lewis (1994), uma observação *outlier* é aquela que parece ser inconsistente frente ao restante do conjunto de dados. Isso revela a importância da detecção e tratamento de *outliers*, já que estes podem agregar inconsistência à análise dos dados.

Diversos fatores podem se relacionar à ocorrência de *outliers*, tais como erros humanos, de instrumentos, desvios em populações, comportamento fraudulento ou falhas ou, simplesmente serem pontos *outliers* (BERTON, 2011). Na modelagem de acidentes de trânsito não é diferente, diversos erros na coleta e anotação de dados podem gerar *outliers*, além de existirem características naturais que, frente ao resto do conjunto de dados, podem ser considerados *outliers* (SUN *et al.*, 2017).

Chandola *et al*. (2009) descrevem sete técnicas de detecção de *outliers*, cada uma caracterizada por diferentes abordagens. O presente estudo foi desenvolvido a partir do uso de técnicas estatística e técnicas baseadas em agrupamento. Tal escolha é justificada por serem as mais comumente utilizadas na literatura e por conduzirem a resultados razoáveis.

2. Materiais e Métodos

2.1. Dados

Os dados de acidentes foram obtidos junto à Agência Nacional de Transportes Terrestres (ANTT), sendo relativos a um trecho de 170 km da BR-116 (rodovia de pista dupla), entre o km 333 e km 163 no estado do Rio de Janeiro.

O horizonte de análise foi de 6 anos, compreendido entre 2011 e 2016. O número total de acidentes registrados no período em análise foi de 28.034, dos quais 5.361 ocorreram em 2011, 4.646 em 2012, 4.920 em 2013, 5.051 em 2014, 4.235 em 2015 e 3.821 em 2016. A base de

dados é mista (contém variáveis categóricas e numéricas), em que cada acidente é caracterizado segundo 16 atributos, conforme apresentado nas Tabela 1 e 2.

2.2. Detecção de *outliers*

Foram utilizados neste trabalho os principais (e de aplicação mais simplificada) métodos estatísticos de detecção de *outliers* e a análise de *cluster*. O esquema metodológico geral está apresentado na Figura 1.

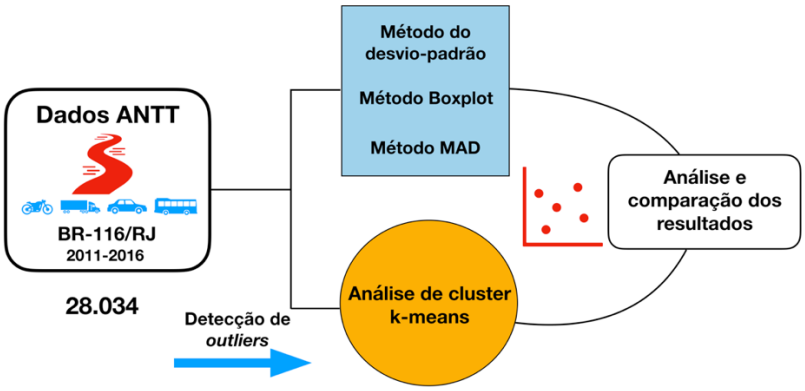

Figura 1 – Procedimento metodológico geral da pesquisa

2.2.1. Técnicas estatísticas

As técnicas adotadas consistem basicamente em gerar um intervalo ou critério para detecção de *outlier* a partir das hipóteses de teste, e todas as observações se estejam fora do intervalo ou critério são consideradas *outliers*.

- **Método do desvio-padrão**

Conforme Howell (1998), o método do desvio-padrão consiste em estabelecer um intervalo de valores, a partir da média, tendo como parâmetro o desvio-padrão. Desta forma, os valores que se encontrarem fora desse intervalo são considerados *outliers*. Neste estudo, para estabelecimento do intervalo se considera a adição/subtração de 3 vezes o desvio-padrão a partir da média, conforme Equação 1.

$$Intervalo = \overline{x} \pm 3.s \qquad (1)$$

Onde: $\overline{x}$ é a média;

s é o desvio-padrão.

- **Método *Boxplot***

Este método consiste na construção do gráfico *Boxplot* para facilitar na identificação dos *outliers*. Foi proposto por Turkey (1977) e leva em consideração a mediana, os valores extremos, o quartil superior e o quartil inferior, conforme Equação 2. Assim, todos os valores fora do intervalo estabelecido são considerados *outliers*.

$$Q1 - 1{,}5.IQR \leq Intervalo \leq Q3 + 1{,}5.IQR \qquad (2)$$

Onde: $Q1$ é o quartil inferior;

$Q3$ é o quartil superior;

IQR é o intervalo interquartil.

