UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

Miriam Karla Rocha

ABORDAGENS MULTIVARIADAS APLICADAS EM DADOS DE SISTEMAS DE TRANSPORTES

Porto Alegre

2020

Miriam Karla Rocha

ABORDAGENS MULTIVARIADAS APLICADAS EM DADOS DE SISTEMAS DE TRANSPORTES

Tese submetida ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal do Rio Grande do Sul como requisito parcial à obtenção do título de Doutora em Engenharia, na área de concentração em Sistemas de Transportes.

Orientador: Michel José Anzanello, *Ph.*D.

Porto Alegre

2020

Miriam Karla Rocha

ABORDAGENS MULTIVARIADAS APLICADAS EM DADOS DE SISTEMAS DE TRANSPORTES

Esta tese foi julgada adequada para a obtenção do título de Doutora em Engenharia e aprovada em sua forma final pelo Orientador e pela Banca Examinadora designada pelo Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal do Rio Grande do Sul.

_____

Prof. Michel José Anzanello, *Ph.*D.
Orientador PPGEP/UFRGS

_____

Prof. Alejandro Germán Frank, Dr.
Coordenador PPGEP/UFRGS

Banca Examinadora:

Professora Helena Beatriz Bettella Cybis, *Ph.*D. (PPGEP/UFRGS)

Professor Alessandro Kahmann, Dr. (IMEF/FURG)

Professor Felipe Caleffi, Dr. (LML/UFSM)

Dedico esta tese à minha família (pai, mãe, irmãos e minha avó). Estes me acolheram da melhor forma que eles poderiam nesta vida, proporcionando-me amor, aprendizado, compreensão e evolução.

AGRADECIMENTOS

Primeiramente, gostaria de agradecer ao PPGEP-UFRGS pela oportunidade e à UFERSA por me fornecer meios para aproveitá-la. Considero este ponto crucial para esta conquista, pois a oportunidade é, geralmente, a alavanca que impulsiona a transformação do destino de um indivíduo. Na sequência, agradeço algumas pessoas que me ajudaram nesta trajetória.

À minha avó, Severina Neri Rocha, que sempre esteve do meu lado com seu incentivo, suas palavras, ensinamentos, sua cantoria e colo. Infelizmente, ela não pode ver, em vida, sequer uma das minhas colações de grau, mas sempre penso nela a cada degrau que eu subo.

Ao meu pai, João de Deus, e minha mãe, Maria de Fátima Silva, que sempre me apoiaram nas minhas empreitadas.

Ao Professor Michel Anzanello, meu orientador, que sempre demonstrou generosidade e paciência em transferir uma parte valiosa do seu conhecimento. Ele teve um papel decisivo na minha evolução como pesquisadora. Sou muito grata e espero ajudar outras pessoas, assim como ele me ajudou.

Aos meus professores da Escola Estatual Professor Edgar Barbosa, especialmente ao Professor Ivanaldo que sempre ressaltou meu potencial e que me fez acreditar num destino diferente (naquela época eu sequer tinha pretensão de fazer uma graduação).

Aos professores Helena Cybis e Felipe Caleffi pela parceria acadêmica.

A engenheira Cristina Albuquerque e arquiteto Reinaldo Germando por abrirem as portas da empresa e me ajudarem na elaboração do artigo 1.

À minha irmã, Simey Rocha, por ser um valioso ponto de apoio dos momentos difíceis.

À Gabrielli Yamashita, minha parceira acadêmica e amiga, pelas diversas conversas sobre métodos, ideias de artigos e, principalmente, pelo apoio durante esse período na UFRGS.

Aos amigos (Thiago Broze, Alessandro Kahmann, Ariane Ávila, Érica Ross e Cíntia Franco) que estiveram presentes e proporcionaram momentos de descontração, tão importantes para aliviar a tensão.

Às amigas (Vanessa Fernandes, Ana Cristina Girão, Hildelana Paiva e Janne Albuquerque) que, apesar da distância, sempre estiveram do meu lado.

A Damian Steppacher, meu namorado, por estar sempre comigo nesses quatro anos. Seu apoio foi essencial para me manter forte (mentalmente e fisicamente) nesse período. Além disso, seu lar também foi um abrigo nos momentos difíceis. Seus pais (Sara e Marcelo Steppacher) demonstraram muito amor e carinho por mim. Também sou grata a eles!

E a todos(as) que contribuíram, direta ou indiretamente, para elaboração deste trabalho.

"Each year it seems to take less time to fly
across the ocean and longer to drive to work."
Autor desconhecido

RESUMO

Esta tese tem por objetivo a proposição de métodos apoiados em ferramentas multivariadas voltados à seleção de variáveis para clusterização e classificação de dados dentro de sistemas de transporte (corredores prioritários de ônibus, conflitos de trânsito e acidentes de trânsito). Para tanto, ela é sustentada por três artigos. O artigo 1 propôs uma nova estrutura para identificar as variáveis mais informativas para agrupar corredores prioritários de ônibus de acordo com suas similaridades (aspectos de sistemas, físicos e operacionais). No artigo 2, conflitos de tráfego foram agrupados usando o *self-organizing maps* (SOM) com base em perfis e características semelhantes que contribuem para a ocorrência de conflitos de tráfego; fim a melhorar a qualidade dos grupos formados, foi desenvolvido um novo índice de importância de variável baseado nos resultados do *nonlinear principal component analysis* (NLPCA). No artigo 3, foram analisados acidentes de trânsito nas áreas rurais e urbanas do Brasil (BR) e da Grã-Bretanha (GB) ocorridos em 2018, com o objetivo de identificar as variáveis mais relevantes para a classificação de acidentes de trânsito em fatais e não fatais. Desta forma, esta tese forneceu contribuições teóricas e práticas. Foram propostas abordagens inéditas, na área de análise multivariada de dados, como um (i) novo índice para mensurar a qualidade da clusterização, e (iii) um novo índice de importância de variáveis baseado nos resultados do NLPCA. Ainda, dentro da área de segurança viária, foi proposto um (iii) método de seleção de variáveis para classificar acidentes fatais e não-fatais (análise similar não foi encontrada na literatura). Em termos práticos, pesquisadores e profissionais podem se beneficiar das proposições desta tese para (i) projetar estratégias de atendimento de corredores prioritários de ônibus, em diferentes cidades ao redor no mundo, com base nas suas características mais relevantes, (ii) gerenciar as condições dos conflitos de trânsito mais suscetíveis a ocorrência de acidentes, e (iii) desenvolver políticas de redução de acidentes com base nas variáveis mais relevantes para discriminar acidentes de trânsito fatais e não-fatais.

Palavras-chave: ferramentas multivariadas, seleção de variáveis, clusterização, classificação, corredores de ônibus, conflitos de trânsito, acidentes de trânsito.

