UNIVERSIDADE DE SÃO PAULO

ESCOLA DE ENGENHARIA DE SÃO CARLOS

Comparação de algoritmos de Aprendizagem de Máquinas para análise desagregada de viagens intermunicipais

ANDREZA DORNELAS DE SOUZA ROMA

ORIENTADORA: PROF[a].ASSOCIADA CIRA SOUZA PITOMBO

SÃO CARLOS

2017

ANDREZA DORNELAS DE SOUZA ROMA

Comparação de algoritmos de Aprendizagem de Máquinas para análise desagregada de viagens intermunicipais

Dissertação apresentada ao Departamento de Engenharia de Transportes (STT/EESC/USP), para obtenção do título de Mestre em Engenharia de Transportes.

Área de concentração: Planejamento e Operação de Sistemas de Transporte.

Orientadora: Prof[a]. Associada Cira Souza Pitombo

SÃO CARLOS

2017

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.

R729c

Roma, Andreza Dornelas de Souza
Comparação de algoritmos de Aprendizagem de Máquinas para análise desagregada de viagens intermunicipais / Andreza Dornelas de Souza Roma; orientadora Cira Souza Pitombo. São Carlos, 2017.

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, 2017.

1. Distribuição de viagens. 2. Modelo Gravitacional. 3. Análise desagregada. 4. Aprendizagem de máquinas. 5. Árvores de decisão. 6. CART. 7. Algoritmos genéticos. I. Título.

FOLHA DE JULGAMENTO

Candidata: Bacharela **ANDREZA DORNELAS DE SOUZA.**

Título da dissertação: "Comparação de algoritmos de aprendizagem de máquinas para análise desagregada de viagens intermunicipais".

Data da defesa: 30.06.2017

Comissão Julgadora:	Resultado:
Profa. Associada **Cira Souza Pitombo (Orientadora)** (Escola de Engenharia de São Carlos/EESC)	Aprovada
Profa. Dra. **Suely da Penha Sanches** (Universidade Federal de São Carlos/UFSCar)	APROVADA
Prof. Dr. **Cassiano Augusto Isler** (Universidade Federal de Santa Catarina/UFSC)	Aprovada

Coordenadora do Programa de Pós-Graduação em Engenharia de Transportes:
Profª. Associada **Ana Paula Camargo Larocca**

Presidente da Comissão de Pós-Graduação:
Prof. Associado **Luis Fernando Costa Alberto**

AGRADECIMENTOS

Agradeço aos meus pais, Luiza e José Renato, pelo incentivo, carinho e esforço durante todo o período do Mestrado.

Aos meus irmãos, Michele e José Renato, e sobrinho Matheus, pela preocupação e ternura.

Ao meu marido, Rodolfo, pelo amor, companheirismo e paciência.

A Izabela, meu maior presente.

À minha professora e orientadora Cira Souza Pitombo pelos ensinamentos, oportunidade e disponibilidade durante a realização deste projeto.

Ao Henrique Guimarães pela contribuição neste estudo através do desenvolvimento de seu Trabalho de Conclusão de Curso.

Aos colegas, amigas e professores do Departamento de Engenharia de Transportes da Escola de Engenharia de São Carlos por terem me proporcionado um ambiente de aprendizagem único ao longo deste período.

Aos funcionários do Departamento de Engenharia de Transportes, pela simpatia e gentileza.

À CAPES pelo apoio financeiro.

Ao CETRAMA (Centro de Estudos de Transportes e Meio Ambiente) pela cessão dos dados utilizados neste trabalho.

Enfim, a todos àqueles que de algum modo contribuíram para a realização deste projeto.

RESUMO

ROMA, A. D. S. **Comparação de algoritmos de Aprendizagem de Máquinas para análise desagregada de viagens intermunicipais**. 2017. 78p. Dissertação (Mestrado). Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2017.

Usualmente, os modelos de distribuição de viagens tradicionais ignoram que as escolhas de destinos são realizadas individualmente e consideram geralmente apenas variáveis agregadas, tais como o emprego e o custo médio de viagem. Este estudo considera as características agregadas das cidades de origem e destino, uma medida de impedância de viagem, além das variáveis desagregadas individuais/ domiciliares.

O objetivo desta pesquisa é avaliar o desempenho de algoritmos de Aprendizagem de Máquinas (AM), CART (*Classification and Regression Tree*) e Algoritmo Genético (AG), em uma análise desagregada de escolhas de destinos para viagens intermunicipais.

A análise baseia-se em um banco de dados de uma Pesquisa Origem-Destino realizada em onze cidades na Bahia, Brasil, em 2012, no qual foram calibrados modelos para escolhas de destinos intermunicipais por meio dos algoritmos mencionados.

O desempenho de cada algoritmo é comparado à abordagem tradicional por meio de modelos gravitacionais estimados pelos métodos *stepwise* e procedimento duplamente restringido, respectivamente. A comparação dos modelos com os modelos gravitacionais foi realizada considerando os seguintes critérios: distribuição das distâncias de viagens, medidas de ajuste e perspectiva qualitativa.

Os resultados evidenciam que os algoritmos CART e Algoritmo Genético apresentam melhores previsões para a escolha de destinos, sendo que, o segundo, apresenta vantagens na obtenção dos parâmetros estimados relativos às variáveis (agregadas e desagregadas) utilizadas no modelo. A principal conclusão é que tais algoritmos podem ser aplicados na modelagem de distribuição de viagens, incorporando o efeito das variáveis desagregadas, sem suposições ou restrições matemáticas rigorosas como independência das alternativas, por exemplo.

Palavras-chave: Distribuição de viagens, Modelo Gravitacional, Análise Desagregada, Aprendizagem de Máquinas, Árvores de Decisão, CART, Algoritmos Genéticos.

ABSTRACT

ROMA, A. D. S. **Comparison of Machine Learning algorithms for disaggregated analysis of intercity trips**. 2017. 78p. Thesis (Master of Science). São Carlos School of Engineering, University of São Paulo, São Carlos, 2017.

Usually, traditional trip distribution models ignore that destination choices are made individually and generally only consider aggregate variables such as employment and average travel cost. This study considers the aggregate characteristics of origin and destination cities, a measure of trip impedance, in addition to individual / household disaggregated variables.

The objective of this research is to evaluate the performance of Machine Learning (ML) algorithms, CART (*Classification and Regression Tree*) and Genetic Algorithm (GA), in a disaggregated analysis of destination choices for intercity trips.

The analysis is based on a database of an Origin-Destination Survey carried out in eleven cities in Bahia, Brazil, in 2012, in which models were calibrated for choices of intermunicipal destinations through the mentioned algorithms.

The performance of each algorithm is compared to the traditional approach using gravitational models estimated by the stepwise and double constrained procedure, respectively. The comparison of the models with the gravitational models was performed considering the following criteria: distribution of travel distances, adjustment measures and qualitative perspective.

The results show that the ML algorithms CART and GA present better predictions for the destinations choices, and the second one presented advantages in obtaining the estimated parameters related to the variables (aggregated and disaggregated) used in the model. The main conclusion is that such algorithms can be applied in trip distribution modeling, incorporating the effect of the disaggregated variables, without rigorous assumptions such as independence of the alternatives, for example.

Keywords: Trip Distribution, Gravitational Model, Disaggregated Analysis, Machine Learning, Decision Trees, CART, Genetics Algorithms.

SUMÁRIO

1 INTRODUÇÃO 10
1.1 APRESENTAÇÃO DO PROBLEMA E JUSTIFICATIVA DO TRABALHO . 10
1.2 APRENDIZAGEM DE MÁQUINAS 13
1.3 OBJETIVOS 15
1.3.1 Objetivos específicos 16
1.4 ESTRUTURA DO TRABALHO 16

2 EMBASAMENTO TEÓRICO 17
2.1 REVISÃO DA LITERATURA 17
2.2 TÉCNICAS UTILIZADAS 21
2.2.1 ABORDAGEM TRADICIONAL: MODELO GRAVITACIONAL 21
2.2.2 ÁRVORE DE DECISÃO 22
2.2.3 ALGORITMOS GENÉTICOS 26

3 MATERIAIS E MÉTODO 29
3.1 MATERIAIS: BANCO DE DADOS E APLICATIVOS 30
3.1.1 Amostra 31
3.1.2 Aplicativos utilizados 36
3.2 MODELO DE OTIMIZAÇÃO 36
3.3 MÉTODO 38

4 RESULTADOS E DISCUSSÕES 40
4.1 DESCRIÇÃO GERAL DOS RESULTADOS 40
4.2 MODELO 1: MODELO GRAVITACIONAL 1 41
4.3 MODELO 2: MODELO GRAVITACIONAL 2 42
4.4 MODELO 3: ALGORITMO CART 43
4.5 MODELO 4: ALGORITMO GENÉTICO 46
4.5.1 Procedimento de escolha da melhor replicação 48
4.6 COMPARAÇÕES DAS ABORDAGENS 53
4.6.1 Distribuição das distâncias de viagens 54
4.6.2 Medidas de qualidade do ajuste 58
4.6.3 Perspectiva qualitativa 59

5 CONSIDERAÇÕES FINAIS E SUGESTÕES PARA TRABALHOS FUTUROS 64
5.1 SÍNTESE DO TRABALHO 64
5.2 CONCLUSÕES 66
5.3 SUGESTÕES PARA TRABALHOS FUTUROS 67

6 REFERÊNCIAS BIBLIOGRÁFICAS 68

7 APÊNDICE I 74

8 APÊNDICE II 76

1 INTRODUÇÃO

Este capítulo inicial apresenta os aspectos introdutórios da presente pesquisa de Mestrado, caracterizando o problema em questão e as motivações para a realização do trabalho. Também fazem parte do conteúdo deste capítulo o objetivo geral do estudo e os objetivos específicos, bem como, o referencial teórico preliminar.

1.1 APRESENTAÇÃO DO PROBLEMA E JUSTIFICATIVA DO TRABALHO

A gestão de transportes e de circulação entre os municípios são práticas obrigatórias de todo Poder Público, tendo em conta seu comprometimento em propiciar à população um sistema de transportes intermunicipal de qualidade. Para garantir uma maior qualidade e eficiência do sistema de transportes intermunicipais é essencial a realização de pesquisas e estudos para conhecimento da demanda atual e futura de usuários.

Apesar da sua ampla relevância, o sistema de transporte é um dos principais problemas urbanos de um município. O longo tempo gasto nas viagens de curtas distâncias e a frequência de saída de veículos mal dimensionada torna inconveniente a utilização do transporte público intermunicipal por grande parte dos usuários.

Em vista disso, o estudo da demanda por viagens intermunicipais é fundamental para a adequação da oferta de transportes entre municípios, fornecendo os principais indicadores para dimensionamento da oferta aos gestores públicos. Tais indicadores baseiam-se em estimativas mais precisas de demanda por viagens intermunicipais.

Existem diversos modelos que auxiliam na estimativa da demanda por viagens, sendo tradicionalmente o mais utilizado o modelo quatro-etapas: (1) Geração de Viagens, (2) Distribuição de Viagens, (3) Escolha Modal e (4) Alocação de Tráfego.

A análise deste trabalho baseia-se na segunda etapa do modelo sequencial tradicional, a Distribuição de Viagens. Os modelos de distribuição de viagens têm o

objetivo de prever as escolhas de destinos dadas as viagens produzidas e atraídas em cada Zona de Tráfego (ZT), as características sociodemográficas agregadas, além de variáveis de impedância de viagem entre cada par de origem e destino (De Grange et al., 2010; Wilson, 1967).

Os primeiros modelos de distribuição de viagens propostos foram baseados em fatores de crescimento (Evans, 1970; Williams, 1976). No entanto, estes modelos são fortemente dependentes da precisão da matriz base anual (Ortúzar e Willumsen, 2011).

Posteriormente, o modelo gravitacional foi amplamente aceito e permanece como um dos modelos de distribuição de viagens mais populares (Cascetta et al., 2007). No entanto, o modelo gravitacional está, geralmente, limitado à modelagem agregada, limitado às dificuldades da inclusão do comportamento individual na análise (Mishra et al., 2013).

As interações entre origens e destinos são complexas e esses modelos comumente contemplam variáveis agregadas. Tradicionalmente, a distribuição de viagens é considerada um problema agregado e os modelos relacionados ignoram o fato de que as escolhas de destinos são realizadas individualmente. Assim, considerar as características dos indivíduos e dos domicílios, além dos fatores agregados, pode tornar as previsões mais eficientes (Ortúzar e Willumsen, 2011).

A análise desagregada para distribuição de viagens foi introduzida a partir do desenvolvimento dos modelos de escolha discreta no início da década de 1980 (Fotheringham, 1983; Ben - Akiva e Lerman, 1985).

Tais modelos ajustam o banco de dados a formulações matemáticas e partem da suposição de estimativa de parâmetros que compõem as funções utilidades das alternativas. No entanto, eles implicam limitações relacionadas a um atributo reconhecido como IIA (Independência das Alternativas).

O atributo IIA envolve a restrição de que os termos de erro aleatório são independentes (sem correlação) e igualmente distribuídos (variância constante) (Koppelman e Wen, 2000).

Tal restrição não faz parte dos algoritmos de Aprendizagem de Máquinas AM), os quais são técnicas semi-paramétricas ou não paramétricas que identificam padrões e classificam indivíduos, dado um conjunto de dados. Este conjunto de algoritmos pode ser útil na análise da demanda por viagens, pois resultam em estimativas convenientes por meio de um modelo com várias funções. Além disso, eles não têm limitações importantes, tais como um conjunto de dados multicolinear, suposições de distribuições populacionais ou IIA. A Figura 1.1 esquematiza uma justificativa para o uso de algoritmos de Aprendizagem de Máquinas (AM) na etapa de distribuição de viagens.

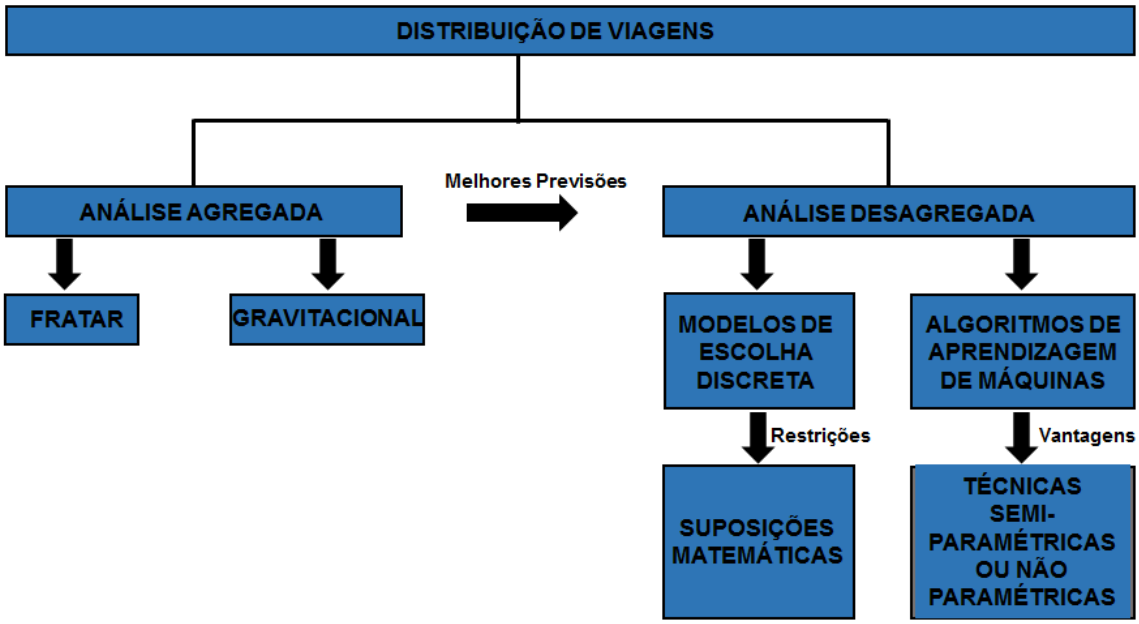

Figura 1.1: Aplicação de algoritmos de Aprendizagem de Máquinas em problemas de escolha de destinos intermunicipais (Adaptado de Pitombo et al., 2017)

Neste contexto, a modelagem do comportamento individual relativo às viagens pode ser descrita formalmente como uma tarefa de reconhecimento de padrões em que vários atributos comportamentais humanos, representados por variáveis explicativas, permitem prever uma escolha entre um conjunto de alternativas (Xie et al., 2003).

O presente estudo baseia-se na pesquisa Origem Destino realizada nos onze municípios do estado da Bahia em 2012 e propõe a aplicação de algoritmos de AM para classificação dos dados segundo a variável dependente (escolha do destino intermunicipal), considerando características desagregadas dos indivíduos, além de variáveis agregadas tradicionais.

Esta pesquisa de Mestrado possui uma contribuição de carácter metodológico a partir dos testes de algoritmos eficazes e não triviais, neste campo de pesquisa, para a previsão de demanda de usuários de viagens intermunicipais.

1.2 APRENDIZAGEM DE MÁQUINAS

O termo Aprendizagem Automáticas (em inglês, *Machine Learning*) atribui-se à implantação de sistemas computacionais capazes de assimilar conhecimento automaticamente, aprimorando seu funcionamento em questões de classificação por meio de testes já realizados (Mitchell, 1997). Aprendizagem de Máquinas é o campo que se dedica ao uso de ferramentas computacionais, aptas a assimilar e modificar conhecimentos e habilidades, aprimorando seus desempenhos baseando-se em experiências bem sucedidas anteriormente (Mitchell, 1997; Alpaydin, 2010).

É possível realizar a segregação do sistema de Aprendizagem de Máquinas em três grupos distintos, aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço, apresentados na Figura 2. A aprendizagem supervisionada se dá por meio do fornecimento de um conjunto de exemplos com a saída conhecida para cada conjunto de dados de entrada apresentado, de modo que o sistema de aprendizado aprenda uma regra geral por meio do mapeamento dos dados de entrada com os de saída (Russel et al., 2002). No aprendizado não supervisionado, o conjunto de dados de saída não é fornecido, permitindo-o que realize sozinho a obtenção do número de classes e reconhecimentos de novos padrões como, por exemplo, na clusterização (Russel et al., 2002). E, por fim, no aprendizado por reforço existe apenas uma medida indicando se uma ação aleatória, realizada pelo sistema, foi satisfatória ou não, mas não informa com clareza qual foi a melhor.

No geral, o campo de Aprendizagem de Máquinas envolve diversos algoritmos inseridos nos três grupos apresentados, tais como Árvore de Decisão (Quinlan, 1993), Redes Neurais Artificiais (Haykin, 2001), Máquinas de Vetores de Suporte (Cortes et al., 1995), *Naive Bayes* (Charniak, 1991) e Aprendizagem Baseada em Instâncias (Aha et al., 1991).

É importante salientar que a escolha do algoritmo de Aprendizagem para ser utilizado em uma análise depende das características da pesquisa, da agilidade com que o algoritmo utilizado realiza a convergência dos resultados (classificação dos dados), além dos objetivos do pesquisador. Por exemplo, os algoritmos de Redes Neurais Artificiais, Árvore de Decisão e Algoritmo Genético são utilizados na identificação de padrões e classificação de indivíduos em diversas áreas distintas. Feng et al. (2011) desenvolveram um método de concentração de ozônio baseado em Algoritmo Genético e Redes Neurais Artificiais. Junguo et al. (2013) utilizaram a técnica de Redes Neurais Artificiais para analisar a distribuição espacial da iluminação solar. Cho et al. (2011) aplicaram Árvore de Decisão para classificar e reduzir a dimensionalidade dos dados de ruídos eletrônicos. E, por fim, Mohanty et al. (2013) também empregaram a técnica de Árvore de Decisão para a classificação de mamografias.

As análises desagregadas, delineadas neste trabalho, pertencem ao primeiro grupo mencionado, Aprendizagem Supervisionada, e englobam dois tipos de algoritmos, os modelos de classificação Árvore de Decisão e Algoritmo Genético.