- **Método MAD**

Popularizado por Hampel (1974), o também denominado Método de *z-score* modificado, utiliza o desvio absoluto da mediana (MAD) como parâmetro para identificação de *outliers*. Neste caso, a mediana, assim como a média nos outros métodos, é a medida de tendência central, mas com a vantagem de ser menos sensível e influenciada pelos *outliers*. Assim, o intervalo fica definindo segundo a Equação 3. Assim como nos métodos anteriores, todos os valores fora do intervalo estabelecido são considerados *outliers*.

$$M - 3.MAD \leq Intervalo \leq M + 3.MAD \quad (3)$$

Onde: M é a mediana;
MAD é o desvio médio absoluto.

2.2.2. Análise de *cluster*

A análise de *cluster* é um método de mineração de dados que objetiva segmentar os elementos dos dados em grupos (*clusters*), a partir da similaridade e dissimilaridade dos dados. Desta forma, tanto a homogeneidade dentro dos *clusters* quanto a heterogeneidade entre *clusters* são maximizadas (FRALEY e RAFTERY, 2002; HAIR *et al.*, 1998). Tal capacidade dos algoritmos de análise de *cluster* também são úteis para detecção de *outliers*.

No presente estudo foi utilizado o algoritmo *k-means* para análise. Este método basicamente mede a proximidade dos grupos usando a distância euclidiana entre os centroides dos grupos. Trata-se de um método de partição em que, inicialmente seleciona-se a partição inicial dos *n* objetos em *k clusters*; calcula-se os centroides para cada um dos *k clusters*; agrupa-se os objetos aos *clusters* cujos centroides se encontram mais próximos, atualizando-se o cálculo dos centroides dos *clusters* até que não ocorra variação significativa na distância mínima de cada objeto da base de dados a cada um dos centroides dos *k clusters* (JOHNSON e WICHERN, 2002).

A escolha do *k-means* é justificada por este algoritmo ser bastante adequado para o processo de análise de *cluster* de grandes bases de dados e por possuir reconhecido desempenho positivo nas análises (MACQUEEN, 1967).

3. Resultados e Discussão

3.1. Estatística descritiva dos dados

A estatística descritiva dos dados, passo inicial da análise, está apresentada na Tabela 1 para as variáveis categóricas e na Tabela 2 para as variáveis numéricas.

Tabela 1 – Distribuição de frequência das variáveis categóricas

Variável	Valores/Frequência
Data	Dia da semana - 19.443 (*69,36%*) / Fim de semana - 8.591 (*30,64%*)
Horário	Amanhecer - 2.613 (*9,32%*) / Pleno Dia - 13.074 (*46,64%*) / Anoitecer - 5.254 (*18,74%*) / Plena Noite - 7.093 (*25,30%*)
Tipo de ocorrência	Acidente sem vítima - 19.743 (*70,42%*) / Acidente com vítima - 7.824 (*27,91%*) / Acidente com morte - 467 (*1,67%*)
Tipo de acidente agrupado	Atropelamento - 762 (*2,72%*) / Colisão - 23.637 (*84,31%*) / Capotamento/Tombamento - 2.792 (*9,96%*) / Saída de pista - 443 (*1,58%*) / Outros - 400 (*1,43%*)
Sentido	Pista Norte - 14.801 (*52,80%*) / Pista Sul - 13.233 (*47,20%*)
Tipo de trecho	Curva acentuada - 5.339 (*19,04%*) / Curva suave - 2.547 (*9,09%*) / Reta - 20.148 (*71,87%*)
Altimetria	Aclive - 3.187 (*11,37%*) / Declive - 4.179 (*14,91%*) / Em nível - 20.668 (*73,72%*)

Tabela 2 – Estatística descritiva das variáveis numéricas

Variável	Média	Mín.	Máx.	Mediana	Desvio-padrão	Quartil inferior (Q1)	Quartil superior (Q3)
NVL	1,0971	0	12	1	0,89678	1	1
NVP	0,4340	0	5	0	0,64781	0	1
NVDR	0,1154	0	4	0	0,33349	0	0
NOV	0,0995	0	5	0	0,31752	0	0
NVI	3,3809	0	197	2	7,63335	1	3
NVF	0,3895	0	35	0	0,89816	0	1
NVFG	0,0227	0	4	0	0,17026	0	0
NVFAT	0,0180	0	4	0	0,1464	0	0

NVL - Nº de veículos leves**; NVP -** Nº de veículos pesados**; NVDR -** Nº de veículos de duas rodas**; NOV -** Nº de outros veículos**; NVI -** Nº de vítimas ilesas**; NVF -** Nº de vítimas feridas**; NVFG -** Nº de vítimas feridas gravemente**; NVFAT -** Nº de vítimas fatais

3.2. Detecção de *outliers* por meio de técnicas estatísticas

Realizada a estatística descritiva dos dados (Tabelas 1 e 2), passa-se para a os métodos de definição intervalar para identificação dos *outliers*.

