TRANSPORTES
www.revistatransportes.org.br
anpet
associação nacional de pesquisa e ensino em transportes
www.anpet.org.br

Análise de desempenho de algoritmos de aprendizagem de máquinas para análise desagregada de viagens intermunicipais

Andreza Dornelas de Souza Roma[1], Cira Souza Pitombo[2], Henrique Stramandinoli Guimarães[3], Luis Henrique Magalhães Costa[4]

[1]Departamento de Engenharia de Transportes, USP, Brasil, andrezadornelas@gmail.com
[2]Departamento de Engenharia de Transportes, USP, Brasil, cirapitombo@usp.br
[3]Departamento de Engenharia de Transportes, USP, Brasil, hstr.guimaraes@gmail.com
[4]Universidade Estadual Vale do Acarau, Brasil, lhenriquemc.uva@gmail.com

Recebido:
9 de março de 2018
Aceito para publicação:
24 de setembro de 2018
Publicado:
4 de novembro de 2018
Editor de área:
Helena Beatriz Cybis

Palavras-chaves:
Distribuição de viagens;
Algoritmos Genéticos;
Árvore de Decisão;
Modelos Gravitacionais.

Keywords:
Trip Distribution;
Genetic Algorithms;
Decision Tree;
Gravitational Model.

DOI:10.14295/transportes.v26i3.1614

RESUMO
Este trabalho propõe uma análise desagregada de escolhas de destinos para viagens intermunicipais, por meio da aplicação de algoritmos de Aprendizagem de Máquinas - AM (*Classification And Regression Tree* - CART e Algoritmos Genéticos - AG). Foi utilizada uma Pesquisa OD, realizada pelo Centro de Estudos de Transportes e Meio Ambiente (UFBA), em 2012/2013 em onze municípios do estado da Bahia. Foi realizada a calibração de um Modelo *Logit Multinomial* a partir do algoritmo AG, trazendo a vantagem de associação das escolhas dos destinos a valores de coeficientes estimados das funções utilidade aleatórias, sem os problemas relativos à calibração dos modelos *logit* tradicionais, tais como erros identicamente distribuídos, seguindo a distribuição de *Gumbel*. O desempenho de cada algoritmo de AM foi comparado à abordagem tradicional (modelo gravitacional). Os resultados evidenciaram que os algoritmos de AM apresentaram melhores previsões para a escolha de destinos, sendo que o AG apresentou vantagens na obtenção dos parâmetros associados às variáveis independentes. A principal conclusão é que tais algoritmos podem ser aplicados na modelagem de distribuição de viagens, incorporando o efeito das variáveis desagregadas, sem suposições matemáticas rigorosas contidas no ajuste de modelos tradicionais desagregados.

ABSTRACT
This paper proposes a disaggregated analysis of intercity destination choices, through the application of Machine Learning (ML) algorithms (Classification And Regression Tree - CART and Genetic Algorithms - GA). An Origin-Destination Survey was carried out by the Center of Transportation and Environmental Studies (UFBA) in 2012/2013 in eleven municipalities in the state of Bahia, Brazil. It was carried out a calibration of a *Multinomial Logit Model* with GA algorithm, bringing the advantage of association of the destination choices to values of estimated coefficients of the random utility functions, without the problems related to the calibration of the traditional *logit* models, such as Irrelevant Alternatives (IIA) assumption. The performance of each ML algorithm was compared to a traditional approach (Gravitational Model). The results showed that the ML algorithms presented better predictions for destination choices, and GA presented advantages in obtaining the estimated parameters related to the covariates. The main conclusion is that such algorithms can be applied in trip distribution step, incorporating the effect of the disaggregated variables, without rigorous assumptions of the traditional disaggregated models.

1. INTRODUCTION

A análise delineada neste artigo 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) – ou variáveis sociodemográficas agregadas, além de variáveis de impedância de viagem entre cada par de origem e destino (De Grange *et al*., 2010; Wilson, 1967).

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 Irrelevantes).

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), seguindo uma distribuição de Gumbel. Tais restrições não fazem 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 não tem limitações importantes, tais como dados multicolineares, suposições de distribuição populacionais ou IIA. A Figura 1 ilustra a justificativa para o uso de algoritmos de AM na etapa de distribuição de viagens.