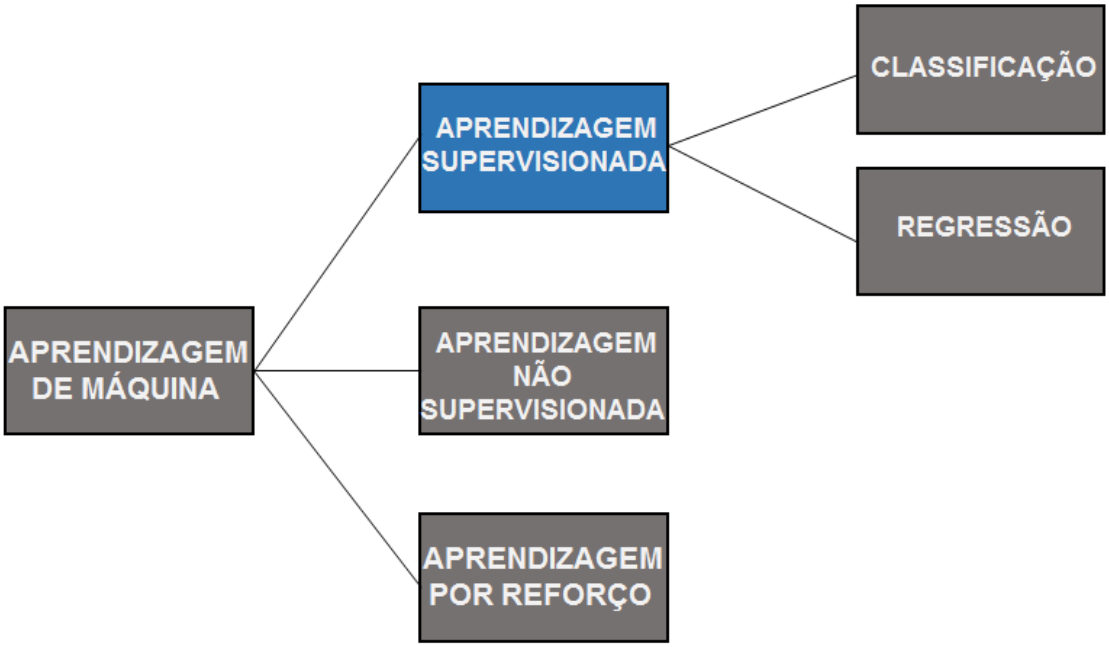

Figura 1.2: Tipos de Aprendizagem de Máquinas (Adaptado de Lacerda, 2013)

A análise desagregada, proposta neste estudo, se deu por meio da introdução do conceito de Aprendizagem de Máquinas, tecnologia do campo da Inteligência Artificial, a qual visa a solução de problemas por meio dos modelos de classificação Árvore de Decisão e Algoritmos Genéticos, presentes na aprendizagem supervisionada.

O modelo de Árvore de Decisão foi escolhido pela facilidade de se trabalhar com um conjunto de dados denso, identificando padrões e classificando indivíduos. Porém, por ser um modelo de classificação que não gera parâmetros, foi necessária a aplicação de outro algoritmo de classificação, Algoritmos Genéticos, o qual também fosse capaz de utilizar bancos de dados extensos e ao mesmo tempo fornecer funções (parâmetros) que podem melhor descrever as influências das variáveis explicativas na escolha do destino pelos indivíduos da amostra.

1.3 OBJETIVOS

O objetivo principal desta pesquisa é avaliar a aplicação de algoritmos de Aprendizagem de Máquinas (CART – *Classification And Regression Tree* e

Algoritmos Genéticos) em uma análise desagregada de escolhas de destinos para viagens intermunicipais.

1.3.1 Objetivos específicos

- Identificar as variáveis independentes importantes para escolha de destinos intermunicipais (agregadas e desagregadas);
- Verificar melhorias na estimativa de distribuição de viagens a partir de uma amostra desagregada e das abordagens propostas neste estudo, comparando-as a modelos tradicionais (Gravitacionais).

1.4 ESTRUTURA DO TRABALHO

Este texto apresenta quatro capítulos, além desta introdução. Os capítulos são: (2) Embasamento Teórico, (3) Materiais e Método, (4) Resultados e Discussões e (5) Considerações finais e sugestões para trabalhos futuros.

Capítulo 2: São definidas todas as técnicas que serão aplicadas no banco de dados para a obtenção dos resultados, bem como, os algoritmos de Aprendizagem de Máquinas empregados. Além disso, apresenta a revisão da literatura e a contribuição e lacuna do presente trabalho.

Capítulo 3: São apresentados o banco de dados da Pesquisa OD, os aplicativos utilizados, além do procedimento metodológico proposto.

Capítulo 4: São expostos os resultados obtidos aplicando algoritmos de Aprendizagem de Máquinas e a abordagem tradicional (Modelos Gravitacionais), além da comparação entre tais abordagens.

Capítulo 5: Este capítulo finaliza a dissertação, apresentando as considerações finais e sugestões para futuros trabalhos.

2 EMBASAMENTO TEÓRICO

Neste capítulo será realizada a revisão bibliográfica e apresentados as particularidades e conceitos relacionados aos modelos abordados neste projeto de Mestrado: Gravitacional, Árvore de Decisão e Algoritmo Genético.

Será realizada também a exposição de quadros-resumo relativos aos algoritmos, sintetizando suas vantagens e desvantagens, descrições, etc.

2.1 REVISÃO DA LITERATURA

Nas últimas décadas, alguns estudos consideraram o reconhecimento de padrões aplicado à modelagem relativa às viagens, fazendo a aplicação de diversos algoritmos de Aprendizagem de Máquinas. Shmueli et al. (1996) exploraram a aplicação de Redes Neurais Artificiais para comparar os padrões de viagens de homens e mulheres em Israel. Strambi e Van de Bilt (1998) analisaram as taxas de geração de viagem por meio do algoritmo CHAID de Árvore de Decisão (AD). Xie et al. (2007) investigaram o desempenho de Árvore de Decisão e Redes Neurais Artificiais para a modelagem de escolha do modo de transportes para viagens com motivo trabalho.

Lu et al. (2012) analisaram a viabilidade do uso de uma ferramenta de Aprendizagem de Máquinas, utilizando o algoritmo C4.5 (Árvore de Decisão), na escolha do motivo da viagem, baseando-se em dados coletados por GPS instalado em veículos na Universidade de Minnesota. Ichikawa et al (2002), Pitombo et al. (2011) e Pitombo et al. (2013) analisaram as relações entre as características socioeconômicas, o uso do solo, a participação em atividades e os padrões de viagens aplicando o algoritmo CART, por meio de uma abordagem de viagens baseadas em atividades.

Pulugurta et al. (2013) utilizaram o conceito de Aprendizagem de Máquinas na análise de escolha do modo de viagem, realizando uma comparação com o Modelo Logit Multinomial. Griffin et al. (2008) empregaram o algoritmo C4.5 (Árvore de Decisão) para obter o motivo da viagem, por meio de 50 viagens geradas aleatoriamente e considerando uma série de suposições. Deng et al. (2010)

aplicaram uma série de atributos na construção da Árvore de Decisão para obter os modos e os motivos de viagens, utilizando 226 registros de viagens, de 36 entrevistados, coletados pelo GPS.

Já Omrani (2015) explorou o conceito de Aprendizagem de Máquinas, utilizando as técnicas Redes Neurais Artificiais, Regressão Logística Multinominal e máquinas de vetor de suporte (*support vector machine*) para prever a escolha do modo de viagem dos indivíduos na cidade de Luxemburgo. Além dos trabalhos mencionados anteriormente, existem alguns estudos recentes que consideram a aplicação e a adequabilidade de Redes Neurais Artificiais na modelagem de distribuição de viagens (Rasouli e Nikras, 2013; Mozolin et al., 2015).

Outra técnica de Aprendizagem de Máquinas, também encontrada em trabalhos recentes da área de transportes, é o Algoritmo Genético. Charypar e Nagel (2005) aplicaram a técnica de Algoritmo Genético na fase de Geração de viagens, por meio da determinação de uma função utilidade, necessária para avaliar o desempenho de cada viagem. Cunha et al. (2009) exploraram a técnica de Algoritmo Genético para calibrar dois modelos de simulação de tráfego, baseados em dados de desempenho de caminhões coletados em rodovias brasileiras. Além disso, encontram-se diversos trabalhos na literatura em que se aplicam Algoritmo Genético em modelos de simulação de tráfego. Rilett et al. (2006) propuseram um sistema de calibração automática de modelos de simulação de tráfego, baseado em Algoritmos Genéticos. Setti e Cunha (2006) calibraram um modelo de locomoção de caminhões do *software* de simulação de tráfego CORSIM, por meio de um Algoritmo Genético. Setti e Bessa Júnior (2011) criaram um método que, por meio de um simulador, produz dados de tráfego sintéticos, fundamentado em um Algoritmo Genético que calibra um modelo de simulação microscópica. No entanto, o seu uso na área de demanda por transportes é raro.

É importante salientar que o emprego dos algoritmos de Aprendizagem de Máquinas ainda é raro na estimativa de escolhas de destinos intermunicipais em um banco de dados desagregados (Yang et al., 2014; LaMondia et al., 2009; Pitombo et al., 2017).

Especificamente para o caso de distribuição de viagens intermunicipais, Pitombo e Guimarães (2016) testaram a adequabilidade do uso das técnicas de Aprendizagem de Máquinas para modelagem de distribuição de viagens intermunicipais, por meio da comparação de modelos obtidos pela técnica de Árvore de Decisão (CART e CHAID) e pela abordagem tradicional de distribuição de viagens.

Assim, a principal lacuna, e consequente contribuição deste trabalho, está focada na apresentação e teste de algoritmos de Aprendizagem de Máquinas para o caso específico de distribuição de viagens intermunicipais. Tal pesquisa traz a comparação de um algoritmo não paramétrico (CART – *Classification And Regression Tree*) e um algoritmo semi-paramétrico (Algoritmo Genético) no intuito de compreender os fatores que influenciam as escolhas de destinos intermunicipais. O Quadro 2.1 apresenta uma sucinta descrição dos algoritmos de Aprendizagem de Máquinas utilizados neste trabalho, e sua tarefa ao serem aplicados em um banco de dados.

Quadro 2.1: Descrição das técnicas utilizadas (Adaptado de Goldschimidt, 2005)

Técnica	Descrição	Tarefas
Árvore de Decisão	Baseada em estágios de decisão (nós) e na identificação de classes organiza os dados de forma hierárquica	Classificação e predição
Algoritmo Genético	Método de busca e otimização, inspirados na Teoria da Evolução, em que a cada nova geração, soluções melhores têm mais chances de gerar “descendentes”.	Classificação e agrupamento

O Quadro 2.2, exposto a seguir, apresenta um resumo com as principais vantagens e desvantagens na utilização dos algoritmos CART e Algoritmo Genético, além do tradicional Modelo Gravitacional.

Quadro 2.2: Quadro-resumo das vantagens e desvantagens encontradas na utilização dos algoritmos (Ortúzar e Willumsen, 2011; Xie et al., 2003; Rodrigues, 2005)

MODELOS	VANTAGENS	DESVANTAGENS
Gravitacional	Técnica paramétrica	Utiliza somente variáveis agregadas
Árvore de Decisão	Capaz de identificar padrões de indivíduos e realocá-los em categorias; Utiliza variáveis agregadas e desagregadas; Não apresenta suposições matemáticas rigorosas e restrições para uso das variáveis de entrada	Técnica não paramétrica
Algoritmo Genético	Técnica semi-paramétrica; Não apresenta suposições matemáticas rigorosas e nem restrição de uso de determinadas variáveis de entrada; Utiliza variáveis agregadas e desagregadas	

2.2 TÉCNICAS UTILIZADAS

2.2.1 ABORDAGEM TRADICIONAL: MODELO GRAVITACIONAL

A abordagem tradicional selecionada para ser comparada aos resultados dos algoritmos de Aprendizagem de Máquinas foi o Modelo Gravitacional. Tal modelo é o mais utilizado na análise agregada de distribuição de viagens. Nesta pesquisa, os métodos utilizados na calibração foram: Método dos Mínimos Quadrados – Modelo Gravitacional 1 e Modelo Duplamente Restringido – Modelo Gravitacional 2.

O Modelo Gravitacional de distribuição viagens foi inicialmente proposto na década de 1950 por Casey (1955), sugerindo uma abordagem para sintetizar viagens por motivo de compras entre as cidades de uma região. Alguns pesquisadores forneceram melhorias notáveis (Murchland, 1966; Evans e Kirby, 1974; Mikkonen e Luoma, 1999) de tal forma a generalizar a função de impedância, acrescentando variáveis como tempo de viagem ou custo generalizado. A Equação 2.1 representa a formulação geral do Modelo Gravitacional.

$$V_{ij} = A_i . O_i B_j . D_j . f\left(c_{ij}\right) \tag{2.1}$$

Em que V_{ij} é o número de viagens estimadas entre a origem *i* e o destino *j*, A_i é a característica agregada da origem *i*, O_i é o total de viagens produzidas na origem *i*, B_j é a característica agregada do destino *j*, D_j é o total de viagens atraídas no destino *j*, c_{ij} é o custo da viagem entre *i* e *j*, e $f(c_{ij})$ é a função generalizada da impedância de viagem, por exemplo, $e^{(-\beta c_{ij})}$ *ou* $c_{ij}^{(-\beta)}$.

No caso do Método Duplamente Restringido, os valores dos fatores de ponderação são representados na Equação 2.2 e Equação 2.3.

$$A_i = \frac{1}{\sum_j B_j D_j f\left(c_{ij}\right)} \tag{2.2}$$

$$B_j = \frac{1}{\sum_i A_i O_i f(c_{ij})} \tag{2.3}$$

Assim, o processo é iterativo. Dado um conjunto de valores para a função $f(c_{ij})$, ele se inicia com todos $B_j=1$. Em seguida, ele deve ser resolvido para A_i e esses valores devem ser usados para reestimar $B_{j's}$. O procedimento é repetido até que ele atinja a convergência (Ortúzar e Willumsen, 2011).

Os parâmetros da função $f(c_{ij})$ são estimados considerando que o modelo é capaz de reproduzir a distribuição dos custos de viagens das observações. Assim, neste trabalho, a distribuição das distâncias de viagens foi utilizada como um dos critérios para medir a precisão ou a eficiência de todos os modelos utilizados.

2.2.2 ÁRVORE DE DECISÃO

Árvore de Decisão representa um conjunto de técnicas não-paramétricas que têm objetivo de prever ou classificar dados. Tais técnicas geram uma estrutura representativa de uma sequência de decisões por meio das quais ocorrerão sucessivas divisões em um conjunto de dados inicial até que o mesmo seja representado por diversas classes (Breiman et al., 1984).

Este procedimento compreende diversos algoritmos de particionamento sequencial que divide, sucessivamente, os dados para criar subconjuntos homogêneos, resultando em uma estrutura hierárquica de regras induzidas (Breiman et al., 1984). Os principais algoritmos de Árvores de Decisão são C4.5 (Quinlan, 1983), CHAID (Kass, 1980) e CART (Breiman et al., 1984) e as suas propriedades são descritas a seguir:

- A estrutura hierárquica resultante é chamada de árvore e cada segmento é denominado nó;
- O nó raiz contém o conjunto de dados completo;
- O nó raiz é dividido sequencialmente criando nós-filhos;
- Quando nenhuma outra subdivisão dos dados é possível, os subconjuntos finais são denominados nós terminais ou folhas;

- Três elementos principais devem ser definidos para executar a Árvore de Decisão: um conjunto de questões que delimita a divisão dos dados, um critério para estabelecer a melhor divisão na produção de nós filhos, e uma regra de parada para as subdivisões (*stop-splitting rule*).

Este estudo aplicou o algoritmo CART para estimar a distribuição de viagens intermunicipais. Tal algoritmo constrói árvores, onde cada nó, não terminal, identifica uma condição de divisão que produza a melhor previsão (de variáveis dependentes contínuas) ou a classificação (para variáveis dependentes categóricas). Assim, este algoritmo pode ser utilizado para analisar problemas de regressão ou de classificação. Esta pesquisa apresenta o algoritmo CART para problemas de classificação, conforme descrito nas seguintes subseções. Posteriormente, foi verificada a adequabilidade do uso de Algoritmos Genéticos (AG) para a mesma finalidade. Conceitos básicos de AG são descritos na seção 2.3.

2.2.2.1 CART

O algoritmo CART (*Classification And Regression Tree*) foi desenvolvido por Breiman et al. (1984) e fundamenta-se na realização de uma sequência de divisões binárias do conjunto de dados inicial até atingir a máxima homogeneidade dentro dos nós terminais. A partição do banco de dados é realizada de forma a minimizar a impureza dos nós filhos. A Figura 2.1, a seguir, esquematiza a divisão binária realizada pelo algoritmo CART, em que t_{Raiz} é denominado nó raiz, x_j é uma variável explicativa j, x_j^R é o melhor valor de divisão da variável x_j, P é a probabilidade de cada nó, e t_{Esq} e t_{Dir} são os dois nós filhos gerados na primeira divisão binária.

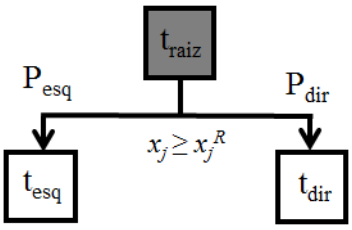

Figura 2.1: Algoritmo de divisão CART (Adaptado de Breiman et al., 1984)

O processo de crescimento da árvore, realizado pelo algoritmo CART, envolve as seguintes etapas descritas, em seguida, segundo Breiman et al. (1984): (1) Perguntas do tipo binárias; (2) Critério de Divisão e Medidas de impurezas; (3) Regras de parada (4) Divisões substitutas e (5) Importância das variáveis explicativas.

2.2.2.1.1 Perguntas do tipo binárias

O algoritmo CART realiza as divisões do conjunto de dados aplicando-se perguntas do tipo “sim” e “não”. Dessa forma, os dados relacionados às respostas “sim” seguem para o nó esquerdo e os referentes ao “não”, para o nó direito. Em cada nó filho gerado forma-se uma nova ligação portando uma parcela dos dados, surgindo um novo teste binário, do qual outro nível da árvore será construído.

Este algoritmo avalia todas as possíveis divisões considerando todas as variáveis explicativas e seleciona aquela que maximiza a homogeneidade do nó filho, a partir do nó raiz, bem como o seu valor de corte. O processo é repetido para os nós subsequentes. A melhor divisão é selecionada pela menor impureza ou maior homogeneidade (Berk, 2008).

No presente algoritmo, a construção da árvore realiza em cada nó um teste baseado em apenas um atributo (variável). Tais divisões são denominadas univariadas. Assim, uma árvore cresce a partir do nó raiz, considerando uma única variável e seu respectivo valor de corte que produzem menor medida de impureza no nó subsequente.

2.2.2.1.2 Critério de divisão e medida de impureza

O melhor critério de divisão do conjunto de dados é verificado quando um determinado atributo (variável) realizar a melhor partição dos dados. Este atributo é selecionado de modo que ocorra um decréscimo da medida de impureza de um determinado nó e, consequentemente, a sua máxima homogeneidade.

Inicialmente, o nó raiz apresenta grau de impureza máximo por ser o nó no qual está contido o conjunto de dados completo. As categorias dentro deste nó serão definidas de acordo com a variável dependente estabelecida para o problema. Dessa forma, considerando que a variável dependente apresente *n* categorias *(1,2,3,i,...,n)*, a probabilidade da categoria *i* aparecer no nó raiz, por exemplo, definido como nó inicial 0, será $p(i/0)$. Cabe ressaltar que a soma das probabilidades de todas as categorias é equivalente a 1.

Após sucessivas divisões irá ocorrer a diminuição da medida de impureza dentro dos nós filhos gerados. A máxima homogeneidade ($Impureza = 0$) em um nó t será alcançada quando um nó contiver uma única categoria com 100% do conjunto de dados ($p(i/t) = 1$) ou quando um critério de parada for atingido.

