UNIVERSIDADE DE SÃO PAULO
Escola de Engenharia de São Carlos
Departamento de Engenharia de Transportes
Programa de Pós-Graduação em Engenharia de Transportes

MURILO CASTANHO DOS SANTOS

**Explorando técnicas para modelagem de dados agregados
de óbitos provenientes de acidentes por automóvel**

São Carlos
2015

MURILO CASTANHO DOS SANTOS

**Explorando técnicas para modelagem de dados agregados
de óbitos provenientes de acidentes por automóvel**

Dissertação submetida à Escola de Engenharia de São Carlos, da Universidade de São Paulo, como parte dos requisitos para a obtenção do título de Mestre em Ciências, Programa de Pós-Graduação em Engenharia de Transportes. Área de Concentração: Planejamento e Operação de Sistemas de Transportes

Orientadora: Prof. Dra. Cira Souza Pitombo

São Carlos

2015

AUTORIZO A REPRODUÇÃO TOTAL OU PARCIAL DESTE TRABALHO,
POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO, PARA FINS
DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE.

S237e Santos, Murilo Castanho
Explorando técnicas para modelagem de dados
agregados de óbitos provenientes de acidentes por
automóvel / Murilo Castanho Santos; orientadora Cira
Souza Pitombo. São Carlos, 2015.

Dissertação (Mestrado) - Programa de Pós-Graduação
em Engenharia de Transportes e Área de Concentração em
Planejamento e Operação de Sistemas de Transporte --
Escola de Engenharia de São Carlos da Universidade de
São Paulo, 2015.

1. Previsão de Acidentes. 2. Árvore de Decisão. 3.
Taxas de óbitos. 4. Classificação. I. Título.

FOLHA DE JULGAMENTO

Candidato: Bacharel **MURILO CASTANHO DOS SANTOS**.

Título da dissertação: "Explorando técnicas para modelagem de dados agregados de óbitos provenientes de acidentes por automóvel".

Data da defesa: 01/10/2015

Comissão Julgadora:

Resultado:

Profa. Dra. **Cira Souza Pitombo**
(Orientadora)
(Escola de Engenharia de São Carlos/EESC)

Aprovado

Profa. Dra. **Simone Becker Lopes**
(Universidade Federal de Santa Catarina/UFSC)

APROVADO

Prof. Dr. **Flávio José Craveiro Cunto**
(Universidade Federal do Ceará/UFC)

Aprovado

Coordenador do Programa de Pós-Graduação em Engenharia de Transportes:
Prof. Associado **Paulo César Lima Segantine**

Presidente da Comissão de Pós-Graduação:
Prof. Associado **Paulo César Lima Segantine**

*Dedicado às pessoas que trabalham
por um trânsito mais seguro.*

AGRADECIMENTOS

Agradeço aos meus pais pela educação, incentivo e apoio ao longo dos anos de estudo, e aos meus irmãos pela colaboração sempre que precisei.

Gostaria de agradecer imensamente a minha orientadora, Professora Cira Pitombo, por não medir esforços em auxiliar da melhor maneira possível, e pela dedicação e disponibilidade ao longo do período de pesquisa.

Quero agradecer ao Professor Flávio Cunto e Professora Simone Lopes pelas recomendações ao trabalho realizado.

Agradeço ao Professor Cassiano Isler, Professor André Cunha e Thaís Guerreiro pela consultoria, habilidade técnica e imenso auxílio neste processo.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) e à Universidade de São Paulo (USP) pelo apoio financeiro.

Aos professores do Departamento de Engenharia de Transportes (STT) por compartilharem experiências e conhecimento, e aos funcionários que estiveram dispostos a atender e ajudar.

Agradeço ao dileto Conrado Plaza pela amizade e estímulo durante este período, e ao amigo Djalma Ferraz pela companhia agradável.

Quero agradecer às amigas Thalita Nascimento e Angélica Oliveira pelos vários momentos alegres, pela influência positiva e por estarem presentes desde o início do mestrado.

Aos companheiros de sala Igor Silveira, Diego Camargo e Miguel Castillo pelos bons encontros e momentos de estudo.

À amiga Monique Gomes pelo carinho, conselhos e apoio, e à amiga Andressa NG pelo cuidado e disposição constante durante este processo.

Ao amigo Bruno Medeiros por ser solícito neste período e pelo dinamismo.

Ao Evert Bacchini pela atenção com minha saúde.

Agradeço à minha avó Zilda Castanho pelo otimismo transmitido, aos demais familiares que torceram e me ajudaram de alguma forma, e à amiga da família Arlete Ouvidor pela preocupação.

Muitas pessoas ainda estiveram presentes nesta jornada, quero destacar os amigos: Paulo Oliveira, Rayanne Estrella, Jéssica Silva, Thiago Ramos, Felipe Bethonico, Lucas Verdade, Heymar Suárez, Andrea Nicola, Artur Paiva, Fernando Piva, Marcela Navarro, Fábio Vieira, Matheus Silva, Júlia Savietto, Leonardo Cadurin, Andrise Klug, Juliane Fürst, Isabela Tagomori, Felipe Furlanetto, Lucas Godoy e Natalia Macias.

O valor das coisas não está no tempo que elas duram,
mas na intensidade com que acontecem.
Por isso existem momentos inesquecíveis,
coisas inexplicáveis e pessoas incomparáveis.

Fernando Pessoa

RESUMO

SANTOS, M. C. **Explorando técnicas para modelagem de dados agregados de óbitos provenientes de acidentes por automóvel.** 119 p. Dissertação de Mestrado. Escola de Engenharia de São Carlos, Universidade de São Paulo, 2015.

Esta dissertação se baseia na exploração de técnicas para modelagem de óbitos provenientes de acidentes por automóvel no estado de São Paulo. A análise foi agregada por área, e utilizou a razão de óbitos por população, por área e por fluxo veicular como variáveis dependentes e as variáveis independentes foram características socioeconômicas, área, frota de veículos, IDHM, fluxo veicular anual e distâncias entre microrregiões. Os dados do ano 2000 foram utilizados na calibração e dados de 2010 na validação dos modelos, com a técnica de mineração de dados (algoritmos de Árvore de Decisão - AD: CART - *Classification And Regression Tree* e CHAID - *Chi-squared Automatic Interaction Detection*) e Regressão Linear Múltipla (RLM) para fins comparativos com os modelos de AD. A partir dos resultados verifica-se que a RLM foi a técnica que obteve melhores erro médio, erro médio absoluto e coeficiente de correlação, e o algoritmo CART da AD o menor erro médio normalizado. Ao comparar as taxas de óbitos, a relação por área apresentou melhor erro médio e coeficiente de correlação, já a relação por população obteve menor erro médio normalizado e erro médio absoluto. Vale ressaltar que os algoritmos de AD são técnicas adequadas para classificação de áreas segundo faixas de valores de variáveis explicativas e valores médios da variável objeto de estudo. Além disso, tais técnicas são mais flexíveis em relação a alguns pressupostos de modelos de regressão. Dessa forma, a principal contribuição deste trabalho consiste na exploração de tais algoritmos para previsão de acidentes e classificação de regiões.

Palavras-chave: Previsão de Acidentes; Árvore de Decisão; Taxas de óbitos; Classificação.

ABSTRACT

SANTOS, M. C. **Exploring techniques for modeling of aggregates data from deaths automobile accidents**. 119 p. São Carlos School of Engineering, University of São Paulo, 2015.

This dissertation is based on techniques exploration for modeling of deaths from automobile accidents on the state of São Paulo. The analysis was aggregated by area, and used the ratio of deaths per population, by area and by vehicle flow as dependent variables and the independent variables were socioeconomic characteristics, area, vehicle fleet, Municipal Human Development Index (MHDI), annual vehicle flow and distances between micro-regions. The 2000 data were used for calibration and 2010 data to validate the models with data mining technique (decision tree - DT algorithms: CART - Classification And Regression Tree and CHAID - Chi-squared Automatic Interaction Detection) and Multiple Linear Regression (MLR) for comparative purposes with the DT models. From the results it appears that the RLM was the technique that achieved better mean error, mean absolute error and correlation coefficient values, while the CART algorithm presented the lowest value of mean normalized error. When comparing death rates, a relation by area showed better mean error and correlation coefficient values, as the ratio by population had lower mean normalized error and mean absolute error values. It is noteworthy that the DT algorithms are suitable techniques for classification of areas in accordance with explanatory variables of value ranges and average values of the variable object of study. Furthermore, such techniques are more flexible compared to some assumptions regression models. Thus, the main contribution of this study is the exploration of such algorithms for prediction of accidents and regions classification.

Keywords: Accident Prediction; Decision Tree; Death Rates; Classification.

LISTA DE FIGURAS

Figura 1.1 - Fluxograma das etapas da pesquisa	27
Figura 3.1 - Exemplo de visualização de uma Árvore de Decisão	42
Figura 3.2 - Exemplo da estrutura de árvore do algoritmo CHAID	43
Figura 3.3 - Exemplo da estrutura de árvore do algoritmo CART.....	44
Figura 4.1 - Microrregiões do estado de São Paulo	47
Figura 4.2 - Esquema da classificação dos acidentes por automóvel na CID-10.....	49
Figura 4.3 - Esquema do cálculo do IDHM (PNUD, 2013)	53
Figura 4.4 - Etapas do processo de obtenção do fluxo veicular sintético	54
Figura 4.5 - Rede rodoviária do estado de São Paulo por triangulação	55
Figura 4.6 - Volume de tráfego estimado no estado de SP no ano 2010	56
Figura 4.7 - Segmentação da rede rodoviária do estado de SP, dentro dos municípios	57
Figura 4.8 - Fluxo de veículos proporcional nos municípios (veic.km)	58
Figura 4.9 - Mapas temáticos dos valores das variáveis dependentes	61
Figura 5.1 - Identificação dos valores de cada nó da Árvore de decisão (<i>output</i> do IBM SPSS 22.0)	67
Figura 5.2 - Árvore de decisão de Treinamento do Algoritmo CHAID para óbitos por população (amostra 2000)	71
Figura 5.3 - Árvore de decisão de Teste do Algoritmo CHAID para óbitos por população (amostra 2010)	72
Figura 5.4 - Árvore de decisão de Treinamento do Algoritmo CHAID para óbitos por área (amostra 2000).....	74
Figura 5.5 - Árvore de decisão de Teste do Algoritmo CHAID para óbitos por área (amostra 2010).....	75
Figura 5.6 - Árvore de decisão de Treinamento do Algoritmo CHAID para óbitos por fluxo (amostra 2000)	77
Figura 5.7 - Árvore de decisão de Teste do Algoritmo CHAID para óbitos por fluxo (amostra 2010).....	78
Figura 5.8 - Árvore de decisão de Treinamento do Algoritmo CART para óbitos por população (amostra 2000)	80

Figura 5.9 - Árvore de decisão de Teste do Algoritmo CART para óbitos por população (amostra 2010)	81
Figura 5.10 - Árvore de decisão de Treinamento do Algoritmo CART para óbitos por área (amostra 2000)	83
Figura 5.11 - Árvore de decisão de Teste do Algoritmo CART para óbitos por área (amostra 2010)	84
Figura 5.12 - Resultados do Algoritmo CHAID para óbitos por população	87
Figura 5.13 - Resultados do Algoritmo CHAID para óbitos por área	89
Figura 5.14 - Resultados do Algoritmo CHAID para óbitos por fluxo	91
Figura 5.15 - Resultados do Algoritmo CART para óbitos por população.....	94
Figura 5.16 - Resultados do Algoritmo CART para óbitos por área.....	97
Figura 5.17 - Análise gráfica de resíduos do modelo para variável dependente óbitos por população	100
Figura 5.18 - Análise gráfica de resíduos do modelo para variável dependente óbitos por área	101
Figura 5.19 - Análise gráfica de resíduos do modelo para variável dependente óbitos por fluxo.....	102

LISTA DE TABELAS

Tabela 2.1 – Resumo de pesquisas anteriores em previsão de acidentes	38
Tabela 3.1 - Técnicas de mineração de dados, adaptado de Goldschmidt e Passos (2005).....	39
Tabela 3.2 - Características dos algoritmos de AD: CHAID e CART	45
Tabela 4.1 - Medidas descritivas dos óbitos por acidente de automóvel nas microrregiões do estado de SP	49
Tabela 4.2 - Medidas descritivas das variáveis demográficas e socioeconômicas das microrregiões do estado de SP	50
Tabela 4.3 - Medidas descritivas da frota veicular das microrregiões do estado de SP	51
Tabela 4.4 - Medidas descritivas da média de IDHM das microrregiões do estado de SP	52
Tabela 4.5 - Medidas descritivas do fluxo veicular anual nas microrregiões do estado de SP	58
Tabela 4.6 - Medidas descritivas das distâncias entre microrregiões do estado de SP	59
Tabela 4.7 - Medidas descritivas das taxas de óbitos por automóvel	60
Tabela 4.8 - Medidas descritivas das taxas de óbitos por automóvel pela média	60
Tabela 4.9 - Medidas descritivas das variáveis independentes pela média	62
Tabela 4.10 - Testes de normalidade na variável transformada.....	64
Tabela 5.1 - Escala da taxa média de óbitos	85
Tabela 5.2 - Classificação das Microrregiões segundo o algoritmo CHAID para óbitos por população.....	86
Tabela 5.3 - Classificação das Microrregiões segundo o algoritmo CHAID para óbitos por área.....	88
Tabela 5.4 - Classificação das Microrregiões segundo o algoritmo CHAID para óbitos por fluxo	90
Tabela 5.5 - Classificação das Microrregiões segundo o algoritmo CART para óbitos por população.....	92
Tabela 5.6 - Classificação das Microrregiões segundo o algoritmo CART para óbitos por área.....	95

Tabela 5.7 - Resultados da Análise de RLM <i>stepwise</i> (ajuste dos modelos).....	98
Tabela 5.8 - Resultados do valor <i>t</i> da Análise de RLM.....	99
Tabela 5.9 - Comparação dos resultados do algoritmo CHAID	103
Tabela 5.10 - Comparação dos resultados do algoritmo CART.....	104
Tabela 5.11 - Comparação dos resultados da RLM	104
Tabela 5.12 - Comparação dos resultados da taxa de óbitos por população	105
Tabela 5.13 - Comparação dos resultados da taxa de óbitos por área	105
Tabela 5.14 - Comparação dos resultados da taxa de óbitos por fluxo	105
Tabela 5.15 - Resumo do resultado final da análise	106

LISTA DE ABREVIATURAS E SIGLAS

AASHTO	<i>American Association of State Highway and Transportation Officials</i>
AD	Árvore de Decisão
CART	<i>Classification and Regression Tree</i>
CHAID	<i>Chi-square Automatic Interaction Detector</i>
CID-10	Décima Revisão da Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde
CMFs	<i>Crash Modification Factors</i>
CONTRAN	Conselho Nacional de Trânsito
DATASUS	Departamento de Informática do SUS
DENATRAN	Departamento Nacional de Trânsito
HSM	<i>Highway Safety Manual</i>
IBGE	Instituto Brasileiro de Geografia e Estatística
IDH	Índice de Desenvolvimento Humano
IDHM	Índice de Desenvolvimento Humano Municipal
K-S	<i>Kolmogorov-Smirnov</i>
MD	Mineração de Dados
MS	Ministério da Saúde
PIB	Produto Interno Bruto
PNUD	Programa das Nações Unidas para o Desenvolvimento
RLM	Regressão Linear Múltipla
RNA	Redes Neurais Artificiais
SIM	Sistema de Informações sobre Mortalidade
SP	São Paulo
SPFs	<i>Safety Performance Functions</i>
S-W	<i>Shapiro-Wilk</i>
VDMA	Volume Diário Médio Anual
WHO	<i>World Health Organization</i>
ZI	<i>Zero-Inflated</i>

SUMÁRIO

1	INTRODUÇÃO	23
1.1	CONSIDERAÇÕES INICIAIS	23
1.2	PROBLEMA DE PESQUISA E JUSTIFICATIVA	24
1.3	OBJETIVOS	25
1.3.1	Objetivos Específicos	26
1.4	SÍNTESE DO MÉTODO E ESTRUTURA DO TRABALHO	26
2	PREVISÃO DE ACIDENTES	29
2.1	CONSIDERAÇÕES SOBRE ACIDENTES DE TRÂNSITO	29
2.2	MODELOS DE PREVISÃO DE ACIDENTES	30
2.2.1	Modelo de Distribuição de Poisson	31
2.2.2	Modelo de Distribuição Binomial Negativa	31
2.2.3	Modelos Inflacionados de Zeros (ZI)	33
2.2.4	Técnicas de Mineração de dados (MD)	33
2.3	PREVISÃO DE ACIDENTES DESAGREGADA	34
2.4	PREVISÃO DE ACIDENTES AGREGADA	36
2.5	SÍNTESE DE CITAÇÕES	38
3	TÉCNICAS ABORDADAS	39
3.1	MINERAÇÃO DE DADOS (MD)	39
3.1.1	Árvore de Decisão (AD)	41
3.2	REGRESSÃO LINEAR MÚLTIPLA	45
3.2.1	Modelo de Regressão Linear Múltipla (RLM)	45
4	DADOS	47
4.1	CARACTERIZAÇÃO DA ÁREA DE ESTUDO	47
4.2	VARIÁVEIS E AMOSTRA FINAL	48
4.2.1	Dados de óbitos	48
4.2.2	Dados do IBGE	49
4.2.3	Dados da frota veicular	50
4.2.4	Dados de IDHM	51

4.2.5	Dados de fluxo veicular anual sintético	53
4.2.6	Dados de distâncias entre Microrregiões	59
4.2.7	Mensuração de óbitos por acidentes de trânsito	59
4.2.8	Testes de normalidade	63
5	RESULTADOS E DISCUSSÕES	67
5.1	ÁRVORES DE DECISÃO (AD).....	68
5.1.1	CHAID	68
5.1.2	CART	78
5.2	CLASSIFICAÇÃO DAS MICRORREGIÕES	85
5.3	REGRESSÃO LINEAR MÚLTIPLA (RLM).....	97
5.3.1	Resíduos	100
5.4	COMPARAÇÃO DOS RESULTADOS.....	102
6	CONCLUSÕES	107
6.1	SÍNTESE DO TRABALHO.....	107
6.2	PRINCIPAIS CONTRIBUIÇÕES E CONCLUSÕES	108
6.3	LIMITAÇÕES METODOLÓGICAS	109
6.4	RECOMENDAÇÕES PARA TRABALHOS FUTUROS.....	110
	REFERÊNCIAS BIBLIOGRÁFICAS	113

1 INTRODUÇÃO

1.1 CONSIDERAÇÕES INICIAIS

A segurança viária é uma das áreas da engenharia de transportes que merece monitoramento contínuo, de forma a mitigar o maior número possível de acidentes de trânsito. E, a perda de vidas tem impacto negativo para a sociedade. Por outro lado, a prevenção dos acidentes contribui para a redução de custos médicos, sendo um benefício para a economia de um país.