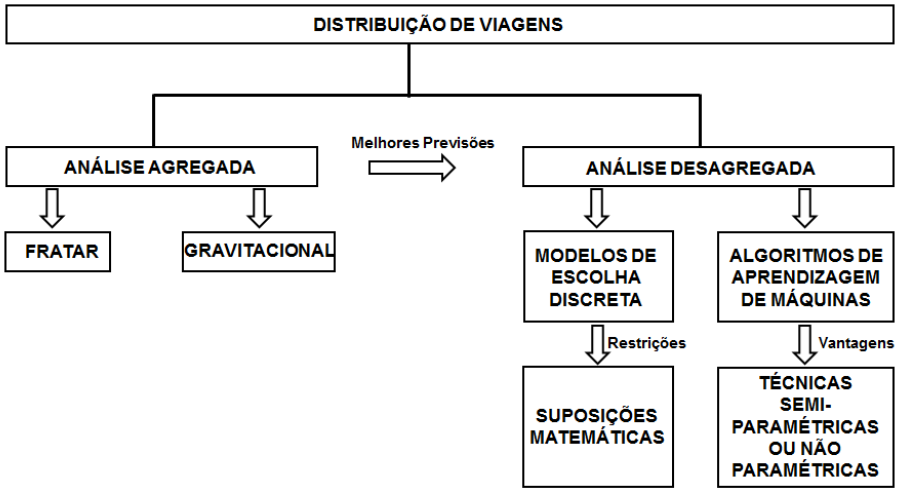

Figura 1. Aplicação de algoritmos de Aprendizagem de Máquinas em problemas de escolha de destinos intermunicipais. Adaptado de Pitombo *et al*., 2017; De Souza, 2017.

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).

Dessa forma, a literatura recente apresenta a aplicação de algoritmos de Aprendizagem de Máquinas a problemas de demanda por transportes (Pulugurta *et al*., 2013; Pitombo *et al*., 2011; Pitombo *et al*., 2013; Ichikawa *et al*., 2002, Omrani, 2015). No entanto, a aplicação de tais técnicas ainda é rara na estimativa de escolhas de destinos intermunicipais em um banco de dados desagregados (Yang e*t al*., 2014; LaMondia *et al*., 2009; Pitombo *et al*., 2017). 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).

A principal lacuna, e consequente contribuição deste trabalho, está focada na apresentação de tais algoritmos de AM para o caso específico de distribuição de viagens intermunicipais. Assim, o objetivo principal deste trabalho é propor uma análise desagregada de escolha de destinos intermunicipais por meio da aplicação da técnica não-paramétrica – CART (*Classification And Regression Tree*), e da técnica semi-paramétrica (Algoritmos Genéticos - AGs). Além disso, o presente trabalho propõe-se a identificar as variáveis mais importantes, agregadas e/ou desagregadas, na escolha de destinos intermunicipais, além de verificar o desempenho de tais algoritmos na estimativa de distribuição de viagens.

Este artigo é formado por cinco seções, incluindo esta introdução. A Seção 2 descreve e conceitua os algoritmos aplicados no trabalho. A Seção 3 apresenta os materiais utilizados (dados e aplicativos) bem como a sequência metodológica proposta. A Seção 4 traz os resultados relativos a cada algoritmo, além da comparação de cada um deles e a abordagem tradicional (modelo gravitacional). Finalmente, a Seção 5 apresenta as principais conclusões obtidas.

2. UMA BREVE DESCRIÇÃO DOS ALGORITMOS DE AM UTILIZADOS

2.1. Árvore de Decisão (AD)

Árvore de Decisão representa um conjunto de técnicas não-paramétricas que têm objetivo de prever (para variáveis dependentes contínuas ou discretas) ou classificar dados (para variáveis dependentes categóricas). Tais técnicas geram uma estrutura representativa de uma sequência de decisões por meio das quais ocorrem sucessivas divisões em um conjunto de dados inicial (nó raiz) até que o mesmo seja representado por diversas classes dentro dos nós filhos gerados (Breiman *et al*., 1984). Quando nenhuma outra subdivisão dos dados é possível, os subconjuntos finais são denominados nós terminais ou folhas.

A aplicação da AD é realizada levando em conta três elementos principais: um conjunto de questões que delimita a divisão dos dados, um critério para estabelecer a melhor divisão na obtenção de nós filhos e uma regra de parada para as subdivisões (*stop-splitting rule*). Os principais algoritmos de Árvores de Decisão são C4.5 (Quinlan, 1983), CHAID (Kass, 1980) e CART (Breiman *et al*., 1984). Neste trabalho optou-se pela aplicação do algoritmo 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.