O cálculo mais comum da heterogeneidade dos dados do algoritmo CART é realizado pelo Índice Gini, dado pela Equação 2.4.

$$G(t) = 1 - \sum_{i=1}^{n} p^2(i/t) \tag{2.4}$$

A diferença entre o Índice Gini para o nó pai e a soma dos valores para os nós filhos, ponderados pela proporção de casos em cada filho, é apresentada na árvore como *aprimoramento*. A escolha da melhor variável explicativa e melhor valor de corte se dá pela combinação (de variável e valor de corte) que produz maior valor de *aprimoramento*.

2.2.2.1.3 Regras de parada

Como foi mencionado anteriormente, o processo de construção da árvore baseia-se em recursivas repartições de um banco de dados até que se atinja máxima homogeneidade no nó filho ou até o momento em que alguma regra de parada (interrupção) do crescimento da árvore seja alcançada. Algumas regras de parada são apresentadas a seguir:

- Se a profundidade da árvore atual atinge o valor limite definido pelo usuário;
- Se o número de observações em um nó filho é menor do que o valor mínimo definido previamente pelo usuário;
- Se a homogeneidade de um nó é menor do que o Índice Gini adotado.

2.2.2.1.4 Divisões substitutas

As divisões substitutas *são* um artifício utilizado pelo algoritmo CART para tratamento de valores desconhecidos dentro de um determinado nó. Este artifício se dá por meio do armazenamento de um *ranking* de variáveis explicativas com comportamento semelhante à variável principal escolhida em determinada divisão dos dados, isto é, aquela que minimiza a medida de impureza dentro do nó. Dessa forma, é possível considerar outra variável explicativa de forma que realize divisões semelhantes, mensuradas através de uma medida positiva de associação.

2.2.2.1.5 Importância das variáveis explicativas

A Medida de Importância de uma variável explicativa em relação à árvore final é definida como a soma (ponderada) das participações de determinada variável, no papel de primária e/ou substituta. A Importância de uma variável é expressa em termos de uma quantidade normalizada com relação à variável que tem a maior medida de importância. Ela varia de 0 a 100, com a variável tendo a maior medida de importância pontuada como 100.

2.2.3 ALGORITMOS GENÉTICOS

Os Algoritmos Genéticos (AGs) foram introduzidos por John Holland (1975) e popularizados por David Goldberg (1989). Esta técnica consiste em métodos de

otimização e busca inspirados em mecanismos de evolução de populações de seres vivos (De Carvalho et al., 1999). O Algoritmo Genético é uma aproximação computacional da maneira como a evolução realiza a busca pelo melhor indivíduo, alterando o gene e, consequentemente, a forma dos indivíduos, a função *fitness* (função aptidão) (Marsland, 2009). Segundo Kononenko e Kukar (2007), Algoritmos Genéticos baseiam-se na evolução e seleção natural, em que cada hipótese corresponde a uma resposta, codificando com uma cadeia de bits, chamada genes.

Tais algoritmos seguem o princípio da teoria evolutiva enunciado pelo fisiologista inglês Charles Darwin em 1859. Segundo Darwin, os organismos mais bem adaptados ao meio possuem maiores chances de sobreviver e gerar descendentes. A questão abordada pela técnica dos Algoritmos Genéticos é localizar candidatos (soluções) em um espaço de busca, identificando a melhor hipótese como aquela que otimiza uma medida numérica predefinida para o problema em questão (Mitchell, 1997).

A otimização consiste em um processo de busca pela melhor solução com o propósito de alcançar os objetivos determinados. As técnicas de busca e otimização devem ser utilizadas quando não existe uma solução simples e diretamente calculável para o problema. Isso geralmente ocorre quando a solução do problema é complexa ou existem milhões de possíveis soluções. Tais técnicas, geralmente, apresentam um espaço de busca, onde estão todas as possíveis soluções para o problema analisado, e uma função objetivo, utilizada para avaliar todas as soluções geradas por meio da atribuição de uma nota para cada uma delas. Vale ressaltar que, na literatura de Algoritmos Genéticos (AGs), esta função objetivo também pode ser chamada de função de aptidão (De Carvalho et al., 1999).

No entanto, em termos matemáticos, o processo de otimização consiste em localizar pontos que maximizam o valor da função de aptidão. Ao final do processo, a melhor solução encontrada é aquela correspondente ao valor máximo (ou mínimo) obtido também denominado “máximo global” (ou “mínimo global).

2.2.3.1 Representação dos parâmetros

Uma possível solução do problema a ser otimizado é representada por um vetor ou sequência de bits (0 ou 1) composta por populações de cromossomos. Cada cromossomo representa um conjunto de parâmetros da função objetivo cuja resposta será maximizada ou minimizada. O espaço de busca é formado pelo conjunto de todas as configurações que o cromossomo pode assumir. Se o cromossomo representa *n* parâmetros de uma função, então o espaço de busca é um espaço com *n* dimensões (De Carvalho et al., 1999).

2.2.3.2 Seleção

O Algoritmo Genético realiza a seleção dos melhores cromossomos da cadeia considerando os maiores valores da função de aptidão da população inicial de cromossomos. Este procedimento é realizado por mecanismos de busca da própria técnica dos Algoritmos Genéticos denominados *crossover* (cruzamento) e mutação. Os cromossomos-pais, selecionados por esses operadores, são realocados formando uma população intermediária, que será composta apenas de indivíduos cuja probabilidade seja proporcional à sua aptidão. Logo, a probabilidade de seleção é dada pela Equação 2.5, a seguir:

$$p_i = \frac{f_i}{\sum_{i=1}^{N} f_i} \tag{2.5}$$

Onde, p_i é a probabilidade de seleção do cromossomo *i*, f_i é a função objetivo do cromossomo *i* e N é o número total de cromossomos.

2.2.3.3 *Crossover* (cruzamento) e mutação

Primeiramente, é aplicado o operador *crossover* em cada par de cromossomos da população inicial, de modo que o cromossomo-filho receba uma parte do primeiro cromossomo e parte do segundo (Marsland, 2009). Este operador aplica uma probabilidade em cada par de cromossomos selecionados que varia

entre 60% e 90%. Os cromossomos-filhos serão iguais aos pais quando esta taxa não for aplicada fazendo com que algumas soluções sejam mantidas.

O operador de mutação é sempre aplicado após o mecanismo de *crossover* por meio de uma probabilidade em cada bit (0 e 1) dos dois cromossomos-filho previamente gerados. Este operador cria alterações genéticas nos cromossomos-filho, que podem ser positivas ou negativas para o processo de seleção. Naturalmente, ela é positiva ao gerar um indivíduo com um melhor valor de aptidão, ou negativa reduzindo este valor. Nos problemas de Algoritmos Genéticos, as mutações são importantes para aumentar a diversidade de soluções na população e evitar uma convergência muito rápida, que é ocasionada pela ocorrência frequente de máximos locais (De Carvalho et al., 1999).

2.2.3.4 Elitismo

É provável que, gerando diversas populações apenas de um único par de cromossomos, faça com que se perca aqueles indivíduos com melhores valores de aptidão gerados anteriormente. O valor da função de aptidão pode ser reduzido, conforme outros cromossomos sejam gerados (Marsland, 2009). Para isso não acontecer, a estratégia do Elitismo é utilizada com frequência. Este procedimento transfere o melhor cromossomo de uma geração para outra sem alterações em sua cadeia de bits.

3 MATERIAIS E MÉTODO

Neste capítulo são apresentados os materiais necessários para a determinação das modelagens propostas: (1) A amostra coletada para a elaboração do banco de dados da Pesquisa OD e as informações necessárias de cada município da análise; (2) Os aplicativos utilizados na determinação das modelagens; e (3) a elaboração do modelo de otimização para a aplicação da técnica do Algoritmo Genético. O método proposto no estudo também foi detalhado nas próximas subseções.

3.1 MATERIAIS: BANCO DE DADOS E APLICATIVOS

O estado da Bahia é um dos 27 estados do Brasil, localizado na região Nordeste, que compreende uma área de 564.733,177 km² e aproximadamente 15,5 milhões de habitantes (7,5% da população total do Brasil), distribuídas em 417 municípios, incluindo Salvador, a capital. A análise deste trabalho foi baseada no banco de dados de uma Pesquisa Origem-Destino de 2012, realizada em 11 municípios da Bahia: Alagoinhas, Catú, Pojuca, Mata de São João, Dias D'Ávila, Camaçari, Simões Filho, Salvador, Candeias, Santo Amaro e Conceição da Feira. A Figura 3.1 ilustra a área e destaca os 11 municípios.

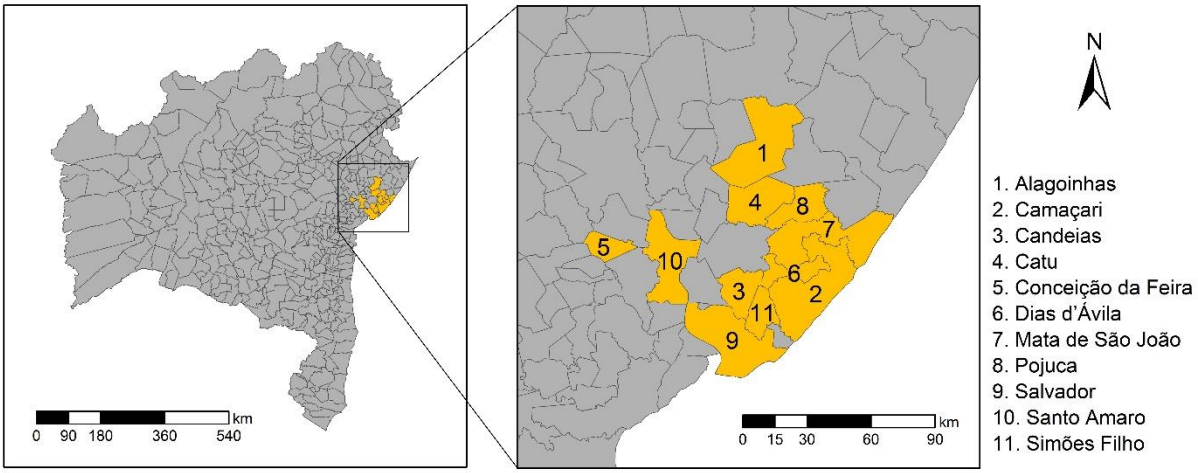

Figura 3.1: Estado da Bahia e os onze municípios estudados (Pitombo et al., 2017)

Dentre todos os pares dos onze municípios envolvidos na pesquisa, estão inclusos importantes polos de atividades com diversas viagens intermunicipais em que o principal motivo da viagem é o trabalho. O estudo de caso envolve centros urbanos de turismo e possui grande importância regional histórica. A Tabela 3.1 apresenta as informações demográficas das onze cidades.

Tabela 3.1: Informações demográficas dos onze municípios

Municípios	População (2010)	PIB per capita	Empregos (saldo)	Área (km²)
Alagoinhas	141.949	12.897,57	667	718,1
Camaçari	242.970	51.221,59	-1.028	784,7
Candeias	83.158	26.384,89	-534	251,6
Catu	51.077	8.127,33	239	416,2
Conceição da Feira	20.391	4.665,71	15	162,9
Dias D'Ávila	66.440	33.239,10	-751	184,2
Mata de São João	40.183	15.264,30	375	633,2
Pojuca	33.066	28.402,54	106	290,1
Salvador	2.675.656	15.229,24	5.962	692,8
Santo Amaro	57.800	7.589,97	67	492,9
Simões Filho	118.047	27.571,93	484	201,6

(Fonte: IBGE, 2010; SEI-BA, 2012).

O espaço amostral estudado foi baseado na pesquisa Origem-Destino, realizada pelo CETRAMA (Centro de Estudos de Transportes e Meio Ambiente-Universidade Federal da Bahia). A Tabela 3.1 apresenta os onze municípios envolvidos na pesquisa, bem como a descrição das variáveis quantitativas referentes a cada cidade. É importante ressaltar que a variável *Empregos* em cada município representa o saldo entre demissões e admissões, em 2012, motivo pelo qual alguns municípios apresentaram valores negativos.

A Pesquisa OD forneceu as informações necessárias para a composição de um banco de dados denso a respeito dos indivíduos entrevistados, no qual foram aplicadas as técnicas de Aprendizagem de Máquinas, CART (*Classification And Regression Tree*) e Algoritmo Genético. Neste trabalho, a área de estudo é representada pelos onze municípios por meio de pares Origem-Destino formados por eles, em que a origem pode ser representada por qualquer uma das onze cidades..

3.1.1 Amostra

A amostra é composta por viajantes intermunicipais (usuários de ônibus nos terminais rodoviários e com motoristas e passageiros de automóveis ou motocicletas nas rodovias intermunicipais). Em vista disso, a coleta ocorreu em polos geradores

dos 11 municípios envolvidos na pesquisa ou em postos rodoviários. A pesquisa envolveu um levantamento do perfil socioeconômico dos entrevistados, bem como informações inerentes à viagem.

A base de dados da pesquisa é composta por pessoas maiores de quinze anos (idade mínima para um indivíduo desacompanhado de um responsável viajar de transportes ferroviários ou rodoviários de passageiros entre as cidades), residentes nos municípios que abrangem a pesquisa. A unidade de seleção amostral é o indivíduo e as informações foram coletadas por meio de um questionário estruturado, elaborado pelos pesquisadores do CETRAMA (UFBA).

Devido à restrição de recursos orçamentários, previstos para o levantamento de campo, foi estabelecido que as entrevistas fossem realizadas nos Pólos Geradores de Tráfego mais representativos em cada uma das cidades localizados, principalmente no entorno das vias e das estações ferroviárias, a uma distância de aproximadamente 800 metros e/ou que tivessem representatividade regional. Tais Polos Geradores de Tráfego englobam os terminais rodoviários intermunicipais, centros comerciais, parques industriais, universidades e centros educacionais em que foi assegurada a aleatoriedade de seleção das unidades amostrais.

O plano amostral escolhido foi o estratificado e cada estrato representa um município da análise. Em cada estrato, amostras aleatórias simples foram realizadas de forma independente. Isto é, a amostra foi constituída por onze estratos. Cada estrato foi proporcional ao número de pessoas com mais de catorze anos que residem nos municípios. Os indivíduos foram alocados de acordo com os critérios de alocação proporcional dada pela Equação 3.1 e Equação 3.2.

$$n_h = n.W_h, \qquad h = 1, 2, 3, \dots, 11 \tag{3.1}$$

Em que, W_h é o peso do estrato h dado pela Equação 3.2.

$$W_h = \frac{N_h}{N}, \qquad h = 1, 2, 3, \dots, 11 \tag{3.2}$$

Em que, N_h é o número de indivíduos com mais de catorze anos no estrato *h* (município) e *N* é o número total de pessoas com mais de catorze anos que residem nos onze municípios. Foi adotado um número mínimo de 100 indivíduos para o tamanho da amostra de cada estrato. A amostra inicial consistiu de 3.300 indivíduos.

Para tratamento da amostra foram utilizadas variáveis socioeconômicas categóricas relativas ao questionário aplicado, além de informações relativas às viagens intermunicipais (motivo, modo de transporte, frequência da viagem, etc.). Além disso, foram adicionados ao banco de dados, variáveis agregadas provenientes do último Censo Demográfico e variáveis de tempo e distância de viagem. Associadas a essas informações estão as escolhas de destino intermunicipais (variável dependente). A variável dependente *Escolha de destinos* refere-se aos onze municípios (Figura 3.1). A Tabela 3.2, Tabela 3.3 e Tabela 3.4 caracterizam as variáveis (dependente e independentes) de acordo com a amostra final de 3.229 indivíduos obtida após tratamento dos dados. O questionário utilizado na Pesquisa OD encontra-se no Apêndice I deste texto.

Tabela 3.2: Características de viagem da amostra

Categorias (municípios) da variável dependente	Número de viagens	Percentagem de viagens (%)
Alagoinhas	339	10,5
Camaçari	588	18,2
Candeias	209	6,5
Catú	134	4,1
Conceição da Feira	33	1,0
Dias D'Ávila	179	5,5
Mata de São João	94	2,0
Pojuca	97	3,0
Salvador	927	28,7
Santo Amaro	160	5,0
Simões Filho	469	14,5

Tabela 3.3: Frequência das categorias das variáveis categóricas (N=3.229)

Variáveis	Variáveis categóricas	Quantidade na amostra	Percentagem (%)
Gênero	Masculino	1.662	51,5
	feminino	1.567	48,5
Idade	até 19	211	6,5
	de 20 a 29	918	28,4
	de 30 a 39	844	26,1
	de 40 a 49	541	16,8
	de 50 a 65	561	17,4
	Acima de 65	144	4,5
Nível de instrução	Sem instrução	142	4,4
	Primeiro grau	976	30,2
	Segundo grau	1.487	46,1
	Terceiro grau	574	17,8
Renda (mensal)*	até 1 SMB**	544	16,8
	entre 1 e 3 SMB**	1.344	41,6
	entre 3 e 5 SMB**	1.010	31,3
	acima 5 SMB**	186	5,8
Ocupação	comércio	956	29,6
	indústria	462	14,3
	serviços	641	19,9
	agricultura	31	1,0
	estudante	276	8,5
	aposentado	247	7,6
	Outra ocupação	550	17,0
Residência	própria	2.565	79,4
	alugada	577	17,9
	cedida	66	2,0
Número de carros por domicílio	nenhum	2.326	72,0
	1	628	19,4
	2	55	1,7
	Mais de 2	18	0,6
Frequência de viagem/Semana	1 dia	1.137	35,2
	2 dias	613	19,0
	3 dias	300	9,3
	4 dias	102	3,2
	5 dias	302	9,4
	6 dias	135	4,2
	Todos os dias	285	8,8
	Outras frequências	224	6,9
Motivo da viagem	Trabalho	1.267	39,2
	Estudo	228	7,1
	Compras	207	6,4
	Lazer	581	18,0

	Saúde	195	6,0
	Visita	572	17,7
	Outros	149	4,6
Tempo de viagem	até 30 min	404	12.5
	entre 30 e 60 min	1.176	36.4
	Acima de 60 min	1.581	49.0
Modo de viagem	ônibus	2.653	82.2
	carro	440	13.6
	van/similar	59	1.8
	a pé	3	0.1
	bicicleta	2	0.1
	other	22	0.7
Custo da viagem por viagem*	Abaixo de US$ 2.56	878	27.2
	entre US$ 2.56 - 5.13	1.016	31.5
	entre US$ 5.13 - 10.26	657	20.3
	Acima de US$ 10.26	612	19.0

* BRL para USD sobre a taxa de câmbio media em 2012.
** SMB: Salário Mínimo Brasileiro em 2012 (aproximadamente 319 US$).

Tabela 3.4: Medidas descritivas das variáveis quantitativas da amostra

Informações	Tamanho da amostra	Mínimo	Máximo	Média	Desvio padrão
População	3.229	20.391	2.675.656	1.331.616,7	1.285.409,1
PIB (R$)	3.229	4.665,7	51.221,6	20.963,11	13.907,92
Saldo Emp Norm	3.229	-1.028	5.962	509	1.883,91

Assim, o banco de dados final foi composto por dois tipos de informações: em um nível agregado e desagregado. Contudo, cada registro refere-se a um indivíduo. Ou seja, um único registro pode ter variáveis agregadas relacionadas à cidade de origem e destino. O próximo registro pode referir-se à mesma cidade de origem, uma vez que ambos os indivíduos são provenientes do mesmo município.

Em outras palavras, a amostra completa contém 3.229 indivíduos (registros), ou seja, 3.229 diferentes observações relacionadas aos indivíduos e às 11 cidades, representando cada cidade de origem e destino na área de estudo. Isto significa que

os valores de variáveis explicativas, em um nível agregado, podem ser repetidos ao longo do banco de dados.

3.1.2 Aplicativos utilizados

As análises realizadas por meio dos algoritmos de Aprendizagem de Máquinas, CART (*Classification And Regression Tree*) e Algoritmo Genético, foram realizadas pelos seguintes aplicativos*, respectivamente:* IBM SPSS 24.0 e MathWorks MATLAB R2016b. O *software* Excel foi utilizado para gerar a modelagem tradicional gravitacional, além de servir como base no tratamento do banco de dados da Pesquisa OD.