Os acidentes de trânsito são estimados como a oitava causa de morte no mundo, com um risco semelhante ao causado por muitas doenças. Em torno de 1,24 milhão de pessoas morrem todos os anos nas rodovias de todo o mundo, e, aproximadamente, entre 20 e 50 milhões sofrem lesões resultantes de colisões no trânsito. Os jovens com idade economicamente ativa, entre 15 e 29 anos, são as vítimas mais comuns (WHO, 2013).

Dados do Ministério da Saúde (MS) mostram que no Brasil, no ano de 2012, o número de ocorrências de mortes no trânsito foi 44.812. O número dessas mortes cresceu a uma taxa média anual de 3,77%, correspondente ao período entre os anos 2000 e 2012 (DATASUS, 2014).

De acordo Ferraz et al. (2012), em um mesmo país, os índices de acidentes de trânsito variam significativamente entre diferentes estados, regiões ou cidades, dependendo, dentre outros fatores, do desenvolvimento econômico e social.

São Paulo é o estado brasileiro com a maior economia e o mais populoso, concentra 21,56% dos habitantes do país (IBGE, 2010). O estado também segue na frente com os índices de mortalidade no trânsito. Em 2012 foram 7.003 acidentes fatais, taxa de 16,71 por cem mil habitantes (DATASUS, 2014).

Conhecer os fatores que influenciam o grande número de mortes provenientes de acidentes de trânsito é uma forma de incentivo em priorizar políticas públicas, voltadas para segurança viária.

O combate à acidentalidade no trânsito com base científica, portanto com maior eficiência e eficácia, no espaço geográfico estudado, somente é possível com a realização adequada das tarefas envolvidas na coleta de dados dos acidentes, constituição do banco de dados e processamento/sistematização/análise das informações (Ferraz et al., 2012).

A melhor compreensão e previsão de acidentes podem ser alcançadas através da estatística clássica e também métodos computacionais mais recentes. A mineração de dados é um exemplo de análise que explora grande conjunto de dados, a fim de encontrar padrões entre as variáveis.

O desenvolvimento de mineradores de dados emergiu a partir da década de 1990 com o objetivo de obter informações úteis a partir de um grande conjunto de dados através de regras que caracterizam padrões de ocorrência (Mannila, 1997). Dentre as técnicas de mineração de dados, as árvores de decisão podem ser usadas em muitos problemas, tanto para a exploração e classificação, quanto para a previsão. A Árvore de Decisão (AD) produz modelos que apresentam, de forma gráfica, as relações das variáveis independentes com a variável dependente e representam cenários de um problema real.

A Árvore de Decisão é uma técnica simples de classificação das relações entre atributos do conjunto de dados. Uma AD permite classificar uma base de dados em um número finito de classes através de regras hierárquicas e da sua divisão em grupos, organizando os dados de maneira compacta e permitindo uma visão geral da natureza do processo (Quinlan, 1983).

1.2 PROBLEMA DE PESQUISA E JUSTIFICATIVA

Os principais fatores de risco relacionados à severidade de acidentes causadores de lesões e mortes no trânsito, que precisam de legislação mais rígida, são os seguintes: velocidade alta, álcool ao volante, não utilização de cinto de segurança, de capacetes pelos motociclistas e de cadeiras de segurança para crianças (WHO, 2013).

Assim, muitos acidentes de trânsito são, em grande parte, causados por erros humanos, ao invés de fatores relacionados à via. A modelagem de um segmento ou interseção é um desafio, por conta de componentes aleatórios de falhas não explicadas pelas características das vias locais. A um nível mais agregado, os acidentes estão relacionados com dados demográficos, características de condutores mais agressivos, etc., sem a necessidade de atribuir falhas aos links ou segmentos das vias (Washington et al., 2006).

No estudo de planejamento de transportes um dos objetivos é melhorar a segurança da rede de transportes e antecipar os problemas antes que surjam. Assim, o planejador seria capaz de minimizar a quantidade de mortes no trânsito.

Dessa forma, o problema da pesquisa é a definição das técnicas estatísticas que identificam os fatores envolvidos nos acidentes fatais, sobretudo considerando análises agregadas, auxiliando decisões estratégicas regionais. Além da importância de incluir novas técnicas capazes de identificar os fatores demográficos e agregados significativos que influenciam as ocorrências de acidentes.

A principal justificativa para o desenvolvimento deste trabalho está em entender a relação entre variáveis agregadas e seus impactos nos óbitos provenientes de acidentes de trânsito. O método procura utilizar variáveis que envolvam dados de fácil acesso. Os resultados podem ser usados no processo de planejamento e implantação de políticas públicas a longo prazo. Além disso, tal método pode ser aplicado a diversas regiões com a finalidade de investigações agregadas.

1.3 OBJETIVOS

O objetivo principal deste trabalho de pesquisa é explorar técnicas de mineração de dados (algoritmos de Árvore de Decisão) para contribuir na identificação das variáveis explicativas de óbitos provenientes de acidentes por automóvel em uma análise agregada.

1.3.1 Objetivos Específicos

Também são objetivos desse trabalho:

- Explorar formas de mensurar óbitos provenientes de acidentes de automóvel numa análise agregada por área.
- Contribuir para o entendimento dos acidentes e compreender a influência das variáveis estudadas nos acidentes;
- Classificar as microrregiões segundo aspectos sociodemográficos e óbitos provenientes de acidentes segundo algoritmos de AD.
- Propor um procedimento para estimação de fluxo veicular anual sintético através de uma rede rodoviária triangulada.

1.4 SÍNTESE DO MÉTODO E ESTRUTURA DO TRABALHO

O presente documento apresenta procedimentos metodológicos e resultados obtidos na tentativa de modelagem de óbitos por acidentes de trânsito, com base em informações socioeconômicas/demográficas e de frota, numa análise agregada. Este texto é formado por essa introdução, e outros seis capítulos.

O capítulo 2 e o capítulo 3 apresentam a revisão da literatura e referencial teórico que serviram de base para a obtenção do conhecimento necessário para o andamento desse trabalho. O primeiro descreve modelos de previsão de acidentes e o segundo define as técnicas abordadas.

Em seguida, o capítulo 4 descreve os dados utilizados na pesquisa, faz a caracterização da área de estudo, identifica as variáveis, amostra final e testes estatísticos realizados previamente à aplicação das técnicas.

No capítulo 5 são apresentados os resultados e análises dos modelos desenvolvidos. O método de análise consiste em calibrar e validar modelos, utilizando a técnica de Árvore de Decisão, e comparar os valores encontrados com modelos lineares tradicionais (RLM).

O capítulo 6 apresenta as conclusões, observações finais e sugestões para trabalhos futuros.

No capítulo 7 são listadas as referências bibliográficas, citadas neste documento.

As etapas da pesquisa, que representam os próximos capítulos desta dissertação, são apresentadas na Figura 1.1.

Figura 1.1 - Fluxograma das etapas da pesquisa

2 PREVISÃO DE ACIDENTES

Conhecer as características relacionadas à acidentalidade e construir um banco de dados consistente são estratégias importantes para o desenvolvimento de projetos que visam à redução dos diversos tipos de acidentes. Este capítulo traz uma breve revisão da literatura relacionada à segurança no trânsito e modelos de previsão de acidentes de trânsito.

2.1 CONSIDERAÇÕES SOBRE ACIDENTES DE TRÂNSITO

Considera-se como acidente de trânsito um evento envolvendo um ou mais veículos, motorizados ou não, em movimento por uma via, que provoca ferimentos em pessoas e/ou danos em veículos e/ou em outros elementos (postes, edificações, sinais de trânsito, etc.) (Ferraz et al., 2012). Os acidentes de trânsito são raros e aleatórios, pois representam apenas uma pequena parcela do número total de eventos que ocorrem no sistema de transporte e são influenciados por vários fatores, parte determinístico (que pode ser controlado) e parte estocástico (aleatório e imprevisível) (AASHTO, 2010).

O tipo de exposição ao risco de um acidente de trânsito é um aspecto que influencia o número de pessoas mortas ou feridas. Medir a exposição ao risco do acidente de uma forma satisfatória, não é tarefa fácil. Um indicador da probabilidade de ocorrência de acidentes é a taxa de acidentes, que indica o risco de acidentes por unidade de exposição. Um fator de risco para acidentes é qualquer fator que aumenta a probabilidade de ocorrência de acidentes. A gravidade de um acidente pode variar desde uma pequena colisão entre veículos, a desastres com vítimas fatais (Elvik et al., 2009). Nesta dissertação, será feita a avaliação de alguns dos fatores que contribuem para os óbitos provenientes de acidentes por automóvel de forma adequada.

Enquanto as taxas de mortalidade no trânsito diminuem em alguns países de alta renda, em países de renda baixa e média o aumento de mortes ainda é constante. A tendência aponta que os acidentes de trânsito se tornem a quinta principal causa de

morte em 2030, com a evidente diferença entre países de alta e de baixa renda (WHO, 2013). Independente do país, o foco deve ser reduzir os óbitos no trânsito.

A base de uma pesquisa sobre segurança viária depende de fontes de dados existentes, e a análise estatística dos dados de acidentes auxilia o desenvolvimento de políticas de segurança no trânsito para salvar vidas e reduzir a gravidade das lesões. Apesar das limitações dos dados, o desenvolvimento de pesquisas tenta melhorar a compreensão dos fatores que influenciam a ocorrência de óbitos e lesões no trânsito (Mannering e Bhat, 2014). Os modelos de previsão de acidentes indicam possíveis resultados futuros de acidentes e quais fatores merecem atenção para o aumento da segurança no trânsito.

2.2 MODELOS DE PREVISÃO DE ACIDENTES

A previsão da ocorrência de acidentes de trânsito é geralmente feita através de modelos matemáticos que, de uma maneira geral, determinam o valor esperado de uma taxa de acidentes, em função de determinadas variáveis que possuem influência significativa para que os acidentes aconteçam (Cardoso e Goldner, 2007). As estimativas de acidentes são necessárias para avaliar em quais locais adotar e aplicar medidas de redução de acidentes.

As perdas para a sociedade resultantes de acidentes de veículos motorizados têm sido pesquisadas continuamente, de forma a obter uma melhor compreensão dos fatores que afetam a ocorrência de acidentes, na esperança de que serão capazes de prever melhor a probabilidade de falhas e guiar políticas e medidas destinadas a reduzir o número de acidentes. Normalmente, dados detalhados dos fatores de risco associados ao comportamento do condutor (alta velocidade, percepção e reação, falta de habilidade, etc.) e ao veículo não estão disponíveis. Com isso, as pesquisas são encaminhadas a estudar os fatores que afetam o número de acidentes que ocorrem em algum espaço geográfico (geralmente um segmento viário ou interseção) durante algum período de tempo especificado (semana, mês, ano, número de anos) (Lord e Mannering, 2010).

Esta dissertação apresenta dados de acidentes baseados em taxas e sugere a aplicação de modelos de previsão de forma agregada em microrregiões (unidades de área). De acordo com a finalidade da pesquisa, a análise de acidentes pode ser feita a um nível mais abrangente, no caso de zonas, ou de forma mais local, como em interseções. A seguir são apresentados alguns dos modelos de previsão desenvolvidos para segurança viária, além de pesquisas realizadas com modelos de forma desagregada e agregada de dados.

2.2.1 Modelo de Distribuição de Poisson

A distribuição de Poisson utiliza dados de contagem, que consistem em valores inteiros não negativos e são encontrados com frequência na modelagem de fenômenos relacionados ao planejamento de transporte, como a ocorrência de acidentes e número de veículos que esperam em uma fila. Uma das condições da distribuição de Poisson é que a média seja igual à variância, sendo ao mesmo tempo uma limitação do modelo. Quando a variância é significativamente maior do que a média, ocorre a superdispersão dos dados (Washington, Karlaftis e Mannering, 2003).

A distribuição de Poisson bivariada é adequada para a modelagem de duas variáveis aleatórias com dependência positiva. O modelo bivariado utiliza dados de contagem, que devem ser correlacionados. Além disso, é capaz de lidar tanto com superdispersão quanto subdispersão dos dados (Lao et al., 2011).

Para dados de contagem que envolvem mais de duas variáveis aleatórias que sejam correlacionadas, utiliza-se a distribuição multivariada. A vantagem deste método é permitir a dependência tanto positiva quanto negativa entre as variáveis. A limitação do modelo ocorre quando há grande quantidade de zeros nas observações (Castro, Paleti e Bhat, 2012).

2.2.2 Modelo de Distribuição Binomial Negativa

A modelagem de acidentes de trânsito mais utilizada pelos analistas de segurança viária é a distribuição de Poisson e a Binomial Negativa. Uma vez que ocorre

superdispersão nos dados, o modelo binomial negativo oferece forma de ajuste. Porém o modelo não é adequado para casos de subdispersão e para amostras de tamanho médio e pequeno (Geedipally e Lord, 2008; Lord e Mannering, 2010).

2.2.2.1 Modelo do *Highway Safety Manual* (HSM)

O *Highway Safety Manual* (Manual de Segurança Viária) publicado em 2010 pela *American Association of State Highway and Transportation Officials* (AASHTO), fornece conhecimentos sobre segurança e ferramentas para facilitar uma melhor tomada de decisão. Parte do HSM é dedicada a modelos de previsão de acidentes e reúne informações e metodologias para medir, estimar e avaliar a frequência de acidentes (número de acidentes por ano) e gravidade dos acidentes (nível de lesões por causa dos acidentes) das rodovias.

Os modelos do HSM são mais aplicáveis quando desenvolvidos para avaliar soluções de um local específico, por um determinado período de tempo. Ainda assim, é possível estimar o número de acidente em situações em que não possuem um histórico de dados ou não existe um modelo de previsão disponível.

Os modelos de previsão do HSM variam de acordo com o tipo de rodovia (pista simples, pista dupla, arteriais urbanas e suburbanas), mas todos têm os mesmos elementos básicos:

- Funções de Desempenho de Segurança (SPFs – *Safety Performance Functions*): modelos estatísticos são utilizados para estimar a frequência média de acidente para um tipo de rodovia com as condições de base;
- Fatores de Modificação de Acidentes (CMFs – *Crash Modification Factors*): CMFs são usados para levar em conta as características que diferem das condições base e as peculiaridades de segmentos específicos;
- Fator de Calibração (C): multiplicado na frequência de acidente previsto pelo SPF para explicar as diferenças entre o período de tempo para o qual foram desenvolvidos os modelos de previsão e período de tempo a que são aplicados os modelos do HSM.

Embora a forma funcional das SPFs varie no HSM, o modelo de previsão para estimar a frequência média de acidente $N_{previsto}$, é geralmente calculado usando a Equação 2.1.

$$N_{previsto} = N_{SPFx} \times (CMF_{1x} \times CMF_{2x} \times \dots \times CMF_{yx}) \times C_x \quad (2.1)$$

em que:

$N_{previsto}$: número previsto de acidentes para um determinado ano e local do tipo x (acidentes/ano);

N_{SPFx} : número médio de acidentes previstos para a condição ideal considerada no desenvolvimento de SPF para o local do tipo x (acidentes/ano);

CMF_{yx} : Fator de Modificação de Acidentes para condição y e local do tipo x;

C_x : Fator de Calibração para ajuste em condições locais em locais do tipo x.

2.2.3 Modelos Inflacionados de Zeros (ZI)

A frequência de acidentes de trânsito em uma seção rodoviária ou em uma interseção (durante algum período de tempo) frequentemente apresenta zeros em excesso. O modelo ZI (do inglês *zero-inflated*) geralmente fornece melhor ajuste estatístico para dados com aparentes excessos de zeros em comparação com os modelos de Poisson e binomial negativo. O predomínio de zeros observados leva a considerar a existência de dois estados: um onde a probabilidade de acidentes é tão baixa que não pode ser estatisticamente distinta de zero e o outro estado que é de contagem normal que inclui zeros e inteiros positivos (Lord, Washington e Ivan, 2005; Lord, Washington e Ivan, 2007; Malyskina e Mannering, 2010).

2.2.4 Técnicas de Mineração de dados (MD)

Diferente de muitos modelos estatísticos, as técnicas de mineração de dados, descritas mais detalhadamente no capítulo subsequente, são capazes de detectar padrões *a priori* desconhecidos. São descritas, em seguida, aplicações de modelagem de acidentes através de duas técnicas de MD: Redes Neurais Artificiais (RNA) e Árvore de Decisão (AD).

A RNA não tem suposições de modelo pré-definidas entre as variáveis dependentes e independentes. Se estes pressupostos são violados, o modelo poderia levar a uma estimativa errada da probabilidade de acidentes. A RNA é uma poderosa ferramenta matemática capaz de estabelecer um mapeamento não linear entre um conjunto de entradas e uma ou várias saídas. A abordagem da rede neural demonstrou ser bem sucedida ao lidar com problemas de previsão e de classificação, desde que existam dados suficientes disponíveis. No entanto, o modelo criado pela rede neural não pode ser aplicado em outro conjunto de dados (Riviere et al., 2006). O estudo de Chang (2005) comparou o desempenho de modelos de previsão de acidentes utilizando regressão binomial negativa e RNA, em uma das rodovias mais importante de Taiwan, com intenção de verificar a relação entre os acidentes com veículos e a geometria da rodovia, as características de tráfego e condições ambientais. Na análise dos dados de treinamento o modelo de RNA apresentou melhor desempenho que o modelo de regressão binomial negativa. De acordo com a pesquisa desenvolvida o desempenho do modelo de RNA pode chegar a aproximadamente 60% de acerto.

Para identificar os principais fatores que contribuem na ocorrência de acidentes graves, Oña, López e Abellán (2013) mostram uma aplicação com a construção de árvores de decisão (AD). A análise de acidentes nas rodovias rurais, na província de Granada (Espanha) entre 2003 e 2009 mostrou que os métodos utilizados na construção de AD, com os algoritmos CART e C4.5, permitem a classificação dos acidentes com base na severidade e que gestores implantem ações prioritárias de acordo com a classificação dos acidentes. Os dois algoritmos indicaram que as mulheres têm maior risco de acidentes severos em condições de iluminação inexistente ou insuficiente. Os modelos de AD são alternativos a modelos paramétricos, pois identificam padrões de dados e podem ser usados para determinar as interações entre as variáveis, que modelos estatísticos tradicionais não encontram.

2.3 PREVISÃO DE ACIDENTES DESAGREGADA

Nesta seção serão apresentados alguns dos trabalhos que desenvolveram modelos de previsão de acidentes de forma desagregada, independente da técnica utilizada.

O objetivo da síntese é possibilitar o conhecimento de estudos que estejam próximos ao desenvolvido na dissertação.

Uma alternativa mais adequada, para Chin e Quddus (2003), seria o modelo de efeito aleatório binomial negativo, capaz de lidar com os efeitos espaciais e temporais dos dados. Os autores examinaram a relação entre a ocorrência de acidentes e as características dos cruzamentos sinalizados em Cingapura. As variáveis encontradas, que são altamente significativas, associadas a uma maior ocorrência de acidentes totais são o volume total nas aproximações, o número de fases por ciclo, faixa da esquerda não controlada e a presença de câmera de vigilância.