Para problemas de classificação (Árvore de classificação – variável dependente categórica), como no caso do presente trabalho, 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* (1,2,3,i,...,n) categorias, 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 da variável dependente em um dado nó 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*). O cálculo mais comum da heterogeneidade dos dados do algoritmo CART é realizado pelo Índice Gini, dado pela Equação 1.

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

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*.

Neste trabalho, a aplicação do algoritmo CART ocorreu como um problema de classificação, já que a variável dependente é categórica, com 11 categorias associadas (municípios). As variáveis explicativas, provenientes do banco de dados são descritas na Seção 3.

2.2. Algoritmos Genéticos

Algoritmos Genéticos (AGs) são um conjunto de algoritmos que consiste em métodos de otimização e busca inspirados em mecanismos de evolução de populações de seres vivos (Carvalho *et al*., 1999). Os AGs são uma aproximação computacional da maneira como a evolução realiza a busca, alterando o gene e, consequentemente, a forma dos indivíduos. 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.

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, da qual se utiliza 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 (Carvalho *et al*., 1999).

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 (Carvalho *et al*., 1999).

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. 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.

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.

Para o procedimento de otimização, aplicável no presente trabalho, cada município (destino) foi associado a uma função utilidade aleatória linear e as probabilidades associadas aos destinos foram funções de tais utilidades. A otimização buscava maximizar a função de aptidão, a qual é composta pelas probabilidades de escolha das alternativas (municípios) corretas. Ao final do processo de otimização, os parâmetros das funções utilidades foram calibrados com base em tais pressupostos. Para a escolha da melhor solução, entre os ótimos locais, foi proposto um procedimento alternativo (AG seguido de CART), descrito em seguida.

3. MATERIAIS E MÉTODO

3.1. Banco de Dados

O estado da Bahia é um dos 26 estados do Brasil, localizado na região nordeste e contém 417 municípios, incluindo Salvador, a capital. A análise deste artigo 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 2 ilustra a área e destaca os 11 municípios.

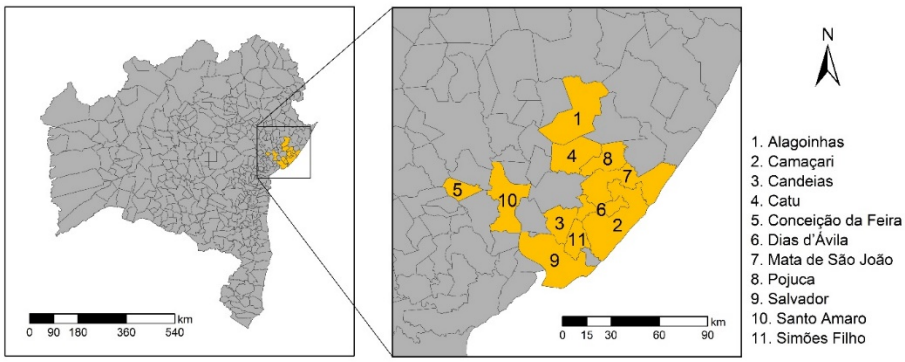

Figura 2. Estado da Bahia e os onze municípios estudados (Pitombo *et al*., 2017).

3.2. Amostra

A amostra é composta por indivíduos maiores de quinze anos, usuários dos ônibus nos terminais rodoviários residentes nos municípios que abrangem a pesquisa e por motoristas e passageiros de automóveis nas rodovias intermunicipais ao longo dos trechos em estudo. As informações foram coletadas por meio de um questionário estruturado, elaborado pelos pesquisadores do Centro de Estudos de Transporte e Meio Ambiente (Universidade Federal da Bahia).

O plano amostral escolhido foi estratificado e os estratos foram constituídos por cada município que compõem a pesquisa (11 estratos/municípios). Selecionando uma amostra aleatória simples sem reposição de indivíduos em cada um dos estratos de forma independente. A alocação da amostra em cada estrato foi proporcional ao tamanho dos estratos, cuja medida de tamanho foi o número de pessoas maiores de catorze anos residentes nos municípios, conforme os dados disponíveis no Censo Demográfico de 2010. Neste trabalho foi arbitrado um número mínimo de cem indivíduos para o tamanho amostral de cada estrato. A amostra inicial consistiu de 3.300 indivíduos.