3.2 MODELO DE OTIMIZAÇÃO

Para a análise desagregada de distribuição de viagens, realizada pelo Algoritmo Genético, foi efetuada, neste trabalho, a execução de um procedimento de otimização que utilizou os princípios conceituais da modelagem desagregada baseada nos modelos de Escolha Discreta.

Os modelos de Escolha Discreta buscam representar as condições em que um indivíduo realiza a sua escolha diante de um conjunto finito de alternativas, em que cada alternativa é representada por uma função utilidade. A utilidade de cada alternativa é descrita por meio de uma função matemática, definida pela combinação das variáveis, agregadas e desagregadas, selecionadas no conjunto de dados, conforme mostrada na Equação 3.3 (Ben-Akiva e Lerman 1985),

$$\mathrm{U_i} = const + \alpha.\mathrm{A} + \beta.\mathrm{B} + \gamma.\mathrm{C} + \xi_\mathrm{i} \tag{3.3}$$

Para o caso deste trabalho, U_i é a função utilidade da alternativa i $(destino\ intermunicipal)$; $const$ é a constante que engloba todas as variáveis que

não estão inclusas no conjunto de dados; A, B e C são as variáveis relacionadas ao individuo (desagregadas) ou ao município (agregadas); e α, β e γ são os parâmetros a serem estimados das variáveis referentes ao indivíduo ou ao município.

Dessa forma, as funções utilidades das onze alternativas (municípios) podem ser representadas pelas Equações 3.4 e 3.5.

$$\begin{aligned} \mathrm{U_i} = \mathrm{ASCi} + \alpha_\mathrm{i}.\,\mathrm{Idade} + \beta_\mathrm{i}.\,\mathrm{Renda} + \cdots + \gamma_\mathrm{i}.\,\mathrm{População_{origem}} \\ + \delta_\mathrm{i}.\,\mathrm{PIB_{origem}} + \mu_\mathrm{i}.\,\mathrm{PIB_{destino}} + \rho_\mathrm{i}.\,\mathrm{Empregos_{destino}} + \theta_\mathrm{i}\,.\,\mathrm{Distância} \end{aligned} \tag{3.4}$$

$$\mathrm{U_{11}} = 0 \tag{3.5}$$

Em que, U_i é a utilidade do município i, com i = 1, 2, 3,...,10; ASC_i é a constante da função U_i, *Idade*, *Renda*,... são as variáveis selecionadas para a calibração do modelo de otimização; α_i, β_i e γ_i,... representam os parâmetros a serem estimados correspondentes às variáveis que compõem U_i. É importante ressaltar que a utilidade de um dos municípios foi fixada como valor zero (referência).

A partir da obtenção da utilidade para cada alternativa é possível calcular as probabilidades de escolha de cada município. A probabilidade de escolha de uma das alternativas é a probabilidade de que a sua utilidade seja maior que a de todas as outras alternativas e pode ser representada pela Equação 3.6.

$$\mathrm{P_i} = \frac{\mathrm{e^{U_i}}}{\sum_{\mathrm{j=1}}^{\mathrm{n}}(\mathrm{e^{U_j}})} \tag{3.6}$$

Em que, P_i é a probabilidade da alternativa (município) i ser escolhida; U_i é a utilidade da alternativa (município) i e o denominador corresponde ao somatório de todas as alternativas (municípios) disponíveis.

Após a obtenção das utilidades de cada alternativa (município), das probabilidades de cada município *i* ser escolhido e dos parâmetros da função utilidade, o Algoritmo Genético maximizou esta função, de forma que o maior valor de probabilidade é representado pelo município realmente escolhido. Assim, a função objetivo, definida pelo Algoritmo Genético, pode ser representada pela Equação 3.7. O Algoritmo Genético gera os coeficientes relativos a cada variável das onze funções utilidade do modelo de otimização quando o valor máximo da função objetivo é atingido.

$$\sum_{n=1}^{11} \ln(P_n) \tag{3.7}$$

3.3 MÉTODO

O procedimento metodológico adotado e todas as etapas realizadas estão ilustrados na Figura 3.2.

Neste estudo, para o caso dos algoritmos de Aprendizagem de Máquinas, a variável dependente é definida como a *escolha da cidade de destino* com 11 categorias e as variáveis independentes utilizadas foram as características socioeconômicas dos indivíduos da amostra, as características de viagem e as variáveis agregadas dos municípios de origem e destino.

O método descrito neste trabalho segue quatro passos principais, apresentados na Figura 3.2. Primeiramente, dois modelos gravitacionais foram estimados, respectivamente, pelo método dos mínimos quadrados e por meio de um procedimento duplamente restringido. Na sequência, foram aplicados os algoritmos de Aprendizagem de Máquinas, o CART (*Classification And Regression Tree*) e o Algoritmo Genético, para a análise desagregada de distribuição de viagens. E, por

fim, os resultados da aplicação do método desagregado (algoritmos de Aprendizado de Máquina) e dos modelos gravitacionais foram comparados segundo três critérios distintos: distribuição das distâncias de viagens, medidas de ajuste e da perspectiva qualitativa.

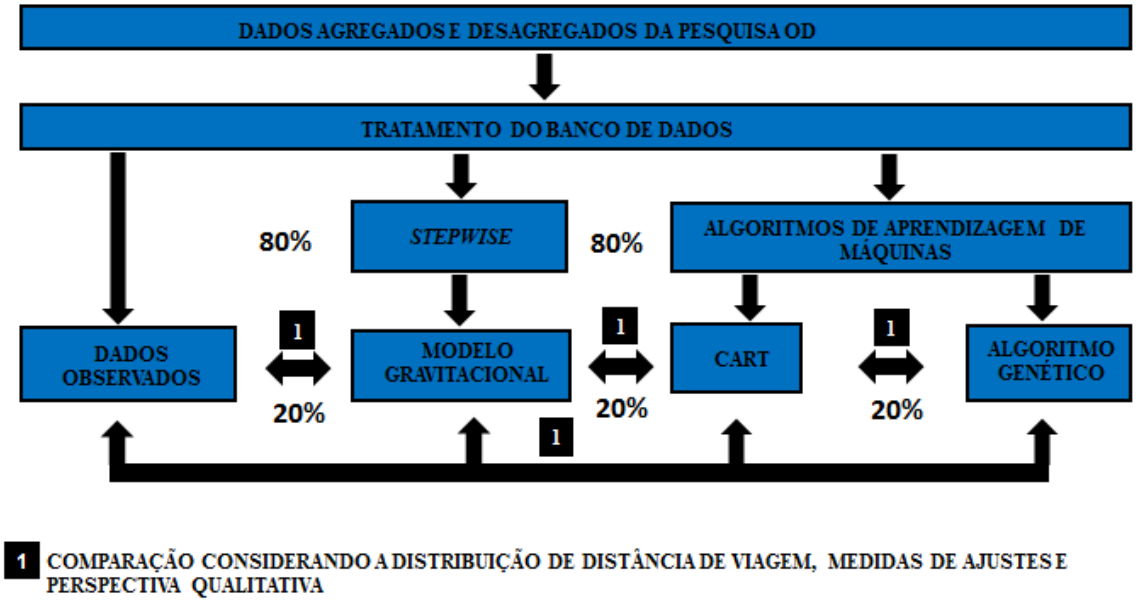

Figura 3.2: Método proposto para estimar a distribuição de viagens intermunicipais (Pitombo et al., 2017)

4 RESULTADOS E DISCUSSÕES

Este capítulo apresenta os resultados das etapas metodológicas apresentadas no capítulo anterior, além de apresentar as principais discussões. Traz os resultados dos quatro modelos propostos (Gravitacional 1, Gravitacional 2, CART e Algoritmo Genético) e a comparação de todas as ferramentas propostas, utilizando três critérios distintos. Finalmente, é comprovada a eficiência e adequabilidade dos algoritmos de Aprendizagem de Máquinas para previsão de viagens intermunicipais por meio de amostra desagregada.

4.1 DESCRIÇÃO GERAL DOS RESULTADOS

Inicialmente, foram realizadas duas análises por meio da calibração de modelos gravitacionais tradicionais, pelo Procedimento *Stepwise* (Modelo Gravitacional 1) e Método Duplamente Restringido (Modelo Gravitacional 2), respectivamente. Na etapa posterior, os algoritmos de Aprendizagem de Máquinas, CART (Modelo 3) e Algoritmo Genético (Modelo 4), foram submetidos a um processo de treinamento e de otimização, respectivamente, usando as variáveis descritas anteriormente. A amostra utilizada continha 2.584 indivíduos, correspondente a 80% da amostra inicial. Tais indivíduos foram escolhidos aleatoriamente a partir do banco de dados original. Os 20% restantes (645 indivíduos) da amostra foram utilizados para validação das ferramentas.

Na análise do algoritmo CART, de Árvore de Decisão, a variável dependente foi definida como a cidade de escolha do destino, com 11 categorias. As variáveis independentes consideradas foram as variáveis socioeconômicas (*Idade*, *Grau de instrução*, *Renda média*, *N° de carros no domicílio* e *Frequência semanal de viagem*), três variáveis referentes às viagens (*Tempo de viagem*, *Custo de viagem* e *Distância de viagem*), além das variáveis agregadas dos municípios (*População da cidade de origem*, *PIB per capita da cidade de origem*, *PIB per capita da cidade de destino* e *Saldo de Empregos Normalizado*). Para a análise de otimização, do Algoritmo Genético, foram utilizadas as mesmas variáveis consideradas no algoritmo CART a fim de se obter a utilidade da escolha de cada alternativa (município).

4.2 MODELO 1: MODELO GRAVITACIONAL 1

O objetivo da calibração dos modelos gravitacionais é a comparação do desempenho dos modelos obtidos pelo algoritmo CART (*Classification And Regression Tree*) e pelo Algoritmo Genético (AG) com modelos clássicos de distribuição de viagens.

O Modelo Gravitacional 1 foi estimado considerando as variáveis independentes agregadas relacionadas à cidade de origem e destino (*População da cidade de origem* e *Saldo de Empregos Normalizado*) e a *Distância de viagem* e a variável dependente utilizada foi o número de viagens entre as onze cidades. Os coeficientes do modelo foram calculados por meio do procedimento dos mínimos quadrados (Procedimento *Stepwise*). Os parâmetros estimados neste modelo estão apresentados na Equação 4.1. Vale ressaltar que foram utilizadas as mesmas amostras de calibração (treinamento) e validação (teste) em todas as etapas metodológicas.

$$V_{ij}^{intermunicipais} = \frac{P_i^{0,406} \cdot E_j^{0,283}}{d_{ij}^{0.110}} \tag{4.1}$$

Onde, $V_{ij}^{intermunicipais}$ é o total de viagens intermunicipais de *i* para *j*, P_i é a população da cidade de origem, E_j é o saldo de empregos normalizado e d_{ij} é a distância entre o município *i* e o município *j*.

Antes da aplicação do procedimento *Stepwise*, foi realizada uma transformação logarítmica das variáveis originais. O Coeficiente de Determinação, R^2, foi de 0,615.

As Tabelas 4.1 e 4.2 apresentam os resultados principais do modelo e a significância dos coeficientes estimados, respectivamente. Ressalta-se ainda que as três variáveis foram selecionadas pelo método *Stepwise* como sendo àquelas associadas aos parâmetros significativos.

Tabela 4.1: ANOVA (Modelo Gravitacional 1)

Modelo	Soma dos Quadrados	Quadrado Médio	F	Sig.
Regressão	1,096	0,365	57,077	0,000
Resíduo	0,685	0,006		
Total	1,781			

Tabela 4.2: Significância dos coeficientes da Regressão Linear (Modelo Gravitacional 1)

Modelo	Coeficientes não padronizados		Coeficientes padronizados	t	Sig.
	B	Erro Padrão	Beta		
Ln População	0,406	0,040	0,675	10,276	0,000
Ln Empregos	0,283	0,042	0,560	6,681	0,000
Ln Dij	-0,110	0,030	-0,331	-3,723	0,000

4.3 MODELO 2: MODELO GRAVITACIONAL 2

O Modelo Gravitacional 2 (Procedimento Duplamente Restringido) foi estimado considerando as mesmas variáveis independentes consideradas no Modelo Gravitacional 1 e a ferramenta de otimização *solver* no *Microsoft Excel*. Os parâmetros estimados no Modelo Gravitacional 2 são mostrados na Equação 4.2.

Pode-se notar que o critério do Modelo Gravitacional 2 minimiza as diferenças entre as distribuições de distâncias de viagens observadas e estimadas enquanto o critério do Modelo Gravitacional 1 minimiza a diferença entre o número de viagens observadas e estimadas entre origens e destinos. Apesar de ter critérios diferentes ao calibrar os modelos, ambos os modelos gravitacionais apresentaram equações semelhantes. As demais variáveis agregadas (*PIB da cidade de origem* e *PIB da cidade de destino*) também não foram consideradas importantes no processo iterativo de otimização utilizado na ferramenta *solver*.

$$V_{ij}^{intermunicipais} = \frac{P_i^{6,24} \cdot E_j^{7,02}}{d_{ij}^{2,13}} \quad (4.2)$$

Onde $V_{ij}^{intermunicipais}$ é o total de viagens intermunicipais de *i* para *j*, P_i é a população da cidade de origem, E_j é o saldo de empregos normalizado e d_{ij} é a distância entre o município *i* e o município *j*.

4.4 MODELO 3: ALGORITMO CART

A partir da base de dados correspondente a 80% da amostra inicial, contida no nó raiz, o algoritmo CART produziu divisões binárias, tendo sido escolhido como critério maximizar a homogeneidade dos nós-filhos, bem como interromper o processo de divisão quando o número de observações em um nó filho fosse menor do que o valor mínimo especificado no *software* (casos mínimos em nó filho é igual a 25 indivíduos). A Figura 4.1 representa esquematicamente o mapa de árvore da etapa de treinamento, com 80% dos registros, em que a caixa superior (número 0) refere-se ao nó-raiz, as caixas subsequentes são os nós-filhos e as últimas caixas (sombreadas em cores cinzentas) referem-se aos seis nós-terminais. A árvore final é composta por cinco níveis.

É importante ressaltar a necessidade da transformação dos valores reais, das variáveis selecionadas pelo algoritmo CART, em valores normalizados, pelo fato dos valores da variável *Saldo de Empregos Normalizado* estarem muito discrepantes dos valores das demais variáveis. O procedimento de normalização adotado considerou os valores máximo e mínimo da variável *Saldo de Empregos Normalizado*.

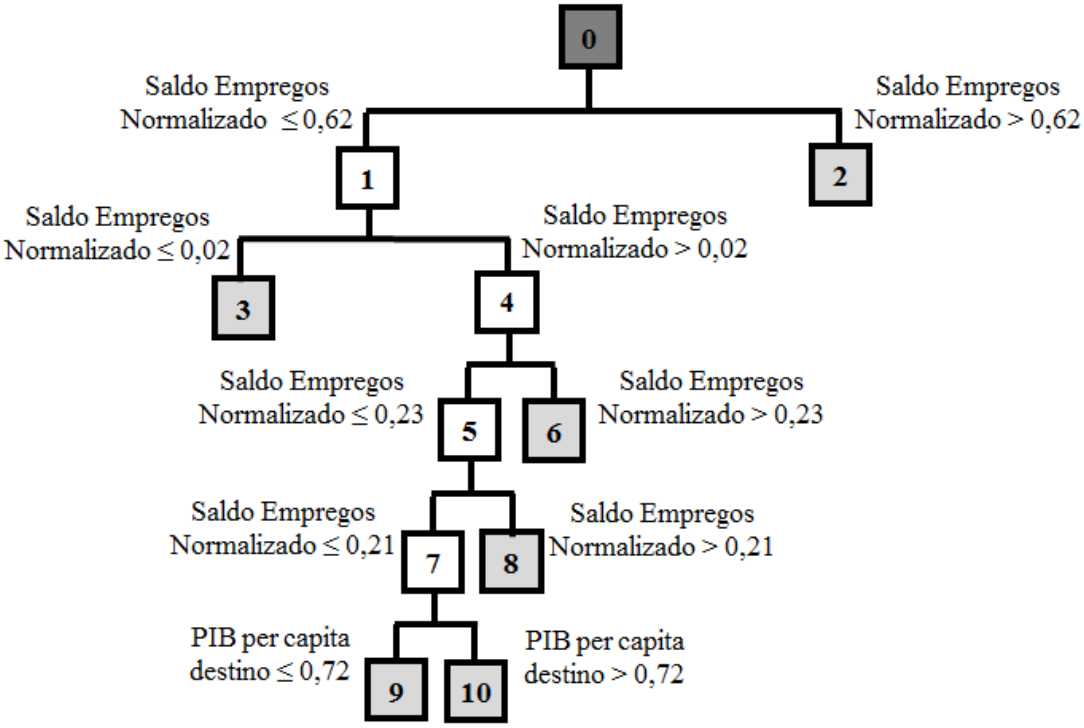

Figura 4.1: Mapa da Árvore de Decisão (Modelo 3: Algoritmo CART)

A Tabela 4.3 apresenta as informações dos nós-terminais, incluindo as três escolhas de destinos mais frequentes selecionadas pelo algoritmo CART.

Algumas conclusões podem ser exploradas a partir do modelo representado na Tabela 4.2 e Figura 4.1. As divisões da amostra de treinamento resultaram em 10 subgrupos, compreendendo todos os nós-filhos (incluindo os nós-terminais), com diferentes combinações de faixa de valores para as variáveis agregadas *Saldo de Empregos Normalizado* e *PIB da cidade de destino*. Observou-se, ainda, homogeneidade total nos nós 2, 3, 6, 8 e 10. Tais folhas apresentaram 100% de observações em apenas uma das onze categorias possíveis da variável dependente.

Tabela 4.3: Municípios de destino mais frequentes e as condições de corte nos nós terminais (Modelo 3: Algoritmo CART)

Nó	Municípios de destinos mais frequentes por folhas	%
2	Salvador (100%)	28,9
3	Camaçari (100%)	18,4
6	Alagoinhas (100%)	10,1
8	Simões Filho (100%)	14,6
9	Dias D'Ávila (23,5%), Santo Amaro (23,0%), Catu (20,7%)	21,7
10	Candeias (100%)	6,3
Nó	**Condições de corte das variáveis explicativas**	**%**
2	Saldo Empregos Normalizado > 0,62	28,9
3	Saldo Empregos Normalizado ≤ 0,02	18,4
6	Saldo Empregos Normalizado ≤ 0,62; Saldo Empregos Normalizado > 0,23	10,1
8	Saldo Empregos Normalizado ≤ 0,23; Saldo Empregos Normalizado > 0,21	14,6
9	Saldo Empregos Normalizado > 0,02; Saldo Empregos Normalizado ≤ 0,21; PIB per capita ≤ 0,72	21,7
10	Saldo Empregos Normalizado > 0,02; Saldo Empregos Normalizado ≤ 0,21; PIB per capita > 0,72	6,3

A primeira variável independente selecionada foi *Saldo de Empregos Normalizado,* com o valor de corte de 0,62. Esta variável independente (e valor de corte selecionado) forneceu os dois grupos (nós filhos seguintes ao nó raiz: Nó 1 e Nó 2) mais homogêneos do banco de dados original, segundo a variável dependente *Escolha de destinos*. Análogo a este processo, os seguintes nós dividiram-se considerando novos valores das variáveis *Saldo de Empregos Normalizado* e *PIB per capita*.

Pode-se notar que Salvador é o único município com *Saldo de Empregos Normalizado* maior que 0,62. Assim, considera-se que os residentes em Salvador compõem o ramo direito da árvore final.

É importante mencionar que, apesar de não ter variáveis desagregadas na estrutura da árvore, o algoritmo CART selecionou uma variável desagregada como divisão substituta: *N° de carros no domicílio*. A Tabela 4.4 mostra as principais divisões e as divisões substitutas (incluindo a variável *N° de carros no domicílio*). A partir da amostra de validação do processo de Árvore de Decisão (AD), o algoritmo CART forneceu uma precisão 85,9% de acertos.