Para explicar que a frequência e a gravidade dos acidentes de diferentes tipos variam de acordo com os fenômenos, Jonsson, Ivan e Zhang (2007) usaram modelagem linear generalizada para estimar o número de acidentes por: (a) distribuição da severidade, (b) dependência de fluxo de tráfego, e (c) variáveis que melhor explicam a variação do local de ocorrência de diferentes tipos de acidentes, em cruzamentos de rodovias rurais de quatro faixas na Califórnia, Estados Unidos. O estudo concluiu que existem diferenças evidentes na severidade, bem como na relação entre fluxo e vários tipos de colisão. Interseções e colisões em sentidos opostos são mais severas que acidentes na mesma direção. Os acidentes com veículos em direções opostas e em mesma direção quando relacionados ao fluxo, apresentaram semelhanças, mas verificam-se diferenças quando comparados a acidentes de um único veículo.

Silva (2012) avaliou os resultados da aplicação do modelo de previsão de acidentes apresentado no HSM, para trechos de rodovias de pista simples, localizados no interior do estado de São Paulo. Os valores resultantes da aplicação do método do HSM calibrado e do método de Bayes, em todas as análises realizadas, aproximaram-se dos valores observados. Em outras regiões do país esses valores podem ser empregados como primeira aproximação, pois é recomendada a realização de estudos específicos regionais para a aplicação do modelo.

O estudo de modelagem feito por Cunto, Castro Neto e Barreira (2012) verificou que para a estimação do número total de acidentes de trânsito em interseções semaforizadas de Fortaleza, utilizando a distribuição binomial negativa, o modelo com as variáveis explicativas volume diário médio anual e número de faixas apresentou desempenho mais adequado. Para a estimação do número de acidentes de trânsito com vítimas feridas e fatais, entretanto, nenhuma das variáveis investigadas foi significativa.

2.4 PREVISÃO DE ACIDENTES AGREGADA

Esta dissertação procura entender melhor a relação entre variáveis agregadas relacionadas aos óbitos provenientes de acidentes de trânsito. Desta forma, tal modelo seria classificado como estratégico, auxiliando o planejamento regional. Com isso, os estudos mostrados a seguir mostram relações de acidentes de trânsito a um nível agregado.

Os estudos de Hadayeghi, Shalaby e Persaud (2003) utilizaram modelagem linear generalizada para previsão de acidentes de zonas de tráfego da cidade de Toronto, Canadá. Foram calibrados modelos para o total de acidentes e por severidade (lesão fatal e não fatal) em função da demanda por viagens e de dados socioeconômicos e demográficos da rede. As variáveis que foram significativas na ocorrência de acidentes foram o número de domicílios, o número de quilômetros das principais rodovias, o número de quilômetros percorridos pelos veículos, a alta densidade em interseção, velocidade permitida, e a relação volume-capacidade.

Modelos de previsão de acidentes para a região metropolitana de Tucson, Estados Unidos, foram calibrados no estudo realizado por Guevara, Washington e Oh (2004). Todos os modelos produzem previsões de acidentes no nível de zona de tráfego. Os dados foram separados em óbito (lesão fatal), ferimentos e danos materiais nos acidentes. Para acomodar superdispersão nos dados, foram aplicados modelos de regressão binomial negativa. Para acomodar a simultaneidade de óbitos e de ferimentos nos resultados, foi realizada a estimativa simultânea dos modelos. As variáveis significativas dos modelos para ferimentos e danos materiais foram densidade populacional, número de pessoas com emprego, a densidade em

interseção, percentagem de milhas da via arterial principal, percentual de milhas de vias arteriais menores, e percentagem de milhas de vias coletoras. Estes modelos podem preencher lacunas nas atividades de nível de planejamento que queiram considerar a segurança futura e sirvam de apoio à criação de programas de incentivo para melhorar a segurança.

Lovegrove e Sayed (2007) investigaram modelos de previsão de acidentes ao nível macro em uma aplicação para melhorar os programas tradicionais de segurança em regiões urbanas e rurais de Vancouver, Canadá. Os autores utilizaram modelos de regressão binomial negativa, identificando que o uso de modelos a nível macro proporcionou atenção antecipada a locais propensos a acidentes em pelo menos um bairro e em outras zonas da cidade foram feitas a identificação e classificação de locais para diagnóstico.

Modelos de previsão de acidentes em zona de tráfego, em função das características de uso do solo, para a cidade de Charlotte, Estados Unidos, foram desenvolvidos por Pulugurtha, Duddu e Kotagiri (2013). Os modelos utilizaram distribuição binomial negativa e, características demográficas/socioeconómicas, tais como população, o número de domicílios e pessoas com emprego, indicadores de tráfego, tais como produções e atrações de viagem e característica da rede como limite de velocidade na faixa central foram observadas para serem correlacionadas ao uso do solo, e, portanto não foram consideradas no desenvolvimento de modelos de previsão de acidente a nível zonal. Os resultados indicam que as características de uso do solo estão associadas e desempenham um papel estatisticamente significativo na estimativa de acidentes.

Zhang et al. (2014) avaliaram a segurança de trânsito em zonas de tráfego, de forma que desenvolveram um modelo de regressão binomial negativa para estimar os fatores estatisticamente significativos em zonas inseguras. Os resultados do modelo indicam que a frequência de acidentes a nível zonal pode ser associada a fatores socioeconômicos, demográficos, e do sistema de transporte. Concluem que a avaliação de segurança e a estimativa de acidentes desempenham um papel importante na identificação de locais críticos, e potenciais melhoras na segurança.

Outros trabalhos exploram ainda outras abordagens para acidentes com dados agregados. A investigação adicional se faz necessária a depender do objetivo de estudo, uma vez que previsão de acidentes agregada é útil na identificação de fatores que influenciam os acidentes e em nível de planejamento da segurança no trânsito de forma regional.

2.5 SÍNTESE DE CITAÇÕES

A Tabela 2.1 apresenta os trabalhos citados, de acordo com a abordagem metodológica de pesquisas anteriores.

Tabela 2.1 – Resumo de pesquisas anteriores em previsão de acidentes

Método de Análise	Referências
Modelos de Previsão de Acidentes	Cardoso e Goldner (2007), Lord e Mannering (2010)
Modelo de Distribuição de Poisson/Modelo Bivariado/Multivariado	Washington, Karlaftis e Mannering (2003), Lao et al., (2011), Castro, Paleti e Bhat, (2012)
Modelo de Distribuição Binomial Negativa	Geedipally e Lord (2008), Lord e Mannering (2010), AASHTO, (2010)
Modelos Inflacionados de Zeros	Lord, Washington e Ivan (2005), Lord, Washington e Ivan (2007), Malyskhina e Mannering (2010)
Técnicas de Mineração de dados	Chang (2005), Riviere et al. (2006), Oña, López e Abellán (2013)
Previsão de Acidentes Desagregada	Chin e Quddus (2003), Jonsson, Ivan e Zhang (2007), Silva (2012), Cunto, Castro Neto e Barreira (2012)
Previsão de Acidentes Agregada	Hadayeghi, Shalaby e Persaud (2003), Guevara, Washington e Oh (2004), Lovegrove e Sayed (2007), Pulugurtha, Duddu e Kotagiri (2013), Zhang et al. (2014)

3 TÉCNICAS ABORDADAS

3.1 MINERAÇÃO DE DADOS (MD)

Uma grande quantidade de dados é armazenada todos os dias, e com os avanços tecnológicos é possível armazenar essa vasta quantidade de dados a um custo relativamente baixo. A necessidade de utilizar grande banco de dados motivou a exploração e análise para descobrir padrões *a priori* desconhecidos.

Mineração de dados (MD) é um conjunto de técnicas de exploração e análise de grandes quantidades de dados para descobrir padrões e regras significativas. Esse processo pode ser utilizado em diversas áreas, como medicina, administração, computação, etc. (Berry e Linoff, 2004).

As pesquisas e desenvolvimento de MD emergiram a partir dos anos 90. As principais técnicas da mineração de dados apresentadas por Goldschmidt e Passos (2005) são: Árvore de Decisão (AD), Redes Neurais, Raciocínio Baseado em Casos, Algoritmos Genéticos, Conjuntos *Fuzzy*, Regras de Indução e Regras de Associação. A Tabela 3.1 contém as técnicas e tarefas utilizadas na mineração de dados.

Tabela 3.1 - Técnicas de mineração de dados, adaptado de Goldschmidt e Passos (2005)

Técnica	Descrição	Tarefas	Algoritmos
Árvore de Decisão	Baseada em estágios de decisão (nós) e na separação de classes e subconjuntos, organiza os dados de forma hierárquica.	- Classificação - Predição	CART, CHAID, C5.0, ID-3
Redes Neurais	Modelos inspirados na fisiologia do cérebro, nos quais o conhecimento é fruto do mapa de conexões neuronais e dos pesos dessas conexões.	- Classificação - Agrupamento - Predição	Perceptron, Rede MLP, Redes de Kohonen, Rede Hopfield, Rede BAM, Rede ART, Rede IAC, Rede LVQ, Rede Counterpropagation, Rede RBF, Rede PNN, Rede Time Delay, Neocognitron, Rede BSB
Raciocínio Baseado em Casos	Baseado no método do vizinho mais próximo combina e compara atributos para estabelecer hierarquia de semelhança.	- Classificação - Agrupamento	BIRCH, CLARANS CLIQUE

Técnica	Descrição	Tarefas	Algoritmos
Algoritmos Genéticos	Métodos gerais de busca e otimização, inspirados na Teoria da Evolução, em que a cada nova geração, soluções melhores têm mais chance de ter "descendente".	- Classificação - Agrupamento	Algoritmo Genético Simples, Genitor, GA-Nuggets, GAPVMINER
Conjuntos Fuzzy	Oferece uma grande vantagem para classificar dados com um alto nível de abstração.	- Classificação - Agrupamento	K-means, FCMdd
Regras de Indução	Processo para obter uma hipótese a partir de dados e fatos já existentes.	- Classificação - Predição	CART, CHAID
Regras de Associação	Estabelece uma correlação estatística entre atributos de dados e conjuntos de dados.	- Associação	Apriori, AprioriTid, AprioriHybrid, AIS, SETM

Recentemente, verifica-se o aumento de trabalhos que descrevem a aplicação de tais técnicas nas áreas de planejamento de transportes e engenharia de tráfego (Xie et al., 2003; Rasouli e Nikras, 2013, Gonçalves et al., 2015; Mozolin et al., 2000). No caso desta pesquisa, o foco é a previsão de acidentes e classificação de microrregiões do estado de São Paulo, segundo variáveis explicativas e óbitos provenientes de acidentes por automóvel através de Árvore de Decisão.

Na MD, os dados são armazenados e a busca é automatizada por computador. Esclarecer os padrões encontrados, de forma inteligente em meio a grande quantidade de dados torna a mineração de dados a base para resolução de problemas complexos.

Berry e Linoff (2004) explicam que um modelo na MD é um algoritmo ou conjunto de regras. A tarefa de modelagem é encontrar regras que explicam os valores conhecidos da variável dependente. As tarefas que podem ser realizadas com as técnicas de MD são:

- Classificação
- Estimação
- Predição
- Regras de Associação
- *Clusters* (Agrupamentos)
- Identificação de perfis

Este trabalho utilizou como técnica principal a Árvore de Decisão, disponível no *software* IBM SPSS 22.0, para obter a relação existente entre as variáveis independentes, apresentadas no próximo capítulo, e os óbitos provenientes de acidentes de automóvel. Segue na próxima seção uma breve descrição da AD e dos algoritmos CHAID e CART, utilizados para classificação de microrregiões do estado de São Paulo e observação de relações entre variáveis explicativas e a variável objeto de estudo.

3.1.1 Árvore de Decisão (AD)

A AD é uma das técnicas de MD baseada na classificação ou predição. O processo é composto por duas etapas: treinamento e teste. Na fase de treinamento, o algoritmo escolhido classifica o conjunto de dados para análise através da divisão em grupos com características comuns e chega ao resultado da árvore. A fase de teste compara os valores previstos pela árvore aos valores previamente observados. A AD é uma técnica exploratória não paramétrica, sem suposições de modelos previamente definidas.

Quando uma Árvore de Decisão é usada para tarefas de classificação, é identificada por árvore de classificação. Quando é utilizado para tarefas de predição, é chamada de árvore de regressão. Quando a variável resposta é categórica, segue uma distribuição multinomial e trata a árvore como modelo de probabilidade. Se a variável resposta for numérica, como no caso deste trabalho, o modelo é conhecido como árvore de regressão e assume uma distribuição normal (gaussiana).

AD são geralmente representadas graficamente como estruturas hierárquicas, tornando-as mais fáceis de interpretar do que outras técnicas (Rokach e Maimon, 2008).

Cada segmento da árvore é chamado nó. O segmento que contém os dados de todos os elementos de análise é o nó raiz. Os nós seguintes, subdivisões do nó raiz, são chamados de nós filhos. Caso os nós não sejam mais divididos é denominado de nó folha ou nó terminal. O algoritmo usado para dividir os dados identifica as

variáveis independentes que fornecem máxima segregação dos dados segundo a variável dependente. Alguns dos algoritmos de árvores de decisão são o C4.5 (Quilan, 1993), CHAID (Kass, 1980), CART (Breiman et al., 1984) e QUEST. A Figura 3.1 mostra um exemplo da estrutura da AD.

Figura 3.1 - Exemplo de visualização de uma Árvore de Decisão

Os critérios para divisão dos dados são dependentes dos diferentes algoritmos de AD. Os algoritmos utilizados nesta pesquisa, para criação de Árvores de Decisão, são o CHAID (*Chi-square Automatic Interaction Detector*) e CART (*Classification and Regression Tree*), contidos no *software* IBM SPSS 22.0. A AD é capaz de apresentar de forma clara os fatores que influenciam as previsões e classificações dos resultados. Uma das vantagens da aplicação de algoritmos de AD é que relações entre variável dependente e variáveis explicativas podem ser facilmente visualizadas através do gráfico acíclico.

3.1.1.1 CHAID

O algoritmo CHAID foi proposto por Kass (1980), originalmente concebido para lidar com apenas atributos nominais. O método estatístico é utilizado na construção de uma árvore de classificação, que avalia todos os valores das variáveis independentes que sejam significativas, e une as categorias que são

estatisticamente homogêneas em relação à variável dependente. O nível de significância para divisão dos nós, desta pesquisa, foi fixado em 5%.

O método de CHAID pode produzir mais de duas ramificações em qualquer nível da árvore. A primeira ramificação da árvore é criada quando a variável independente, que possui a maior interação com a variável dependente, é selecionada. Cada nó possui valores homogêneos de casos segundo variáveis selecionadas. O processo é realizado em todas as variáveis independentes, para descobrir o melhor número de classes (Rokach e Maimon, 2008).

As variáveis de entrada podem ser de todos os tipos. Se a variável dependente for contínua, o teste estatístico utilizado será o teste F (teste de Fisher) e se for categórica, o teste será o Qui-Quadrado. A Figura 3.2 apresenta um exemplo da estrutura de árvore usada pelo algoritmo CHAID.

Figura 3.2 - Exemplo da estrutura de árvore do algoritmo CHAID

3.1.1.2 CART

As árvores construídas pelo algoritmo CART, desenvolvido por Breiman et al. (1984), são indicadas para problemas não lineares, segundo os autores, e alcança resultados satisfatórios tanto para variáveis numéricas ou categóricas. O

crescimento da árvore é de forma binária, ou seja, cada nó tem duas ramificações, de maneira que os valores da variável dependente sejam mais homogêneos que a divisão anterior. Dentro de uma árvore, existem muitas subárvores mais simples, então a árvore obtida tem a possibilidade de ser podada uma vez que o processo esteja finalizado, conforme mostrado na Figura 3.3.

Figura 3.3 - Exemplo da estrutura de árvore do algoritmo CART

O processo é totalmente iterativo e para quando já não é possível que os valores tenham homogeneidade segundo os critérios de parada definidos pelo pesquisador. Para o caso de variável dependente numérica, as folhas informam um valor de previsão numérico. A previsão em cada folha é baseada na média ponderada para o nó. O método do algoritmo procura minimizar o erro quadrático médio da previsão.

A Tabela 3.2 apresenta uma breve comparação dos algoritmos utilizados na pesquisa.

Tabela 3.2 - Características dos algoritmos de AD: CHAID e CART

	CHAID	CART
Vantagem	<ul style="list-style-type: none"> • Utiliza variável dependente de qualquer tipo • Não precisa seguir parâmetros preestabelecidos • Não tem tratamento de poda 	<ul style="list-style-type: none"> • Utiliza variável dependente de qualquer tipo • Não precisa seguir parâmetros preestabelecidos • Gera árvores binárias
Desvantagem	<ul style="list-style-type: none"> • Requer grande quantidade de dados para resultados satisfatórios 	<ul style="list-style-type: none"> • A árvore gerada possui muitos níveis

3.2 REGRESSÃO LINEAR MÚLTIPLA

A técnica escolhida para comparar os resultados obtidos com a AD foi a Regressão Linear Múltipla (RLM), embora não seja uma técnica adequada para análise de previsão de acidentes, já que a variável objeto de estudo em geral é discreta e positiva. No entanto, neste trabalho, decidiu-se trabalhar com taxas e utilizar RLM apenas para testar a aplicação dos algoritmos de AD através de uma técnica paramétrica e confirmatória. A seguir o modelo de RLM é apresentado.

3.2.1 Modelo de Regressão Linear Múltipla (RLM)

Na análise de Regressão Linear Múltipla uma variável dependente (ou resposta) depende de duas ou mais variáveis independentes ou explicativas. O objetivo da análise de regressão múltipla é prever os valores da variável dependente através dos valores conhecidos das variáveis independentes. A Equação 3.2 mostra a formulação básica.

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon \quad (3.2)$$

em que:

Y é a variável dependente;

α é um parâmetro que representa o intercepto;

$\beta_1, \beta_2, \dots, \beta_n$ são parâmetros estimados correspondentes às variáveis independentes (coeficientes de regressão);

$X_1, X_2, \dots, X_n$ são as variáveis independentes;

ε é o erro aleatório.

4 DADOS

4.1 CARACTERIZAÇÃO DA ÁREA DE ESTUDO

Localizado na região sudeste do Brasil, o estado de São Paulo (SP) é um importante polo industrial do país, possui aproximadamente 44 milhões de habitantes e se destaca em atração e produção de viagens em seus 645 municípios e 248 km² de extensão (IBGE, 2014). No ano de 2012, foi registrado no estado 7.003 óbitos por acidentes de transportes terrestres (DATASUS, 2014).