3.2.1. Método do desvio-padrão

Mediante aplicação da Equação 1, foram obtidos os resultados dispostos na Tabela 3.

Tabela 3 – Resultados da aplicação do método do desvio-padrão

Variável	Média	Desvio-padrão	Limite inferior		Limite superior	
NVL	1,0971	0,89678	-1,14485	***(0)***	3,78744	***(4)***
NVP	0,4340	0,64781	-1,185525	***(0)***	2,37743	***(3)***
NVDR	0,1154	0,33349	-0,718325	***(0)***	1,11587	***(2)***
NOV	0,0995	0,31752	-0,6943	***(0)***	1,05206	***(2)***
NVI	3,3809	7,63335	-15,702475	***(0)***	26,28095	***(27)***
NVF	0,3895	0,89816	-1,8559	***(0)***	3,08398	***(4)***
NVFG	0,0227	0,17026	-0,40295	***(0)***	0,53348	***(1)***
NVFAT	0,0180	0,1464	-0,348	***(0)***	0,4572	***(1)***

O valor entre parênteses nas colunas "Limite inferior" e "Limite superior" correspondem aos valores finais do intervalo de valores típicos de cada variável. Note-se que, como no evento em análise, o acidente de trânsito, é um número sempre inteiro e positivo, o arredondamento resulta na configuração apresentada na Tabela 3.

Como é notório, as técnicas estatísticas utilizadas são para análise univariada de *outliers*, implicando na necessidade de associação dos resultados para identificação e tratamento de todos os valores discrepantes.

Iniciando pelo "Nº de veículos leves", tem-se que os valores compreendidos entre 0 e 4 são considerados típicos, ao passo que 5 ou mais veículos leves envolvidos são considerados potenciais *outliers*. Aplicando o intervalo para essa variável, são identificados 166 valores discrepantes.

Mediante aplicação do limite correspondente à variável "Nº de veículos pesados" são identificados 15 *outliers*, 6 valores discrepantes com o limite da variável "Nº de veículos de duas rodas", 14 *outliers* com o limite da variável "Nº de outros veículos", 638 *outliers* com o limite da variável "Nº de vítimas ilesas", 130 *outliers* com o limite da variável "Nº de vítimas

feridas", 54 *outliers* com o limite da variável "Nº de vítimas feridas gravemente" e 29 valores discrepantes com o limite da variável "Nº de vítimas fatais".

Além dos limites resultantes da análise de valores de cada variável, tem-se a execução combinada dos limites de valores típicos, chegando-se ao conjunto final de *outliers*, já excluídas as sobreposições. Aplicando todos os limites obtidos pelo método do desvio-padrão, são identificados 1.003 *outliers* em toda a base de dados analisada (3,6% da amostra).

3.2.2. Método *Boxplot*

Na Tabela 4 estão apresentados os resultados do método *Boxplot*, após aplicação da Equação 2.

Tabela 4 – Resultados da aplicação do método *Boxplot*

Variável	Mín.	Máx.	Q1	Q3	IQR	Limite inferior	Limite superior
NVL	0	12	1	1	0	1 ***(1)***	1 ***(1)***
NVP	0	5	0	1	1	-1,5 ***(0)***	2,5 ***(3)***
NVDR	0	4	0	0	0	0 ***(0)***	0 ***(0)***
NOV	0	5	0	0	0	0 ***(0)***	0 ***(0)***
NVI	0	197	1	3	2	-2 ***(0)***	6 ***(6)***
NVF	0	35	0	1	1	-1,5 ***(0)***	2,5 ***(3)***
NVFG	0	4	0	0	0	0 ***(0)***	0 ***(0)***
NVFAT	0	4	0	0	0	0 ***(0)***	0 ***(0)***

Nota-se que este método gera resultados muito mais conservadores que o anterior e ainda, conduz à limites que admitem apenas um valor. Isso é explicado pelo mecanismo de funcionamento do método, que é baseado na distância interquartil. Contudo, o método conduz a resultados pouco razoáveis para a análise do problema real. De toda forma, são apresentados, na sequência, os valores de *outliers* identificados mediante aplicação dos limites obtidos.