Para tratamento da amostra foram utilizadas variáveis socioeconômicas, além de informações relativas às viagens intermunicipais (motivo, modo de transporte, frequência da viagem, etc.) – *variáveis categóricas ordinais*. Além disso, foram adicionados ao banco de dados, variáveis agregadas dos municípios de origem e destino, provenientes do último Censo Demográfico e variáveis de tempo e distância de viagem (*variáveis contínuas normalizadas*). Associadas a essas informações estão as escolhas de destino intermunicipais (variável dependente). A variável dependente de escolhas de destinos refere-se aos onze municípios (Figura 2). O tamanho final da amostra resultou em 3.229 indivíduos, sendo que, a amostra de calibração com 80% e a de validação com 20% dos dados, continham 2.584 e 645 indivíduos, respectivamente. Ressalta-se que a amostra de validação (20%) foi utilizada para comparação entre as abordagens. As tabelas, em seguida, apresentam a análise exploratória dos dados relativos a amostra final completa (3.229). A Tabela 1 apresenta a frequência de ocorrência das categorias da variável dependente. Já a Tabela 2 mostra as frequências das variáveis independentes categóricas. Finalmente, a Tabela 3 traz as principais medidas descritivas das variáveis numéricas.

Tabela 1: 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 2: 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**	1.01	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

Tabela 2: Frequência das categorias das variáveis categóricas (N=3.229) (continuação)

Variáveis	Variáveis Categóricas	Quantidade na Amostra	Percentagem (%)
Residência	estudante	276	8,5
	aposentado	247	7,6
	outra ocupação	550	17,0
	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
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

*BRL para USD sobre a taxa de câmbio media em 2012.

**SMB: Salário Mínimo Brasileiro em 2012 (aproximadamente 319 US$).

Tabela 3: Medidas descritivas das variáveis quantitativas (N=3.229)

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

*Saldo Emp Norm = Saldo de Empregos Normalizado = saldo entre demissões e admissões, em 2012 (SEI-BA, 2012).

3.3. Modelo de otimização e procedimento auxiliar (AG seguido de CART)

Para a análise desagregada de distribuição de viagens, realizada pelo Algoritmo Genético, foi efetuada a execução de um procedimento de otimização baseada na modelagem desagregada dos 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. A utilidade de cada alternativa é descrita por meio de uma função matemática, definida pela combinação das variáveis que carac-

terizam o indivíduo e as alternativas (Ben-Akiva e Lerman 1985). Neste trabalho, as funções utilidade possuem variáveis relativas aos indivíduos, domicílios, municípios, além das distâncias de viagens (Equação 2). Vale ressaltar que a função utilidade de um dos municípios foi fixada com o valor zero no intuito de redução de quantidade de parâmetros a serem estimados. O valor fixo de uma função utilidade não restringe a análise, considerando que, na modelagem, o fator de interesse seria as diferenças entre as utilidades das alternativas.

$$U_i = const_i + \alpha_i . A + \beta_i . B + \gamma_i . C + \cdots \quad (2)$$

Para o caso deste artigo, U_i é a função utilidade da alternativa i ($destino\ intermunicipal$); $const_i$ é a constante, relativa à alternativa i que engloba todas as variáveis que não estão inclusas no conjunto de dados; A, B e C são as variáveis relacionadas ao indivíduo (desagregadas) ou ao município (agregadas); e α_i, β_i e $\gamma_i$são os parâmetros a serem estimados das variáveis referentes ao indivíduo ou ao município.

Para cada município será gerada uma equação da função utilidade (Equação 2). Foram consideradas doze variáveis relacionadas ao indivíduo e ao município (*A,B,C,..*), resultando em 13 parâmetros a serem estimados ($const_i$, α_i, β_i e γ_i, ...). O vetor solução contem esses parâmetros para todos os municípios. Como são considerados 10 municípios com funções utilidades (já que um tem o valor dessa função igual a zero), o vetor solução (cromossomo) terá 130 elementos (genes). A Figura 3, a seguir, ilustra a representação de um cromossomo.