Tabela 4.4: Divisões principais e substitutas

Nó	Divisões principais	Divisões substitutas	Aprimoramento	Associação
0	Saldo de Empregos Normalizado	População da origem	0,097	0,193
		PIB per capita do destino	0,158	0,181
		PIB per capita da origem	0,051	0,177
		Distância de viagem	0,017	0,090
		Carros no domicílio	0,001	0,001
1	Saldo de Empregos Normalizado	PIB per capita do destino	0,130	0,655
		Distância de viagem	0,014	0,025
4	Saldo de Empregos Normalizado	Distância de viagem	0,033	0,352
		PIB per capita da origem	0,011	0,111
		População da origem	0,020	0,107
5	Saldo de Empregos Normalizado	Distância de viagem	0,067	0,641
		PIB per capita do destino	0,067	0,069
		Frequência semanal de viagem	0,004	0,056
		Tempo de viagem	0,006	0,032
		PIB per capita da origem	0,001	0,013
		Custo de viagem	0,009	0,008
7	PIB per capita do destino	Saldo de Empregos Normalizado	0,050	0,183
		PIB per capita da origem	0,001	0,018
		Frequência semanal de viagem	0,002	0,018

4.5 MODELO 4: ALGORITMO GENÉTICO

Para a obtenção do modelo de otimização, por meio da técnica do Algoritmo Genético, foi utilizado o *software* MathWorks MATLAB R2016B, o qual possui sua própria biblioteca (parâmetros) de otimização. Tais parâmetros representam as considerações que o Algoritmo Genético (AG) realiza o processo de otimização: definição da população inicial, seleção, *crossover* e mutação e o critério de parada. Porém, existe a possibilidade de alteração manual desses parâmetros pelo usuário. A Tabela 4.5 apresenta os valores destes parâmetros considerados neste trabalho.

Tabela 4.5: Parâmetros considerados pelo Algoritmo Genético (AG)

Parâmetros	Valor
Tamanho da população	200
Alcance da população inicial	[-100;100]
Contador de elite	10
Fração de *crossover*	0,8
Mutação	Função Gaussiana (adiciona um número aleatório de uma distribuição Gaussiana dos valores de cada indivíduo pai)
Cruzamento	Função *Dispersa* (cria um vetor binário aleatório, onde possuir valor 1 neste vetor, retira o gene do pimeiro pai, onde possuir 0, retira o gene do segundo pai)
Critério de parada	1.Número máximo de gerações ≤ 1300
	2.Número de gerações consecutivas com diferença menor que $10^{-4} > 10$

A partir destas considerações, o Algoritmo Genético calibrou os 130 parâmetros das doze variáveis (e da constante) consideradas nas funções utilidades de cada município da análise, conforme já definidas pelas Equações 3.4 e 3.5, no Capítulo 3.

Analogamente ao algoritmo CART, a calibração dos coeficientes se deu por meio da utilização da amostra de 80% dos indivíduos da amostra, obtidas aleatoriamente, enquanto que, os 20% restantes da amostra foram utilizadas na fase de testes (validação).

A partir daí, por meio da utilização do *software* MATLAB e considerando os parâmetros presentes na biblioteca do *software*, foram obtidas oito replicações distintas dos 130 parâmetros, provenientes de testes, relativas aos destinos escolhidos. Todas estas oito replicações tiveram percentuais de acertos altos, obtidos a partir da amostra de validação, entre aproximadamente 94% e 96%, conforme Tabela 4.6. Todos os parâmetros estimados das variáveis, relativos às

funções utilidade de cada município nas oito replicações, encontram-se no Apêndice II deste texto.

Tabela 4.6: Porcentagem de acertos nas replicações geradas pelo Algoritmo Genético (AG)

Testes	Porcentagem de acertos (%)
1	95,35
2	95,81
3	94,73
4	94,42
5	95,81
6	95,35
7	94,26
8	94,73

4.5.1 Procedimento de escolha da melhor replicação

Na definição do melhor conjunto de parâmetros do Algoritmo Genético (AG) foi utilizado novamente o algoritmo CART de Árvore de Decisão, considerando os sinais e magnitude de tais parâmetros estimados nas onze funções utilidade. O algoritmo CART facilitou tal escolha, pois, fixada cada uma das variáveis independentes como a primeira para divisão dos dados, pôde-se perceber a sua influência (negativa ou positiva) na escolha de cada uma das onze cidades, definindo então o sinal do parâmetro estimado. Desta forma, foi proposto o seguinte procedimento:

(1) Primeiramente, fixa-se cada uma das variáveis independentes como primeira divisão a partir do nó raiz (como *Idade* e *Grau de Instrução*, por exemplo);

(2) Observa-se se o percentual de escolha de cada município diminuiu ou aumentou com o aumento da variável fixada em (1). Por exemplo, as escolhas para o município de Camaçari aumentam com o aumento do *Grau de Instrução*, logo o sinal desta variável para a utilidade do município de Camaçari deve ser positivo;

(3) Adota-se o sinal do parâmetro relativo à variável investigada em (2) para cada uma das funções utilidade;

(4) Repetem-se os passos de (1) a (3) para as demais variáveis independentes;

(5) Escolhe-se a replicação do Algoritmo Genético (AG) que possui maior número de coerências de sinais de parâmetros estimados.

Alguns resultados obtidos pelo algoritmo CART são observados nos esquemas a seguir, Figura 4.2 e Figura 4.3, fixando as variáveis *Idade* e *Grau de Instrução* como primeira divisão, respectivamente.

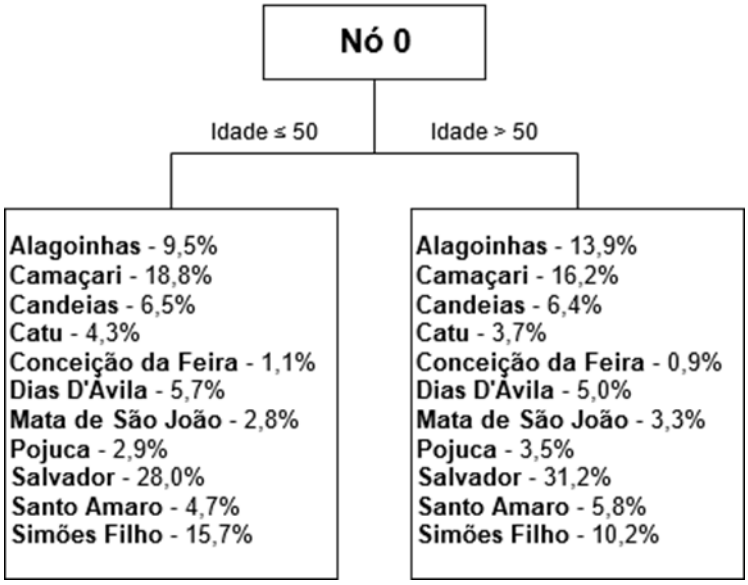

Figura 4.2: Algoritmo CART com primeira variável *Idade* fixada

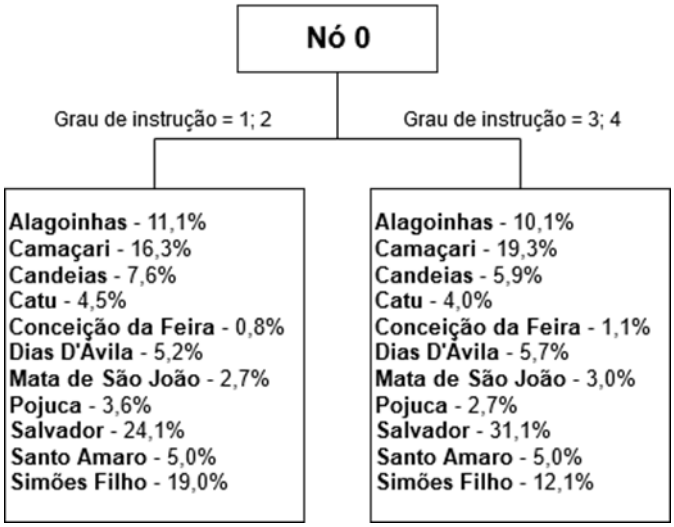

Figura 4.3: Algoritmo CART com primeira variável *Grau de Instrução* fixada (1-Sem Instrução; 2-Pimeiro Grau; 3-Segundo Grau; e 4-Terceiro Grau)

Observando a Figura 4.2, por exemplo, verifica-se que pessoas com idade maiores que 50 anos têm maior propensão de escolher o município de Alagoinhas, quando comparadas às pessoas com idade inferior ou igual a 50 anos. Espera-se então, por exemplo, que o parâmetro relativo à variável idade na função utilidade de Alagoinhas (U_1) tenha sinal positivo.

Já verificando a Figura 4.3, percebe-se que a escolha do destino Salvador é realizada por pessoas com *Grau de Instrução* mais alto. Desta forma, espera-se também sinal positivo para o parâmetro relativo à variável *Grau de Instrução* na função utilidade referente ao município de Salvador (U_9).

Assim, a partir destes resultados, foi analisada a influência de cada variável na escolha de cada município como destino da viagem. Tal influência foi comparada com o sinal de cada parâmetro obtido pelas replicações no Algoritmo Genético (AG), visto que as variáveis que aumentavam o número de viagens atraídas por certo município deveriam possuir coeficiente com sinal positivo no modelo de otimização e as variáveis que diminuíam o número de viagens, deveriam possuir coeficiente com sinal negativo. Desta forma, foi feita uma análise mais cuidadosa dos resultados obtidos nas oito replicações pelo Algoritmo Genético (AG), selecionando a replicação 5 como o melhor conjunto de parâmetros, a qual caracterizou as funções

utilidade, apresentadas na Tabela 4.7, de acordo com os seguintes valores de parâmetros.

Tabela 4.7: Funções utilidade de cada município do Modelo 5

$U_1 = -3{,}52 + 0{,}14.\text{Idade} - 0{,}11.\text{Renda} - 0{,}16.\text{Instrução}$ $- 0{,}06.\text{Carros no domicílio}$ $+ 0{,}26.\text{Frequencia semanal de viagem}$ $- 0{,}67.\text{Tempo de viagem} + 0{,}63.\text{Custo da viagem}$ $- 6{,}34.\text{Pop da cidade de origem}$ $- 2{,}89.\text{PIB da cidade de orige}$ $- 11{,}52.\text{PIB da cidade de destino}$ $+ 33{,}69.\text{Empregos} + 14{,}47.\text{Distância}$	U_1 Alagoinhas
$U_2 = -0{,}58 + 0{,}26.\text{Idade} - 0{,}24.\text{Renda} + 0{,}33.\text{Instrução}$ $- 0{,}34.\text{Carros no domicílio}$ $+ 0{,}26.\text{Frequencia semanal de viagem}$ $- 0{,}16.\text{Tempo de viagem} + 0{,}32.\text{Custo da viagem}$ $- 2{,}88.\text{Pop da cidade de origem}$ $- 3{,}41.\text{PIB da cidade de origem}$ $+ 9{,}35.\text{PIB da cidade de destino}$ $- 44{,}87.\text{Empregos} + 1{,}73.\text{Distância}$	U_2 Camaçari
$U_3 = -4{,}91 - 0{,}02.\text{Idade} - 0{,}42.\text{Renda} - 0{,}05.\text{Instrução}$ $- 0{,}76.\text{Carros no domicílio}$ $+ 0{,}20.\text{Frequencia semanal de viagem}$ $- 0{,}63.\text{Tempo de viagem} + 0{,}29.\text{Custo da viagem}$ $- 3{,}11.\text{Pop da cidade de origem}$ $- 5{,}19.\text{PIB da cidade de origem}$ $+ 17{,}55.\text{PIB da cidade de destino}$ $- 9{,}00.\text{Empregos} + 3{,}68.\text{Distância}$	U_3 Candeias

$U_4 = +2{,}09 + 0{,}54.\,\text{Idade} - 0{,}13.\,\text{Renda} + 0{,}43.\,\text{Instrução}$ $+ 0{,}02.\,\text{Carros no domicílio}$ $+ 0{,}34.\,\text{Frequencia semanal de viagem}$ $- 0{,}49.\,\text{Tempo de viagem} + 0{,}69.\,\text{Custo da viagem}$ $- 3{,}48.\,\text{Pop da cidade de origem}$ $- 0{,}77.\,\text{PIB da cidade de origem}$ $- 21{,}89.\,\text{PIB da cidade de destino}$ $- 4{,}50.\,\text{Empregos} + 7{,}55.\,\text{Distância}$	U_4 Catu
$U_5 = -3{,}70 + 0{,}32.\,\text{Idade} + 0{,}04.\,\text{Renda} - 0{,}30.\,\text{Instrução}$ $- 0{,}05.\,\text{Carros no domicílio}$ $+ 0{,}34.\,\text{Frequencia semanal de viagem}$ $- 0{,}60.\,\text{Tempo de viagem} + 0{,}72.\,\text{Custo da viagem}$ $- 3{,}14.\,\text{Pop da cidade de origem}$ $+ 3{,}81.\,\text{PIB da cidade de origem}$ $- 21{,}67.\,\text{PIB da cidade de destino}$ $- 11{,}86.\,\text{Empregos} + 19{,}17.\,\text{Distância}$	U_5 Conceição da Feira
$U_6 = 2{,}79 + 0{,}36.\,\text{Idade} + 0{,}01.\,\text{Renda} + 0{,}55.\,\text{Instrução}$ $- 0{,}53.\,\text{Carros no domicílio}$ $+ 0{,}17.\,\text{Frequencia semanal de viagem}$ $+ 0{,}01.\,\text{Tempo de viagem} + 0{,}21.\,\text{Custo da viagem}$ $- 2{,}90.\,\text{Pop da cidade de origem}$ $- 1{,}36.\,\text{PIB da cidade de origem}$ $- 2{,}94.\,\text{PIB da cidade de destino}$ $- 28{,}17.\,\text{Empregos} + 6{,}26.\,\text{Distância}$	U_6 Dias D'Ávila
$U_7 = -1{,}15 + 0{,}37.\,\text{Idade} - 0{,}03.\,\text{Renda} + 0{,}25.\,\text{Instrução}$ $- 0{,}05.\,\text{Carros no domicílio}$ $+ 0{,}26.\,\text{Frequencia semanal de viagem}$ $+ 0{,}21.\,\text{Tempo de viagem} + 0{,}45.\,\text{Custo da viagem}$ $+ 1{,}15.\,\text{Pop da cidade de origem}$ $+ 1{,}20.\,\text{PIB da cidade de origem}$ $- 23{,}56.\,\text{PIB da cidade de destino}$ $+ 10{,}83.\,\text{Empregos} - 1{,}23.\,\text{Distância}$	U_7 Mata de São João

$U_8 = -3{,}21 + 0{,}06.\text{Idade} - 0{,}12.\text{Renda} - 0{,}03.\text{Instrução} - 0{,}43.\text{Carros no domicílio} + 0{,}10.\text{Frequencia semanal de viagem} - 0{,}11.\text{Tempo de viagem} + 0{,}42.\text{Custo da viagem} - 5{,}56.\text{Pop da cidade de origem} - 4{,}80.\text{PIB da cidade de origem} + 7{,}15.\text{PIB da cidade de destino} + 4{,}93.\text{Empregos} + 11{,}77.\text{Distância}$	U_8 Pojuca
$U_9 = -12{,}11 + 0{,}44.\text{Idade} - 0{,}73.\text{Renda} - 0{,}28.\text{Instrução} - 0{,}14.\text{Carros no domicílio} + 0{,}17.\text{Frequencia semanal de viagem} - 1{,}16.\text{Tempo de viagem} + 0{,}69.\text{Custo da viagem} - 7{,}92.\text{Pop da cidade de origem} - 3{,}76.\text{PIB da cidade de origem} - 17{,}25.\text{PIB da cidade de destino} + 56{,}85.\text{Empregos} + 11{,}06.\text{Distância}$	U_9 Salvador
$U_{10} = 0{,}48 + 0{,}38.\text{Idade} + 0{,}25.\text{Renda} - 0{,}04.\text{Instrução} + 0{,}18.\text{Carros no domicílio} + 0{,}36.\text{Frequencia semanal de viagem} - 0{,}14.\text{Tempo de viagem} + 0{,}81.\text{Custo da viagem} + 1{,}75.\text{Pop da cidade de origem} + 5{,}11.\text{PIB da cidade de origem} - 26{,}65.\text{PIB da cidade de destino} - 19{,}48.\text{Empregos} + 5{,}01.\text{Distância}$	U_{10} Santo Amaro
$U_{11} = 0$	U_{11} Simões Filho

4.6 COMPARAÇÕES DAS ABORDAGENS

Para comparação das abordagens e mensuração da eficiência dos algoritmos de Aprendizagem de Máquinas na estimativa de distribuição de viagens intermunicipais, são propostos os critérios descritos a seguir: (1) Distribuição de distâncias de viagens; (2) Medidas de qualidade de ajuste; e (3) Perspectiva qualitativa.

4.6.1 Distribuição das distâncias de viagens

A distribuição das distâncias das viagens foi considerada como uma medida de desempenho para avaliar e comparar os modelos previamente estimados. Os histogramas relacionados às distâncias de viagens dos dados observados, dos dados estimados pelos Modelos Gravitacionais (Modelo 1 e Modelo 2) e dos dados estimados pelos algoritmos de Aprendizagem de Máquinas (Modelo 3 e Modelo 4) estão ilustrados na Figura 4.4. Este critério foi escolhido, pois a calibração do Modelo Duplamente Restringido (Modelo Gravitacional 2) baseou-se na minimização das diferenças entre as distâncias observadas e estimadas.

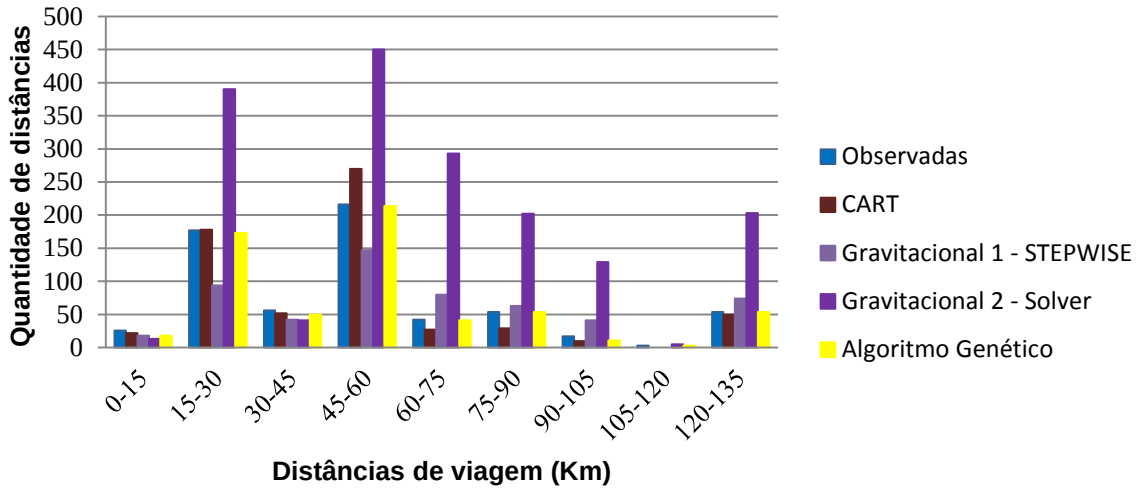

Figura 4.4: Histograma das distâncias de viagens

A Figura 4.4 indica uma forte associação entre a distribuição das distâncias de viagens dos valores observados e os estimados pelos algoritmos de Aprendizagem de Máquinas. Como pode ser observado no histograma, o Algoritmo Genético apresentou o melhor desempenho para qualquer faixa de distância destacada. Com relação ao desempenho do algoritmo CART, foi melhor para viagens menores que 30 km, enquanto o Modelo Gravitacional 1 melhor estima as viagens com distância menores que 15 km e distâncias entre 75 e 90 km.