Na ordem apresentada na Tabela 4, mediante aplicação de cada limite, tem-se os seguintes valores discrepantes: 12.491, 15, 3.115, 2.658, 1.915, 253, 563 e 466. Já de forma combinada, o conjunto total de *outliers* é de 15.987 valores, o que equivale a quase 60% da base de dados inicial.

3.2.3. Método MAD

Esse método baseia-se no desvio absoluto a partir da mediana para estabelecer o intervalo de valores típicos. Dada tal condição, as variáveis que apresentarem mediana igual a zero terão um intervalo nulo, como é o caso de todas as variáveis, à exceção do "Nº de veículos leves" e "Nº de vítimas ilesas". A primeira delas, embora apresente mediana igual a um, tem MAD igual a zero, também gerando intervalo nulo. Por fim, o "Nº de vítimas ilesas" resultou no intervalo de -3,45 (0) a 5,45 (6).

Assim como o método *Boxplot*, os resultados gerados para o conjunto de dados em questão, embora amparados em premissas estatísticas, não conduzem a resultados razoáveis e consistentes para o problema. De todo modo, são apresentados os valores de *outliers* identificados por essa abordagem.

Na ordem apresentada na Tabela 4, mediante aplicação de cada limite, tem-se os seguintes valores discrepantes pelo método MAD: 21.870, 9.916, 3115, 2.658, 1.915, 7.552, 563 e 466. Já de forma combinada, todo o conjunto de dados foi identificado como *outlier*.

3.2.4. Discussão

O método do desvio-padrão, provavelmente o mais utilizado e simples método de detecção de *outliers*, conduziu a resultados aceitáveis. Ele foi capaz de separar apenas a massa de dados que tinha frequência muito reduzida ou rara, sem reduzir substancialmente o tamanho da base de dados original, importante aspecto no processo de modelagem.

Os outros dois métodos, por outro lado, foram ineficazes na determinação dos valores discrepantes. O método *Boxplot* resultou em quatro intervalos nulos de valores típicos, um intervalo unitário, dois intervalos 0-3 e um intervalo 0-6. De forma prática, tais critérios excluem todos os acidentes que envolveram, pelo menos, um veículo de duas rodas (bicicleta ou motocicleta) ou outro tipo de veículo. Ou ainda, todos os acidentes em que, pelo menos, uma pessoa se feriu gravemente ou faleceu. Além disso, apenas os acidentes envolvendo um, e somente um, veículo leve foram considerados como típicos.

O método MAD apresentou resultados inaceitáveis, uma vez que todos os intervalos foram nulos, exceto o intervalo para "Nº de vítimas ilesas". O resultado, quando aplicados todos os limites univariados de forma combinada, é totalmente inadmissível e descabido, já que considerou todos os dados como *outliers*. Ressalta-se ainda que, mesmo com a aplicação de apenas um intervalo univariado, tem-se, na maioria das variáveis, redução significativa e desproporcional da base de dados original, mediante exclusão dos *outliers* identificados.

O fraco desempenho dos métodos pode ser devido à violação do pressuposto de normalidade de distribuição dos dados. Foi realizado o teste *Kolmogorov-Smirnov* para todas as variáveis, tendo sido refutada, em todos os casos, a hipótese de normalidade dos dados, com nível de significância de 0,05. Ainda assim, é importante referir que, mesmo com a normalização dos dados, caso do teste *Z-score* e MAD, não é garantida boa capacidade de detecção de *outliers* para grandes bases de dados. Para o caso do teste *Z-score*, por exemplo, em amostras com mais de 1.000 dados, valores de *Z-score* inferiores a -3,3 ou superiores a 3,3 que, tipicamente indicam *outliers* em pequenas bases de dados, não podem ser utilizados como parâmetro para identificação de valores discrepantes.

Devido às características dos dados de registros de acidentes, eles possivelmente violam a maioria das premissas de técnicas estatísticas de detecção de *outliers*, como foi observado neste caso. Diante disso, é de interesse explorar outras alternativas que se adequem melhor aos dados, como a análise de *cluster*, abordada neste estudo.

3.3. Detecção de *outliers* com recurso à análise de *cluster*

Como trata-se de uma base de dados mistos (categóricos e numéricos), o passo inicial que precedeu a análise de *cluster* foi a criação de uma variável *dummy*, com valores binários, para cada opção das variáveis categóricas. Isso permitiu que o algoritmo *k-means* reconhecesse todas as variáveis e lidasse com todas de forma conjunta.