ABSTRACT

This thesis aimed to proposed methods supported by multivariate tools that integrate a variable selection with clustering and classification within transportation systems (priority bus corridors, traffic conflicts, and road accidents). For this, this research is supported on three papers. Paper 1 proposed a novel framework to identify the most informative variables for clustering bus priority corridors according to their similarities (system, physical, and operational aspects). Paper 2 developed 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). Paper 3 aimed to identify the most relevant variables for the classification of road accidents as fatal and non-fatal; for that matter, data reporting accidents in rural and urban areas of Brazil (BR) and Great Britain (GB) in 2018 were analyzed. Thus, this thesis provided theoretical and practical contributions. New approaches were proposed, in the area of multivariate data analysis, as an (i) index to measure the quality of clustering and a (iii) new variable importance index based on the outputs of the NLPCA. Also, within the area of road safety, a (iii) variable selection method was proposed to classify fatal and non-fatal accidents (similar analysis was not found in the literature). Besides, researchers and professionals can benefit from the results of this thesis. For example, to (i) design service strategies for priority bus corridors, in different cities around the world, based on their most relevant variables; (ii) manage the conditions of traffic conflicts that are more susceptible to accidents; and, (iii) develop accident reduction policies based on the most relevant variables to discriminate between fatal and non-fatal road accidents.

Key words: multivariate tools, variable selection, clustering, classification, priority bus corridors, traffic conflicts, road accidents.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Devido ao desenvolvimento das tecnologias *Web*, dispositivos móveis e sensores, a quantidade de dados gerados/coletados aumentou substancialmente (L'HEUREUX et al., 2017). Diante desse cenário surgiu o termo *Big Data*, em português "Grande Base de Dados", que se caracteriza pelo crescimento rápido de volume, variedade e velocidade dos dados; consequentemente, técnicas e tecnologias existentes de processamento de dados podem não ser adequadas para o tratamento de tais dados (SUTHAHARAN, 2014). Na área de transportes, o aumento do volume de dados é um desafio para pesquisadores e especialistas (VLAHOGIANNI; PARK; VAN LINT, 2015). Com isto, surge uma potencial oportunidade no tratamento e análise de bancos de dados de forma a trazer benefícios ao planejamento e operação de sistemas de transportes. Assim, o aprendizado de máquina por ser útil para o tratamento de informações extraídas desta área (QIU et al., 2016).

Nesse sentindo, ferramentas multivariadas, como clusterização e classificação, podem fornecer importante suporte na compreensão de informações oriundas de sistemas de transporte. Percebe-se aumento do volume de dados (e variáveis) que descrevem, por exemplo, as características de corredores de prioridade de ônibus, bem como dados que representam a condição de uma rodovia no momento de um conflito ou acidente de trânsito. Estes dados podem ser (i) supervisionados, compostos por variáveis independentes e uma ou mais variáveis de respostas/dependentes; ou (ii) não supervisionados, descritos apenas por variáveis independentes, não trazendo variáveis de resposta para "supervisionar" o aprendizado do modelo (COOK; HOLDER; KETKAR, 2006; NERUDA; PILÁT; MOUDŘÍK, 2017).

A partir de conjuntos de dados não supervisionados, a clusterização objetiva agrupar as observações ou eventos que sejam similares dentro de um mesmo *cluster*, mas

diferentes em relação a outro(s) *cluster(s)* (HAIR et al., 2009). Além do *k-means*, uma das técnicas de clusterização mais difundidas (ORTIZ et al., 2012), o *self-organizing maps* (SOM) também foi utilizado nesta pesquisa (KOHONEN, 1995). Já na classificação, é necessário uma definição prévia dos grupos, ou seja, uma supervisão por meio de uma ou mais variáveis de respostas (RENCHER, 2003). Os classificadores *k-Nearest Neighbor* (KNN), *Support Vector Machine* (SVM), *Decision Tree* (DT) e *Random Forest* (RF) foram utilizados para construir os modelos de classificação nesta pesquisa.

Ressalta-se que ferramentas multivariadas podem ter seu desempenho comprometido quando aplicadas a conjuntos de dados compostos por um elevado número de variáveis ruidosas e correlacionadas que pode prejudicar a análise. Para contornar tal limitação, o uso de técnicas de seleção de variáveis tem se mostrado uma maneira adequada de aumentar a eficiência das técnicas estatísticas, enquanto se geram modelos mais simples e mais fáceis de interpretar (ANZANELLO et al., 2014). Os procedimentos de seleção de variáveis (VS) consistem em selecionar variáveis relevantes ou eliminar as menos relevantes em um conjunto de dados, reduzindo um conjunto inicial de variáveis *J*. Dessa forma, é definido um subconjunto de variáveis *p* mais relevantes (como $p < J$), transportando a maioria das informações importantes do processo com o mínimo de ruído (ANZANELLO; FOGLIATTO, 2014).

As técnicas de VS podem ser divididas em três categorias principais de método: *filter*, *wrapper* e *embedded* (ALONSO-ATIENZA et al., 2012). O *wrapper* utiliza o aprendizado de máquina de interesse como uma caixa preta para pontuar subconjuntos de variáveis de acordo com seu poder preditivo; o *filter* seleciona subconjuntos de variáveis como uma etapa de pré-processamento; e o método *embedded* incorpora a seleção de variáveis dentro do processo de treinamento do algoritmo (IGUYON; ELISSEEFF,

2003). Tanto o Índice de Importância de Variável (IIV), quanto a técnica "*omit-one-variable-out-at-a-time*" (O1AT), ambos categorizados como método *wrapper*, se destacam na área de seleção de variáveis (AKARACHANTACHOTE; CHADCHAM; SAITHANU, 2017; ANZANELLO; FOGLIATTO, 2011; BOMBARDIER; WENDLING; SCHMITT, 2011; CALEFFI; ANZANELLO; CYBIS, 2017).

Nesse sentido, esta tese se utiliza de abordagens inovadoras de seleção de variáveis para classificar/clusterizar dados associados a sistemas de transporte. Dado este contexto, emergem as questões que norteiam esta pesquisa:

(i) A clusterização poderia ajudar no entendimento das diferenças físicas e operacionais de corredores prioritários de ônibus ao redor do mundo?

(ii) A NLPCA poderia ser útil na criação de um novo índice de importância de variáveis tendo em vista a seleção de variáveis mais relevantes em um banco de dados não supervisionado de conflitos de tráfego?

(iii) Ainda em relação aos conflitos de trânsito, a clusterização ajudaria na compreensão dos grupos que estão mais propensos à acidentes?

(iv) No que se refere aos dados supervisionados, técnicas de seleção de variáveis seriam úteis para aprimorar a acurácia da classificação entre acidentes trânsito fatais e não-fatais?