A pesquisa desta dissertação investiga os óbitos causados por acidentes de automóvel, agregados pelas áreas das microrregiões, no estado de SP. De acordo com a Divisão Territorial Brasileira em 2006, o estado possui 63 microrregiões, conforme ilustrado na Figura 4.1 (IBGE, 2014).

Figura 4.1 - Microrregiões do estado de São Paulo

4.2 VARIÁVEIS E AMOSTRA FINAL

4.2.1 Dados de óbitos

Para a obtenção regular de dados sobre mortalidade no país, o DATASUS criou o Sistema de Informações sobre Mortalidade (SIM). As Secretarias de Saúde Estaduais e Municipais coletam as Declarações de Óbitos dos cartórios e alimentam o SIM, com as informações nelas contidas. Uma das informações é a causa básica de óbito, a qual é codificada a partir do declarado pelo médico, segundo regras estabelecidas pela Organização Mundial de Saúde (DATASUS, 2014).

O acesso aos dados de mortalidade é feito pelo aplicativo TABNET, um tabulador desenvolvido pelo DATASUS, que gera e organiza informações das bases de dados do SUS. O aplicativo também disponibiliza dados de nascidos vivos, informações epidemiológicas, inquéritos e pesquisas, indicadores de saúde, assistência à saúde, informações demográficas e socioeconômicas, morbidade e cadastros da rede assistencial.

As variáveis selecionadas no TABNET para esta pesquisa foram os óbitos de ocupante de um automóvel traumatizado em um acidente de transporte, ocorridos nas microrregiões do estado de SP, em 2000 para calibração dos modelos obtidos neste trabalho e 2010 para validação dos mesmos. Não foram considerados faixa etária, cor/raça, escolaridade, sexo e estado civil dos falecidos.

As declarações de óbitos são codificadas utilizando-se a Décima Revisão da Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde (CID-10), publicada pela OMS. A CID-10 permite o armazenamento de informações de diagnóstico para fins clínicos e epidemiológicos, e fornece suporte para a elaboração de estatísticas de mortalidade e morbidade (WHO, 2013).

Os óbitos provenientes de acidentes por automóvel são encontrados na CID-10 de acordo com o esquema da Figura 4.2.

Figura 4.2 - Esquema da classificação dos acidentes por automóvel na CID-10

A Tabela 4.1 apresenta as medidas descritivas dos óbitos provenientes de acidentes por automóvel nas microrregiões do estado de SP nos anos 2000 e 2010 segundo o registro do DATASUS. Observa-se um aumento médio de 16% de óbitos por acidentes de automóvel ao longo da década.

Tabela 4.1 - Medidas descritivas dos óbitos por acidente de automóvel nas microrregiões do estado de SP

Ano	Média	Mínimo	Máximo	Desvio Padrão	Primeiro quartil	Terceiro quartil
2000	7,25	0	38	8	2	9
2010	21,84	0	186	26,15	8	27

4.2.2 Dados do IBGE

Os censos populacionais, realizados no Brasil pelo IBGE, produzem informações para a definição de políticas públicas e a tomada de decisões de investimento, e são fonte de referência da situação de vida da população nos municípios (IBGE, 2014).

A calibração dos modelos obtidos nesta pesquisa utilizou dados do Censo 2000, e a validação dados do Censo 2010. As variáveis demográficas e socioeconômicas das microrregiões do estado de SP selecionadas foram: população residente, PIB a preços correntes em mil reais, pessoas com rendimento e pessoas ocupadas. As medidas descritivas dessas variáveis são mostradas na Tabela 4.2.

O crescimento médio da população ao longo dos dez anos foi de 1%, do PIB 10%, as pessoas com rendimento e pessoas ocupadas 3%, ao considerar os valores da média do estado de SP.

As áreas em quilômetros quadrados (km²) das microrregiões foram extraídas do mapa do estado de SP, fornecido na área de *download* do *website* do IBGE.

Tabela 4.2 - Medidas descritivas das variáveis demográficas e socioeconômicas das microrregiões do estado de SP

Variável	Ano	Média	Mínimo	Máximo	Desvio Padrão	Primeiro quartil	Terceiro quartil
População (Pessoas)	2000	587815,9	25452	12788974	1603896	149138	470146,5
	2010	654955,5	26281	13804831	1738577	159413,5	529117
PIB (Mil Reais)	2000	6732719	74469	1,9E+08	23999897	823589	3937140
	2010	19803110	248140	5,28E+08	66962292	2541231	11735448
Rendimento (Pessoas)	2000	297329,8	12787	6552241	819933,4	80159,5	247560
	2010	392600,4	15637	8178751	1028777	100675,5	330483
Ocupados (Pessoas)	2000	239200,6	10001	5303833	663914,7	63781,5	197274,5
	2010	317480,5	11583	6785986	854075,2	75495	261223,5
Área (km ²)	–	3949,9	601,53	17558,3	2797,45	2269,24	4800,33

4.2.3 Dados da frota veicular

O Departamento Nacional de Trânsito (DENATRAN) é o órgão que tem por finalidade, dentre outras, a coordenação, supervisão e fiscalização da política do Programa Nacional de Trânsito (CONTRAN, 2004). Além disso, o órgão disponibiliza periodicamente o número da frota nacional de veículos.

As frotas de automóvel, caminhão, motocicleta e ônibus foram escolhidas como algumas das variáveis independentes dos modelos desta pesquisa. A partir do ano de 2002 o número da frota de veículos, até então divulgado por estado, passou a ser divulgado por município. Dessa forma, é possível agregar os municípios em microrregiões.

Contudo, a calibração dos modelos dessa pesquisa utiliza dados do ano 2000, não disponíveis pelo DENATRAN. A solução para adequar os dados da calibração foi estimar as frotas de automóvel, caminhão, motocicleta e ônibus, dos municípios no ano 2000, através de regressão linear simples para cada frota. O período dos anos 2002 a 2010 foi a referência para essa regressão, que teve o ano como variável independente e cada frota de veículos como variável dependente.

A frota de veículos corrigida para o ano 2000 e a frota do ano de 2010, foram agregadas em microrregiões e utilizadas na calibração e validação dos modelos, respectivamente. A Tabela 4.3 apresenta as medidas descritivas da frota de veículos das microrregiões do estado de SP.

A taxa média de crescimento anual da frota de veículos, nos dez anos, foi de 6% de automóvel, 4% de caminhão, 18% de motocicleta e 5% de ônibus.

Tabela 4.3 - Medidas descritivas da frota veicular das microrregiões do estado de SP

Frota	Ano	Média	Mínimo	Máximo	Desvio Padrão	Primeiro quartil	Terceiro quartil
Automóvel	2000	111616,6	2263	3338557	416463,8	19251,5	71673
	2010	211664,7	4368	5567040	698057,9	38985,5	157932
Caminhão	2000	6198,71	167	129689	16123,03	1783	6289,5
	2010	9250,17	296	162587	20570,82	2721,5	9346,5
Motocicleta	2000	8296,57	77	111391	14909,14	1976	9276
	2010	52738,79	1261	874231	111907,1	16218	54651,5
Ônibus	2000	1167,81	10	31404	3912,56	223	976,5
	2010	2013,97	54	46647	5842,44	410,5	1763

4.2.4 Dados de IDHM

Desenvolvimento Humano é um processo de ampliação das escolhas das pessoas para que elas tenham capacidades e oportunidades para serem aquilo que desejam ser. Em alternativa ao PIB, o IDH é uma medida média das conquistas de desenvolvimento humano básico em um país. Os três pilares que constituem o IDH são: longevidade (saúde), educação e renda (PNUD, 2013).

O IDHM brasileiro adequa a metodologia global do IDH ao contexto brasileiro e à disponibilidade de indicadores nacionais, para avaliar o desenvolvimento dos municípios e regiões metropolitanas brasileiras (Atlas Brasil, 2015).

O índice varia de 0 a 1. Quanto mais próximo de 1, maior o desenvolvimento humano da área de estudo. O cálculo é feito de acordo com a Figura 4.3.

Através da plataforma de consulta ao IDHM, o Atlas do Desenvolvimento Humano no Brasil, obteve-se os índices dos municípios, posteriormente agregados em microrregiões, e somados. Depois da soma dos índices agregados, foi extraída a média aritmética, para cada uma das microrregiões. Assim, a média dos índices do ano 2000 foi utilizada na calibração, e do ano de 2010 na validação dos modelos desta pesquisa. Os valores das medidas descritivas do índice são mostrados na Tabela 4.4. O crescimento médio do IDHM, ao longo do período analisado, foi de 1%.

Tabela 4.4 - Medidas descritivas da média de IDHM das microrregiões do estado de SP

Ano	Média	Mínimo	Máximo	Desvio Padrão	Primeiro quartil	Terceiro quartil
2000	0,645	0,54	0,713	0,029	0,638	0,66
2010	0,74	0,685	0,793	0,02	0,733	0,749

Figura 4.3 - Esquema do cálculo do IDHM (PNUD, 2013)

4.2.5 Dados de fluxo veicular anual sintético

Os dados descritos até o presente momento caracterizam os residentes das microrregiões em estudo. No entanto, acredita-se que boa parte dos óbitos

acontecem com pessoas que estejam de passagem em determinada microrregião. Desta forma, uma variável que mensure a passagem de pessoas pelas microrregiões é importante. Para o caso desse trabalho, esta variável foi o fluxo veicular anual sintético. O procedimento para obtenção deste fluxo é proposto e descrito em seguida.

Uma vez que não existe uma base de dados que informe o número de viagens intermunicipais realizadas por automóvel no estado de SP, esta pesquisa admite a hipótese de que os padrões de viagens por automóvel repetem-se nas mesmas proporções que as viagens intermunicipais por ônibus entre pares de origens e destinos da rede rodoviária do estado de SP. A Figura 4.4 resume o processo de obtenção do fluxo veicular sintético, explicado em seguida.

Figura 4.4 - Etapas do processo de obtenção do fluxo veicular sintético

Por não ter tido acesso, até o presente momento, a uma rede rodoviária georreferenciada do estado de SP, foi utilizada a alternativa de criar uma rede de rodovias, para estimativa dos valores de fluxo veicular sintético, baseada na triangulação dos centroides de cada município, com auxílio do *software* QGIS 2.8.1. A triangulação inicialmente liga todos os municípios, Figura 4.5a. Como esta pesquisa está voltada apenas para o estado de SP, e algumas das rodovias criadas extrapolam os limites do estado, foi necessário fazer um ajuste na rede, deixando apenas as rodovias que não ultrapassam esses limites, Figura 4.5b.

Figura 4.5 - Rede rodoviária do estado de São Paulo por triangulação

Para obtenção da matriz O/D de viagens intermunicipais para o estado de SP, foi utilizado o modelo gravitacional calibrado por Isler (2015). O modelo gravitacional proposto por Isler (2015) considera o número de viagens intermunicipais por ônibus, de acordo com parâmetros socioeconômicos e de características das viagens, conforme a Equação 4.1, sem intercepto e com variáveis independentes relativas à população (POP), renda (REND), pessoal ocupado assalariado (OCUP) e distância rodoviária (d) entre cidades em quilômetros.

$$\ln(V_{ij}) = 0,65 \cdot \ln(POP_i \cdot POP_j) - 0,267 \cdot \ln(OCUP_i \cdot OCUP_j) - 0,528 \cdot \ln(d) + 0,289 \cdot \ln(RENDA_i \cdot RENDA_j) \quad (4.1)$$

As estimativas do número de viagens anuais por automóveis entre cidades da Região Sudeste foram estabelecidas segundo a Equação (4.1) proposta por Isler (2015), que considera o Volume Diário Médio Anual (VDMA) na rede definida pelo Plano Nacional de Logística e Transportes (PNLT, 2015).

Com o auxílio do *software* de Sistema de Informação Geográfica (SIG) TransCAD 5.0, é estimada uma matriz O/D baseada nos volumes obtidos através do modelo proposto por Isler (2015). A alocação de tráfego foi feita através do método “Tudo-ou-Nada”, onde o volume de tráfego estimado é alocado na rota de menor custo entre dois pontos. A Figura 4.6 mostra um exemplo do volume de tráfego estimado para o estado de SP, no ano 2010, dentro das microrregiões. Vale ressaltar que o uso do método “Tudo-ou-Nada” ocorreu devido à falta de dados de restrições de capacidade viária.

Figura 4.6 - Volume de tráfego estimado no estado de SP no ano 2010

Após alocar no TransCAD os fluxos estimados pelo modelo gravitacional, foi feita a segmentação das rodovias por trechos, dentro de cada município, com auxílio do *software* QGIS 2.8.1, a Figura 4.7 mostra um exemplo dessa segmentação em duas microrregiões.

Figura 4.7 - Segmentação da rede rodoviária do estado de SP, dentro dos municípios

Então, foi calculada a proporção de fluxo na rodovia, pelo comprimento do tramo dentro de cada município e agregada por microrregião. Um exemplo ilustrativo de como o cálculo é feito, está na Figura 4.8. Supondo que a Rodovia 1, que liga os municípios A e B, possui 30 km de extensão e fluxo anual de 10 milhões de veículos. Sabe-se que 12km de extensão do trecho pertencem ao município A e o restante ao município B. Desta forma, proporcionalmente, o município A terá um fluxo equivalente de veículos igual a 4 milhões. Assim acontece com a Rodovia 2, que se estende a mais de um município, e com as demais rodovias criadas pela rede triangulada.

Figura 4.8 - Fluxo de veículos proporcional nos municípios (veic.km)

A Tabela 4.5 apresenta as medidas descritivas dos fluxos de veículos estimados para a calibração, ano 2000, e validação, ano 2010, dos modelos. A média diária do fluxo de veículos foi em torno de 2,95 milhões no ano 2000 e 5,34 milhões em 2010. Ao longo da década pesquisada, o crescimento médio do fluxo de veículos foi de 6% ao ano.

Tabela 4.5 - Medidas descritivas do fluxo veicular anual nas microrregiões do estado de SP

Ano	Média	Mínimo	Máximo	Desvio Padrão	Primeiro quartil	Terceiro quartil
2000	1,08E+09	14606677	9,63E+09	1,81E+09	1,56E+08	1,05E+09
2010	1,95E+09	27115132	1,77E+10	3,31E+09	2,67E+08	2,02E+09

Uma limitação importante deste dado foi a falta de validação do procedimento proposto. Para ocorrência de validação, é necessário repetir o método para a rede georreferenciada e comparar tais dados ao VDMA proveniente de contagens de tráfego, sobretudo em praças de pedágio. O grande problema é realmente a dificuldade de informação de fluxo veicular em rodovias não concessionadas ou em trechos urbanos.

4.2.6 Dados de distâncias entre Microrregiões

Dentre as microrregiões do estado de SP mais populosas e com maior fluxo de automóvel estimado, destacam-se: São Paulo, Osasco, Campinas, Santos, Sorocaba e Piracicaba. A partir dessa informação utilizou-se o *software* TransCAD 5.0 para obter a distância, em linha reta, entre os centroides das microrregiões citadas para todas as outras microrregiões. A intenção de obter essas distâncias é observar uma correlação entre o número de óbitos por automóvel e as distâncias até as microrregiões mais populosas ou com maior fluxo de automóvel. As medidas descritivas das distâncias utilizadas nos modelos são apresentadas na Tabela 4.6.

Tabela 4.6 - Medidas descritivas das distâncias entre microrregiões do estado de SP

Variável	Média	Mínimo	Máximo	Desvio Padrão	Primeiro quartil	Terceiro quartil
Dist_Osasco (km)	242,97	0	537,74	153,09	117,47	358,96
Dist_São Paulo (km)	267,41	0	577,44	166,44	134,05	396,38
Dist_Campinas (km)	217,77	0	489,58	128,52	112,65	297,46
Dist_Sorocaba (km)	230,18	0	493,71	133,87	116,69	333,24
Dist_Piracicaba (km)	205,85	0	427,15	104,83	125,79	276,65
Dist_Santos (km)	295,61	0	617,48	175,82	137,34	431,42

4.2.7 Mensuração de óbitos por acidentes de trânsito

Um dos objetivos específicos desta pesquisa é explorar formas de mensurar óbitos provenientes de acidentes de automóvel numa análise agregada por área. Neste trabalho foram testadas três formas de mensuração da variável objeto de estudo (dependente): (1) óbitos por população (obt/hab); (2) óbitos por fluxo (obt/veic/ano); (3) óbitos por área (obt/ha). A Tabela 4.7 apresenta as medidas descritivas dessas taxas.

Tabela 4.7 - Medidas descritivas das taxas de óbitos por automóvel

Variável	Ano	Média	Mínimo	Máximo	Desvio Padrão	Primeiro quartil	Terceiro quartil
Óbito/população	2000	$2,53 \times 10^{-5}$	0	$1,39 \times 10^{-4}$	$2,82 \times 10^{-5}$	$4,71 \times 10^{-6}$	$3,98 \times 10^{-5}$
	2010	$6,29 \times 10^{-5}$	0	$3,13 \times 10^{-4}$	$5,21 \times 10^{-5}$	$2,43 \times 10^{-5}$	$7,61 \times 10^{-5}$
Óbito/fluxo	2000	$2,65 \times 10^{-8}$	0	$2,88 \times 10^{-7}$	$5,31 \times 10^{-8}$	$1,86 \times 10^{-9}$	$2,43 \times 10^{-8}$
	2010	$4,32 \times 10^{-8}$	0	$4,92 \times 10^{-7}$	$8,21 \times 10^{-8}$	$8,47 \times 10^{-9}$	$3,53 \times 10^{-8}$
Óbito/área	2000	$2,17 \times 10^{-3}$	0	$9,37 \times 10^{-3}$	$2,32 \times 10^{-3}$	$5,08 \times 10^{-4}$	$3,11 \times 10^{-3}$
	2010	$7,41 \times 10^{-3}$	0	$7,89 \times 10^{-2}$	$1,10 \times 10^{-2}$	$2,58 \times 10^{-3}$	$8,49 \times 10^{-3}$

Os resultados encontrados tinham alta variabilidade entre microrregiões. Desta forma, buscou-se a alternativa de transformar os dados e obter uma menor dispersão, através da divisão pela média do conjunto de dados de cada variável. A Tabela 4.8 mostra a nova configuração de valores das variáveis.

Tabela 4.8 - Medidas descritivas das taxas de óbitos por automóvel pela média

Variável	Ano	Máximo	Desvio Padrão	Primeiro quartil	Terceiro quartil
Óbito/população	2000	5,5	1,11	0,19	1,5
	2010	4,97	0,83	0,39	1,21
Óbito/fluxo	2000	10,86	2,01	0,07	0,92
	2010	11,4	1,9	0,2	0,82
Óbito/área	2000	4,32	1,07	0,23	1,43
	2010	10,64	1,48	0,35	1,14

Foram gerados mapas das microrregiões do estado de SP, conforme mostrado na Figura 4.9, com os novos valores das taxas, divididos pela média, das variáveis dependentes. A visualização dos valores extremos das taxas de óbitos por população (Figura 4.9a) e fluxo de veículos (Figura 4.9b) indica que nas fronteiras do estado de SP estão concentradas as maiores taxas. Um padrão espacial diferenciado acontece com as taxas de óbitos por área (Figura 4.9c), onde a distribuição apresentada no estado de SP é mais heterogênea, com altos valores próximos à microrregião de São Paulo.