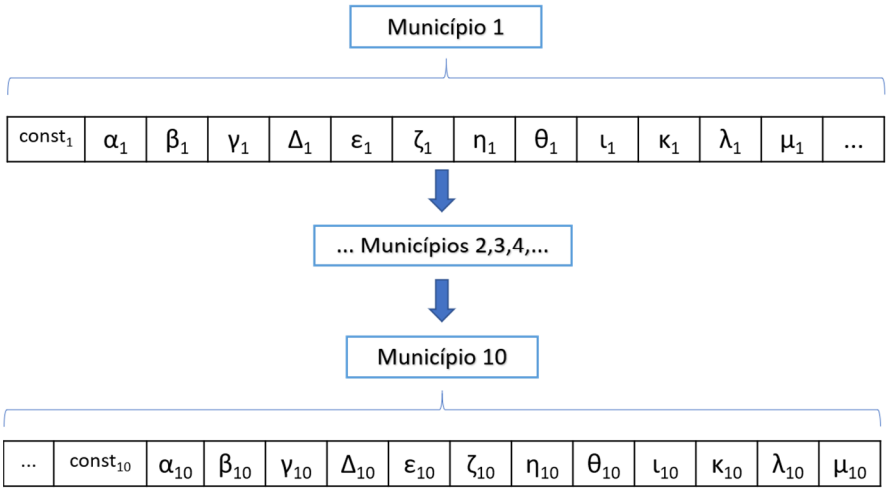

Figura 3. Representação de um cromossomo

A partir da obtenção da utilidade para cada alternativa, é possível calcular as probabilidades de escolha de cada município, representada pela Equação 3.

$$P_i = \frac{e^{U_i}}{\sum_{j=1}^{n}(e^{U_j})} \quad (3)$$

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

O problema de otimização deste artigo é classificado por um problema de maximização sem restrição. A função objetivo (Equação 4), adotada, retorna o maior valor de probabilidade para o município realmente escolhido.

$$\sum_{n=1}^{11} \ln\left(P_n\right) \quad (4)$$

De uma forma geral, o AG possui dois tipos de parâmetros: qualitativos e quantitativos. Os principais parâmetros qualitativos são os tipos dos operadores genéticos. No cruzamento foi utilizado o tipo uniforme, onde os genes dos filhos são gerados a partir das informações dos genes da mesma posição dos cromossomos de seus pais, garantindo que não haja troca de informações entre os genes. Essa estratégia foi utilizada já que para cada município os parâmetros a serem estimados possuem definições e variações diferentes. A mutação do tipo gaussiana foi utilizada, onde a substituição de cada gene é definida por um número aleatório de uma distribuição normal.

Já para os parâmetros quantitativos foram adotados uma população de 200 indivíduos, com taxa de elitismo de 5% (grupo de elite com 10 indivíduos), uma taxa de 80% para o cruzamento e quantidade máxima de gerações de 13 mil (100 multiplicado pela quantidade de variáveis). Também foi inserido um segundo critério de parada para o AG que ocorre quando 10 gerações consecutivas tenham a diferença entre seus valores da função objetivo (do melhor indivíduo) menor que uma tolerância de 10^{-5}.

Neste trabalho, cada resultado do AG é relativo à estimação de 130 parâmetros (13 por função utilidade e 11 funções, sendo a última fixada em valor zero). Como mencionado anteriormente, foram gerados alguns ótimos locais (8 no caso deste trabalho) e a questão principal baseia-se na escolha do melhor resultado, considerando percentuais de acertos similares entre os resultados.

Na definição do melhor conjunto de coeficientes estimados nos oito testes (oito replicações do Algoritmo Genético), deve ser considerado o modelo que se adeque melhor levando-se em conta os sinais e magnitude dos coeficientes estimados nas onze funções utilidade.

A escolha do conjunto de funções utilidades, diante das oito replicações, não é trivial. Para saber se a influência de determinada variável na escolha do destino é positiva ou negativa (definindo então o sinal do parâmetro estimado), os autores, do presente trabalho, propuseram o seguinte procedimento, com auxílio do algoritmo CART, ilustrado na Figura 4:

1) Fixa-se uma das variáveis independentes como primeira divisão a partir do nó raiz do CART (como grau de instrução, por exemplo);
2) Observa-se se o percentual de escolha de cada cidade diminuiu ou aumentou com o aumento da variável escolhida em *1* (as escolhas para a cidade de Camaçari aumentam com o aumento do grau de instrução – sinal positivo para utilidade de Camaçari);
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 de AG que possui maior número de coerências de sinais de parâmetros estimados.

Assim, para a variável grau de instrução, tem-se a seguinte influência nas escolhas dos diferentes destinos: (1- Alagoinhas: negativa; 2 – Camaçari : positiva; 3 – Candeias : negativa; 4 – Catu : negativa; 5 – Conceição da feira : positiva; 6 – Dias Davila : positiva; 7 – Mata de São João: positiva; 8 – Pojuca: negativa; 9 – Salvador: positiva; 10 – Santo amaro: nenhuma; 11 – Simões Filho : negativa).