1.1 Tema da Tese

O nível de desenvolvimento de infraestruturas de transporte é um importante indicador econômico e de urbanização de uma região (MAPARU; MAZUMDER, 2017). A necessidade de locomoção e o crescimento do poder aquisitivo de uma parcela da sociedade, associados a uma oferta precária do serviço de transporte público, promovem o aumento da motorização. Segundo a *Organisation for Economic Co-operation and*

Development (OECD) no *International Transport Forum* (2012), a frota global de veículos chegará a 2 bilhões em 2050, 1 bilhão a mais do que em 2012.

A comodidade que um veículo motorizado oferece é uma das principais vantagens conferida aos potenciais usuários. Por outro lado, uma parcela da população que não pode adquirir um veículo motorizado, ou que opta por não o utilizar, também sofre as consequências do aumento da motorização. Uma série de impactos negativos está associada a esse crescimento, incluindo congestionamentos e acidentes de trânsito (LEVINSON; LOMAX; TURNER, 1997; PAULOZZI et al., 2007).

Congestionamentos de tráfego são um problema complexo e um fenômeno multidimensional, difíceis de serem investigados e contornados (OECD; EUROPEAN, 2007). Nesse sentido, cidades ao redor do mundo têm investido em sistemas de transporte público como forma de reduzir os congestionamentos e mitigar os impactos negativos advindos do aumento da frota de veículos dos grandes centros. Os corredores ou faixas que dão prioridades aos ônibus são um exemplo de priorização do transporte público. Nesse sentindo, os sistemas de *Bus Rapid Transit* (BRT) e sua versão europeia, *Bus with High Level of Service* (BHLS), têm sido adotados em diferentes localidades como forma de priorizar o ônibus, reduzindo congestionamentos e poluição, e aumentando a segurança e conforto dos usuários do transporte público (HERES; JACK; SALON, 2014).

Outra consequência, ainda mais grave, inerente à alta taxa de motorização são os acidentes de trânsito (ALDMAN; THORSON, 1971; AL-REESI et al., 2013; BENER, 2009; QUDDUS; WANG; ISON, 2010) os quais provocam prejuízos inestimáveis para as vítimas e suas famílias, bem como perdas econômicas para a sociedade como um todo. Essas perdas contemplam desde o custo do tratamento das vítimas até a perda da produtividade decorrente de mortos ou inválidos. Também acarretam prejuízos aos

familiares, os quais precisam se ausentar do trabalho ou escola para cuidar dos seus entes feridos (WHO, 2018).

Nesse contexto, esta tese tem como foco selecionar as variáveis mais relevantes para auxiliar no entendimento de características e de problemas relacionados a sistemas de transporte. Para um melhor entendimento, a figura 1.1 apresenta um mapa conceitual que descreve a motivação desta pesquisa. Tanto os corredores prioritários de ônibus quanto conflitos e acidentes de trânsito são descritos um conjunto de dados multivariados. Desta forma, as técnicas de redução de dimensionalidade e ferramentas multivariadas podem ajudar na compreensão de problemas, como os mencionados acima, ligados a sistemas de transporte.

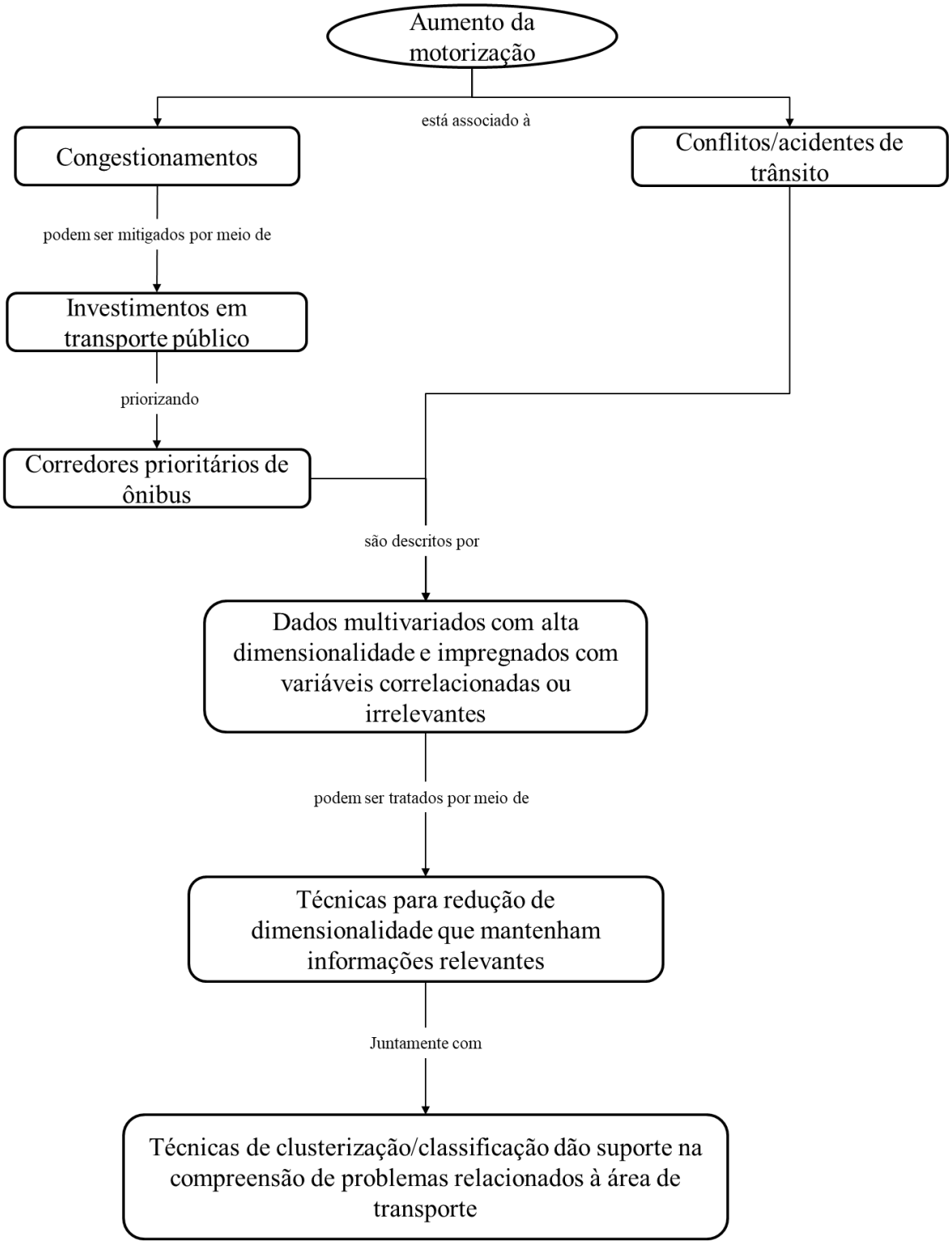

Figura 1.1 - Motivação do tema da tese

1.2 Objetivo da Tese

O objetivo geral desta tese consiste em propor métodos inovadores apoiados em ferramentas multivariadas que integrem a seleção de variáveis à clusterização ou classificação de dados associados a sistemas de transporte. Em relação aos objetivos específicos, como esta tese foi desenvolvida por meio de três artigos, os quais se referem a contextos diferentes. Para facilitar a compreensão de tais objetivos, a tabela 1.1 apresenta as questões de pesquisa e objetivos específicos que cada artigo busca responder, bem como o método de pesquisa e a contribuição inédita de cada um deles.