(a) Taxa de óbitos por população

(b) Taxa de óbitos por fluxo de veículos

(c) Taxa de óbitos por área

Figura 4.9 - Mapas temáticos dos valores das variáveis dependentes

As demais variáveis, ditas independentes, desta pesquisa também foram divididas pela média das observações, para diminuir a variabilidade dos dados. Os novos valores são apresentados na Tabela 4.9.

Tabela 4.9 - Medidas descritivas das variáveis independentes pela média

Variável	Ano	Mínimo	Máximo	Desvio Padrão	Primeiro quartil	Terceiro quartil
População	2000	0,04	21,76	2,73	0,25	0,85
	2010	0,04	21,08	2,65	0,24	0,81
PIB	2000	0,01	28,27	3,56	0,12	0,63
	2010	0,01	26,68	3,38	0,13	0,59
Rendimento	2000	0,04	22,04	2,76	0,27	0,87
	2010	0,04	20,83	2,62	0,26	0,84
Ocupados	2000	0,04	22,17	2,78	0,27	0,88
	2010	0,04	21,37	2,69	0,24	0,82
Área	–	0,15	4,45	0,71	0,57	1,22
Frota de Automóvel	2000	0,02	29,91	3,73	0,17	0,68
	2010	0,02	26,3	3,3	0,18	0,75
Frota de Caminhão	2000	0,03	20,92	2,6	0,29	1,01
	2010	0,03	17,58	2,22	0,29	1,01
Frota de Motocicleta	2000	0,01	13,43	1,8	0,24	1,12
	2010	0,02	16,58	2,12	0,31	1,04
Frota de Ônibus	2000	0,01	26,89	3,35	0,19	0,87
	2010	0,03	23,16	2,9	0,2	0,88
IDHM	2000	0,84	1,11	0,05	0,99	1,02
	2010	0,93	1,07	0,03	0,99	1,01
Fluxo	2000	0,01	8,94	1,68	0,14	0,99
	2010	0,01	9,07	1,7	0,14	1,04
Distância a Osasco	–	0	2,21	0,63	0,48	1,45
Distância a São Paulo	–	0	2,16	0,62	0,5	1,46
Distância a Campinas	–	0	2,25	0,59	0,52	1,36
Distância a Sorocaba	–	0	2,14	0,58	0,51	1,44
Distância a Piracicaba	–	0	2,08	0,51	0,61	1,32
Distância a Santos	–	0	2,09	0,59	0,46	1,44

4.2.8 Testes de normalidade

Uma distribuição normal, caracterizada por uma curva em forma de sino, depende de dois parâmetros: a média (μ), ponto central da distribuição e o desvio padrão (σ), que mede a dispersão da distribuição.

O pressuposto de normalidade estabelece que os valores da variável dependente são normalmente distribuídos para cada valor da variável independente. Alguns testes verificam se os dados de uma população seguem uma distribuição Normal, tais como o teste de Kolmogorov-Smirnov (K-S) e o teste de Shapiro-Wilk (S-W).

Diversos programas estatísticos realizam testes específicos, como os testes de K-S e S-W. Dentre os programas, está o IBM SPSS versão 22.0, que auxilia na análise de dados, gera gráficos e tabelas dos resultados obtidos, além de procedimentos mais avançados como análise multivariada de dados.

Os testes de K-S e S-W foram feitos para as variáveis dependentes e os resultados obtidos não atenderam à suposição de normalidade. A não normalidade dos dados pode ser corrigida ao realizar algumas transformações nas variáveis. As transformações testadas nos dados da pesquisa foram a padronização, a normalização e o logaritmo.

A padronização leva as variáveis transformadas a terem média 0 (zero) e desvio padrão 1 (um), através do valor z. A Equação 4.2 define a padronização.

$$z = \frac{x - \mu}{\sigma} \quad (4.2)$$

Em que z: valor padronizado; x: valor observado; μ : média; σ : desvio padrão. O valor z conhecido como valor padronizado é uma medida relativa. Mede o quanto x se afasta da média (μ), em unidade de desvio padrão (σ) (Barbetta, 2014).

O propósito da normalização é diminuir os problemas relacionados às dispersões heterogêneas. Os dados depois de transformados variam de 0 (zero) a 1 (um), de acordo com a Equação 4.3.

$$y = \frac{x - \min}{\max - \min} \quad (4.3)$$

Em que y: valor normalizado; x: valor observado; min: valor mínimo das observações; max: valor máximo das observações.

A transformação em logaritmo não pode ser realizada com valor zero ou números negativos, por isso, se os seus dados tendem a zero ou possuir números negativos é preciso adicionar uma constante a todos os dados antes de fazer a transformação (Field, 2013). Foi utilizado o $\log(x + 1)$ em todas as variáveis, onde x é o valor observado.

A Tabela 4.10 apresenta um resumo dos testes realizados nas variáveis, e identifica se atende ou não a suposição de normalidade.

Tabela 4.10 - Testes de normalidade na variável transformada

Variável	Normalidade	
	K-S	S-W
Taxa	Não atende	Não atende
Padronizada	Não atende	Não atende
Normalizada	Não atende	Não atende
Logaritmo	Não atende	Não atende

Com o objetivo de reduzir os resíduos, será utilizada como técnica paramétrica a Regressão Linear, apesar de nenhuma das transformações das variáveis dependentes terem atendido aos testes de normalidade. As análises estatísticas serão feitas com base nessa consideração.

Vale ressaltar que os testes de normalidade foram realizados considerando as suposições de uma das ferramentas utilizadas para calibração dos modelos, a Regressão Linear Múltipla (RLM).

Neste trabalho são testados algoritmos de Árvore de Decisão (CART e CHAID). Tais técnicas são mais robustas quando a suposição de normalidade não é atendida.

Recomenda-se, futuramente, a utilização de Modelos de Regressão Generalizados como ferramenta paramétrica para comparação com algoritmos de Árvores de Decisão.

5 RESULTADOS E DISCUSSÕES

Neste capítulo são apresentados os resultados das análises das taxas de óbitos provenientes de acidentes de automóvel no estado de São Paulo (SP), a partir dos algoritmos CHAID (*Chi-squared Automatic Interaction Detection*) e CART (*Classification and Regression Tree*) das Árvores de Decisão (AD), além de resultados comparativos obtidos através da aplicação da Regressão Linear Múltipla (RLM). Para fins comparativos entre os algoritmos são utilizados valores estimados para o ano de 2010 a partir de modelos calibrados para o ano de 2000. Importante ressaltar que os valores estimados nos algoritmos de AD são as médias de taxa de óbitos por cada classe (nó terminal) enquanto que os valores estimados na RLM são valores obtidos a partir de equação previamente calibrada.

Conforme mencionado no Capítulo 4, as variáveis independentes foram divididas pelos respectivos valores médios, e as variáveis dependentes foram as taxas de óbitos por população, área e fluxo, também divididas pelos seus valores médios. Desta forma, as taxas dos óbitos possuem média 1,00, sendo um parâmetro para identificar se cada um dos nós representa valores altos ou baixos de óbitos. Cada nó gerado pelas árvores apresenta algumas características inerentes a cada classe, identificadas na Figura 5.1.

Figura 5.1 - Identificação dos valores de cada nó da Árvore de decisão (*output* do IBM SPSS 22.0)

Vale ressaltar que, para o caso de variáveis numéricas, o valor previsto corresponde à média da variável dependente relativa a cada classe (nó).

5.1 ÁRVORES DE DECISÃO (AD)

Foram geradas dez ADs, descritas mais detalhadamente nas próximas subseções: (1) CHAID para óbitos por população (treinamento e teste); (2) CHAID para óbitos por área (treinamento e teste); (3) CHAID para óbitos por fluxo veicular anual (treinamento e teste); (4) CART para óbitos por população (treinamento e teste); (5) CART para óbitos por área (treinamento e teste).

5.1.1 CHAID

As análises foram realizadas através do algoritmo CHAID, contido no pacote estatístico IBM SPSS 22.0. Conforme mencionado no Capítulo 3, o método CHAID é baseado nos testes de associação qui-quadrado e particiona o conjunto de dados em subconjuntos mutuamente exclusivos que melhor descrevem a variável resposta exaustivamente. As divisões não são necessariamente binárias como no caso do algoritmo CART.

Utilizaram-se como variáveis dependentes as taxas de óbitos por população, área e fluxo veicular anual, divididas pelos seus valores médios. Como variáveis independentes foram utilizadas as seguintes características socioeconômicas: população residente, PIB a preços correntes em mil reais, pessoas com rendimento e pessoas ocupadas, área, frotas de automóvel, caminhão, motocicleta e ônibus, IDHM, fluxo veicular anual e distâncias entre microrregiões. Algumas variáveis independentes variam de acordo com a variável dependente escolhida. Se, por exemplo, a variável dependente for óbito por população, a variável independente população residente não entra na análise. O mesmo ocorre para as variáveis independentes área e fluxo e as variáveis dependentes óbitos por área e óbitos por fluxo, respectivamente. Os critérios adotados para a sua classificação foram: o mínimo de 2 observações/microrregiões por nó terminal e nível de significância de 5% para cada divisão. Além disso, as árvores de treinamentos foram geradas a partir da amostra relativa ao ano de 2000, enquanto que os modelos de AD foram validados (árvores de teste) a partir da amostra relativa ao ano de 2010.

5.1.1.1 CHAID para óbitos por população

A forma gráfica da árvore de treinamento (CHAID) para a variável dependente óbitos por população é apresentada na Figura 5.2. Os valores apresentados na figura nos nós correspondem ao valor médio e desvio padrão de taxa de óbitos encontrados em cada classe de microrregiões. Cada nó terminal ou folha representa uma classe de microrregiões agrupadas segundo características de variáveis independentes e valores médios de taxas de óbitos.

A variável de maior importância para segmentação dos dados é Frota de Caminhões. A partir daí, a árvore se ramifica em três grupos: (1) Microrregiões com frota de caminhões $\leq 0,307$ (1907 veículos) – Nó 1; (2) Microrregiões com frota de caminhões entre 0,307 (1907 veículos) e 0,355 (2206 veículos) – Nó 2; (3) Microrregiões com frota de caminhões $> 0,355$ (2206 veículos). Em seguida, a variável “escolhida” para continuidade da segmentação dos dados é IDHM.

Observam-se 5 nós terminais na Figura 5.2, ou seja, 5 classes de microrregiões do estado com características distintas e seus respectivos valores médios de taxa de óbitos por população. Cada nó terminal é ilustrado pelos retângulos onde não há mais subdivisões.

Analisando as cinco classes (nós terminais) obtidas, observa-se que o grupo de microrregiões que possui a maior média de óbitos do estado, equivalente a 2,626, é aquele cuja frota de caminhão está entre 1907 (0,307) e 2206 (0,355) veículos. Esta classe corresponde ao nó terminal ou folha 2. Fazem parte do nó 2 seis microrregiões, são elas: Jales, Avaré, Botucatu, Amparo, Adamantina e Itapetininga.

Considerando, em seguida, os nós terminais que apresentam frota de caminhão menor igual a 1907 (0,307) veículos, valores de IDHM são diferenciadores das classes. O nó 4 é aquele formado apenas por 3 microrregiões e IDHM menor igual a 0,603 (0,935). Esta classe que poderia ser classificada como grupo de microrregiões com baixa frota de caminhões e também baixo IDHM, apresenta uma média alta de óbitos, 2,482 – Nó 4. Fazem parte deste nó as microrregiões Capão Bonito, Bananal e Paraibuna/Paraitinga.

As microrregiões com frota de caminhão ≤ 1907 (0,307) veículos e IDHM $> 0,603$ (0,935) – Nó 5, tiveram a menor média de óbitos, apenas 0,591. O maior número de microrregiões (24), com média de taxa de óbitos 1,038 é classificado no nó 6, que apresenta a frota de caminhão > 2206 (0,355) veículos e IDHM $\leq 0,658$ (1,02).

Pelos resultados verifica-se que a frota de caminhões nem sempre se relaciona a maiores taxas médias de óbitos. Já valores de IDHM, relativos ao desenvolvimento de regiões, parecem assumir relações inversas à média de óbitos.

A validação do modelo de AD da taxa de óbitos por população, Figura 5.3, mostrou que 19% das microrregiões foram classificadas em classes diferentes, com valores previstos também diferentes. Foram elas: Andradina, Araçatuba, Avaré, São Carlos, Rio Claro, Pirassununga, Moji Mirim, Amparo, Adamantina, Bananal, Paraibuna/Paraitinga e Franco da Rocha, ilustradas na seção 5.1.3.

Figura 5.2 - Árvore de decisão de Treinamento do Algoritmo CHAID para óbitos por população (amostra 2000)

Figura 5.3 - Árvore de decisão de Teste do Algoritmo CHAID para óbitos por população (amostra 2010)

5.1.1.2 CHAID para óbitos por área

Mantendo os mesmos critérios de segmentação apresentados anteriormente, a forma gráfica da árvore de treinamento (CHAID) para a variável dependente óbitos por área é apresentada na Figura 5.4. Cabe lembrar que a variável independente área não foi incluída nesta análise.

A variável de maior importância, neste caso, para segmentação dos dados é Frota de Ônibus. A partir daí, a árvore se ramifica em três grupos: (1) Microrregiões com frota de ônibus $\leq 0,454$ (530 veículos) – Nó 1; (2) Microrregiões com frota de ônibus

entre 0,454 (530 veículos) e 1,336 (1560 veículos) – Nó 2; (3) Microrregiões com frota de caminhões > 1,336 (1560 veículos).

Em seguida, a variável “escolhida” para continuidade da segmentação dos dados é Distância para Campinas, seguida de fluxo veicular anual. Foram observados 5 nós terminais.

Os maiores índices médios de óbitos na relação por área, média de 3,143 no ano 2000, estão nas microrregiões que possuem frota de ônibus > 1560 (1,336) veículos – Nó 3. A menor média de óbitos observada, dentre todos os nós terminais, é a média 0,366 – Nó 6. Esta classe de microrregiões apresenta menores valores de frota de ônibus, maiores distâncias para Campinas e menores fluxos veiculares anuais. Os óbitos por área têm uma relação direta com frotas de ônibus e fluxo veicular e relação inversa com a distância à cidade de Campinas, SP.

Em termos de validação ou teste da árvore de treinamento (Figura 5.5), 14% das microrregiões foram classificadas diferentemente da amostra de 2000. As microrregiões que foram classificadas em classe diferente são: Assis, Ourinhos, São Joaquim da Barra, Tatuí, Sorocaba, Bragança Paulista, Guaratinguetá, Franco da Rocha e Santos, localizadas na subseção 5.1.3.

Figura 5.4 - Árvore de decisão de Treinamento do Algoritmo CHAID para óbitos por área (amostra 2000)

Figura 5.5 - Árvore de decisão de Teste do Algoritmo CHAID para óbitos por área (amostra 2010)

5.1.1.3 CHAID para óbitos por fluxo

Seguindo os mesmos critérios anteriores, o modelo CHAID, obtido para a variável dependente óbitos por fluxo veicular é apresentado na Figura 5.6. Este modelo não

considera a variável independente fluxo veicular anual. Neste caso a variável de maior importância é Frota de Automóveis. A partir daí, a árvore se ramifica em dois grupos: (1) Microrregiões com frota de automóveis $\leq 0,135$ (15058 veículos) – Nó 1; (2) Microrregiões com frota de automóveis $> 0,135$ (15058 veículos).

Em seguida, é selecionada a variável Distância à microrregião de São Paulo, gerando uma AD com 3 nós terminais. Analisando as três classes obtidas, verifica-se que o grupo de microrregiões que possui a maior média de óbitos do estado, 2,583 é aquele cuja frota de automóveis é menor ou igual a 15058 (0,135) veículos. Esta classe corresponde ao nó terminal ou folha 1.

Considerando, em seguida, os nós terminais que apresentam frota de automóveis superior a 15058 (0,135) veículos, valores de distâncias para a microrregião de São Paulo distinguem as duas classes restantes. O nó 4 é aquele formado por 3 microrregiões e distância a São Paulo maior que 510,57 (1,909) km. A classe de microrregiões com alta frota de automóveis, mais próximas da microrregião de São Paulo, é aquela com menor taxa média de óbitos por fluxo 0,520 – Nó 3.

Resultados neste modelo corroboram com a hipótese de que localizações com altos fluxos veiculares estariam menos propícias a ocorrência de acidentes severos. Portanto, não possuem uma relação direta.

O resultado do teste da AD, que relaciona os óbitos por fluxo, Figura 5.7, mostra que as microrregiões Fernandópolis, Dracena, Registro e Itanhaém, 6% do total, não foram classificadas como na amostra de 2000. As demais permaneceram com as mesmas características.

Figura 5.6 - Árvore de decisão de Treinamento do Algoritmo CHAID para óbitos por fluxo (amostra 2000)

Figura 5.7 - Árvore de decisão de Teste do Algoritmo CHAID para óbitos por fluxo (amostra 2010)

5.1.2 CART

O algoritmo CART torna os subconjuntos resultantes cada vez mais homogêneos em relação à variável resposta, mediante sucessivas divisões binárias no conjunto de dados. Foram gerados quatro modelos de CART, conforme descrição em seguida.

Seguindo o mesmo critério de segmentação de tamanho mínimo de nós terminais (2), foi gerado o modelo CART, com desvio mínimo – aprimoramento de 0,0001, para a variável dependente óbitos por população, área e fluxo e as variáveis independentes população residente, PIB a preços correntes em mil reais, pessoas

com rendimento e pessoas ocupadas, área, frotas de automóvel, caminhão, motocicleta e ônibus, IDHM, fluxo veicular anual e distâncias entre microrregiões. Assim como no algoritmo CHAID, a variável independente associada à taxa da variável dependente foi retirada do modelo. Em todos os modelos de AD foram considerados critérios pequenos para divisões das classes devido ao tamanho reduzido da amostra de treinamento e teste (63 microrregiões).

5.1.2.1 CART para óbitos por população

Desta vez, os grupos são gerados aos pares considerando valores de corte de variáveis independentes. A partir da variável IDHM, a árvore se ramifica em dois grupos: (1) Microrregiões com $IDHM \leq 0,959$ (0,618) – Nó 1; (2) Microrregiões com $IDHM > 0,959$ (0,618) – Nó 2. Comparando apenas as classes relativas aos nós 1 e 2, observa-se que microrregiões com valores maiores de IDHM são mais propensas a ter menores médias de óbitos por acidente de automóvel.

Em seguida, são selecionadas variáveis como Distância a Osasco, Distância a Piracicaba e Área para agrupamento de microrregiões e formação de classes. Foi obtido um total de sete nós terminais, tal como ilustrado na Figura 5.8.