O algoritmo foi inicializado com o mínimo de 2 e máximo de 20 *clusters* a serem gerados. Diante das diversas configurações geradas, alguma medida de desempenho deve ser elegida para seleção da configuração final mais adequada. Para tanto, foi utilizado a Largura Média da *Silhouette* (*Average Silhouette Width*, em inglês) proposta por Rousseeuw (1987). Esse indicador descreve quão bem cada objeto de um conjunto de dados se encaixa no *cluster* em que foi atribuído. O procedimento de cálculo baseia-se nas Equações 4 e 5.

$$ASW = \frac{1}{N}\sum_{i=1}^{N} S \qquad (4)$$

$$S = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \qquad (5)$$

Onde: $b(i)$ é a distância mínima do dado *i* a todos os demais dados que não pertencem ao seu grupo;
$a(i)$ é a distância média do dado *i* a todos os demais dados do seu grupo;
S é a largura média de *silhouette* de cada dado;
N é o número de dados do agrupamento;
ASW é a largura média da *silhouette* do agrupamento.

Os valores possíveis estão compreendidos entre -1 e 1. Valores menores ou iguais a 0,25 indicam que estrutura de agrupamento gerada não é consistente. Valores entre 0,26 e 0,50 indicam uma estrutura fraca de agrupamento; de 0,51 a 0,70 uma estrutura razoável e de 0,71 a 1,00 uma estrutura forte e consistente de clusterização.

As estruturas com 2, 3 e 4 *clusters* apresentaram *ASW* superior a 0,95, indicando a descoberta de uma estrutura consistente de agrupamento. A partir de 5 *clusters*, houve uma redução significativa dos valores de *ASW*, passando a patamares inferiores a 0,30. Diante disso, a estrutura escolhida como mais adequada foi de 4 *clusters*, uma vez que apresentou *ASW* muito próximo à 1 e ainda, pelo número de grupos formados, permitiria a descoberta de padrões e extração de conhecimento.

Devido a grande variabilidade do "Nº de vítimas ilesas" e, considerando o procedimento de cálculo de distâncias aos centroides de cada *cluster*, essa variável condicionou a clusterização executada pelo algoritmo *k-means*. Os resultados da divisão dos grupos estão apresentados na Tabela 5.

Os dados contidos nos *clusters* 1, 2 e 3 podem ser considerados como *outliers*, totalizando 1.102 dados. Tais *clusters* apresentam valores atípicos de número de vítimas ilesas, além de que, o alto valor de ASW indica a forte associação entre os valores atribuídos ao *cluster* 0.

Na estrutura de 2 e 3 *clusters*, o *cluster* 0 se manteve inalterado, apresentando 26.932 dados. Tal fato significa que este é o *cluster* principal, que contém todos os dados típicos, ao passo que, quanto mais *clusters* são gerados, tem-se apenas a divisão dos valores discrepantes em mais *clusters*.

Também foi possível verificar a associação do "Nº de veículos leves" e "Nº de vítimas ilesas", produzindo os grupos gerados. A Figura 2 ilustra esse relacionamento.

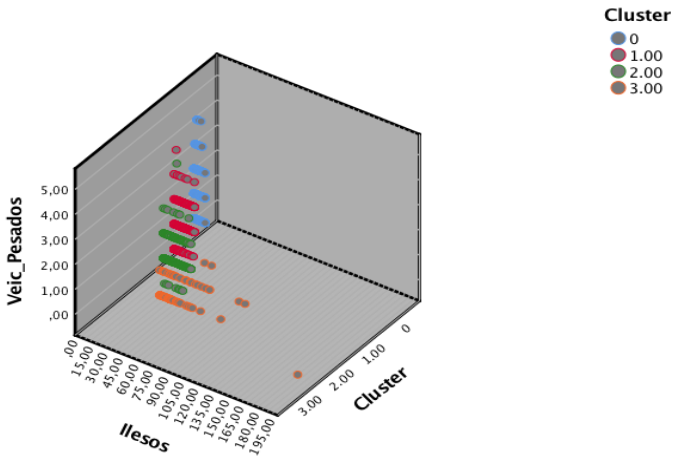

Figura 2 – *Clusters* gerados e relação com "Nº de veículos leves" e "Nº de vítimas ilesas"

Esse comportamento é razoável, já que acidentes envolvendo veículos pesados, especialmente ônibus, implicam, na maioria dos casos, em grande número de pessoas envolvidas. Ademais,

dos envolvidos tem-se reduzido número de gravemente feridos ou mortos, apresentando, por outro lado muitos ilesos.