Tabela 1.1: Estrutura das etapas da pesquisa desenvolvida

Artigo	Questões de pesquisa	Objetivos específicos	Método de pesquisa	Contribuição teórica (inédita)
Artigo 1 [(a)]	(i) A clusterização poderia ajudar no entendimento das diferenças físicas e operacionais de corredores prioritários de ônibus ao redor do mundo?	a) Entender como métodos de clusterização podem promover uma melhor compreensão sobre corredores de prioridade de ônibus; b) Propor e validar um novo índice de mensuração de qualidade de clusters baseado em dois outros índices existentes: Silhouette *index* (*S*) e Davies-Bouldin *index* (*DB*);	Pesquisa quantitativa: (i) Coletar e pré-processar dados que descrevam corredores de ônibus; (ii) Definir o número de clusters a serem gerados; (iii) Agrupar iterativamente corredores de ônibus e eliminar variáveis de agrupamento menos relevantes.	Proposição de um novo índice de mensuração de qualidade de clusters.
Artigo 2 [(b)]	(ii) A NLPCA poderia ser útil na criação de um novo índice de importância de variáveis tendo em vista a seleção de variáveis mais relevantes em um banco de dados não supervisionado de conflitos de tráfego? (iii) Ainda em relação ao conflitos de trânsito, a clusterização ajudaria na compreensão dos grupos que estão mais propensos à acidentes?	c) Avaliar o desempenho de um método de clusterização não usual dentro da área de segurança de tráfego, SOM, em conflitos de tráfego de uma rodovia brasileira; d) Propor e validar um novo índice de importância baseado nos parâmetros do NLPCA para selecionar variáveis mais relevantes; e	Pesquisa quantitativa: (i) Coletar e preparar os dados que descrevem os eventos de conflito de tráfego; (ii) Definir o número de clusters a serem gerados; (iii) Gerar o índice de importância baseado no NLPCA para guiar a remoção de variáveis menos relevantes; (iv) Clusterizar os conflitos de trânsito e eliminar as variáveis menos relevantes.	Proposição de um novo índice de importância baseado nos resultados da NLPCA.

Artigo	Questões de pesquisa	Objetivos específicos	Método de pesquisa	Contribuição teórica (inédita)
Artigo 3 [c]	(iv) No que se refere aos dados supervisionados, técnicas de seleção de variáveis seriam úteis para aprimorar a acurácia da classificação entre acidentes trânsito fatais e não-fatais?	e) Avaliar o desempenho da abordagem de seleção de variáveis no contexto de segurança viária com vistas à classificação de acidentes de trânsito.	Pesquisa quantitativa: (i) Preparar os dados que descrevem os acidentes; (ii) Classificar as observações e eliminar as variáveis menos relevantes iterativamente; (iii) Avaliar o desempenho do modelo de classificação e avaliar qualitativamente as variáveis retidas.	Utilização de uma sistemática de seleção de variáveis para identificar as características mais relevantes para discriminar acidentes fatais e não-fatais em zonas rurais e urbanos.

a) (a) Artigo submetido ao periódico *Transport Policy* (segunda rodada de revisão).

b) (b) Artigo publicado no periódico *Accident Analysis and Prevention*.

c) (c) Artigo a ser submetido ao periódico *Safety Science*.

1.3 Justificativa do tema e dos objetivos

Separar faixas prioritárias de ônibus em grupos que apresentam perfis comparáveis em relação às condições físicas e operacionais permite estender medidas bem-sucedidas de uma faixa específica para um grupo de faixas apresentando semelhanças entre si. Em uma perspectiva gerencial, tais informações constituem-se em um recurso valioso para a identificação de grupos de corredores, que podem se beneficiar de ações implementadas com sucesso com vistas no *benchmarking* dos pertencentes ao mesmo grupo (ou mesmo ações empregadas em outros grupos após alguma adaptação). Para tanto, um novo índice de mensuração de qualidade de *clusters* baseado em dois outros índices *S* e *DB* foi proposto para ser integrado à técnica O1AT.

No que tange aos conflitos tráfego, compreender as causas que dão origem a diferentes tipos de eventos de conflito, bem como suas características, pode ajudar pesquisadores e autoridades de trânsito a elaborar estratégias adaptadas para reduzir a ocorrência de colisões. Nesse sentido, o agrupamento de conflitos pode ser um recurso útil, pois permite inserir eventos de conflito de tráfego em grupos que apresentam características semelhantes e destacar fatores que contribuem para a ocorrência de conflitos. Esses grupos podem então ser usados para desenvolver estratégias conjuntas destinadas a reduzir a probabilidade de ocorrência de acidentes. Tais benefícios justificam as abordagens aqui propostas em termos práticos.

Embora a literatura ofereça vários índices para identificar as variáveis mais informativas/relevantes, eles são tipicamente baseados em modelagem linear e propensos à perda de informação (ANZANELLO et al., 2016; ANZANELLO; FOGLIATTO; ROSSINI, 2011). Ademais, técnicas não-lineares podem melhorar o manuseio e a descrição de relações mais complexas entre variáveis, que normalmente tendem a acontecer em aplicações do mundo real (CLAVERÍA; POLUZZI, 2017). Desta forma,

um novo índice de importância baseado nos resultados do NLPCA para selecionar variáveis mais relevantes se justificaria do ponto de vista acadêmico.

Ainda no âmbito de justificação teórica, percebeu-se a inexistência de aplicação do SOM a dados associados à segurança de tráfego. Esta técnica de inteligência artificial, também conhecida como mapas auto-organizáveis, fornece uma separação visual dos diferentes clusters por meio de um gráfico de similaridade (KOHONEN, 1995). Embora seja uma ferramenta multivariada amplamente utilizada em diversas áreas com o propósito de formar de grupos, apenas cinco trabalhos foram encontrados aplicando o SOM na área de segurança de tráfego (LIU; BUCKNALL, 2018; PRATO; KAPLAN, 2013; PRIETO; ALLEN, 2009; STOICA; SEVERI; DE ABREU, 2015; WANG; CHAI; WU, 2014). Estes estudos oferecem contribuições para reduzir a ocorrência de acidentes, porém nenhum deles agrupou e analisou os conflitos de tráfego com o SOM, ou propôs estruturas de seleção de variáveis para aprimorar a análise.