Pelos resultados obtidos, verifica-se que microrregiões mais distantes de Osasco, com maiores valores de IDHM, são mais propensas a maiores taxas de óbitos (Nó 4). Além disso, observa-se uma relação inversa entre distância a Piracicaba e maiores taxas de óbitos, além de relação direta entre tamanho da área total da microrregião e taxas de óbitos.

Figura 5.8 - Árvore de decisão de Treinamento do Algoritmo CART para óbitos por população (amostra 2000)

Os valores previstos para o ano de 2010, Figura 5.9, mostram que 19% das microrregiões foram classificadas diferentemente: Barretos, Franca, Ribeirão Preto,

Araçatuba, Avaré, São Carlos, Pirassununga, Moji Mirim, Amparo, Ourinhos, Itapetininga, Guaratinguetá.

Figura 5.9 - Árvore de decisão de Teste do Algoritmo CART para óbitos por população (amostra 2010)

5.1.2.2 CART para óbitos por área

A partir do mesmo algoritmo e critérios de partição, foi gerado o modelo CART para a variável dependente óbitos por área e demais variáveis independentes: população residente, PIB a preços correntes em mil reais, pessoas com rendimento e pessoas ocupadas, frotas de automóvel, caminhão, motocicleta e ônibus, IDHM, fluxo veicular anual e distâncias entre microrregiões, conforme ilustrado na Figura 5.9 (CART de treinamento).

A partir da variável Frota de ônibus, a árvore se ramifica em dois grupos: (1) Microrregiões com Frota de ônibus $\leq 1,383$ (1615 veículos) – Nó 1; (2) Microrregiões com Frota de ônibus $> 1,383$ (1615 veículos) – Nó 2. Comparando apenas as classes relativas aos nós 1 e 2, observa-se que microrregiões com valores maiores de Frotas de caminhões são mais propensas a ter maiores médias de óbitos por acidente de automóvel por área.

Em seguida, são selecionadas variáveis como Distância a Osasco, Distância a Santos, IDHM e Fluxo veicular anual para agrupamento de microrregiões e formação de classes. Foi obtido um total de dez nós terminais, tal como ilustrado na Figura 5.10.

As relações encontradas pela AD de treinamento foram: (1) Relações inversas de taxas médias de óbitos por área e distâncias a Osasco e Santos; (2) Relações inversas entre a variável dependente e valores de IDHM e (3) Relações diretas entre fluxos veiculares e taxas médias de óbitos por área.

Validando a AD calibrada para o ano de 2000, a AD de teste da taxa de óbitos por área no ano de 2010, Figura 5.10, identificou a mudança de classificação de 8% das microrregiões. As microrregiões que sofreram mudanças foram: Votuporanga, Ituverava, Lins, Adamantina e Sorocaba.

Figura 5.10 - Árvore de decisão de Treinamento do Algoritmo CART para óbitos por área (amostra 2000)

Figura 5.11 - Árvore de decisão de Teste do Algoritmo CART para óbitos por área (amostra 2010)

5.1.2.3 CART para óbitos por fluxo

A AD da taxa de óbitos por fluxo não selecionou nenhuma das variáveis do banco de dados desta pesquisa. Nenhuma das variáveis independentes foi significativa para

agrupamento de classes segundo a taxa de óbitos por fluxo anual e critérios de partição do algoritmo CART.

5.2 CLASSIFICAÇÃO DAS MICRORREGIÕES

Nesta seção é apresentado um quadro resumo de cada uma das classes (nós terminais), que identifica a qual nó pertence cada microrregião, no ano de 2000 e 2010, assim como a análise visual dos resultados, através dos mapas do estado de SP, para os anos da calibração e validação dos modelos. Nos mapas são localizadas as microrregiões pertencentes a cada nó terminal nos dois anos e identificadas as que foram classificadas diferentemente na validação dos modelos gerados pela AD.

A taxa média dos óbitos desta classificação foi feita de acordo com a escala da Tabela 5.1. A variável de distâncias foi identificada como próximo, distância intermediária e distante. O restante das variáveis independentes foi caracterizado em três níveis: baixo, médio e alto.

Tabela 5.1 - Escala da taxa média de óbitos

Taxa média de óbitos	
$\leq 0,5$	Muito baixa
$0,5 - 1,0$	Baixa
$1,0 - 1,5$	Moderada
$1,5 - 2,0$	Alta
$> 2,0$	Muito Alta

Tabela 5.2 - Classificação das Microrregiões segundo o algoritmo CHAID para óbitos por população

Nó Terminal	Variáveis	Taxa média de óbitos	Microrregiões 2000	Microrregiões 2010	Nomenclatura proposta
2	Frota de caminhão entre 1907 e 2206	2,626	Jales, Avaré, Botucatu, Amparo, Adamantina, Itapetininga.	Jales, Andradina, Botucatu, Itapetininga	Frota de caminhões média e taxa de óbitos muito alta
4	Frota de caminhão $\leq$ 1907 IDHM $\leq$ 0,603	2,482	Capão Bonito, Bananal, Paraibuna/Paraitinga	Capão Bonito	Frota de caminhão baixa, IDHM baixo e taxa de óbitos muito alta
5	Frota de caminhão $\leq$ 1907 IDHM $>$ 0,603	0,591	Fernandópolis, Votuporanga, Auriflama, Nhandeara, Novo Horizonte, Barretos, Ituverava, Batatais, Andradina, Lins, Dracena, Tupã, Campos do Jordão, Caraguatatuba, Itanhaém, Franco da Rocha	Fernandópolis, Votuporanga, Auriflama, Nhandeara, Novo Horizonte, Barretos, Ituverava, Batatais, Lins, Avaré, Dracena, Adamantina, Tupã, Campos do Jordão, Bananal, Paraibuna/Paraitinga, Caraguatatuba, Itanhaém	Frota de caminhão baixa, IDHM médio e taxa de óbitos baixa
6	Frota de caminhão $>$ 2206 IDHM $\leq$ 0,658	1,038	São José do Rio Preto, Catanduva, São Joaquim da Barra, Franca, Jaboticabal, Ribeirão Preto, Birigui, Bauru, Jaú, Araraquara, São João da Boa Vista, Presidente Prudente, Marília, Assis, Ourinhos, Itapeva, Tatuí, Piedade, Sorocaba, Bragança Paulista, Guaratinguetá, Registro, Itapeçerica da Serra, Mogi das Cruzes	São José do Rio Preto, Catanduva, São Joaquim da Barra, Franca, Jaboticabal, Ribeirão Preto, Araçatuba, Birigui, Bauru, Jaú, Araraquara, São Carlos, Rio Claro, Pirassununga, São João da Boa Vista, Moji Mirim, Amparo, Presidente Prudente, Marília, Assis, Ourinhos, Itapeva, Tatuí, Piedade, Sorocaba, Bragança Paulista, Guaratinguetá, Registro, Franco da Rocha, Itapeçerica da Serra, Mogi das Cruzes	Frota de caminhão alta, IDHM médio e taxa de óbitos moderada
7	Frota de caminhão $>$ 2206 IDHM $>$ 0,658	0,388	Araçatuba, São Carlos, Rio Claro, Limeira, Piracicaba, Pirassununga, Moji Mirim, Campinas, Jundiaí, São José dos Campos, Osasco, Guarulhos, São Paulo, Santos	Limeira, Piracicaba, Campinas, Jundiaí, São José dos Campos, Osasco, Guarulhos, São Paulo, Santos	Frota de caminhão alta, IDHM alto e taxa de óbitos muito baixa

(a) Ano 2000

(b) Ano 2010

Figura 5.12 - Resultados do Algoritmo CHAID para óbitos por população

Verifica-se na Figura 5.12 que as microrregiões mais problemáticas, com maiores taxas de óbitos são aquelas classificadas nos nós 2 (verde) e 4 (amarelo). São identificadas também as microrregiões que foram classificadas erroneamente na amostra de validação.

Tabela 5.3 - Classificação das Microrregiões segundo o algoritmo CHAID para óbitos por área

Nó Terminal	Variáveis	Taxa média de óbitos	Microrregiões 2000	Microrregiões 2010	Nomenclatura proposta
2	Frota ônibus entre 530 e 1560	1,083	São José do Rio Preto, Catanduva, Jaboticabal, Araçatuba, Bauru, Jaú, Araraquara, Limeira, Piracicaba, São João da Boa Vista, Moji Mirim, Presidente Prudente, Sorocaba, Jundiaí, Guaratinguetá, Franco da Rocha, Itapeceira da Serra, Mogi das Cruzes, Santos	São José do Rio Preto, Catanduva, São Joaquim da Barra, Jaboticabal, Araçatuba, Bauru, Jaú, Araraquara, Limeira, Piracicaba, São João da Boa Vista, Moji Mirim, Presidente Prudente, Assis, Ourinhos, Tatuí, Jundiaí, Bragança Paulista, Itapeceira da Serra, Mogi das Cruzes	Frota de ônibus média e taxa de óbitos moderada
3	Frota de ônibus > 1560	3,143	Ribeirão Preto, Campinas, São José dos Campos, Osasco, Guarulhos, São Paulo	Ribeirão Preto, Campinas, Sorocaba, São José dos Campos, Osasco, Guarulhos, São Paulo, Santos	Frota de ônibus alta e taxa de óbitos muito alta
4	Frota de ônibus $\leq$ 530 Distância a Campinas $\leq$ 96,7	1,972	Rio Claro, Amparo, Tatuí, Bragança Paulista	Rio Claro, Amparo, Franco da Rocha	Frota de ônibus baixa, próximo a Campinas e taxa de óbitos alta
6	Frota de ônibus $\leq$ 530 Distância a Campinas > 96,7 Fluxo $\leq$ 684900214	0,366	Jales, Fernandópolis, Votuporanga, Auriflama, Nhandeara, Novo Horizonte, Barretos, São Joaquim da Barra, Ituverava, Franca, Batatais, Andradina, Birigui, Lins, São Carlos, Pirassununga, Dracena, Adamantina, Tupã, Marília, Assis, Itapeva, Itapetininga, Capão Bonito, Piedade, Campos do Jordão, Bananal, Paraibuna/Paraitinga, Caraguatatuba, Registro, Itanhaém	Jales, Fernandópolis, Votuporanga, Auriflama, Nhandeara, Novo Horizonte, Barretos, Ituverava, Franca, Batatais, Andradina, Birigui, Lins, São Carlos, Pirassununga, Dracena, Adamantina, Tupã, Marília, Itapeva, Itapetininga, Capão Bonito, Piedade, Campos do Jordão, Guaratinguetá, Bananal, Paraibuna/Paraitinga, Caraguatatuba, Registro, Itanhaém	Frota de ônibus baixa, distante de Campinas, baixo fluxo de veículos e taxa de óbitos muito baixa
7	Frota de ônibus $\leq$ 530 e Distância a Campinas > 96,7 Fluxo > 684900214	1,446	Avaré, Botucatu, Ourinhos	Avaré, Botucatu	Frota de ônibus baixa, distante de Campinas, alto fluxo de veículos e taxa de óbitos moderada

(a) Ano 2000

(b) Ano 2010

Figura 5.13 - Resultados do Algoritmo CHAID para óbitos por área

A Figura 5.13 localiza as classes de microrregiões agrupadas pelo algoritmo CHAID segundo as taxas de óbitos por áreas. As microrregiões mais problemáticas são aquelas agrupadas nos nós 3 e 4. Tais microrregiões possuem taxas de óbitos altas ou muito alta. Além disso, são identificadas na figura as microrregiões classificadas diferentemente na amostra de validação.

Tabela 5.4 - Classificação das Microrregiões segundo o algoritmo CHAID para óbitos por fluxo

Nó Terminal	Variáveis	Taxa média de óbitos	Microrregiões 2000	Microrregiões 2010	Nomenclatura proposta
1	Frota de automóvel $\leq$ 15058	2,583	Fernandópolis, Auriflama, Nhandeara, Novo Horizonte, Ituverava, Dracena, Capão Bonito, Campos do Jordão, Bananal, Paraibuna/Paraitinga, Registro, Itanhaém	Auriflama, Nhandeara, Novo Horizonte, Ituverava, Capão Bonito, Campos do Jordão, Bananal, Paraibuna/Paraitinga	Frota de automóvel baixa e taxa de óbitos muito alta
3	Frota de automóvel $>$ 15058 Distância a São Paulo $\leq$ 510,57	0,520	Votuporanga, São José do Rio Preto, Catanduva, Barretos, São Joaquim da Barra, Franca, Jaboticabal, Ribeirão Preto, Batatais, Araçatuba, Birigui, Lins, Bauru, Jaú, Avaré, Botucatu, Araraquara, São Carlos, Rio Claro, Limeira, Piracicaba, Pirassununga, São João da Boa Vista, Moji Mirim, Campinas, Amparo, Adamantina, Tupã, Marília, Assis, Ourinhos, Itapeva, Itapetininga, Tatuí, Piedade, Sorocaba, Jundiaí, Bragança Paulista, São José dos Campos, Guaratinguetá, Caraguatatuba, Osasco, Franco da Rocha, Guarulhos, Itapeceira da Serra, São Paulo, Mogi das Cruzes, Santos	Votuporanga, São José do Rio Preto, Catanduva, Barretos, São Joaquim da Barra, Franca, Jaboticabal, Ribeirão Preto, Batatais, Araçatuba, Birigui, Lins, Bauru, Jaú, Avaré, Botucatu, Araraquara, São Carlos, Rio Claro, Limeira, Piracicaba, Pirassununga, São João da Boa Vista, Moji Mirim, Campinas, Amparo, Adamantina, Tupã, Marília, Assis, Ourinhos, Itapeva, Itapetininga, Tatuí, Piedade, Sorocaba, Jundiaí, Bragança Paulista, São José dos Campos, Guaratinguetá, Caraguatatuba, Registro, Itanhaém, Osasco, Franco da Rocha, Guarulhos, Itapeceira da Serra, São Paulo, Mogi das Cruzes, Santos	Frota de automóvel alta, próximo de São Paulo e taxa de óbitos baixa
4	Frota de automóvel $>$ 15058 Distância a São Paulo $>$ 510,57	2,348	Jales, Andradina, Presidente Prudente	Jales, Fernandópolis, Andradina, Dracena, Presidente Prudente	Frota de automóvel alta, distante de São Paulo e taxa de óbitos muito alta

(a) Ano 2000

(b) Ano 2010

Figura 5.14 - Resultados do Algoritmo CHAID para óbitos por fluxo

A Figura 5.14 ilustra e localiza os três nós terminais obtidos pelo algoritmo CHAID segundo taxa de óbitos por fluxos. Estão sinalizadas as microrregiões classificadas em nós diferentes na amostra de validação. Os nós 1 e 4 são aqueles que compõem as microrregiões com maiores taxas.

Tabela 5.5 - Classificação das Microrregiões segundo o algoritmo CART para óbitos por população

Nó Terminal	Variáveis	Taxa média de óbitos	Microrregiões 2000	Microrregiões 2010	Nomenclatura proposta
1	IDHM $\leq 0,618$	2,087	Franca, Avaré, Ourinhos, Itapeva, Itapetininga, Capão Bonito, Piedade, Bananal, Paraibuna/Paraitinga, Registro	Itapeva, Capão Bonito, Piedade Bananal, Paraibuna/Paraitinga, Registro	IDHM baixo e taxa de óbitos muito alta
4	IDHM $> 0,618$ Distância a Osasco $> 509,53$	2,001	Jales, Andradina, Dracena, Presidente Prudente	Jales, Andradina, Dracena, Presidente Prudente	IDHM médio, distante de Osasco e taxa de óbitos muito alta
5	IDHM $> 0,618$ Distância a Osasco $\leq 509,53$ Distância a Piracicaba $\leq 62,65$	1,579	Botucatu, Rio Claro, Limeira, Piracicaba, Tatuí	Botucatu, Rio Claro, Limeira, Piracicaba, Tatuí	IDHM médio, próximo de Osasco, próximo de Piracicaba e taxa de óbitos alta
9	IDHM $> 0,618$ Distância a Osasco $\leq 509,53$ Distância a Piracicaba $> 62,65$ IDHM $\leq 0,658$ IDHM $\leq 0,657$	0,631	Fernandópolis, Votuporanga, São José do Rio Preto, Catanduva, Auriflama, Novo Horizonte, São Joaquim da Barra, Ituverava, Jaboticabal, Batatais, Birigui, Lins, Bauru, Jaú, Araraquara, São João da Boa Vista, Adamantina, Tupã, Marília, Assis, Sorocaba, Bragança Paulista, Campos do Jordão, Itanhaém, Franco da Rocha, Itapecerica da Serra, Mogi das Cruzes	Fernandópolis, Votuporanga, São José do Rio Preto, Catanduva, Auriflama, Novo Horizonte, Barretos, São Joaquim da Barra, Ituverava, Franca, Jaboticabal, Ribeirão Preto, Batatais, Araçatuba, Birigui, Lins, Bauru, Jaú, Avaré, Araraquara, São Carlos, Pirassununga, São João da Boa Vista, Moji Mirim, Amparo, Adamantina, Tupã, Marília, Assis, Ourinhos, Itapetininga, Sorocaba, Bragança Paulista, Campos do Jordão, Guaratinguetá, Itanhaém, Franco da Rocha, Itapecerica da Serra, Mogi das Cruzes	IDHM médio, próximo de Osasco, distante de Piracicaba e taxa de óbitos baixa
10	IDHM $> 0,618$ Distância a Osasco $\leq 509,53$ Piracicaba $> 62,65$ IDHM $\leq 0,658$ IDHM $> 0,657$	1,729	Ribeirão Preto, Amparo, Guaratinguetá		IDHM médio, próximo de Osasco, distante de Piracicaba e taxa de óbitos alta

Nó Terminal	Variáveis	Taxa média de óbitos	Microrregiões 2000	Microrregiões 2010	Nomenclatura proposta
11	IDHM > 0,618 Distância a Osasco ≤ 509,53 Piracicaba > 62,65 IDHM > 0,658 Área ≤ 3623,20	0,192	Nhandeara, Barretos, São Carlos, Pirassununga, Moji Mirim, Campinas, Jundiaí, Caraguatatuba, Osasco, Guarulhos, São Paulo, Santos	Nhandeara, Campinas, Jundiaí, Caraguatatuba, Osasco, Guarulhos, São Paulo, Santos	IDHM alto, próximo de Osasco, distante de Piracicaba, área pequena e taxa de óbitos muito baixa
12	IDHM > 0,618 Distância a Osasco ≤ 509,53 Piracicaba > 62,65 IDHM > 0,658 Área > 3623,20	0,862	Araçatuba, São José dos Campos	São José dos Campos	IDHM alto, próximo de Osasco, distante de Piracicaba, área grande e taxa de óbitos baixa

(a) Ano 2000

(b) Ano 2010

Figura 5.15 - Resultados do Algoritmo CART para óbitos por população

A Figura 5.15 ilustra as microrregiões do estado, classificadas em seis nós terminais. Os grupos mais problemáticos em termos de taxa de óbitos são nós 1 e 4.