Tabela 5 – Resultados da análise de *cluster* – base de dados original

Cluster	**Número de dados**	**Intervalo de "Nº de vítimas ilesas"**
0	26.932	0 a 11
1	576	12 a 32
2	427	33 a 60
3	99	61 a 197

Buscando investigar capacidade de detecção dos *outliers* em subconjunto de dados, procedeu-se separação dos dados por tipo de acidente, sendo gerados 4 subconjuntos: Atropelamento, Colisão, Capotamento/Tombamento e Saída de pista. Note-se que, o subconjunto "Outros", contendo 400 ocorrências não foi considerado. Em todos os casos, a estrutura com 2 e 3 *clusters* apresentaram os melhores resultados (maiores valores de *ASW*), pelo que são apresentados os resultados destes agrupamentos na Tabela 6.

Tabela 6 – Resultados da análise de *cluster* – subconjuntos de dados

Estrutura		**Atropelamento (762)**	**Colisão (23.637)**	**Capotamento/Tombamento (2.792)**	**Saída de pista (443)**
***k* = 2**	0	743	22.732	2.783	442
	1	19	905	9	1
***k* = 3**	0	728	22.732	1.757	240
	1	19	298	9	202
	2	15	607	1.026	1

Com raciocínio análogo, tem-se que o *cluster* 0 contém todos os valores típicos do conjunto de dados. Desta forma, para $k = 2$, o total de *outliers* é 934 e para $k = 3$, tem-se 2.177. Na estrutura com apenas 2 *clusters*, tem-se o número de *outliers* similar ao encontrado na análise do conjunto único de dados, entretanto, para $k = 3$, o número de valores discrepantes é quase dobro do identificado no conjunto único. Embora os maiores valores de *ASW* correspondam à estrutura de $k = 2$, é interessante investigar o motivo da diferença na identificação dos *outliers*, sendo apresentado na Tabela 7 a média de cada variável dos agrupamentos com $k = 3$.

Tabela 7 – Média das varáveis, por subconjunto e para cada *cluster*

Cluster	Veículos leves	Veículos pesados	Veículos de duas rodas	Outros	Ilesos	Feridos	Feridos graves	Mortos
Atropelamento								
0	0,49	0,13	0.36	0.30	1.23	0.69	0.19	0.29
1	0,00	1,00	0.47	0.00	44.79	0.79	0.16	0.32
2	0,47	0,93	0.20	0.00	14.20	0.53	0.27	0.27
Colisão								
0	1,22	0,43	0.07	0.11	2.39	0.31	0.01	0.01
1	0,82	1,30	0.06	0.08	58.47	0.69	0.02	0.01
2	1,02	1,24	0.03	0.07	28.48	0.66	0.03	0.01
Capotamento/Tombamento								

0	0,34	0,17	0.57	0.03	0.23	1.30	0.06	0.02
1	0,44	1,44	0.22	0.22	34.67	3.89	0.33	0.22
2	0,65	0,31	0.22	0.03	1.60	0.14	0.02	0.02
Saída de pista								
0	0,85	0,17	0.01	0.02	1.53	0.05	0.01	0.00
1	0,79	0,22	0.02	0.01	0.16	1.37	0.14	0.08
2	0,00	1,00	0.00	0.00	39.00	5.00	0.00	0.00

É perceptível a influência condicionante do número de ilesos na divisão dos grupos. Isso pode ser notado pelo fato de, apenas essa variável apresentar diferentes valores médios entre todos os *clusters*, enquanto outras variáveis, em diferentes *clusters*, apresentarem valores médios muito similares. Para exemplificar, tomemos a comparação entre o *cluster* 1 e 2 do subconjunto "Colisão": os valores médios, em termos absolutos, para todas as variáveis, não diferem mais que 0,06; a exceção é número de ilesos, que apresenta diferença de mais de 30 entre os dois *clusters*. Nitidamente, esta foi a condição para criação de um novo grupo, agregando as ocorrências com número de ilesos similares.

Assim, é suposto que o maior número total de *outliers* identificados por meio da análise de *cluster* ($k = 3$) dos subconjuntos, é resultado do melhor ajuste do algoritmo a cada subconjunto, especialmente aos três menores subconjuntos. A exemplo, destaca-se o subconjunto "Colisão" que resulta no número de ocorrências discrepantes compatível com o obtido na análise considerando toda a base de dados. O subconjunto, por representar cerca de 85% dos registros da base de dados original e, portanto, possuir a maior quantidade de valores extremos de todas as variáveis, é pouco sensível. Já no caso dos subconjuntos menores, por se tratarem de tipos mais singulares de acidentes, implicam em maior semelhança dos dados, o que faz com que o algoritmo lide melhor com as dissimilaridades e crie um número maior de *clusters*, identificando melhor os valores discrepantes.

4. Conclusões

O objetivo deste trabalho foi avaliar, por meio de dados reais de acidente de trânsito, o potencial de detecção de *outliers* por técnicas estatísticas e por análise de *cluster*. A identificação e tratamento de valores discrepantes é essencial para reduzir erros e ruídos no processo de modelagem da segurança viária, essencialmente baseada nos dados de acidentes.