No que diz respeito a acidentes de trânsito, alguns estudos já propuseram sistemáticas de seleção de variáveis em bancos de dados daquela natureza (WANG et al., 2005; WEI; WU; KOU, 2011; GEETHA RAMANI; SELVARAJ, 2014; ÇELIK; SENGER, 2014; LIN; WANG; SADEK, 2015; CALEFFI; ANZANELLO; CYBIS, 2017). Estes estudos apresentam contribuições relevantes; no entanto, não foram encontrados trabalhos que selecionem variáveis mais relevantes para classificar acidentes de trânsito em fatais e não fatais, o que se constitui em mais uma justificativa para o presente estudo.

1.4 Delineamento do Estudo

Definidos os objetivos e justificativa desta tese, esta seção apresenta o enquadramento da pesquisa do ponto de vista metodológico. Traz ainda o método de trabalho para alcançar os objetivos propostos e responder as questões de pesquisas levantadas nesta tese.

1.4.1 Método de Pesquisa

Do ponto de vista da natureza e abordagem, esta pesquisa é classificada como aplicada e quantitativa, respectivamente. Uma pesquisa aplicada gera conhecimentos para aplicação prática com vistas a soluções de problemas específicos (SILVA; MENEZES, 2005). Já a quantitativa se baseia em métodos lógico-dedutivos para explicar relações de causa/efeito e, por meio da generalização, possibilitar replicações (BERTO; NAKANO, 1999). Assim sendo, requer o uso de recursos e técnicas estatísticas para traduzir em números informações com a finalidade de analisá-las e classificá-las (SILVA; MENEZES, 2005).

Em relação aos objetivos da pesquisa, esta tese é classificada como exploratória. Pesquisas exploratórias têm por finalidade desenvolver, esclarecer e modificar conceitos e ideias, considerando a formulação de problemas mais precisos ou hipóteses pesquisáveis; além disso, este tipo pesquisa é busca proporcionar uma visão geral acerca de determinado fato, sendo assim, é realizada especialmente quando o tema escolhido é pouco explorado e torna-se difícil formular, sobre ele, hipóteses precisas e operacionalizáveis (GIL, 2008).

1.4.2 Método de Trabalho

A partir de três artigos aplicados a diferentes contextos em sistema de transporte, busca-se atingir o objetivo geral da tese. Cada artigo cumpre um (ou mais) objetivo(s)

específico(s); e, apresenta pelo menos uma contribuição inédita na literatura acadêmica. Abaixo são descritos os três artigos que compõem esta tese. Vale salientar que os artigos são apresentados no formato de submissão aos periódicos internacionais; assim sendo, estão escritos em língua inglesa.

O Artigo 1 - *Selecting the most relevant variables towards clustering bus priority corridors* (Selecionando as variáveis mais relevantes para agrupar os corredores prioritários de ônibus) – propõe uma nova estrutura para identificar as variáveis mais informativas para agrupar corredores prioritários de ônibus de acordo com suas similaridades (aspectos de sistemas, físicos e operacionais). Embora cada corredor de ônibus tenha suas peculiaridades, entender as semelhanças entre corredores de diferentes regiões do mundo pode ajudar pesquisadores e especialistas em trânsito a elaborar estratégias adaptadas para melhorar o tráfego nas cidades. Para tanto, uma nova métrica para medir a qualidade dos agrupamentos formados foi integrada ao procedimento de seleção "*omit-one-variable-out-at-a-time*" (O1AT). O método proposto baseia-se em 3 etapas: (i) coletar e pré-processar dados que descrevam corredores de ônibus; (ii) definir o número de clusters a serem gerados com base em uma abordagem hierárquica; e (iii) iterativamente agrupar corredores de barramento e eliminar variáveis de agrupamento menos relevantes. Quando aplicado a um conjunto de dados composto de 296 corredores prioritários de ônibus de 45 países e descrito por 44 variáveis, a estrutura proposta reteve apenas 4 variáveis (marca e/ou logotipo, espaçamento de estação, estações aprimoradas e velocidade de operação) responsáveis pela melhor estratificação de corredores. Quatro grupos foram formados e avaliados qualitativamente em relação às suas similaridades em termos de aspectos sistêmico, físico e operacional. Os corredores foram agrupados em corredores básicos (cluster 1), corredores melhorados (cluster 2), sistemas *Bus Rapid*

Transit (BRT) e *Bus with High Level of Service* (BHLS) (cluster 3) e serviços expressos de parada limitada (cluster 4).

O Artigo 2 - *A multivariate-based variable selection framework integrated to self-organized maps for clustering traffic conflicts in a Brazilian freeway* (Um método de seleção de variáveis integrado a mapas auto-organizados para agrupar conflitos de tráfego em uma rodovia brasileira) – objetiva agrupar conflitos de tráfego via SOM com base em perfis e característica semelhantes que contribuem para a ocorrência de conflitos de tráfego. A fim de melhorar a qualidade dos grupos formados, foi desenvolvido um novo índice de importância de variável baseado nos resultados do NLPCA. Esse índice orienta um procedimento *backward* de seleção de variáveis, no qual variáveis menos relevantes são removidas uma a uma; após cada remoção, a qualidade de agrupamento é avaliada através do índice de Davies-Bouldin (DB). A abordagem proposta foi aplicada a um conjunto de dados em tempo real coletados em uma rodovia brasileira visando alocar conflitos de tráfego em grupos que apresentavam perfis semelhantes. As variáveis selecionadas sugerem que velocidades médias mais baixas, que são tipicamente verificadas durante eventos de congestionamento, contribuem para a ocorrência de conflitos. A maior variabilidade na velocidade (denotada pelo alto desvio padrão e os coeficientes de velocidade dos níveis de variação nessa variável), que também é percebida na via expressa avaliada próximo aos períodos de congestionamento, também contribui para os conflitos.

O Artigo 3 - *Variable selection techniques for classifying traffic accidents into fatal and non-fatal* (Técnicas de seleção de variáveis para classificar acidentes de trânsito em fatal e não fatal) – analisa acidentes trânsito nas áreas rurais e urbanas do Brasil (BR) e da Grã-Bretanha (GB), no ano de 2018, com o objetivo de identificar as variáveis mais relevantes para a classificação de acidentes de trânsito em fatais e não fatais. Para tanto,

após o pré-tratamento dos dados, foi integrada a técnica “omit-one-variable-at-a-time” (O1AT) no procedimento para identificação de variáveis mais relevantes para classificar os acidentes de trânsito. Para escolher o melhor método de classificação, também foram utilizados métodos de seleção de variáveis baseados em IIV’s. A técnica O1AT se destacou como o melhor método de seleção de variáveis na porção de treino. Além disso, na avaliação final do modelo de classificação, na porção de teste, a O1AT mostrou a maior precisão nos conjuntos de dados avaliados. Assim, a O1AT superou os IIV’s para discriminar os acidentes de trânsito em fatais e não fatais. Sobre as variáveis retidas, no conjunto de dados BR, a variável “Accident_Type_BR” foi selecionada nos dois conjuntos de dados; também foi selecionado “Number_of_Casualties_BR” para classificar a área rural e a variável binária “Weekend_BR” foi selecionada para classificar a área urbana. No conjunto de dados GB, destacamos a variável retida “Pedestrian_Crossing-Human_Control_GB” que foi selecionada para áreas urbanas e rurais.