Além da interpretação visual, testes estatísticos não paramétricos de *Mann-Whitney* e *Kolmogorov-Smirnov*, foram realizados a fim de comparar as distribuições das distâncias de viagens estimadas pelas diferentes abordagens. A hipótese nula é que duas amostras são delineadas a partir da mesma distribuição de probabilidade. Por isso, dividiu-se a análise em grupos. O Grupo 1 refere-se à comparação das distribuições das distâncias de viagens observadas com as distribuições das distâncias de viagens estimadas por cada uma das abordagens. O Grupo 2 refere-se à comparação das distribuições das distâncias de viagens estimadas pelo CART com as distribuições das distâncias de viagens estimadas pelas demais abordagens. O Grupo 3 refere-se à comparação das distribuições das distâncias de viagens estimadas pelo Modelo Gravitacional 1 com as distribuições das distâncias de viagens estimadas pelas demais abordagens. O Grupo 4 refere-se à comparação das distribuições das distâncias de viagens estimadas pelo Modelo Gravitacional 2 com as distribuições das distâncias de viagens estimadas pelas demais abordagens. E, por fim, o Grupo 5 refere-se à comparação das distribuições das distâncias de viagens estimadas pelo Algoritmo Genético (AG) com as distribuições das distâncias de viagens estimadas pelas demais abordagens. Além disso, o teste da mediana foi realizado para testar se as amostras têm medidas centrais semelhantes.

Diante disso, as Tabelas 4.8 a 4.12 apresentam os resultados dos testes estatísticos ao comparar a frequência de distâncias nas situações mencionadas. A Tabela 4.13 mostra os valores das médias e medianas das distâncias de viagens para cada abordagem.

Com base nos resultados, pode-se afirmar que, para o Grupo 1, o Algoritmo Genético, seguido pelo CART, forneceram as estimativas mais precisas. Observando os testes não paramétricos, pode-se notar que ambos os algoritmos de Aprendizagem de Máquinas têm a mesma distribuição de probabilidade dos valores observados, bem como, as mesmas medidas centrais. No entanto, para ambos os modelos gravitacionais, esta hipótese foi rejeitada.

Tabela 4.8: Teste de distribuição de probabilidade para distâncias de viagem (Grupo 1)

Comparação	Teste mediana		*Mann-Whitney*		Teste Kolmogorov-Smirnov	
	Sig.	Hipótese nula	Sig.	Hipótese nula	Sig.	Hipótese nula
CART x Obs.	0,892	Não rejeitada	0,579	Não rejeitada	0,027	Rejeitada
AG x Obs.	0,856	Não rejeitada	0,770	Não Rejeitada	1,000	Não Rejeitada
Gravit. 1 x Obs.	0,000	Rejeitada	0,000	Rejeitada	0,000	Rejeitada
Gravit. 2 x Obs.	0,000	Rejeitada	0,000	Rejeitada	0,000	Rejeitada

Para o Grupo 2, que compara a distribuição das distâncias de viagens estimadas pelo CART com as distribuições das distâncias de viagens estimadas pelas demais abordagens, o AG apresentou os melhores resultados para o teste não paramétrico Mann-Whitney e o Teste da Mediana, para os quais a hipótese nula não foi rejeitada.

Tabela 4.9: Teste de distribuição de probabilidade para distâncias de viagem (Grupo 2)

Comparação	Teste mediana		*Mann-Whitney*		Teste Kolmogorov-Smirnov	
	Sig.	Hipótese nula	Sig.	Hipótese nula	Sig.	Hipótese nula
Gravit. 1 x CART	0,000	Rejeitada	0,000	Rejeitada	0,000	Rejeitada
Gravit. 2 x CART	0,000	Rejeitada	0,000	Rejeitada	0,000	Rejeitada
AG x CART	0,994	Não Rejeitada	0,387	Não Rejeitada	0,026	Rejeitada

Para o Grupo 3, que compara a distribuição das distâncias de viagens estimadas pelo Modelo Gravitacional 1 com as distribuições das distâncias de viagens estimadas pelas demais abordagens, o Modelo Gravitacional 2 apresentou os melhores resultados para o teste não paramétrico Mann-Whitney e o Teste da Mediana, para os quais a hipótese nula não foi rejeitada.

Tabela 4.10: Teste de distribuição de probabilidade para distâncias de viagem (Grupo 3)

Comparação	Teste mediana		*Mann-Whitney*		Teste Kolmogorov-Smirnov	
	Sig.	Hipótese nula	Sig.	Hipótese nula	Sig.	Hipótese nula
Gravit. 2 x Gravit. 1	0,427	Não rejeitada	0,596	Não Rejeitada	0,024	Rejeitada
CART x Gravit. 1	0,000	Rejeitada	0,000	Rejeitada	0,000	Rejeitada
AG x Gravit. 1	0,000	Rejeitada	0,000	Rejeitada	0,000	Rejeitada

Para o Grupo 4, que compara a distribuição das distâncias de viagens estimadas pelo Modelo Gravitacional 2 com as distribuições das distâncias de viagens estimadas pelas demais abordagens, o Modelo Gravitacional 1 apresentou os melhores resultados para o teste não paramétrico Mann-Whitney e o Teste da Mediana, para os quais a hipótese nula não foi rejeitada.

Tabela 4.11: Teste de distribuição de probabilidade para distâncias de viagem (Grupo 4)

Comparação	Teste mediana		*Mann-Whitney*		Teste Kolmogorov-Smirnov	
	Sig.	Hipótese nula	Sig.	Hipótese nula	Sig.	Hipótese nula
Gravit. 1 x Gravit. 2	0,427	Não Rejeitada	0,596	Não Rejeitada	0,024	Rejeitada
CART x Gravit. 2	0,000	Rejeitada	0,000	Rejeitada	0,000	Rejeitada
AG x Gravit. 2	0,000	Rejeitada	0,000	Rejeitada	0,000	Rejeitada

Para o Grupo 5, que compara a distribuição das distâncias de viagens estimadas pelo Algoritmo Genético (AG) com as distribuições das distâncias de viagens estimadas pelas demais abordagens, as viagens observadas forneceram as estimativas mais precisas para todos os testes, seguida pelo CART, quando a amostra só foi rejeitada pelo teste não paramétrico *Kolmogorov-Smirnov*.

Tabela 4.12: Teste de distribuição de probabilidade para distâncias de viagem (Grupo 5)

Comparação	Teste mediana		*Mann-Whitney*		Teste Kolmogorov-Smirnov	
	Sig.	Hipótese nula	Sig.	Hipótese nula	Sig.	Hipótese nula
Obs. x AG	0,856	Não Rejeitada	0,770	Não Rejeitada	1,000	Não Rejeitada
Gravit. 1 x AG	0,000	Rejeitada	0,000	Rejeitada	0,000	Rejeitada
Gravit. 2 x AG	0,000	Rejeitada	0,000	Rejeitada	0,000	Rejeitada
CART x AG	0,994	Não Rejeitada	0,387	Não Rejeitada	0,026	Rejeitada

Tabela 4.13: Valores de médias e medianas para cada método analisado

Análise	Média	Mediana
Observados	51,7	46,4
CART	49,7	46,4
AG	52,1	46,4
Gravit. 1	62,4	56,1
Gravit. 2	62,6	58,3

4.6.2 Medidas de qualidade do ajuste

Para o critério da qualidade do ajuste, algumas medidas relativas ao erro foram avaliadas, como as mostradas nas Equações 4.3 a 4.6, que, respectivamente, referem-se ao Erro Médio Quadrático, Raiz do Erro Médio Quadrático, Raiz Quadrada Média do Erro, o Erro Médio Absoluto e Correlação de *Pearson*.

$$\frac{1}{N}\sum(x_i - y_i)^2 \tag{4.3}$$

$$\sqrt{\frac{1\sum(x_i - y_i)^2}{N}} \tag{4.4}$$

$$\frac{1}{N}\sum(x_i - y_i) \tag{4.5}$$

$$\frac{1}{N-1}.\sum\frac{(x_i-\bar{x})(y_i-\bar{y})}{\sigma_x\sigma_y} \tag{4.6}$$

Onde x_i é a medida estimada; y_i é a medida observada; N é o número de medidas, $\bar{x}$ e $\bar{y}$ são as médias das amostras; σ_x e σ_y são os desvios-padrão da amostra.

Um modelo é considerado com melhor desempenho do que outro se a sua qualidade do ajuste for melhor. A Tabela 4.14 apresenta os valores das medidas propostas. Observando as medidas estatísticas, pode-se notar que o procedimento de otimização do Algoritmo Genético (AG), seguido pelo CART, proporcionou melhores resultados do que ambos os modelos gravitacionais. O Modelo Gravitacional 1 (*STEPWISE*) forneceu melhores resultados do que o Modelo Gravitacional 2.

Tabela 4.14: Medidas estatísticas da qualidade do ajuste

Método	Erro médio quadrático	Raiz do erro médio quadrático	Erro médio absoluto	Correlação de Pearson
CART	36,74	6,06	1,59	0,91
AG	0,68	0,83	0,25	1,00
Modelo Gravit. 1	80,61	8,98	4,52	0,73
Modelo Gravit. 2	770,30	27,75	11,98	0,79

4.6.3 Perspectiva qualitativa

4.6.3.1 Árvore de Decisão

O processo de divisão do algoritmo CART selecionou apenas uma variável desagregada, *N° de carros no domicílio,* como divisor substituto. Tal variável foi escolhida pelo fato de haver maior semelhança com relação à variável principal, minimizando a impureza de determinado nó.

O algoritmo CART fornece a medida de importância de uma variável explicativa em relação à árvore final. Esta medida é expressa em termos de quantidade normalizada, em relação à variável com a maior dimensão de importância. A Tabela 4.15 apresenta a importância normalizada das variáveis selecionadas por meio do algoritmo CART, sendo quase a sua totalidade formada pelas variáveis agregadas.

Tabela 4.15: Medidas de importância das variáveis explicativas

Variável explicativa	Medida de Importância
Empregos	100,0%
PIB da cidade de destino	63,4%
Distância	20,0%
População da cidade de origem	17,9%
PIB da cidade de origem	9,9%
Custo de viagem	1,4%
Frequência semanal de viagem	1,0%
Tempo de viagem	0,9%
Carros no domicílio	0,1%

É possível verificar que a variável agregada *Empregos* especialmente também selecionada pelos modelos gravitacionais, foi a variável explicativa mais importante. Além disso, é relevante mencionar que o algoritmo CART não perde significativamente a precisão do modelo (taxas de acertos) quando a variável desagregada é utilizada como divisor substituto a partir do nó raiz.

Outra análise detalhada do divisor substituto é ilustrada na Figura 4.5, a qual apresenta as diferenças nas escolhas de destinos quando a variável desagregada é selecionada como divisor substituto no nó raiz (nó 0). Observando a Tabela 4.4, a única variável desagregada que poderia ser usada como divisor substituto no nó raiz foi o *N° de carros no domicílio*. Esta variável é categórica, em que a primeira categoria indica que o indivíduo não possui automóveis em seu domicílio, na segunda categoria o indivíduo possui um automóvel, na terceira possui dois automóveis e na quarta o entrevistado possui mais de dois automóveis.

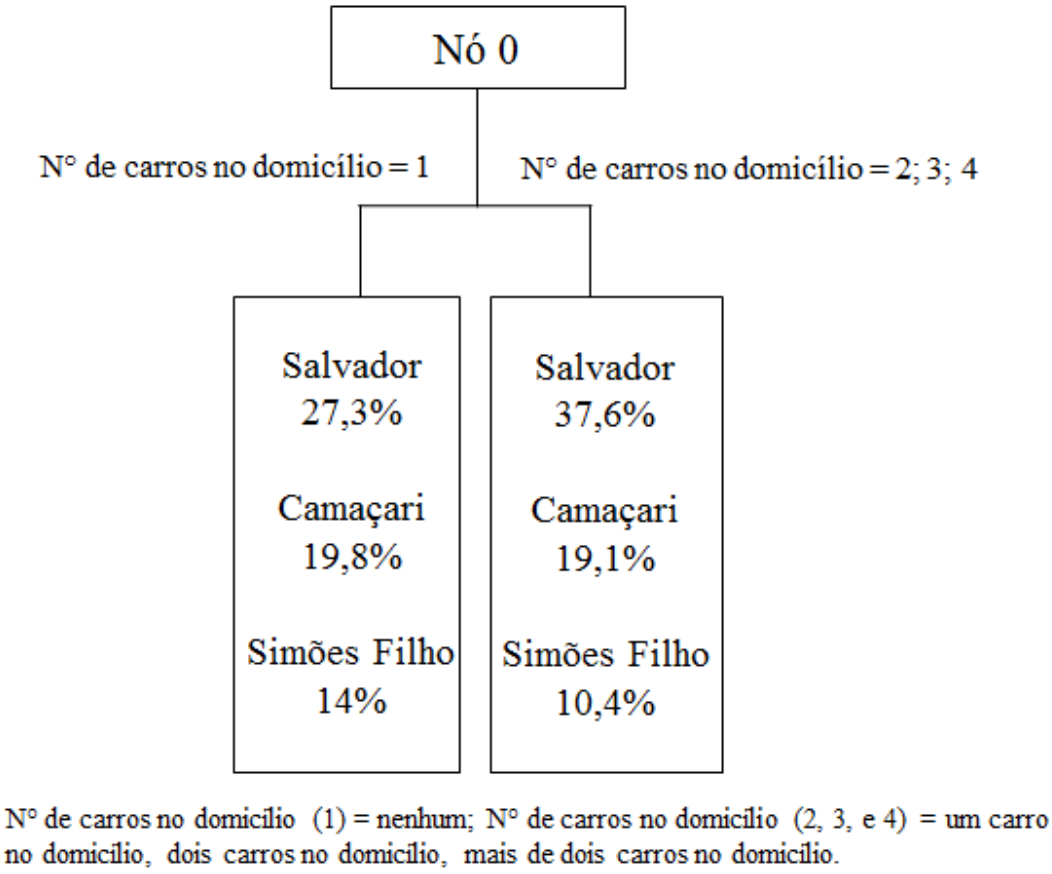

Figura 4.5: Variável *N° de carros no domicílio* utilizada como divisor substituto no nó raiz

Dessa forma, pode-se observar que os indivíduos que possuem pelo menos um carro em seu domicílio são mais propensos a escolher Salvador como município de destino e menos dispostos a escolher Camaçari e Simões Filho quando comparados com os indivíduos que não possuem automóveis, por exemplo.

4.6.3.2 Algoritmo Genético

A análise qualitativa do Algoritmo Genético (AG) se deu por meio da avaliação dos valores dos coeficientes estimados das funções utilidade, podendo compreender a influência positiva ou negativa de cada variável, inclusive das desagregadas, na escolha de determinado município. Tal análise não foi possível através do algoritmo

CART, já que, além deste algoritmo ser não paramétrico, selecionou apenas uma variável desagregada como influente na escolha dos destinos intermunicipais.

A Tabela 4.16 sintetiza a influência (positiva ou negativa) de cada variável (agregada ou desagregada) na escolha do destino intermunicipal, a partir da calibração das funções utilidade pelo Algoritmo Genético (Replicação 5).

Tabela 4.16: Influência de cada variável na escolha de destino intermunicipal

	Alagoinhas	Camaçari	Candeias	Catu	Conceição da Feira	Dias D'Ávila	Mata de São João	Pojuca	Salvador	Santo Amaro	Simões Filho
Idade	+	-	+	-	-	-	+	+	+	+	**REF**
Grau de Instrução	+	+	+	-	-	+	+	-	-	-	**REF**
Renda	+	-	+	+	+	+	+	-	-	+	**REF**
N° carros no domicílio	+	-	+	+	-	-	-	-	-	+	**REF**
População da cidade de origem	+	-	-	+	-	-	+	-	+	+	**REF**
PIB da cidade de origem	+	+	+	-	-	+	-	+	-	-	**REF**
PIB da cidade de destino	+	+	+	+	+	-	+	+	+	+	**REF**
Empregos na cidade de destino	-	+	+	+	+	+	-	-	+	+	**REF**
Distância de viagem	+	-	+	-	+	-	+	+	+	+	**REF**
Custo de viagem	+	-	-	+	+	-	-	+	+	+	**REF**
Tempo de viagem	-	+	+	+	-	-	+	-	-	-	**REF**
Frequência semanal de viagem	-	+	+	-	+	-	-	-	-	+	**REF**

5 CONSIDERAÇÕES FINAIS E SUGESTÕES PARA TRABALHOS FUTUROS

Este capítulo final abrange os principais desdobramentos das técnicas aplicadas na análise de distribuição de viagens intermunicipais. Primeiramente, é realizada uma síntese de todo o desenvolvimento deste estudo, de forma que reproduza os principais pontos abordados. Em seguida, são apresentadas as conclusões das aplicações das abordagens, assim como as sugestões para a continuidade deste tema.

5.1 SÍNTESE DO TRABALHO

O objetivo do presente trabalho foi realizar uma comparação entre o uso de algoritmos de Aprendizagem de Máquinas na análise desagregada de viagens intermunicipais e a abordagem tradicional de distribuição de viagens (Modelos Gravitacionais). Além disso, foi realizada ainda a comparação entre a aplicação dos dois algoritmos de Aprendizagem de Máquinas: *Classification And Regression Tree* (CART) e Algoritmos Genéticos (AG). A análise comparativa foi realizada considerando três critérios distintos:

- Distribuição de distâncias de viagens;
- Medidas estatísticas de desempenho;
- Perspectiva qualitativa.

Os algoritmos de Aprendizagem de Máquinas foram aplicados no banco de dados desagregado, obtido a partir de uma Pesquisa Origem-Destino realizada em 2012 pelo CETRAMA (Universidade Federal da Bahia) em onze municípios do estado da Bahia.

Em relação ao método proposto, primeiramente, dois Modelos Gravitacionais foram estimados e, na sequência, os algoritmos de Aprendizagem de Máquinas foram aplicados. Em seguida, as quatro abordagens foram comparadas segundo os critérios mencionados anteriormente.

Para calibração dos modelos gravitacionais, foi considerada uma agregação do banco de dados. A variável dependente foi o *N° de viagens entre os onze pares de cidades da análise* e as variáveis independentes foram a *População da cidade de*

origem, *Empregos da cidade de destino* e *Distância entre cidades de origem e destino*. Para calibração dos modelos, foi utilizado o método de Regressão Linear, *Stepwise* e o método duplamente restringido.

Já para a execução dos algoritmos de Aprendizagem de Máquinas, a variável dependente foi determinada como a *Escolha da cidade de destino*, com 11 categorias representadas pelos municípios, e as variáveis independentes utilizadas se referem às características socioeconômicas dos indivíduos, às características de viagens e às variáveis agregadas dos municípios.

O algoritmo CART, não paramétrico, selecionou para divisões primárias apenas variáveis agregadas relativas aos municípios: *Empregos da cidade de destino* e *PIB da cidade de destino*. A única variável desagregada selecionada pelo modelo foi *N° de carros no domicílio*, selecionada como divisor substituto. Fixando tal variável desagregada como primeiro divisor a partir do nó raiz, é possível verificar a sua influência na escolha de cada um dos onze destinos possíveis. Além disso, para fins de validação, o algoritmo CART apresentou aproximadamente 85% de acertos.

O modelo de otimização desenvolvido para ser executado no Algoritmo Genético foi baseado no conceito de modelos de escolha discreta e função utilidade. Assim, foram propostas funções utilidades referentes a cada um dos municípios. Tais funções eram lineares, baseadas nas variáveis agregadas e desagregadas do banco de dados. Considerando a função objetivo como a representação da soma dos logaritmos das probabilidades das alternativas realmente escolhidas, fazendo-se a maximização de tal função, foi possível a calibração das funções utilidades relativas a oito diferentes testes. O modelo escolhido baseou-se na magnitude e sinais dos parâmetros estimados das funções utilidade. O AG obteve ótimo desempenho, com taxa de acertos relativa à amostra de validação em torno de 95%.

Considerando os critérios utilizados na análise comparativa entre as quatro ferramentas, os melhores resultados observados foram provenientes da calibração do Algoritmo Genético, seguido pela Árvore de Decisão. Os piores resultados segundo a distribuição das distâncias de viagens (histograma), assim como, nos resultados dos testes de distribuição de probabilidade e pela qualidade do ajuste foram obtidos pela abordagem tradicional (Modelo Gravitacional 1 e Modelo Gravitacional 2).