Para tanto, optou-se pela utilização dos três métodos estatísticos mais comumente utilizados: método do desvio-padrão, método *Boxplot* e método MAD. A maior limitação de tais métodos é a necessidade de os dados apresentarem distribuição normal. Os dados reais analisados não cumprem essa condição, o que explica o fraco desempenho nos últimos dois métodos. Um, por se basear nos quartis inferior e superior, e outro na mediana, apresentaram intervalos – de valores típicos – nulos ou unitários, na maioria das variáveis. Limites como os obtidos não fazem sentido para a análise da segurança viária, uma vez que restringem os dados à tipos específicos de ocorrências, além da drástica redução no volume de dados. Também deve-se referir que esses métodos não lidam bem com grandes bases de dados, geralmente superiores a 1.000 observações.

No caso do método do desvio-padrão, os resultados foram considerados razoáveis, embora o conjunto de dados não apresente distribuição normal. Tal fato carece atenção, uma vez que a principal premissa do método foi violada. Ademais, isso ainda demanda cautela no emprego do método, já pode conduzir a resultados pouco razoáveis em outras bases de dados.

Outra limitação desses métodos estatísticos é a análise univariada, requerendo análise adicional de como aplicar os intervalos de valores típicos. Em alguns casos isso pode ser intuitivo e fácil, no entanto, existem situações em que essa é uma tarefa difícil e que, se executada de forma equivocada, pode conduzir a viés nos resultados. Neste trabalho, apresentou-se os limites e *outliers* identificados para cada variável e procedeu-se a utilização combinada de todos os limites, garantindo que todos os dados que estivessem fora de algum dos intervalos estabelecidos fossem considerados valores discrepantes. Isso, no entanto, levou a resultados inaceitáveis, como para os limites obtidos pela aplicação do método MAD. Naquela situação, após aplicação combinada de todas as restrições, todos os registros da base de dados foram considerados *outliers*.

Dada a natureza dos dados de acidentes de trânsito, com alto potencial de não normalidade e, necessidade de análise multivariada de *outliers*, investigou-se o potencial da análise de *cluster* para identificação de *outliers*. Além disso, por meio da análise de *cluster*, é a possível analisar, de forma conjunta, dados categóricos e numéricos.

Para isso, recorreu-se ao algoritmo *k-means*, mais conhecido e indicado para grandes bases de dados, identificando-se como melhor estrutura de agrupamento, 4 *clusters*. Tal configuração teve indicador *ASW* superior a 0,95, indicando ser uma estrutura de agrupamento forte e consistente. O resultado foi, basicamente, de um *cluster* com cerca de 96% dos dados e outros três pequenos *clusters* contendo valores atípicos.

Notou-se a alta dependência da clusterização aos valores da variável "Nº de vítimas ilesas", que contém a maior variabilidade. Isso se deve ao mecanismo de cálculo baseado no cálculo de distâncias até o centroide de cada *cluster*. O total de *outliers* identificado foi, em números absolutos, similar ao identificado pelo método do desvio-padrão. Não se tratam, entretanto, dos mesmos dados, uma vez que no método estatístico fez-se a aplicação combinada de todos os limites gerados ao passo que, na análise de *cluster*, os resultados já levavam em conta a atuação conjunta de todas as variáveis.

Foi avaliado, ainda, o comportamento do algoritmo aplicado a subconjuntos de dados. Para tanto, fez-se a divisão da base original em 4 subconjuntos de dados, por tipo de acidente. As melhores configurações de agrupamento, foram com 2 e 3 *clusters* para todos os subconjuntos. Os resultados revelaram mais uma vez a dependência dos agrupamentos aos valores de ilesos envolvidos. No maior subconjunto (Colisão), os resultados obtidos foram similares aos encontrados na análise da base original. Já nos menores subconjuntos (Atropelamento, Capotamento/Tombamento e Saída de pista) foi possível uma melhor identificação das dissimilaridades e divisão de grupos, o que se deu pelo fato desses subconjuntos reunirem tipos mais específicos de acidentes e por, supostamente, apresentarem maior semelhança dos dados.

Dessa forma a análise de *cluster* apresenta-se como uma alternativa para a detecção de *outliers* em dados de acidentes. Vale ressaltar que o método não tem rígidas limitações a pressupostos estatísticos, é adequado a grandes bases de dados, permite a avaliação multivariada dos dados e ainda a análise combinada de dados categóricos e numéricos.