1.5 Delimitações do Estudo

Ferramentas de clusterização foram priorizadas nos dois primeiros artigos; tendo a seleção de variáveis o objetivo de aprimorar a qualidade dos clusters gerados. No artigo 3, uma vez que o banco de dados é composto por variáveis de acidentes de trânsito que os categorizam como fatais e não-fatais, análise de classificação se mostrou mais adequada para este caso. Desta forma, esta tese delimitou à clusterizar e classificar dados de sistemas de transporte, ou seja, técnicas de predição não foram utilizadas nesta tese. Por fim, foram utilizadas técnicas de seleção de variáveis (IIV e O1AT) da categoria *wrapper*.

1.6 Estrutura da Tese

Esta tese foi organizada em cinco capítulos. O primeiro forneceu uma contextualização do problema de pesquisa, bem como, tema, objetivos, justificativa da tese; além do delineamento da pesquisa por meio do método de pesquisa e trabalho; e, finalizou com as delimitações do estudo e estrutura da tese. Na sequência, os capítulos 2, 3 e 4 apresentaram os artigos 1, 2 e 3, respectivamente. Por fim, o capítulo 5 abordou as conclusões desta tese, bem como novas direções de pesquisa.

1.7 Referências

AKARACHANTACHOTE, N.; CHADCHAM, S.; SAITHANU, K. Variable importance index based on the partial least squares and boxplot cutoff threshold for variable selection. **International Journal of Data Analysis Techniques and Strategies**, v. 9, n. 1, p. 34–45, 2017.

ALDMAN, B.; THORSON, J. Motorization and traffic mortality in Sweden. **Accident Analysis & Prevention**, v. 3, n. 3, p. 215–221, 1 out. 1971.

ALONSO-ATIENZA, F. et al. Feature selection using support vector machines and bootstrap methods for ventricular fibrillation detection. **Expert Systems with Applications**, v. 39, n. 2, p. 1956–1967, 1 fev. 2012.

AL-REESI, H. et al. Economic Growth, Motorization, and Road Traffic Injuries in the Sultanate of Oman, 1985–2009. **Traffic Injury Prevention**, v. 14, n. 3, p. 322–328, 25 fev. 2013.

ANZANELLO, M. et al. Wavelength selection framework for classifying food and pharmaceutical samples into multiple classes. **Journal of Chemometrics**, v. 30, n. 6, p. 346–353, 2016.

ANZANELLO, M. J. et al. Selecting relevant Fourier transform infrared spectroscopy wavenumbers for clustering authentic and counterfeit drug samples. **Science & Justice**, v. 54, n. 5, p. 363–368, 2014.

ANZANELLO, M. J.; FOGLIATTO, F. S. **Clustering variables selection in mass customized scenarios affected by workers' learning**. . In: IEEE INTERNATIONAL CONFERENCE ON INDUSTRIAL ENGINEERING AND ENGINEERING MANAGEMENT. 2011

ANZANELLO, M. J.; FOGLIATTO, F. S. A review of recent variable selection methods in industrial and chemometrics applications. **European Journal of Industrial Engineering**, v. 8, n. 5, p. 619–645, 2014.

ANZANELLO, M. J.; FOGLIATTO, F. S.; ROSSINI, K. Data mining-based method for identifying discriminant attributes in sensory profiling. **Food Quality and Preference**, v. 22, n. 1, p. 139–148, 2011.

BENER, A. Emerging trend in motorisation and the epidemic of road traffic crashes in an economically growing country. **International Journal of Crashworthiness**, v. 14, n. 2, p. 183–188, 30 abr. 2009.

BERTO, R. M. V. S.; NAKANO, D. N. A produção científica nos anais do encontro nacional de engenharia de produção: um levantamento de métodos e tipos de pesquisa. **Production**, v. 9, n. 2, p. 65–75, dez. 1999.

BOMBARDIER, V.; WENDLING, L.; SCHMITT, E. **Feature selection based on importance and interaction idexes-hierarchical fuzzy rule classifier application**. 2011

CALEFFI, F.; ANZANELLO, M. J.; CYBIS, H. B. B. A multivariate-based conflict prediction model for a Brazilian freeway. **Accident Analysis & Prevention**, v. 98, p. 295–302, 2017.

ÇELIK, A. K.; SENGER, Ö. RISK FACTORS AFFECTING FATAL VERSUS NON-FATAL ROAD TRAFFIC ACCIDENTS: THE CASE OF KARS PROVINCE, TURKEY. **International Journal for Traffic and Transport Engineering**, v. 4, n. 3, p. 339–351, set. 2014.

CLAVERÍA, Ó.; POLUZZI, A. Positioning and clustering of the world's top tourist destinations by means of dimensionality reduction techniques for categorical data. **Journal of Destination Marketing & Management, 2016, vol. 6, num. 1, p. 22-32**, 2017.

COOK, D. J.; HOLDER, L. B.; KETKAR, N. Unsupervised and Supervised Pattern Learning in Graph Data. In: **Mining Graph Data**. [s.l.] John Wiley & Sons, Ltd, 2006. p. 159–181.

GEETHA RAMANI, R.; SELVARAJ, S. A pragmatic approach for refined feature selection for the prediction of road accident severity. **Studies in Informatics and Control**, v. 23, n. 1, p. 41–52, 2014.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 6. ed ed. São Paulo: Editora Atlas S.A., 2008.

HAIR, J. F. et al. **Análise multivariada de dados**. [s.l.] Bookman Editora, 2009.

HERES, D. R.; JACK, D.; SALON, D. Do public transport investments promote urban economic development? Evidence from bus rapid transit in Bogotá, Colombia. **Transportation**, v. 41, n. 1, p. 57–74, 1 jan. 2014.

IGUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. **Journal of Machine Learning Research**, v. 3, p. 1157–1182, 2003.

KOHONEN, T. **Self-Organizing Maps**. 3. ed. Berlin Heidelberg: Springer-Verlag, 1995.

LEVINSON, H. S.; LOMAX, T. J.; TURNER, S. **Traffic congestion - past - present - future**. . In: PROCEEDINGS OF THE CONFERENCE ON TRAFFIC CONGESTION AND TRAFFIC SAFETY IN THE 21ST CENTURY. 1997

L'HEUREUX, A. et al. Machine Learning With Big Data: Challenges and Approaches. **IEEE Access**, v. 5, p. 7776–7797, 2017.