Dessa forma, pode-se comprovar que os algoritmos de Aprendizagem de Máquinas, CART e Algoritmo Genético, apresentaram maior eficiência na análise desagregada de distribuição de viagens. A vantagem encontrada está no uso do Algoritmo Genético (AG) devido à sua capacidade de se obter parâmetros estimados de todas as variáveis, agregadas e desagregadas, que foram selecionadas para o modelo de otimização, enquanto que o algoritmo CART selecionou apenas uma variável desagregada, além das agregadas.

5.2 CONCLUSÕES

O estudo comparativo das técnicas de Aprendizagem de Máquinas e da abordagem tradicional de distribuição de viagens permitiu a comprovação de que a análise desagregada pode ser altamente eficaz na previsão de escolhas de destino, utilizando os algoritmos CART e Algoritmo Genético.

Além disso, os algoritmos de Aprendizagem de Máquinas, aqui utilizados, não partem de suposições matemáticas rigorosas, sendo possível utilizar qualquer tipo de variável como dados de entrada, obtendo-se assim a probabilidade de escolha de cada destino como o principal dado de saída.

Os melhores resultados foram obtidos pelo Algoritmo Genético devido à sua capacidade de determinar os parâmetros estimados de todas as variáveis (agregadas e desagregadas) selecionadas, assim como as elevadas taxas de acertos obtidas do processo de validação dos modelos. Para a replicação 5, escolhida como o de melhor desempenho, a taxa de acertos resultou em 95,8%

É importante ressaltar que a fixação das variáveis independentes pelo algoritmo CART como divisões primárias foi um artifício criado que auxiliou no procedimento para determinação do melhor modelo do Algoritmo Genético, a fim de escolher o modelo condizente com o tipo de influência de cada variável na escolha da cidade de destino.

Finalmente, este trabalho trouxe uma contribuição metodológica eficaz para estimativa de distribuição de viagens intermunicipais, a partir de características dos

indivíduos, das cidades, além dos fatores relacionados às viagens realizadas. Foram apresentadas ferramentas não triviais na área, que são mais flexíveis ao uso, considerando limitações de ferramentas tradicionais de distribuição de viagens.

5.3 SUGESTÕES PARA TRABALHOS FUTUROS

A utilização dos algoritmos de Aprendizagem de Máquinas na previsão da escolha do destino, proposta neste trabalho, apresentou os melhores resultados quando comparadas à análise agregada (Modelos Gravitacionais). Sendo assim, pode ser levada em consideração em estudos futuros em que as características individuais sejam importantes no processo de escolha.

Para fins de continuidade da presente pesquisa, são sugeridas as seguintes análises como tema para futuros trabalhos:

- Calibração do modelo logit multinomial para comparação com o Algoritmo Genético (AG) aqui proposto;
- Uso de algoritmos de Aprendizagem de Máquinas para distribuição de viagens urbanas com maior número de destinos possíveis;
- Investigação da aplicabilidade de tais técnicas para o problema de escolha de rotas;
- Teste da adequabilidade dos algoritmos aqui utilizados no processo de escolhas simultâneas de transportes (modo de viagem, horário de partida, atividade, etc. – abordagem de viagens baseadas em atividade).

6 REFERÊNCIAS BIBLIOGRÁFICAS

Aha, D. W.; Kibler, D.; Albert, M. K. Instance-based learning algorithms. *Machine Learning*, v. 6, p. 37-66, 1991.

Alpaydin, E; *Introduction to Machine Learning*. MIT Press, 2ª edição, 2010. ISBN10 026201243X.

Arentze, T.; Timmermans, H. (2007) Parametric action decision trees: Incorporating continuous attribute variables into rule-based models of discrete choice. *Transportation Research Part B: Methodological*, v. 41, n. 7, p. 772-783. doi:10.1016/j.trb.2007.01.001

Ben-Akiva, M.E.; Lerman, S.R. (1985) *Discrete Choice Analysis: Theory and Application to Travel Demand*. The MIT Press, Cambridge, MA.

Berk, R. A. (2008) *Statistical learning from a regression perspective*. Springer Science & Business Media.

Biggs, D., B. de Ville, Suen, E. (1991). A method of choosing multiway partitions for classification and decision trees. *Journal of Applied Statistics*, 18, 49–62.

Breiman, L.; Friedman, J.H; Olshen, R.A.; Stone, C.J. (1984) *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA.

Bessa Jr, J. E.; Setti, J. R. (2011) Produção de dados de tráfego sintéticos através de algoritmo genético e simulação microscópica, v.18, revistatransportes.org.br

Carvalho, A. C. P. L. F.; Galvão, C. O.; Lacerda, E. G. M.; Diniz, L. S.; Valença, M. J. S.; Ludermir, T. B.; Vieira, V. P. P. B. (1999). Sistemas inteligentes: Aplicações a recursos hídricos e ambientais. Porto Alegre: Editora Universidade/ UFRGS/ ABRH. ISBN 8570255276.

Cascetta, E.; Pagliara, F.; Papola, A. (2007) Alternative Approaches to Trip Distribution Modelling: A Retrospective Review and Suggestions for Combining Different Approaches. *Papers in Regional Science,* v. 86 n.4, 597–620. doi: 10.1111/j.1435-5957.2007.00135.x

Casey, H.J. (1955) Applications to traffic engineering of the law of retail gravitation. *Traffic Quarterly*, v.9, n.1, p. 23–25.

CETRAMA Centro de estudos de transportes e meio ambiente (2013) Relatório parcial – Análise de demanda - Propensão à mudança modal.

Charniak, E. Bayesian networks without tears. *Artificial Inteligence Magazine*, v. 12, p. 50-63, 1991.

Charypar, D. (2008) Efficient Algorithms for the Microsimulation of Travel Behavior in Very Large Scenarios. 151 f. Dissertação (Master of Scienses in Computer Sciense), ETH Zurich.

Charypar, D.; Nagel, K. (2005) Generating complete all-day activity plans with genetic algorithms. Transportation 2005, v. 32, p. 369-397, doi: 10.1007/s11116-004-8287-y

Cho, J. H.; Kurup, P.U. Decision Tree approach for classification and dimensionality reduction of electronic noise data. Sensors and Actuators, B: Chemical, v. 160, p. 542-548, 2011.

Cunha, A. L. et al. (2009) Genetic Algorithm for the Calibration of Vehicle Performance Models of Microscopic Traffic Simulators.

De Grange, L.; Fernández, E.; de Cea, J. (2010) A consolidated model of trip distribution. *Transportation Research Part E: Logistics and Transportation Review*, v. 46, n.1, p. 61–75. doi: 10.1016/j.tre.2009.06.001

Deng, Z.; Ji, M. (2010) Deriving Rules for Trip Purpose Identification from GPS Travel Survey Data and Land Use Data: A Machine Learning Approach. Traffic and Transportation Studies, p. 768-777.

Evans, A.W. (1970) Some Properties of Trip Distribution Methods. Transportation Research, v.4, n.1, p. 19–36. doi: 10.1016/0041-1647(70)90072-9

Evans, S.P.; Kirby, H.R. (1974) A three-dimensional furness procedure for calibrating gravity models. *Transportation Research*, v.8, n.2, p. 105–122. doi: 10.1016/0041-1647(74)90037-9

Feng, Y. et al. Ozone concentration forecast method based on genetic algorithm optimized back propagation neural networks and support vector machine data classification. *Atmospheric environment*, v. 45, p. 1979-1985, 2011.

Fotheringham, A.S. (1983) Some theoretical aspects of destination choice and their relevance to production-constrained gravity models. *Environment and Planning A*, v.15, n.8, p. 1121–1132. doi: 10.1068/a151121

Griffin, T; Huang, Y. (2005) A Decision Tree Based Classification Model to Automate Trip Purpose Derivation. In the Proceedings of the 18th International

Conference on Computer Applications in Industry and Engineering, Honolulu, Hawaii.

Goldschimidt, R.; Passos, E. (2005) Data Mining: Um guia prático. Rio de Janeiro.

Haykin, S. Redes Neurais: Princípios e Prática. Porto Alegre: Bookman, 2001.

Holland, J. (1975) Adaptation in Natural and Artificial Systems. 211 pp. ISBN: 9780262275552.

IBGE - Instituto Brasileiro de Geografia e Estatística (2010) Demographic Census 2010.

Ichikawa, S.M., Pitombo, C.S., Kawamoto, E. (2002) Aplicação de Minerador de dados na obtenção de relações entre padrões de viagens encadeadas e características socioeconômicas. Anais do XVI do Congresso de Pesquisa e Ensino em Transportes, Anpet, Natal (RN), vol2, p175-186.

Jones, I.S.; Nichols, A.J. (1983) The demand for inter-city rail travel in the United Kingdom: some evidence. *Journal of Transport Economics and Policy*, v.17, n.2, p. 133-153.

Junguo, H.; Guomoa, Z.; Xiaojuna, X. Using an improvement back propagation neural network to study spatial distribution of sunshine illumination from sensor network data. Ecological Modelling, v. 266, p. 86-96, 2013.

Kass, G.V. (1980) An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, v.29, p. 119–127. doi: 10.2307/2986296

Kononenko, I; Kukar, M. (2007) *Machine Learning and Data Mining: Introduction to Principles and Algorithms*. Horwood Publishing. Chichester, UK.

Koppelman, F. S.; Wen, C.H. (2000) The paired combinatorial logit model: properties, estimation and application. *Transportation Research Part B: Methodological*, v. 34, n. 2, p. 75-89. doi: 10.1016/S0191-2615(99)00012-0

LaMondia, J.; Snell, T.; Bhat, C.R. (2009) Traveler Behavior and Values Analysis in the Context of Vacation Destination and Travel Mode Choices: A European Union Case Study. *Transportation Research Record: Journal of the Transportation Research Board*, n. 2156, p. 140-149. doi: 10.3141/2156-16

Lu, Y. et al. (2012) *A Machine Learning Approach to Trip Purpose Imputation in GPS-Based Travel Surveys*. Department of Civil and Environmental Engineering, University of Maryland.

Marsland, S. (2009) Machine Learning: An Algorithmic Perspective. CRC Press. Cambridge, UK.

Mikkonen, K.; Luoma, M. (1999) The parameters of the gravity model are changing - how and why? *Journal of Transport Geography*, v.7, n.4, p.277–283. doi: 10.1016/S0966-6923(99)00024-1

Mishra, S.; Wang, Y.; Zhu, X.; Moecke, R.; Mahaparta, S. (2013) Comparison between gravity and destination choice models for trip distribution in Maryland. *Transportation Research Board* 92nd Annual Meeting. n.13-3196, p. 1-22.

Mitchell, Tom M. (1997) *Machine Learning*. McGraw-Hill. ISBN 0070428077.

Mohanty, A. K. et al. Texture-based features for classification of mammograms using decision tree. *Neural Computing and Applications*, v. 23, p. 1011-1017, 2013.

Mozolin, M.; Thill, J.C.; Linn, U.E. (2005) Trip distribution forecasting with multilayer perceptron neural networks: A critical evaluation. *Transportation Research Part B: Methodological*, v. 34, p.53-73. doi: 10.1016/S0191-2615(99)00014-4

Murchland, J.D. (1966) Some Remarks on the Gravity Model of Traffic Distribution, and an Equivalent Maximization Formulation. London School of Economics and Political Science, *Transport Network Theory Unit*, London.

Omrani, H. (2015) Predicting travel mode of individuals by machine learning. 18th Euro Working Group on Transportation, EWGT 2015, p. 840-849.

Ortúzar, J. D.; Willumsen, L. G. (2011) *Modelling Transport*. Wiley, London.

Pitombo, C.S.; Kawamoto, E.; Sousa, A.J. (2011) An exploratory analysis of relationships between socioeconomic, land use, activity participation variables and travel patterns. Transport Policy, v. 18, 347-357. doi: 10.1016/j.tranpol.2010.10.010

Pitombo, C.S.; Kawamoto, E.; Sousa, A.J. (2013) Linking activity participation, socioeconomic characteristics, land use and travel patterns: a comparison of industry and commerce sector workers. *JTL-RELIT*, v.7, p. 59-86. doi: 10.1590/s2238-10312013000300004

Pitombo, C. S. et al. (2017) Comparing decision tree algorithms to estimate intercity trip distribution. *Transportation Research Part C* 77, p. 16-32

Pulugurta, et al. (2013) Use of Artificial Intelligence for Mode Choice Analysis and Comparison with Traditional Multinomial Logit Model, v. 104, p. 583-592. doi: 10.1016/ j. sbspro.2013.11.152

Pitombo, C. S.; Guimarães, H. S. (2016) Uma análise desagregada na modelagem de distribuição de viagens intermunicipais. XXIX Congresso nacional de Pesquisa em Transporte da ANPET, Ouro Preto, 9 a 13 de novembro de 2015.

Quinlan, R. (1983) Learning efficient classification procedures and their application to chess end-games. *Machine Learning: An Artificial Intelligence Approach*, Tioga, Palo Alto, pp. 463-482.

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

Ripley, B.D. (1996) *Pattern Recognition and Neural Networks*. Cambridge University Press.

Rodrigues, M. A. S. (2005) Árvores de Classificação. Monografia. 30 f. Departamento de Matemática. Universidade de Açores.

Russel, S.J.; Norvig, P. Artificial Intelligence: A Modern Approach. 2ª edição. Prentice-Hall, 2002. ISBN10 0137903952

SEI-BA – Superintendência de Estudos Econômicos e Sociais da Bahia (2012). Tabelas e Gráficos CAGED – Variação mensal do emprego formal

Setti, J. R.; Cunha, A. L. (2006) Avaliação do impacto da medida de desempenho no equivalente veicular de caminhões. *Transportation Systems Planning and Operation*. doi: 10.11606/D.18.2007.tde-27112007-094400

Shmueli, D.; Salomon, I.; Shefer, D. (1996). Neural network analysis of travel behavior: Evaluating tools for prediction. *Transportation Research Part C: Emerging Technologies*, v.4, n.3, p.151–166. doi: 10.1016/S0968-090X(96)00007-1

Souza, G. P. K. (2014) Otimização de funções reais multidimensionais utilizando algoritmo genético contínuo. Universidade Federal de Santa Catarina, Departamento de Informática e Estatística.

Strambi, O.; Van de Bilt, K. (1998) Trip Generation Modeling Using CHAID, a Criterion-Based Segmentation Modeling Tool. *Transportation Research*

Record: Journal of the Transportation Research Board, n. 1645, p. 24-31. doi: 10.3141/1645-04

Timofeev, R. (2004) *Classification and regression trees (CART) theory and applications* (Master of Art Thesis) Humboldt University, Berlin. 40p.

Williams, I. (1976) A comparison of some calibration techniques for doubly constrained models with an exponential cost function. *Transportation Research*, v. 10, n. 2, 91-104. doi: 10.1016/0041-1647(76)90045-9

Wilson, A.A. (1967) Statistical Theory of Spatial Distribution Models. *Transportation Research*, v.1, p. 253-269. doi: 10.1016/0041-1647(67)90035-4

Wirasinghe, S.C.; Kumarage, A.S. (1998) An aggregate demand model for intercity passenger travel in Sri Lanka*. Transportation*, v.5, n.1, p. 77-98.

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. doi: 10.3141/1854-06

Yang, C.; Tsai, M.; Chang, C, 2014. Investigating the joint choice behavior of intercity transport mode and high-speed rail cabin with a strategy map. *Journal of Advanced Transportation*. doi: 10.1002/atr.1264

7 APÊNDICE I

TREM REGIONAL DE PASSAGEIROS – UNIVERSIDADE FEDERAL DA BAHIA/ MINISTÉRIO DOS TRANSPORTE
Questionário A – Entrevistas nos Terminais Rodoviários e PGTs – Questionário nº |___|___|___|

BLOCO I: IDENTIFICAÇÃO DA PESQUISA

Data: ____/____/____ Hora:_____ Supervisor:_____

Pesquisador: _____ Trecho:_____

Município: _____Local / Pólo_____

BLOCO II: INFORMAÇÕES SOBRE O ENTREVISTADO

1- Identif. do entrevistado (Perguntar só no final da entrevista)

1- Nome: _____ 2- Telefone:_____

2 -Sexo

1-☐ Masculino 2 -☐ Feminino

3 - Idade

1 -☐ Até 19 anos 2 -☐ de 20 a 29 anos 3 - ☐ de 30 a 39 anos

4 -☐ de 40 a 49 anos 5- ☐ de 50 a 65 anos 6 -☐ acima de 65 anos

4 - Instrução

1 -☐ Sem instrução 2 -☐ 1º Grau 3 -☐ 2º Grau 4 -☐ 3º Grau

5 - Renda

1 -☐ Até 1 S.M 2 -☐ Entre 1 a 3 S.M

3 -☐ Entre 3 a 5 S.M. 4 -☐ Acima de 5 S.M

6 - Ocupação

1 -☐ Comércio 2 - ☐ Indústria 3 - ☐ Serviços 4 -☐ Agricultura 5 - Estudante☐

6 -☐ Aposentado 7 -☐ Outra - Qual? _____

7 - Residência

1 -☐ Própria 2 -☐ Alugada 3 - ☐ Cedida

8 - Carros no Domicílio

1 -☐ Zero 2 -☐ 1 3 -☐ 2 4-☐ Mais de 2

BLOCO III: INFORMAÇÕES SOBRE A VIAGEM DO ENTREVISTADO

1 – Semanalmente você viaja para uma dessas cidades? (Mostrar Mapa)

1 -☐ Sim 2 ☐ Não(Neste caso encerrar entrevista aqui)

VIAGEM I

1 - Origem

1 - Bairro_____ 2 - Distrito _____

3 - Município _____

2 - Destino

1 - Bairro_____ 2 - Distrito _____

3 - Município _____

3 - Frequência Semanal de Viagem

1 -☐ 1 dia 2 -☐ 2 dias 3 -☐ 3 dias 4 -☐ 4 dias 5 - ☐ 5 dias 6 -☐ 6 dias

7 - ☐ Todos os dias 8 - ☐ Outra - Qual? _____

4 - Motivo da Viagem

1- ☐ Trabalho 2 - ☐ Estudo 3 - ☐ Compras 4 - ☐ Lazer

5- ☐ Saúde 6 - ☐ Visita 7 - ☐ Outros

5 - Tempo de Viagem

1 - ☐ Até 30 min. 2 - ☐ Entre 30 a 60 min. 3 - ☐ Acima de 60 min.

6 - Modo da Viagem

1 - ☐ Ônibus 2 - ☐ Carro 3 - ☐ Van/Similar 4 - ☐ A pé

5 - ☐ Bicicleta 6 - ☐ Motocicleta 7- ☐ Outro_____

7 - Custo da Viagem

1 - ☐ Até R$ 5,00 2 - ☐ De R$ 5,00 a R$ 10,00

3 - ☐ De R$ 10,00 a R$ 20,00 4 - ☐ Acima de R$ 20,00

8 - Forma de Pagamento

1 - ☐ Vale Transporte 2 - ☐ Dinheiro 3 - ☐ Outro

VIAGEM II

1 - Origem

1 - Bairro_____ 2 - Distrito _____

3 - Município _____

2 - Destino

1 - Bairro_____ 2 - Distrito _____

3 - Município _____

3 - Frequência Semanal de Viagem

1 -☐ 1 dia 2 - ☐ 2 dias 3 - ☐ 3 dias 4 - ☐ 4 dias 5 ☐ 5 dias

6 - ☐ 6 dias 7 ☐ Todos os dias 8 - ☐ Outra - Qual?

_____

4 - Motivo da Viagem

1- ☐ Trabalho 2 - ☐ Estudo 3 - ☐ Compras 4 - ☐ Lazer

5- ☐ Saúde 6 - ☐ Visita 7 - ☐ Outros

5 - Tempo de Viagem

1 - ☐ Até 30 min. 2 - ☐ Entre 30 a 60 min. 3 - ☐ Acima de 60 min.