Tabela 5.6 - Classificação das Microrregiões segundo o algoritmo CART para óbitos por área

Nó Terminal	Variáveis	Taxa média de óbitos	Microrregiões 2000	Microrregiões	Nomenclatura proposta
3	Frota de ônibus $\leq$ 1615 Distância a Osasco $\leq$ 71,22	2,039	Sorocaba, Jundiaí, Bragança Paulista, Franco da Rocha, Itapeverica da Serra	Jundiaí, Bragança Paulista, Franco da Rocha, Itapeverica da Serra	Frota de ônibus baixa, próximo de Osasco e taxa de óbitos muito alta
5	Frota de ônibus $>$ 1615 Distância a Santos $\leq$ 121,25	3,751	São José dos Campos, Osasco, Guarulhos, São Paulo	São José dos Campos, Osasco, Guarulhos, São Paulo	Frota de ônibus alta, próximo de Santos e taxa de óbitos muito alta
6	Frota de ônibus $>$ 1615 Distância a Santos $>$ 121,25	1,928	Ribeirão Preto, Campinas	Ribeirão Preto, Campinas, Sorocaba	Frota de ônibus alta, distância a Santos intermediária e taxa de óbitos alta
12	Frota de ônibus $\leq$ 1615 Distância a Osasco $>$ 509,53	0,790	Jales, Andradina, Dracena, Presidente Prudente	Jales, Andradina, Dracena, Presidente Prudente	Frota de ônibus baixa, distante de Osasco e taxa de óbitos baixa
13	Frota de ônibus $\leq$ 1615 Distância a Osasco $>$ 71,22 Distância a Osasco $\leq$ 335,18 Distância a Santos $\leq$ 137,34 Fluxo $\leq$ 351453053	0,379	Paraibuna/Paraitinga, Caraguatatuba, Itanhaém	Paraibuna/Paraitinga, Caraguatatuba, Itanhaém	Frota de ônibus baixa, distância a Osasco intermediária, distância a Santos intermediária, fluxo de veículos baixo e taxa de óbitos muito baixa
14	Frota de ônibus $\leq$ 1615 Distância a Osasco $>$ 71,22 Distância a Osasco $\leq$ 335,18 Distância a Santos $\leq$ 137,34 Fluxo $>$ 351453053	0,140	Piedade, Campos do Jordão, Mogi das Cruzes, Santos	Piedade, Campos do Jordão, Mogi das Cruzes, Santos	Frota de ônibus baixa, distância a Osasco intermediária, distância a Santos intermediária, fluxo de veículos alto e taxa de óbitos muito baixa

Nó Terminal	Variáveis	Taxa média de óbitos	Microrregiões 2000	Microrregiões	Nomenclatura proposta
15	Frota de ônibus ≤ 1615 Distância a Osasco > 71,22 Distância a Osasco ≤ 335,18 Distância a Santos > 137,34 Distância a Santos ≤ 177,01	2,655	Amparo, Guaratinguetá	Amparo, Guaratinguetá	Frota de ônibus baixa, distância a Osasco intermediária, distância a Santos intermediária e taxa de óbitos muito alta
16	Frota de ônibus ≤ 1615 Distância a Osasco > 71,22 Distância a Osasco ≤ 335,18 Distância a Santos > 177,01	0,870	Catanduva, Franca, Jaboticabal, Batatais, Bauru, Jaú, Avaré, Botucatu, Araraquara, São Carlos, Rio Claro, Limeira, Piracicaba, Pirassununga, São João da Boa Vista, Moji Mirim, Ourinhos, Itapeva, Itapetininga, Tatuí, Capão Bonito, Bananal, Registro	Catanduva, Franca, Jaboticabal, Batatais, Bauru, Jaú, Avaré, Botucatu, Araraquara, São Carlos, Rio Claro, Limeira, Piracicaba, Pirassununga, São João da Boa Vista, Moji Mirim, Ourinhos, Itapeva, Itapetininga, Tatuí, Capão Bonito, Bananal, Registro	Frota de ônibus baixa, distância a Osasco intermediária, distante de Santos e taxa de óbitos baixa
17	Frota de ônibus ≤ 1615 Distância a Osasco > 71,22 Distância a Osasco > 335,18 Distância a Osasco ≤ 509,53 IDHM ≤ 0,639	0,568	Ituverava, Birigui, Lins, Tupã	Votuporanga, Birigui, Adamantina, Tupã	Frota de ônibus baixa, distância a Osasco intermediária, IDHM baixo e taxa de óbitos baixa
18	Frota de ônibus ≤ 1615 Distância a Osasco > 71,22 Distância a Osasco > 335,18 Distância a Osasco ≤ 509,53 IDHM > 0,639	0,124	Fernandópolis, Votuporanga, São José do Rio Preto, Auriflama, Nhandeara, Novo Horizonte, Barretos, São Joaquim da Barra, Araçatuba, Adamantina, Marília, Assis	Fernandópolis, São José do Rio Preto, Auriflama, Nhandeara, Novo Horizonte, Barretos, São Joaquim da Barra, Ituverava, Araçatuba, Lins, Marília, Assis	Frota de ônibus baixa, distância a Osasco intermediária, IDHM alto e taxa de óbitos muito baixa

Figura 5.16 - Resultados do Algoritmo CART para óbitos por área

A Figura 5.16 ilustra os nós terminais obtidos pelo algoritmo CART segundo taxa de óbitos por área. São sinalizadas as microrregiões classificadas diferentemente na amostra de validação (2010).

5.3 REGRESSÃO LINEAR MÚLTIPLA (RLM)

As equações desta seção foram calibradas com o banco de dados do ano 2000, tal como na geração das árvores de treinamento. Foram utilizadas como variáveis dependentes na RLM *stepwise* as taxas de óbitos por automóvel pela população,

área e fluxo, divididas pelas médias. As variáveis independentes foram população residente, PIB a preços correntes em mil reais, pessoas com rendimento e pessoas ocupadas, área, frotas de automóvel, caminhão, motocicleta e ônibus, IDHM, fluxo veicular anual e distâncias entre microrregiões, mostradas no capítulo 4 desta dissertação, e variam de acordo com a escolha da variável dependente, sendo excluída da análise a mesma que estiver no denominador da taxa de óbito. A Tabela 5.7 apresenta quais as variáveis independentes com maior contribuição para cada modelo, selecionadas considerando os óbitos provenientes de acidentes por automóvel. Cada modelo passou pelas etapas de análise de RLM (ajuste, significância das variáveis e principais suposições).

Tabela 5.7 - Resultados da Análise de RLM *stepwise* (ajuste dos modelos)

Variável dependente	Variáveis independentes significativas	R ²	Z	Sig.
Óbito por população	IDHM	0,145	10,356	0,002
Óbito por área	Fluxo	0,400	20,038	0,000
	Distância_São Paulo			
Óbito por fluxo	Distância_Piracicaba	0,199	7,462	0,001
	IDHM			

A precisão do modelo de regressão pode ser melhorada com adição de uma constante. Se todos os valores das variáveis independentes forem nulos, a variável dependente terá o mesmo valor da constante. Os modelos de óbitos por acidentes de automóvel analisados apresentaram constante em todas as taxas, ou seja, a constante era estatisticamente significativa.

O R² (coeficiente de determinação) é uma medida da proporção da variação da variável dependente em torno de sua média que pode ser explicada pelas variáveis independentes, o valor está no intervalo entre 0 e 1. A taxa que obteve o maior coeficiente de determinação foi de óbito proveniente de acidente de automóvel por área.

Os modelos selecionaram poucas variáveis independentes significativas para cada taxa, o máximo de variáveis independentes que explicaram um modelo foi 2. Dentre as variáveis independentes, a média de IDHM esteve presente em dois dos três resultados analisados.

Todas as variáveis independentes selecionadas foram consideradas significativas, pois o procedimento *stepwise* seleciona apenas aquelas significativas. Os valores obtidos dos coeficientes das variáveis para calibração dos modelos foram considerados adequados, assim como os sinais obtidos e relações esperadas.

As equações dos modelos calibrados foram:

Óbitos provenientes de automóvel por população:

$$y = 10,37 - 9,37IDHM$$

Óbitos provenientes de automóvel por área:

$$y = 1,17 + 0,29Fluxo - 0,46Distância_SãoPaulo$$

Óbitos provenientes de automóvel por fluxo:

$$y = 10,07 + 1,39Distância_Piracicaba - 10,46IDHM$$

A Tabela 5.8 apresenta o valor *t* da constante e de cada variável independente selecionada em cada modelo da RLM. Hair et al. (2009) explicam que o valor *t* ajuda a determinar se alguma variável deve ser descartada da equação uma vez que outra variável independente tenha sido acrescentada.

Tabela 5.8 - Resultados do valor *t* da Análise de RLM

Variável dependente	Variáveis independentes significativas/constante	t
Óbito por população	Constante	3,558
	IDHM	-3,218
Óbito por área	Constante	4,486
	Fluxo	4,061
	Distância_São Paulo	-2,365
Óbito por fluxo	Constante	1,924
	Distância_Piracicaba	3,041
	IDHM	-2,028

5.3.1 Resíduos

A diferença entre o valor observado da variável dependente e seu valor previsto gera o resíduo. A análise dos resíduos pode ser feita através de gráficos que verificam as seguintes suposições: linearidade, homocedasticidade e a normalidade da RLM. A linearidade da relação entre variáveis dependente e independentes mede quanto uma influencia na outra de forma linear e constante. O pressuposto de normalidade estabelece que os valores da variável dependente são normalmente distribuídos para cada valor da variável independente. A homocedasticidade pode ser definida como a variância constante dos erros para diferentes observações. Os gráficos das figuras seguintes foram gerados pelo *software* IBM SPSS 22.0.

Figura 5.17 - Análise gráfica de resíduos do modelo para variável dependente óbitos por população

(a) Histograma dos resíduos

(b) Gráfico de probabilidade normal acumulada

(c) Análise de resíduos padronizados

Figura 5.18 - Análise gráfica de resíduos do modelo para variável dependente óbitos por área

(a) Histograma dos resíduos

(b) Gráfico de probabilidade normal acumulada

(c) Análise de resíduos padronizados

Figura 5.19 - Análise gráfica de resíduos do modelo para variável dependente óbitos por fluxo

A análise de resíduos para examinar se o modelo de regressão está adequado foi feita, através de uma verificação visual dos gráficos. Segundo observado nos histogramas das Figuras 5.17a, 5.18a e 5.19a, apenas a análise da relação de óbitos por área se aproxima da curva normal. Ao observar os gráficos de probabilidade normal nas Figuras 5.17b, 5.18b e 5.19b, percebe-se um desvio significativo da normalidade, ainda sim as variáveis continuaram na análise, a efeito de comparação dos resultados. Nas Figuras 5.17c, 5.18c e 5.19c nota-se um aumento da dispersão a medida que aumenta o valor da variável dependente. Observa-se que os gráficos apresentam heterocedasticidade.

Conforme dito no capítulo 4, as variáveis objeto de estudo não apresentam mesma distribuição normal. A RLM, neste caso, foi utilizada apenas para fins comparativos dos modelos de AD com uma técnica paramétrica.

5.4 COMPARAÇÃO DOS RESULTADOS

A comparação dos resultados foi baseada na validação dos modelos (amostra de 2010), das técnicas de AD e RLM, através das 63 observações das taxas de óbitos provenientes de acidentes por automóvel por população, área e fluxo.

As medidas de desempenho calculadas para análise dos erros foram o erro médio, erro médio normalizado, erro médio absoluto e o coeficiente de correlação, de

acordo com as Equações 5.1, 5.2, 5.3 e 5.4, respectivamente. Onde x_i é o valor estimado, y_i o valor observado, N o número de observações, $\bar{x}$, $\bar{y}$ médias e σ_x , σ_y desvio padrão.

$$E_M = \frac{1}{N} \sum_{i=1}^N (x_i - y_i) \quad (5.1)$$

E_M = Erro médio

$$E_{MN} = \frac{1}{N} \sum_{i=1}^N \frac{(x_i - y_i)}{y_i} \quad (5.2)$$

E_{MN} = Erro médio normalizado

$$E_{MA} = \frac{1}{N} \sum_{i=1}^N |x_i - y_i| \quad (5.3)$$

E_{MA} = Erro médio absoluto

$$r = \frac{1}{N-1} \sum_{i=1}^N \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y} \quad (5.4)$$

r = Coeficiente de correlação

As medidas de desempenho calculadas para os algoritmos CHAID e CART da AD e para a RLM são apresentadas em seguida nas tabelas, e compara as taxas de óbitos por população, área e fluxo.

Tabela 5.9 - Comparação dos resultados do algoritmo CHAID

Relação de óbitos	Erro médio	Erro médio normalizado	Erro médio absoluto	Coef. Correlação
População	-0,05888	0,575878	0,614176	0,132076
Área	0,056921	0,893264	0,670641	0,528139
Fluxo	-0,07295	1,834381	0,911487	0,372391

A técnica de AD pelo algoritmo CHAID, Tabela 5.9, encontrou melhor resultado de erro médio e coeficiente de correlação na taxa de óbitos por área. A taxa de óbitos

por população esteve melhor que as outras taxas para as medidas erro médio normalizado e erro médio absoluto.

Tabela 5.10 - Comparação dos resultados do algoritmo CART

Relação de óbitos	Erro médio	Erro médio normalizado	Erro médio absoluto	Coef. Correlação
População	-0,12055	0,380228	0,63225	0,297889
Área	-0,00177	0,709537	0,624866	0,641523
Fluxo	-	-	-	-

Foram observadas três melhores medidas de desempenho na taxa de óbitos por área e uma na taxa de óbitos por população, Tabela 5.10, utilizando o algoritmo CART da AD. Nesta técnica a taxa de óbitos por fluxo não selecionou nenhuma das variáveis independentes na análise.

Tabela 5.11 - Comparação dos resultados da RLM

Relação de óbitos	Erro médio	Erro médio normalizado	Erro médio absoluto	Coef. Correlação
População	-3,41E-15	0,811299	0,550186	0,329412
Área	-5,78E-16	0,915434	0,571171	0,703989
Fluxo	-5,01E-15	1,138756	0,951121	0,440021

A Tabela 5.11 apresenta os resultados da RLM e indica que a taxa de óbitos por área se destaca pelo menor erro médio e maior coeficiente de correlação. A taxa de óbitos por população possui melhor resultado de erro médio normalizado e erro médio absoluto.

Observando os resultados apresentados na Tabela 5.9, 5.10 e 5.11, verifica-se que as taxas de óbitos por população e por área são mais razoáveis para representação do objeto de estudo.

Comparando o desempenho das técnicas, o algoritmo CART e a Regressão Linear Múltipla geraram menores erros para as variáveis dependentes taxas de óbitos por população e por área, respectivamente.

Tabela 5.12 - Comparação dos resultados da taxa de óbitos por população

Técnica	Erro médio	Erro médio normalizado	Erro médio absoluto	Coef. Correlação
CHAID	-0,05888	0,575878	0,614176	0,132076
CART	-0,12055	0,380228	0,63225	0,297889
Regressão	-3,41E-15	0,811299	0,550186	0,329412

Tabela 5.13 - Comparação dos resultados da taxa de óbitos por área

Técnica	Erro médio	Erro médio normalizado	Erro médio absoluto	Coef. Correlação
CHAID	0,056921	0,893264	0,670641	0,528139
CART	-0,00177	0,709537	0,624866	0,641523
Regressão	-5,78E-16	0,915434	0,571171	0,703989

A taxa de óbitos por fluxo obteve resultados em apenas duas das técnicas. A RLM assim como nas outras taxas, obteve melhor resultado para três das quatro medidas de desempenho. Com exceção do erro médio normalizado, as demais medidas estiveram melhores na regressão.

Tabela 5.14 - Comparação dos resultados da taxa de óbitos por fluxo

Técnica	Erro médio	Erro médio normalizado	Erro médio absoluto	Coef. Correlação
CHAID	-0,07295	1,834381	0,911487	0,372391
CART	-	-	-	-
Regressão	-5,01E-15	1,138756	0,951121	0,440021

Observou-se que a técnica de AD teve melhor desempenho ao considerar o cálculo do erro médio normalizado. Já a RLM teve melhor resultado de erro médio, erro médio absoluto e coeficiente de correlação entre os valores estimados e observados. O coeficiente de correlação demonstra proporção entre valores observados e estimados e não necessariamente igualdade. É de se esperar valores mais altos de coeficiente de correlação para o caso do modelo linear. Ao analisar as taxas de óbitos por automóvel, a relação por população e área mostraram melhores resultados. A Tabela 5.15 resume quais técnicas e variáveis mostraram melhor desempenho das medidas de ajuste ao final da análise.

Tabela 5.15 - Resumo do resultado final da análise

	Erro médio	Erro médio normalizado	Erro médio absoluto	Coef. Correlação
Técnica	RLM	CART	RLM	RLM
Taxa de óbito	Área	População	População	Área

Vale ressaltar ainda que os algoritmos CART e CHAID também apresentaram erros relativamente pequenos para os parâmetros Erro médio e Erro médio absoluto. O coeficiente de correlação não seria um parâmetro de desempenho adequado e sim uma medida auxiliar que relaciona valores observados e estimados.

Além disso, conforme mencionado anteriormente, a RLM é mais rígida quanto aos pressupostos de normalidade e homocedasticidade, os quais não são atendidos neste estudo.

6 CONCLUSÕES

6.1 SÍNTESE DO TRABALHO

O objetivo deste trabalho é explorar técnicas de mineração de dados (algoritmos de Árvore de Decisão - AD) para identificar as variáveis envolvidas nos óbitos em uma análise agregada por microrregiões do estado de São Paulo (SP). Associado a este objetivo, uma das metas do trabalho é explorar formas de mensurar óbitos provenientes de acidentes de automóvel numa análise agregada por área. Assim, foram testadas três variáveis dependentes em forma de taxas: (1) Razão de óbitos por população; (2) Razão de óbitos por área; (3) Razão de óbitos por fluxo veicular.

A área de estudo foi o estado de SP, dividido em microrregiões e os dados utilizados nos modelos de calibração e validação foram retirados do DATASUS, IBGE, DENATRAN, Atlas Brasil e da pesquisa realizada por Isler (2015), para os anos de 2000 na calibração dos modelos e 2010 na validação.

Neste trabalho, foram utilizados dois algoritmos de AD, CHAID e CART, além do modelo de Regressão Linear Múltipla (RLM) na calibração dos modelos. A escolha da técnica de RLM deu-se para fins comparativos do desempenho dos modelos de AD, levando-se em conta também o seu caráter paramétrico.

Os resultados dos modelos concluíram que a taxa de óbito por população apresentou melhor resultado comparada às taxas por área e fluxo. Quanto ao desempenho das técnicas, mostram-se adequadas para as análises realizadas, pois apresentaram resultados próximos. Observou-se que a RLM foi a técnica que obteve melhores erro médio, erro médio absoluto e coeficiente de correlação, e o algoritmo CART da AD teve destaque com o menor erro médio normalizado. Ao comparar as taxas de óbitos, a relação por área apresentou melhor erro médio e coeficiente de correlação, já a relação por população obteve menor erro médio normalizado e erro médio absoluto.