Ainda assim, o emprego do método deve ser feito de forma a tirar proveito dos pontos fortes da técnica e minimizar suas limitações. A exemplo, no caso estudado, ficou nítido que a permanência da variável "Nº de vítimas ilesas" na base de dados, influenciaria fortemente a clusterização, mesmo em subconjuntos. Diante disso, observa-se que os resultados da análise de *cluster* são adequados, inclusive, para esse propósito: nortear a redução de dimensionalidade dos dados. A clusterização, é útil, portanto, como passo inicial da exploração dos dados, passível de conduzir à resultados consistentes e robustos de divisão de grupos e identificação segmentada de *outliers*.

Agradecimentos

O primeiro autor agradece ao apoio financeiro do Instituto Federal Goiano (IFGoiano). Refere-se ainda que o presente trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001 e da Fundação para a Ciência e Tecnologia- Portugal - (FCT) por meio do projeto "Mobilidade Urbana Sustentável e Segura", no qual este trabalho está inserido. Os autores também agradecem à ANTT (Agência Nacional de Transportes Terrestres) pela disponibilização dos dados.

Referências

BARNETT, V.; LEWIS, T. *Outliers in Statistical Data, Wiley Series in Probability and Mathematical Statistics. v. 3. John Wiley & Sons, 1994.*

BERTON, L. *Caracterização de classes e detecção de outliers em redes complexas. Dissertação de Mestrado. Programa de Pós-Graduação em Ciências de Computação e Matemática Aplicada. Universidade de São Paulo. São Carlos, 2011.*

CHANDOLA, V; BANERJEE, A.; KUMAR, V. *Outlier detection-A survey. ACM Computing Surveys, 2009.*

DATASUS. *Departamento de Informática do SUS. Disponível em: http://datasus.saude.gov.br/. Acesso em 23 jul. 2018.*

FRALEY, C.; RAFTERY, A. E. *Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association, v. 97, n. 458, p. 611-631, 2002.*

GUO, J.; HUANG, W.; WILLIAMS, B. M. *Real time traffic flow outlier detection using short-term traffic conditional variance prediction. // Transportation Research Part C: Emerging Technologies, n. 50, p. 160–172, 2015.*

GRUBBS, F. E. *Procedures for detecting outlying observations in samples. Technometrics, v. 11, n. 1, p. 1–21, 1969.*

HAIR, J. F; ANDERSON, T. E.; TATHAM, R. L.; BLACK, W.C. *Multivariate data analysis, 5 ed., Prentice-Hall, New Jersey, 1998.*

HAMPEL, F. R. *The influence curve and its role in robust estimation. Journal of the American Statistical Association, v. 69, n. 346, p. 383–393, 1974.*

HAWKINS, D. *Identification of Outliers. London: Chapman and Hall, 1980.*

HOWELL, D. C. *Statistical methods in human sciences. New York: Wadsworth, 1998.*

JOHNSON, R. A.; WICHERN, D. W. *Applied Multivariate Statistical Analysis, 5 ed. New Jersey: Prentice Hall, 2002.*

MACQUEEN, J. B. *Some methods for classification and analysis of multivariate observations. In: CAM, L. M. L.; NEYMAN, J. (Ed.). Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability. [S.l.]: University of California Press, v. 1, p. 281–297, 1967.*

OMS. *Organización Mundial de la Salud. Informe sobre la situación mundial de la seguridad vial: es hora de pasar a la acción. Genebra. 287 p., 2010.*

PRASAD, N.; KUMAR, P.; NAIDU, M. M. *An Approach to Prediction of Precipitation Using Gini Index in SLIQ Decision Tree. In: 4th International Conference on Intelligent Systems, Modelling and Simulation. Bangkok, 2013.*

ROUSSEEUW, P. J.; LEROY, A. M. *Robust Regression and Outlier Detection. John Wiley: 1987.*

SUN, B.; CHENG, W.; BAI, G.; GOSWAMI, P. *CORRECTING AND COMPLEMENTING FREEWAY TRAFFIC ACCIDENT DATA USING MAHALANOBIS DISTANCE BASED OUTLIER DETECTION. Technical Gazette, n. 24, v. 5, p. 1597-1607, 2017.*

TURKEY, J. *Exploratory Data Analysis. Addison-Wesley: 1977.*

WHO. *World Health Organization. GLOBAL STATUS REPORT ON ROAD SAFETY 2015. Genebra. 323 p., 2015.*

XUESONG, W.; QIANG, G.; SHANSHAN, L.; RONGFEI, C. *Design and Implementation of School Hospital Information Analysis and Mining System. Applied Science, Materials Science and Information Technologies in Industry, n. 513, p. 498–501, 2014.*