LIN, L.; WANG, Q.; SADEK, A. W. A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction. **Transportation Research Part C: Emerging Technologies**, Engineering and Applied Sciences Optimization (OPT-i) - Professor Matthew G. Karlaftis Memorial Issue. v. 55, p. 444–459, 1 jun. 2015.

LIU, Y.; BUCKNALL, R. Efficient multi-task allocation and path planning for unmanned surface vehicle in support of ocean operations. **Neurocomputing**, v. 275, p. 1550–1566, 2018.

MAPARU, T. S.; MAZUMDER, T. N. Transport infrastructure, economic development and urbanization in India (1990–2011): Is there any causal relationship? **Transportation Research Part A: Policy and Practice**, v. 100, p. 319–336, 1 jun. 2017.

NERUDA, R.; PILÁT, M.; MOUDŘÍK, J. **Unsupervised and Supervised Activity Analysis of Drone Sensor Data**. (J. C. Figueroa-García et al., Eds.)Applied Computer Sciences in Engineering. **Anais**...: Communications in Computer and Information Science.Cham: Springer International Publishing, 2017

OECD; EUROPEAN, C. OF M. OF T. (ECMT). **Managing urban traffic congestion**. [s.l: s.n.]. v. 9789282101506

OECD/ITF. Transport Outlook 2012: Seamless Transport for Greener Growth. **Organisation for Economic Co-operation and Development. Accessed June**, v. 20, p. 2015, 2012.

ORTIZ, R. S. et al. Physical profile of counterfeit tablets Viagra® and Cialis®. **Brazilian Journal of Pharmaceutical Sciences**, v. 48, n. 3, p. 487–495, 2012.

PAULOZZI, L. J. et al. Economic development's effect on road transport-related mortality among different types of road users: A cross-sectional international study. **Accident Analysis & Prevention**, v. 39, n. 3, p. 606–617, 1 maio 2007.

PRATO, C. G.; KAPLAN, S. Promoting safe transit: analyzing bus accident patterns. **Journal of Risk and Governance**, v. 4, n. 1, p. 13, 2013.

PRIETO, M. S.; ALLEN, A. R. Using self-organising maps in the detection and recognition of road signs. **Image and Vision Computing**, v. 27, n. 6, p. 673–683, 2009.

QIU, J. et al. A survey of machine learning for big data processing. **EURASIP Journal on Advances in Signal Processing**, v. 2016, n. 1, p. 67, 28 maio 2016.

QUDDUS, M. A.; WANG, C.; ISON, S. G. Road Traffic Congestion and Crash Severity: Econometric Analysis Using Ordered Response Models. **Journal of Transportation Engineering**, v. 136, n. 5, p. 424–435, 1 maio 2010.

RENCHER, A. C. **Methods of Multivariate Analysis**. [s.l.] John Wiley & Sons, 2003.

SILVA, E. L. DA; MENEZES, E. M. **Metodologia da Pesquisa e Elaboração de Dissertação**. 4. ed ed. Florianópolis: Universidade Federal de Santa Catarina - UFSC, 2005.

STOICA, R.-A.; SEVERI, S.; DE ABREU, G. T. F. **Learning the vehicular channel through the self-organization of frequencies**. 2015 IEEE Vehicular Networking Conference (VNC). **Anais**...IEEE, 2015

SUTHAHARAN, S. **Big data classification: problems and challenges in network intrusion prediction with machine learning**Association for Computing Machinery, , 17 abr. 2014. Disponível em: <https://doi.org/10.1145/2627534.2627557>. Acesso em: 1 abr. 2020

VLAHOGIANNI, E. I.; PARK, B. B.; VAN LINT, J. W. C. Big data in transportation and traffic engineering. **Transportation Research Part C: Emerging Technologies**, Big Data in Transportation and Traffic Engineering. v. 58, p. 161, 1 set. 2015.

WANG, H. et al. **Variable selection and ranking for analyzing automobile traffic accident data**. . In: PROCEEDINGS OF THE ACM SYMPOSIUM ON APPLIED COMPUTING. 2005

WANG, J.; CHAI, R.; WU, Q. **Changing lane probability estimating model based on neural network**. The 26th Chinese Control and Decision Conference (2014 CCDC). **Anais**...IEEE, 2014

WEI, J.-T.; WU, H.-H.; KOU, K.-Y. **Using feature selection to reduce the complexity in analyzing the injury severity of traffic accidents**. . In: PROCEEDINGS - 2011 INTERNATIONAL JOINT CONFERENCE ON SERVICE SCIENCES, IJCSS 2011. 2011

WHO, WORLD HEALTH ORGANIZATION. **Road traffic injuries**. Disponível em: <http://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries>. Acesso em: 7 nov. 2018.

5 CONCLUSÕES E NOVAS DIREÇÕES DE PESQUISA

Este capítulo apresenta as conclusões desta tese, bem como novas direções de pesquisa. A discussão apresentada se divide em quatro partes: (i) conclusões do método de pesquisa, (ii) conclusões dos resultados, (iii) contribuições teóricas e práticas e, por último, as (iv) limitações e direções de pesquisa.

5.1 Conclusões do método de pesquisa

Cada artigo apresentado possui um método de pesquisa diferente. No entanto, todos eles possuem uma similaridade: a seleção de variáveis com vistas no aprimoramento dos resultados de ferramentas multivariadas (clusterização e classificação). Assim, de forma genérica, o método abordado é compreendido em quatro etapas. Iniciando por (i) preparar os dados; este passo inclui a análise de *outliers*, preenchimento ou retirada de *missing values*, análise de variáveis altamente correlacionadas, conversão de variáveis contínuas em categóricas e/ou outras técnicas que se fizerem necessárias. Se for análise de clusterização, é necessário (ii) definir o número de agrupamentos; para tanto, pode-se contar com o auxílio de um dendrograma e análise subjetiva de especialistas, por exemplo.

Na sequência, é necessário (iii) eliminar as variáveis menos relevantes; as técnicas de seleção de variáveis como IIV e O1AT podem ser utilizadas para esta eliminação. Se IIV for utilizado, deve-se gerar o IIV para guiar a remoção (*backward*) ou inclusão (*forward*) de variáveis na clusterização ou classificação. Já a técnica O1AT precisa ser incorporada à ferramenta multivariada escolhida (ver artigo 1 ou 3 para mais detalhes). Na análise de classificação, é aconselhável separar os dados em partições de treino e teste. Na partição de treino, o modelo é construído e validado; já na partição de teste o modelo é avaliado com dados não utilizados na construção do modelo. Por fim, deve-se (iv) avaliar qualitativamente as variáveis selecionadas.