6.2 PRINCIPAIS CONTRIBUIÇÕES E CONCLUSÕES

Esta dissertação contribuiu para o estudo de técnicas de modelagem e formas de mensurar óbitos de acidentes por automóvel em análise agregada, assim como identificar o comportamento das variáveis independentes que possuem influência nos óbitos. Ressalta-se que tais modelos incluem informações socioeconômicas e demográficas e são aplicados ao nível estratégico de planejamento urbano, municipal ou estadual.

Assim, foi atendido o objetivo principal da pesquisa de explorar técnicas de mineração de dados, além da técnica tradicional, para identificar as variáveis envolvidas nos óbitos em uma análise agregada. As ferramentas utilizadas no processo de análise e a elaboração do banco de dados possibilitaram alcançar o objetivo.

Os resultados do capítulo anterior indicam que os algoritmos de AD se mostraram úteis para a classificação das microrregiões considerando variáveis demográficas, de frota, distâncias e médias de óbitos ocorridos em cada classe de microrregiões do estado de São Paulo. Além da sua característica de ferramenta de classificação, os algoritmos de AD também têm o seu poder de previsão, sendo o valor estimado associado a valores médios da variável dependente relativo a cada grupo ou classe.

A utilização de uma rede triangulada e do fluxo veicular anual sintético permitiu a estimativa de uma variável importante, artificialmente, em análises de acidentes de trânsito, sendo uma contribuição metodológica da dissertação. O Volume Médio Diário Anual (VDMA) é uma variável explicativa nos modelos de previsão de acidentes. Encontra-se com bastante frequência tal variável nos modelos obtidos no cenário nacional (Cardoso, 2006; Claude, 2012; Barbosa e Costa, 2011) e na literatura internacional (Sayed e Rodrigues, 1999; Drummer et al., 2004). Entretanto, muitas vezes não há informação disponível relativa a fluxos veiculares. Na literatura da Engenharia de Tráfego, observam-se diversos trabalhos que visam à obtenção de fluxos sintéticos através de simulação microscópica (Bessa e Setti, 2011; Lin et al., 2005; Araújo e Setti, 2007). No entanto, para geração de dados sintéticos, tais técnicas necessitam de dados de detectores, por exemplo. Levando-se em conta

uma pesquisa regional, em âmbito municipal, estadual ou nacional, por exemplo, muitas vezes dados de matrizes O/D e alocação de fluxos nos diversos tramos da rede rodoviária, não se encontram disponíveis. Neste contexto, o presente trabalho propôs também um procedimento simplificado para obtenção de fluxo veicular sintético (VDMA Sintético) para auxílio na modelagem de acidentes de trânsito.

6.3 LIMITAÇÕES METODOLÓGICAS

- O uso da RLM foi útil para efeito de comparação de resultados de um modelo paramétrico e confirmatório com modelos não paramétricos e exploratórios (CART e CHAID). Importante mencionar que a Regressão Linear não é uma técnica adequada para o caso estudado, uma vez que os acidentes de trânsito não seguem uma distribuição normal. Acidentes de trânsito são variáveis de contagem, não negativas, com hipótese de distribuição dos erros seguindo distribuição Binomial Negativa ou de Poisson. Neste caso a utilização de Modelos Lineares Generalizados para fins comparativos seria, sem dúvida, mais apropriada.
- As técnicas exploratórias apresentadas mostram-se adequadas para as análises, embora o pouco número de observações (63) seja uma limitação para tratar com as árvores de decisão, que apresenta melhores resultados com bancos de dados maiores. Daí a necessidade de utilizar critérios de partição pequenos (mínimo de duas observações por classe) para tornar possível a partição dos dados.

Pode-se discutir ainda a importância da inclusão de fatores diferentes dos analisados neste trabalho. O fenômeno de acidentes de trânsito severos é bastante complexo e multidimensional, neste trabalho foram investigados aspectos demográficos e de fluxo sintético, observa-se ainda possíveis influências de: (1) fatores da geometria das vias; (2) extensão de rede viária urbana ou rodoviária por microrregião; (3) extensão de rodovias com pistas simples; (4) extensão de rodovias com pista dupla; (5) Fatores de uso do solo, etc.

- A incorporação de outras variáveis, além daquelas apresentadas nesta dissertação pode gerar melhores resultados, uma vez que demais fatores estão envolvidos nos óbitos de acidente por automóvel. A inclusão de novas variáveis pode contribuir nas análises realizadas, e ainda representar bom desempenho.
- Os dados de mortalidade disponíveis no DATASUS não identificam se o óbito ocorreu em vias urbanas ou em rodovias, além de não estarem georreferenciados, fato que impediu que outros tipos de análises fossem realizadas. Além disso, os óbitos do banco de dados da dissertação utiliza a microrregião de ocorrência, apesar de muitos registros acontecerem com pessoas que estejam de passagem em determinada microrregião. Adicionalmente, as variáveis explicativas utilizadas são inerentes à população residente da microrregião, descaracterizando os acidentes da população “de passagem”.
- A opção de ter usado a rede de veículos com base na triangulação se deu devido ao não acesso à rede rodoviária real georreferenciada até o momento de finalização deste trabalho.
- A falta de validação do procedimento proposto para o fluxo de veículos é outra limitação importante. Seria necessária a repetição do método para a rede de veículos georreferenciada e a comparação de tais dados ao VDMA proveniente de contagens de tráfego, sobretudo em praças de pedágio.

6.4 RECOMENDAÇÕES PARA TRABALHOS FUTUROS

Espera-se que os resultados apresentados nesta dissertação possam orientar novos estudos que aprimorem o método apresentado, bem como análises de resultados. Pode-se verificar, também, o potencial de outras técnicas de mineração de dados, como redes neurais, por exemplo, para o objeto de estudo.

A mortalidade no trânsito engloba outras categorias, além de óbitos por automóvel, como pedestre, ciclista, motociclista, ocupante de triciclo, caminhonete, veículo de

transporte pesado e ônibus. Essas categorias, se utilizadas como variáveis dependentes, permitem gerar um comparativo aos óbitos por automóvel. O uso de drogas e álcool, relevo, clima e sinalização como variáveis independentes podem gerar previsões mais detalhadas.

Além disso, a aplicação do método em outros estados do Brasil pode ser importante para medir o desempenho de cada um ao longo dos anos e no planejamento de ações com objetivo de redução de óbitos no trânsito. E, recomenda-se o uso de séries históricas para incremento de análises, além de incorporação de variáveis que caracterizem o sistema viário urbano e rodoviário pertencente a cada microrregião.

REFERÊNCIAS BIBLIOGRÁFICAS

American Association of State and Highway Transportation Officials. (2010) *Highway Safety Manual* (1ª ed.). Washington.

Araújo, J. J.; Setti J. R. (2007) *Análise do efeito do tráfego de veículos pesados sobre uma ponte através de simulação microscópica*. XXI ANPET - Associação Nacional de Pesquisa e Ensino em Transportes. Rio de Janeiro, RJ.

Atlas do Desenvolvimento Humano no Brasil. (2015) *Atlas Brasil*. Disponível em: <www.atlasbrasil.org.br> Acesso em: 25 abr. 2015.

Barbetta, P. A. (2012) *Estatística aplicada às Ciências Sociais* (8ª ed. rev.). Editora da UFSC, Florianópolis.

Barbosa, H. M.; Costa, T. G. (2011) *Modelos de previsão de acidentes de trânsito em vias urbanas de Belo Horizonte*. XXV ANPET – Congresso de Pesquisa e Ensino em Transportes. Panorama Nacional da Pesquisa em Transportes 2011, p. 1959-1970.

Berry, M. J. A.; Linoff, G. S. (2004) *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management* (2ª ed.). Wiley Publishing, Inc., Indianapolis, IN, USA.

Bessa, J. E.; Setti, J. R. (2011) *Derivation of ATS and PTSF Functions for Two-lane, Rural Highways in Brazil*. 6th International Symposium on Highway Capacity and Quality of Service. Procedia - Social and Behavioral Sciences, v. 16, p. 282-292.

Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. (1984). *Classification and regression trees*. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, USA.

Cardoso, G. (2006) *Modelos para Previsão de Acidentes de Trânsito em Vias Arteriais Urbanas*. Tese de Doutorado, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS.

Cardoso, G.; Goldner, L. G. (2007) *Desenvolvimento e aplicação de modelos para previsão de acidentes de trânsito*. Transportes, v. 15, n. 2, p. 43-51.

Castro, M.; Paleti, R.; Bhat, C. R. (2012) *A latent variable representation of count data models to accommodate spatial and temporal dependence: Application to predicting crash frequency at intersections*. Transportation Research Part B: Methodological, v. 46, n. 1, p. 253-272.

Chang, L. (2005) *Analysis of freeway accident frequencies: Negative binomial regression versus artificial neural network*. Safety Science, v. 43, n. 8, p. 541-557.

Chin, H. C.; Quddus, M. A. (2003) *Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections*. Accident Analysis & Prevention, v. 35, n. 2, p. 253-259.

Claude, G. F. M. (2012) *Previsão da ocorrência de acidentes de trânsito em interseções de vias arteriais urbanas – O caso de Taguatinga/DF*. Dissertação de Mestrado. Universidade de Brasília, Brasília, DF.

Cunto, F. J. C.; Castro Neto, M. M.; Barreira, D. S. (2012) *Modelos de previsão de acidentes de trânsito em interseções semaforizadas de Fortaleza*. Transportes, v. 20, n. 2, p. 55-62.

Departamento de Informática do SUS. Sistema de Informações sobre Mortalidade. (2010) *Estatísticas Vitais*. Disponível em: <tab-net.datasus.gov.br>. Acesso em: 15 abr. 2014.

Departamento Nacional de Trânsito. (2010) *Frota de veículos*. Disponível em: <www.denatran.gov.br>. Acesso em: 29 abr. 2014.

Drummer, O. H.; Gerostamoulos, J.; Batziris, H.; Chu, M.; Caplehorn, J.; Robertson, M. D.; Swann, P. (2004) *The involvement of drugs in drivers of motor vehicles killed in Australian road traffic crashes*. Accident Analysis & Prevention, v. 36, n. 2, p. 239-248.

Elvik, R.; Høy, A.; Vaa, T.; Sørensen, M. (2009) *The Handbook of Road Safety Measures* (2^a ed.). Emerald Group Publishing Limited, Bingley, UK.

Ferraz, A. C. P.; Raia Júnior, A. A.; Bezerra, B. S.; Bastos, J. T.; Silva, K. C. R. (2012) *Segurança Viária* (1^a ed.). Suprema Gráfica e Editora, São Carlos, SP.

Field, A. (2013) *Discovering Statistics using IBM SPSS Statistics* (4^a ed.). SAGE Publications Inc., Thousand Oaks, CA, USA.

Geedipally, S. R.; Lord, D. (2008) *Effects of Varying Dispersion Parameter of Poisson–Gamma Models on Estimation of Confidence Intervals of Crash Prediction Models*. Transportation Research Record: Journal of the Transportation Research Board, n. 2061, p. 46–54. Transportation Research Board of the National Academies, Washington, DC, USA.

Goldschmidt R.; Passos, E. (2005) *Data Mining: Um Guia Prático* (1^a ed.). Elsevier, Rio de Janeiro, RJ.

Gonçalves, D. N. S.; Silva, M. A.; d'Agosto, M. A. (2015) *Procedimento para uso de Redes Neurais Artificiais no planejamento estratégico de fluxo de carga no Brasil*. Journal of Transport Literature, v. 9, n.1, p. 45-49.

Guevara, F. L.; Washington, S. P.; Oh, J. (2004) *Forecasting Crashes at the Planning Level: Simultaneous Negative Binomial Crash Model Applied in Tucson, Arizona*. Transportation Research Record: Journal of the Transportation Research Board, n. 1897, p. 191–199. Transportation Research Board, National Research Council, Washington, DC, USA.

Hadayeghi, A.; Shalaby, A. S.; Persaud, B. N. (2003) *Macrolevel Accident Prediction Models for Evaluating Safety of Urban Transportation Systems*. Transportation Research Record: Journal of the Transportation Research Board, n. 1840, p. 87-95. Transportation Research Board, National Research Council, Washington, DC, USA.

Instituto Brasileiro de Geografia e Estatística. (2010) *Censo Demográfico 2010*. Disponível em: <www.ibge.gov.br>. Acesso em: 15 abr. 2014.

Isler, C. A. (2015) *Avaliação socioeconômica de uma rede ferroviária regional para o transporte de passageiros*. Tese de Doutorado, Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, SP.

Jonsson, T.; Ivan, J. N.; Zhang, C. (2007) *Crash Prediction Models for Intersections on Rural Multilane Highways: Differences by Collision Type*. Transportation Research Record: Journal of the Transportation Research Board, n. 2019, p. 91–98. Transportation Research Board of the National Academies, Washington, DC, USA.

Kass, G.V. (1980) *An Exploratory Technique for Investigating Large Quantities of Categorical Data*. Journal of the Royal Statistical Society. Series C (Applied Statistics), v. 29, n. 2, p. 119-127. Wiley for the Royal Statistical Society.

Lao, Y.; Wu, Y.; Corey, J.; Wang, Y. (2011) *Modeling animal-vehicle collisions using diagonal inflated bivariate Poisson regression*. Accident Analysis & Prevention, v. 43, n. 1, p. 220-227.

Lin, S. M.; Haney, R. P.; Campa, M. J.; Fitzgerald, M. C.; Patz, E. F. (2005). *Characterising phase variations in MALDI-TOF data and correcting them by peak alignment*. Cancer Informatics, v. 1, n. 1, p. 32-40.

Lord, D.; Mannering, F. (2010) *The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives*. Transportation Research Part A: Policy and Practice, v. 44, n. 5, p. 291-305.

Lord, D.; Washington, S.; Ivan, J. N. (2007) *Further notes on the application of zero-inflated models in highway safety*, Accident Analysis & Prevention, v. 39, n. 1, p. 53-57.

Lord, D.; Washington, S. P.; Ivan, J. N. (2005) *Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory*. Accident Analysis & Prevention, v. 37, n. 1, p. 35-46.

Lovegrove, G. R.; Sayed, T. (2007) *Macrolevel Collision Prediction Models to Enhance Traditional Reactive Road Safety Improvement Programs*. Transportation Research Record: Journal of the Transportation Research Board, n. 2019, p. 65-73. Transportation Research Board of the National Academies, Washington, DC, USA.

Malyshkina, N. V.; Mannering, F. L. (2010) *Zero-state Markov switching count-data models: An empirical assessment*. Accident Analysis & Prevention, v. 42, n. 1, p. 122-130.

Mannering, F. L.; Bhat, C. R. (2014) *Analytic methods in accident research: Methodological frontier and future directions*. Analytic Methods in Accident Research, v. 1, p. 1-22.

Mannila, H. (1997) *Methods and problems in data mining*. Database Theory — ICDT '97, p. 41–55. Springer Berlin Heidelberg.

Mozolin, M.; Thill, J.-C.; Usery, E. L. (2000) *Trip distribution forecasting with multilayer perceptron neural networks: A critical evaluation*. Transportation Research Part B: Methodological, v. 34, n. 1, p. 53-73.

Oña, J.; López, G.; Abellán, J. (2013) *Extracting decision rules from police accident reports through decision trees*. Accident Analysis & Prevention, v. 50, p. 1151-1160.

Pulugurtha, S. S.; Duddu, V. R.; Kotagiri, Y. (2013) *Traffic analysis zone level crash estimation models based on land use characteristics*. Accident Analysis & Prevention, v. 50, p. 678-687.

Plano Nacional de Logística e Transportes – PNLT. (2015) Disponível em: <www2.transportes.gov.br/bit/01-inicial/pnlt.html>. Acesso em: 13 mai. 2015.

Programa das Nações Unidas para o Desenvolvimento. (2013) *O Índice de Desenvolvimento Humano Municipal Brasileiro: Série Atlas do Desenvolvimento Humano no Brasil 2013*. Disponível em: <www.pnud.org.br>. Acesso em: 25 abr. 2015.

Quinlan, J. R. (1983) *Learning Efficient Classification Procedures and Their Application to Chess End Games*. Machine Learning: An Artificial Intelligence Approach, p. 463-482. Springer Berlin Heidelberg.

Rasouli, M.; Nikraz, H. (2013) *Trip Distribution Modelling Using Neural Network*. Australasian Transport Research Forum 2013 Proceedings. Brisbane, Australia.

Resolução n. 166. (2004, 15 de setembro). Aprova as diretrizes da Política Nacional de Trânsito. CONSELHO NACIONAL DE TRÂNSITO – CONTRAN. Disponível em <www.denatran.gov.br/download/resolucoes/resolucao166_04.doc>. Acesso em: 20 de abril de 2014.

Riviere, C.; Lauret, P.; Ramsamy, J. F. M.; Page, Y. (2006) *A Bayesian Neural Network approach to estimating the Energy Equivalent Speed*. Accident Analysis & Prevention, v. 38, n. 2, p. 248-259.

Rokach, L.; Maimon, O. (2008) *Data Mining with Decision Trees: Theory and Applications*. Series in Machine Perception and Artificial Intelligence, v. 69. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, USA.

Sayed T.; Rodriguez, F. (1999) *Accident Prediction Models for Urban Unsignalized Intersections in British Columbia*. Transportation Research Record, n. 1665, p. 93-99.

Silva, K. C. R. (2012) *Aplicação do modelo de previsão de acidentes do HSM em rodovias de pista simples do estado de São Paulo*. Dissertação de Mestrado, Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, SP.

Washington, S.; Schalkwyk, I. V.; Mitra, S.; Meyer, M.; Dumbaugh, E.; Zoll, M. (2006). *Incorporating Safety into Long-Range Transportation Planning*. National Cooperative Highway Research Program Report, n. 546. Transportation Research Board, Washington, DC, USA.

Washington, S. P.; Karlaftis, M. G.; Mannering, F. L. (2003) *Statistical and Econometric Methods for Transportation Data Analysis* (1ª ed.). Chapman and Hall/CRC, Boca Raton, FL, USA.

World Health Organization. (2013) *Global status report on road safety 2013: supporting a decade of action*. Disponível em: <www.who.int>. Acesso em: 21 ago. 2014.

Xie, C.; Lu, J.; Parkany, E. (2003) *Work Travel Mode Choice Modeling with Data Mining: Decision Trees and Neural Networks*. Transportation Research Record: Journal of the Transportation Research Board, n. 1854, p. 50-61. Transportation Research Board, National Research Council, Washington, DC, USA.

Zhang, C.; Yan, X.; Ma, L.; An, M. (2014) *Crash Prediction and Risk Evaluation Based on Traffic Analysis Zones*. Mathematical Problems in Engineering, v. 2014.