6 - Modo da Viagem

1 - ☐ Ônibus 2 - ☐ Carro 3 - ☐ Van/Similar 4 - ☐ A pé 5 - ☐ Bicicleta

6 - ☐ Motocicleta 7- ☐ Outro_____

7 - Custo da Viagem

1 - ☐ Até R$ 5,00 2 - ☐ De R$ 5,00 a R$ 10,00

3 - ☐ De R$ 10,00 a R$ 20,00 4 - ☐ Acima de R$ 20,00

8 - Forma de Pagamento

1 - ☐ Vale Transporte 2 - ☐ Dinheiro 3 - ☐ Outro

VIAGEM III

1 - Origem

1 - Bairro_____ 2 - Distrito _____

3 - Município _____

2 - Destino

1 - Bairro_____ 2 - Distrito _____

3 - Município _____

3 - Frequência Semanal de Viagem

1 - ☐ 1 dia 2 - ☐ 2 dias 3 - ☐ 3 dias 4 - ☐ 4 dias 5 - ☐ 5 dias

6 - ☐ 6 dias 7 - ☐ Todos os dias 8 - ☐ Outra - Qual?

_____

4 - Motivo da Viagem

1- ☐ Trabalho 2 - ☐ Estudo 3 - ☐ Compras 4 - ☐ Lazer

5- ☐ Saúde 6 - ☐ Visita 7 - ☐ Outros

5 - Tempo de Viagem

1 - ☐ Até 30 min. 2 - ☐ Entre 30 a 60 min. 3 - ☐ Acima de 60 min.

6 - Modo da Viagem

1 - ☐ Ônibus 2 - ☐ Carro 3 - ☐ Van/Similar 4 - ☐ A pé 5 ☐ Bicicleta

6 - ☐ Motocicleta 7- ☐ Outro_____

7 - Custo da Viagem

1 - ☐ Até R$ 5,00 2 - ☐ De R$ 5,00 a R$ 10,00

3 - ☐ De R$ 10,00 a R$ 20,00 4 - ☐ Acima de R$ 20,00

8 - Forma de Pagamento

1 - ☐ Vale Transporte 2 - ☐ Dinheiro 3 - ☐ Outro

VIAGEM IV

1 - Origem

1 - Bairro_____ 2 - Distrito _____

3 - Município _____

2 - Destino

1 - Bairro_____ 2 - Distrito _____

3 - Município _____

3 - Frequência Semanal de Viagem

1 - ☐ 1 dia 2 - ☐ 2 dias 3 - ☐ 3 dias 4 - ☐ 4 dias 5 - ☐ 5 dias 6 - ☐ 6 dias

7 - ☐ Todos os dias 8 - ☐ Outra - Qual? _____

4 - Motivo da Viagem

1- ☐ Trabalho 2 - ☐ Estudo 3 - ☐ Compras 4 - ☐ Lazer

5- ☐ Saúde 6 - ☐ Visita 7 - ☐ Outros

5 - Tempo de Viagem

1 - ☐ Até 30 min. 2 - ☐ Entre 30 a 60 min. 3 - ☐ Acima de 60 min.

6 - Modo da Viagem

1 - ☐ Ônibus 2 - ☐ Carro 3 - ☐ Van/Similar 4 - ☐ A pé 5 - ☐ Bicicleta

6 - ☐ Motocicleta 7- ☐ Outro_____

7 - Custo da Viagem

1 - ☐ Até R$ 5,00 2 - ☐ De R$ 5,00 a R$ 10,00

3 - ☐ De R$ 10,00 a R$ 20,00 4 - ☐ Acima de R$ 20,00

8 - Forma de Pagamento

1 - ☐ Vale Transporte 2 - ☐ Dinheiro 3 - ☐ Outro

Comentários:

_____

_____

_____

_____

BLOCO IV - TRANSFERÊNCIA MODAL

1 - Você mudaria seu modo de transporte atual para fazer esta mesma viagem de trem?

1 - ☐ Sim 2 - ☐ Não (Neste caso encerrar entrevista aqui).

2 - Quanto você gasta atualmente na viagem?
R$_____

3 - Quanto tempo você gasta na viagem?
_____min.

4 - Quanto você estaria disposto a pagar para viajar no trem, considerando que o tempo total de viagem seria o mesmo da sua viagem atual?

1 - ☐ Até R$ 5,00 2 - ☐ De R$ 5,00 a R$ 10,00 3 - ☐ De R$ 10,00 a R$15,00 4 - ☐ De R$ 15,00 a R$ 20,00 5 - ☐ Acima de R$ 20,00

5 - E no caso da viagem de trem ser um pouco mais rápida que sua viagem atual, quanto você estaria disposto a pagar pela viagem?

1 - ☐ Até R$ 5,00 2 - ☐ De R$ 5,00 a R$ 10,00 3 - ☐ De R$ 10,00 a R$15,00 4 - ☐ De R$ 15,00 a R$ 20,00 5 - ☐ Acima de R$ 20,00

6 - Caso a viagem de trem seja um pouco mais demorada que sua viagem atual, você aceitaria viajar de trem?

1- ☐ Sim Obs.: Caso a resposta seja "Não", encerrar a entrevista aqui.

7 - Quanto estaria disposto a pagar neste caso?

1 - ☐ Até R$ 5,00 2 - ☐ De R$ 5,00 a R$ 10,00 3 - ☐ De R$ 10,00 a R$15,00 4 - ☐ De R$ 15,00 a R$ 20,00 5 - ☐ Acima de R$ 20,00

8 - Quanto tempo a mais de viagem você poderia aceitar para viajar de trem?

_____min.

9 - O que você considera que deva existir em uma Estação de Trem? Escolha três opções abaixo, que você considera como as mais importantes?

1 - ☐ Segurança 2 - ☐ Sanitário 3 - ☐ Acessibilidade (ex. elevador, escada rolante) 4 - ☐ Bancos nas plataformas 5 - ☐ Estacionamento para automóvel 6 - ☐ Bicicletário com cobertura 7 ☐ Outros:_____

Comentários:

_____

_____

8 APÊNDICE II

Município	1	2	3	4	5	6	7	8	9	19	11
	Alagoinhas	Camaçari	Candeias	Catu	Conceição da Feira	Dias D'Ávila	Mata de S	Pojuca	Salvador	Santo Amaro	Simões Filho

MODELO 1

Modelo 1	1	2	3	4	5	6	7	8	9	10	11
Erro	654,138	654,138	654,138	654,138	654,138	654,138	654,138	654,138	654,138	654,138	REF
Constante	-4,077	-4,230	-8,785	1,929	-6,008	0,523	-2,642	-4,017	-10,837	-0,643	REF
Idade	-0,263	-0,059	-0,383	0,020	-0,149	0,072	-0,003	-0,262	-1,010	-0,017	REF
Grau de Instrução	-0,423	0,014	-0,587	-0,019	-0,374	0,376	0,164	-0,341	-1,621	-0,294	REF
Renda	0,785	0,100	0,123	0,765	0,988	0,255	0,639	0,126	0,610	1,039	REF
Carros no Domicílio	-0,459	-0,016	-0,434	-0,346	-0,522	-0,280	-0,189	-0,259	-0,967	-0,240	REF
Frequencia Semanal de Viagem	0,168	0,149	0,068	0,263	0,278	0,078	0,205	0,032	0,306	0,299	REF
Tempo de Viagem	-0,840	0,080	-0,289	-0,588	0,119	0,105	0,033	-0,062	-1,529	-0,210	REF
Custo de Viagem	0,505	-0,007	-0,040	0,530	0,539	-0,023	0,281	0,278	1,377	0,655	REF
População da cidade de origem	-6,297	-2,195	-2,538	-3,211	-4,096	-2,182	2,013	-5,098	-8,309	1,740	REF
PIB da cidade de origem	-1,764	-1,382	-3,303	0,556	3,352	0,046	2,306	-3,665	-1,721	6,227	REF
PIB da cidade de destino	-10,754	12,420	21,369	-15,931	-18,937	-1,481	-23,477	8,181	-15,238	-25,510	REF
Empregos da cidade de destino	36,189	-28,707	3,637	-3,228	-15,928	-19,793	16,180	10,034	61,730	-14,986	REF
Distância de viagem	16,729	4,613	6,243	9,983	22,820	8,250	-0,357	13,248	13,858	7,738	REF

MODELO 2

Modelo 2	1	2	3	4	5	6	7	8	9	10	11
Erro	599,515	599,515	599,515	599,515	599,515	599,515	599,515	599,515	599,515	599,515	REF
Constante	-3,994	-2,691	-8,272	3,710	-3,288	2,537	-0,582	-2,946	-12,475	-0,814	REF
Idade	-0,307	-0,041	-0,290	-0,047	-0,154	0,087	-0,017	-0,217	-1,170	-0,033	REF
Grau de Instrução	-0,015	0,458	0,013	0,326	0,161	0,670	0,560	-0,005	-0,986	0,277	REF
Renda	0,054	0,069	-0,041	0,150	0,250	0,166	0,030	0,081	0,161	0,366	REF
Carros no Domicílio	-0,374	-0,515	-0,999	-0,232	-0,400	-0,657	-0,256	-0,696	-0,575	-0,174	REF
Frequencia Semanal de Viagem	0,208	0,133	0,071	0,262	0,295	0,082	0,199	0,028	-0,032	0,300	REF
Tempo de Viagem	0,044	0,331	-0,160	0,280	0,414	0,452	0,786	0,387	-1,124	0,879	REF
Custo de Viagem	0,293	-0,076	-0,072	0,294	0,399	-0,062	0,092	0,161	0,641	0,434	REF
População da cidade de origem	-7,124	-3,173	-3,344	-4,634	-5,143	-3,073	0,042	-5,795	-7,298	1,327	REF
PIB da cidade de origem	-3,257	-3,541	-5,252	-1,465	1,083	-1,393	0,144	-5,494	0,268	5,151	REF
PIB da cidade de destino	-12,998	11,468	21,343	-20,718	-24,763	-3,358	-21,404	7,152	-10,564	-31,171	REF
Empregos da cidade de destino	41,166	-30,806	0,571	-4,941	-18,623	-25,888	10,539	7,397	69,251	-14,396	REF
Distância de viagem	14,326	3,306	5,153	8,628	21,215	7,012	-0,171	11,489	11,338	4,937	REF

MODELO 3

Modelo 3	1	2	3	4	5	6	7	8	9	10	11
Erro	666,851	666,851	666,851	666,851	666,851	666,851	666,851	666,851	666,851	666,851	REF
Constante	-3,102	-2,505	-7,259	3,165	-4,647	1,963	-1,835	-2,136	-11,058	-0,062	REF
Idade	0,126	0,194	-0,116	0,434	0,329	0,372	0,388	-0,050	-0,505	0,397	REF
Grau de Instrução	-0,154	0,032	-0,458	0,345	-0,324	0,405	0,337	-0,324	-1,127	0,055	REF
Renda	0,069	-0,069	-0,109	0,046	0,431	-0,017	0,116	-0,093	-0,011	0,281	REF
Carros no Domicílio	-0,281	-0,102	-0,668	-0,101	-0,434	-0,228	-0,121	-0,201	-0,413	0,155	REF
Frequencia Semanal de Viagem	0,123	0,100	0,011	0,212	0,189	0,044	0,178	-0,024	-0,079	0,232	REF
Tempo de Viagem	-0,135	0,192	-0,329	0,086	-0,130	0,386	0,892	0,377	-0,313	0,478	REF
Custo de Viagem	-0,615	-0,418	-0,452	-0,583	-0,447	-0,379	-0,658	-0,214	-0,864	-0,392	REF
População da cidade de origem	-5,432	-1,989	-2,271	-2,775	-2,200	-2,285	0,876	-4,801	-6,749	2,009	REF
PIB da cidade de origem	-2,938	-2,431	-4,102	-0,996	6,003	-0,541	1,128	-4,581	-6,038	4,939	REF
PIB da cidade de destino	-13,718	10,894	20,317	-29,563	-12,423	-3,534	-21,303	5,786	-12,742	-26,887	REF
Empregos da cidade de destino	37,951	-27,570	-0,103	0,274	-20,156	-23,707	12,569	6,291	65,517	-14,591	REF
Distância de viagem	16,038	4,137	6,613	9,022	23,527	8,569	1,869	12,560	12,737	7,755	REF

MODELO 4

Modelo 4	1	2	3	4	5	6	7	8	9	10	11
Erro	707,775	707,775	707,775	707,775	707,775	707,775	707,775	707,775	707,775	707,775	REF
Constante	-1,816	-0,754	-6,208	2,548	-1,493	2,065	-0,123	-4,567	-9,298	-1,126	REF
Idade	0,146	0,204	-0,076	0,478	0,193	0,443	0,395	0,076	-0,304	0,434	REF
Grau de Instrução	-0,779	-0,367	-0,891	-0,299	-1,039	0,352	-0,271	-0,343	-1,558	-0,563	REF
Renda	-0,061	-0,107	-0,144	-0,042	0,331	0,000	-0,010	-0,168	-0,635	0,206	REF
Carros no Domicílio	-0,239	-0,625	-1,242	-0,109	-0,538	-0,733	-0,115	-0,534	-0,465	0,024	REF
Frequencia Semanal de Viagem	0,149	0,086	0,019	0,258	0,233	0,045	0,173	0,027	0,000	0,253	REF
Tempo de Viagem	-0,265	-0,353	-0,853	0,045	0,180	0,018	0,564	0,145	-0,522	0,528	REF
Custo de Viagem	0,910	0,464	0,527	0,958	1,000	0,353	0,708	0,638	0,547	1,088	REF
População da cidade de origem	-6,785	-2,175	-2,075	-3,531	-3,662	-2,330	0,161	-5,127	-8,118	1,947	REF
PIB da cidade de origem	-2,837	-1,784	-3,201	-0,561	3,111	-0,182	0,724	-3,847	-4,765	6,587	REF
PIB da cidade de destino	-13,434	10,183	20,417	-23,747	-21,778	-3,844	-22,733	7,547	-13,070	-24,918	REF
Empregos da cidade de destino	33,997	-25,498	-1,593	1,148	-18,315	-22,368	11,922	10,581	63,214	-15,247	REF
Distância de viagem	13,251	3,169	4,152	6,018	18,319	6,833	-1,146	11,392	9,699	4,929	REF

MODELO 6

Modelo 6	1	2	3	4	5	6	7	8	9	10	11
Erro	656,971	656,971	656,971	656,971	656,971	656,971	656,971	656,971	656,971	656,971	REF
Constante	-2,450	-1,142	-7,240	2,875	-2,194	2,107	0,010	-3,481	-11,963	0,234	REF
Idade	-0,006	-0,060	-0,340	0,289	0,094	0,167	0,242	-0,115	-0,260	0,270	REF
Grau de Instrução	-0,266	-0,269	-0,703	0,087	-0,404	0,177	0,215	-0,342	-0,566	-0,156	REF
Renda	-0,235	-0,179	-0,161	-0,242	-0,050	-0,085	-0,163	-0,239	-0,525	0,027	REF
Carros no Domicílio	-0,835	-0,212	-0,639	-0,702	-1,107	-0,492	-0,732	-0,530	-0,977	-0,608	REF
Frequencia Semanal de Viagem	-0,217	-0,063	-0,138	-0,145	-0,178	-0,099	-0,169	-0,108	-0,483	-0,142	REF
Tempo de Viagem	-0,022	0,432	-0,062	0,104	0,306	0,426	0,932	0,498	-0,153	0,512	REF
Custo de Viagem	1,141	0,372	0,368	1,310	1,254	0,327	0,895	0,578	0,864	1,420	REF
População da cidade de origem	-6,549	-2,244	-2,236	-3,165	-3,141	-2,237	0,743	-5,346	-9,631	1,500	REF
PIB da cidade de origem	-2,799	-2,432	-3,761	-0,335	3,730	-0,356	1,522	-4,242	-3,676	5,197	REF
PIB da cidade de destino	-10,884	10,964	20,981	-22,367	-19,028	-1,819	-20,251	7,427	-8,349	-26,048	REF
Empregos da cidade de destino	33,101	-33,308	-0,511	0,012	-14,676	-21,205	5,064	9,587	58,280	-13,576	REF
Distância de viagem	13,180	1,698	3,296	5,613	17,268	6,087	-1,895	10,956	10,895	3,938	REF

MODELO 7

Modelo 7	1	2	3	4	5	6	7	8	9	10	11
Erro	679,034	679,034	679,034	679,034	679,034	679,034	679,034	679,034	679,034	679,034	REF
Constante	-2,211	-1,155	-7,011	4,152	-3,508	3,379	-1,405	-0,004	-8,052	0,170	REF
Idade	-0,373	-0,236	-0,495	-0,128	-0,174	-0,046	-0,025	-0,383	-0,992	-0,143	REF
Grau de Instrução	-0,333	-0,014	-0,516	-0,075	-0,145	0,267	0,333	-0,475	-1,632	-0,182	REF
Renda	-0,202	-0,244	-0,293	-0,157	-0,102	-0,063	-0,181	-0,166	-0,694	0,113	REF
Carros no Domicílio	-0,306	-0,159	-0,604	-0,138	-0,513	-0,402	-0,266	-0,516	-0,618	-0,143	REF
Frequencia Semanal de Viagem	0,007	-0,006	-0,047	0,082	0,073	-0,028	0,042	-0,036	-0,256	0,105	REF
Tempo de Viagem	-0,186	0,124	-0,270	-0,038	-0,236	0,226	0,512	0,114	0,276	0,724	REF
Custo de Viagem	0,727	0,055	-0,017	0,697	0,979	0,056	0,493	0,305	0,475	0,892	REF
População da cidade de origem	-6,414	-2,025	-2,323	-3,122	-3,001	-2,188	1,115	-5,058	-11,569	1,575	REF
PIB da cidade de origem	-2,195	-2,037	-3,639	-0,233	5,380	-0,146	2,055	-4,186	-4,613	5,578	REF
PIB da cidade de destino	-11,685	11,456	21,620	-19,463	-16,548	-2,712	-19,838	5,455	-4,721	-23,508	REF
Empregos da cidade de destino	36,234	-23,484	0,894	-0,288	-19,771	-23,205	14,446	5,276	62,263	-14,371	REF
Distância de viagem	13,656	2,386	4,839	6,687	20,382	6,711	-0,913	11,432	9,114	4,835	REF

MODELO 8

Modelo 8	1	2	3	4	5	6	7	8	9	10	11
Erro	670,080	670,080	670,080	670,080	670,080	670,080	670,080	670,080	670,080	670,080	REF
Constante	-2,910	-0,431	-6,226	3,135	-5,123	3,768	0,059	-4,186	-10,225	0,160	REF
Idade	-0,254	-0,379	-0,644	0,074	-0,094	-0,151	0,013	-0,391	-1,337	0,036	REF
Grau de Instrução	-0,141	-0,196	-0,692	0,358	-0,128	0,147	0,389	-0,368	-1,659	0,192	REF
Renda	-0,478	-0,294	-0,288	-0,511	-0,263	-0,143	-0,496	-0,224	-0,519	-0,287	REF
Carros no Domicílio	-0,629	-0,614	-1,156	-0,408	-0,869	-0,623	-0,347	-0,674	-0,447	-0,385	REF
Frequencia Semanal de Viagem	0,137	0,012	-0,065	0,212	0,210	-0,039	0,140	-0,060	0,089	0,218	REF
Tempo de Viagem	-0,348	0,034	-0,531	-0,208	-0,065	0,116	0,333	0,301	-0,749	0,216	REF
Custo de Viagem	0,614	0,350	0,332	0,665	0,871	0,325	0,448	0,542	0,737	0,834	REF
População da cidade de origem	-5,499	-1,614	-1,951	-2,682	-3,059	-1,501	1,915	-4,740	-7,684	2,366	REF
PIB da cidade de origem	-1,773	-1,287	-3,102	0,480	4,741	0,726	2,052	-3,516	-3,112	6,139	REF
PIB da cidade de destino	-9,816	11,482	22,140	-21,573	-18,171	-2,967	-22,528	9,374	-6,346	-24,264	REF
Empregos da cidade de destino	36,562	-25,204	-0,353	-0,991	-11,033	-22,923	10,155	11,704	65,595	-13,518	REF
Distância de viagem	13,894	2,255	4,848	7,070	20,745	5,989	-1,455	11,138	11,070	4,674	REF