5.2 Conclusões dos resultados

No decorrer desta tese, foi proposto e validado um novo índice de mensuração de qualidade de clusters baseado em dois outros índices existentes: Silhouette index (S) e Davies-Bouldin index (DB). Também foi proposto e validado um novo índice de importância baseado nos resultados do NLPCA para selecionar variáveis mais relevantes. Ainda, diferentes técnicas de seleção de variáveis foram utilizadas para selecionar variáveis mais relevantes para classificar acidentes de trânsito em fatais e não fatais.

Apesar da essência quantitativa da pesquisa, os três artigos se utilizaram de uma análise qualitativa para interpretação do clusters formados, bem como na avaliação nas variáveis mais relevantes para classificar acidentes fatais e não fatais. No artigo 1, quatro de quarenta e quatro variáveis foram consideradas relevantes para a clusterização dos corredores de ônibus: presença de marca e/ou logotipo, distância entre estações, presença de estações melhoradas e velocidade da operação. Do ponto de vista qualitativo, há forte relação entre as dimensões descritas por tais variáveis e a definição de BRT fornecida pelo BRT Standard e ALC-BRT. As variáveis supracitadas foram usadas para agrupar 285 corredores prioritários de ônibus de 206 cidades em 4 clusters; os agrupamentos gerados foram então avaliados qualitativamente.

No artigo 2, três variáveis foram retidas ($Av.Speed_5$, $Coeff.Var.Speed_5$ e $Std.Dev.Speed_5$) e usadas pelo SOM para agrupar os conflitos de tráfego. O novo índice proposto melhorou os resultados do SOM. Estas três variáveis também sugerem que velocidades médias mais baixas, que são tipicamente verificadas durante eventos de congestionamento, contribuem para a ocorrência de conflitos. A maior variabilidade na velocidade (denotada pelo alto desvio padrão e os coeficientes de velocidade dos níveis de variação nessa variável), que também são percebidos na via expressa avaliada próximo aos períodos de congestionamento, também contribui para os conflitos.

No artigo 3, foram comparadas diferentes abordagens de seleção de variáveis com o propósito de selecionar variáveis mais relevantes para classificar acidentes de trânsito em fatais e não fatais nas zonas rurais e urbanas do Brasil e Grã-Bretanha. Ao final, a técnica O1AT demonstrou eficácia uma vez que, na porção de treino, foi obtida a melhor acurácia nos quatro conjuntos de dados analisados em comparação com a classificação com todas as variáveis (ou seja, 100% das variáveis retidas). Já na porção de teste, o conjunto de variáveis retidas foi melhor na classificação em três dos quatro bancos analisados; sendo que o único que a seleção de variáveis apresentou resultados piores (GB-Rural), a diferença foi menor que 3% de acurácia.

O objetivo geral desta pesquisa, que consistia em propor métodos apoiados em ferramentas multivariadas que integrassem a seleção de variáveis à clusterização e classificação de dados relacionados a sistemas de transporte, foi alcançado. Assim como os objetivos específicos, mencionado a seguir, também foram respondidos.

a) Entender como métodos de clusterização podem promover uma melhor compreensão sobre corredores de prioridade de ônibus (artigo 1);
b) Propor e validar um novo índice de mensuração de qualidade de *clusters* baseado em dois outros índices existentes: Silhouette *index* (*S*) e Davies-Bouldin *index* (*DB*) (artigo 1);
c) Avaliar o desempenho de um método de clusterização não usual dentro da área de segurança de tráfego, SOM, em conflitos de tráfego de uma rodovia brasileira (artigo 2);
d) Propor e validar um novo índice de importância baseado nos parâmetros do NLPCA para selecionar variáveis mais relevantes (artigo 2); e

e) Avaliar o desempenho da abordagem de seleção de variáveis no contexto de segurança viária com vistas à classificação de acidentes de trânsito (artigo 3).

Além disso, as questões de pesquisas que nortearam esta tese foram respondidas: (i) a clusterização ajudou no entendimento das diferenças físicas e operacionais de corredores prioritários de ônibus ao redor do mundo; (ii) um novo índice de importância baseado nos resultados do NLPCA se mostrou para seleção de variáveis mais relevantes de conflitos de tráfego; assim como, (iii) a clusterização ajudou na compreensão deste conflitos; e, por fim, (iv) abordagens de seleção de variáveis aprimoraram a acurácia da classificação de acidentes de trânsito fatais e não-fatais.

5.3 Contribuições teóricas e práticas

Algumas lacunas teóricas foram preenchidas ao longo do desenvolvimento desta pesquisa. Foram propostas abordagens inéditas, na área de análise multivariada de dados, como um (i) índice para mensurar a qualidade da clusterização e um (iii) índice de importância de variáveis baseado nos *outputs* do NLPCA. Ainda, dentro da área de segurança viária, foi proposta um (iii) método de seleção de variáveis para classificar acidentes fatais e não-fatais (análise similar não foi encontrada na literatura).

Além disso, almejou-se fornecer contribuições práticas em uma área com um "déficit" de pesquisas com um foco em análises multivariadas. Sistemas de transporte se mostrou uma área promissora neste sentido, apesar da existência de diversos estudos que se utilizam de um arcabouço de técnicas multivariadas. Pesquisadores e profissionais podem se beneficiar com os resultados desta tese. Por exemplo, para (i) projetar estratégias de atendimento de corredores prioritários de ônibus, em diferentes cidades ao redor no mundo, com base nas suas características mais relevantes; (ii) gerenciar as condições dos conflitos de trânsito mais suscetíveis à ocorrência de acidentes; e, (iii)

desenvolver políticas de redução de acidentes com base nas variáveis mais relevantes para discriminar acidentes de trânsito fatais e não-fatais.

5.4 Limitações e direções de pesquisa

Esta tese propôs abordagens de seleção de variáveis com vistas no aprimoramento de análises de clusterização e classificação. Assim, para trabalhos futuros, poder-se-ia avançar em novas técnicas de seleção de variáveis (incluindo os métodos *filter* e *embedded*) para predizer informações ainda não exploradas dentro de sistemas de transporte ou em outras áreas que possibilitem tal análise.

Nesta tese, mais precisamente no artigo 3, foram analisados dados de acidentes de trânsito do Brasil e Grã-Bretanha no ano de 2018. Ao longo do desenvolvimento deste trabalho, foi percebido a existência de um volume considerável de dados abertos de acidentes de trânsito. Somando-se a isto, trabalhos com enfoque em seleção de variáveis mais relevantes para clusterizar acidentes fatais não foram encontrados. Desta forma, uma nova sistemática de seleção de variáveis, para clusterizar acidentes de trânsito fatais, poderia ajudar no entendimento de variáveis mais relevantes para separar os diferentes tipos de acidentes que levam a mortalidade; ainda, forneceria uma contribuição à área de análise multivariada de dados.