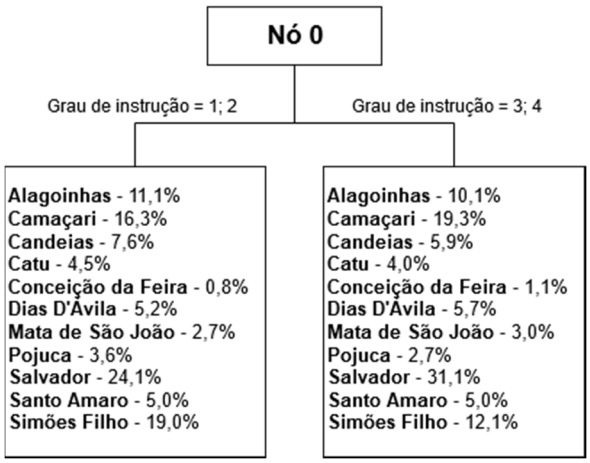

Figura 4. CART com primeira variável fixada *Grau de Instrução*, onde: 1=sem instrução; 2=1° grau; 3=2° grau; 4=3° grau

3.4. Método

A Figura 5 apresenta o método proposto neste artigo. Primeiramente, foi estimado um modelo gravitacional pelo método dos mínimos quadrados (*stepwise*) e, na sequência, foram aplicados os algoritmos de AM (CART e AG) para a análise desagregada de distribuição de viagens. Finalmente, tais modelagens foram comparadas (a partir da amostra de validação – 20%) com os dados observados segundo três critérios distintos: distribuição das distâncias de viagens, medidas de ajuste e da perspectiva qualitativa.

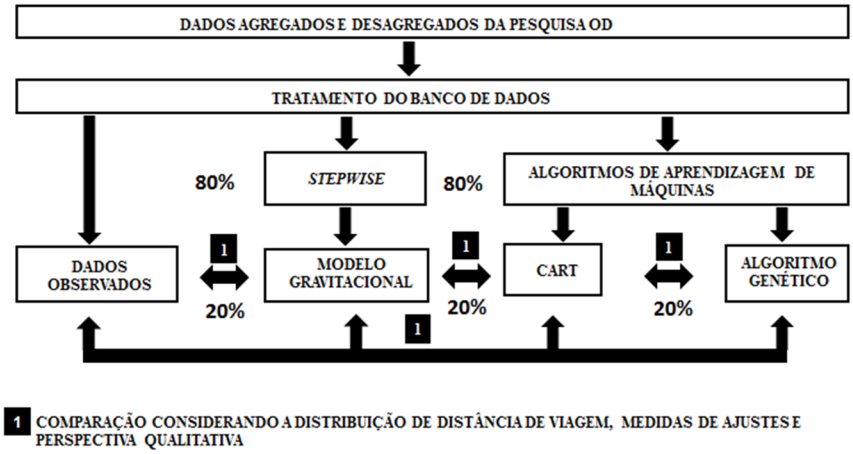

Figura 5. Método proposto para estimar a distribuição de viagens intermunicipais (Pitombo *et al*., 2017; De Souza, 2017)

4. RESULTADOS E DISCUSSÕES

Esta seção descreve os principais resultados provenientes da calibração ou treinamento das três abordagens propostas, bem como comparação da adequabilidade e eficiência das ferramentas, utilizando a amostra de validação.

4.1. Modelo 1: Modelo Gravitacional

Os coeficientes do Modelo Gravitacional foram calculados por meio do procedimento dos mínimos quadrados (Procedimento *Stepwise*). Os parâmetros estimados neste modelo estão apresentados na Equação 5.

$$V_{ij}^{intermunicipais} = \frac{P_i^{0,406} \cdot E_j^{0,283}}{d_{ij}^{0.110}} \quad (5)$$

Em que, $V_{ij}^{intermunicipais}$ é o total de viagens intermunicipais de i para j; P_i é a população da cidade de origem; E_j é a variável saldo de empregos na cidade de destino; 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. Ressalta-se ainda que as três variáveis foram selecionadas pelo método *Stepwise* como sendo aquelas associadas aos parâmetros estimados estatisticamente significativos.

4.2. Modelo 2: Algoritmo CART

A Figura 6 representa, esquematicamente, o mapa de árvore da etapa de treinamento obtida pelo algoritmo CART.

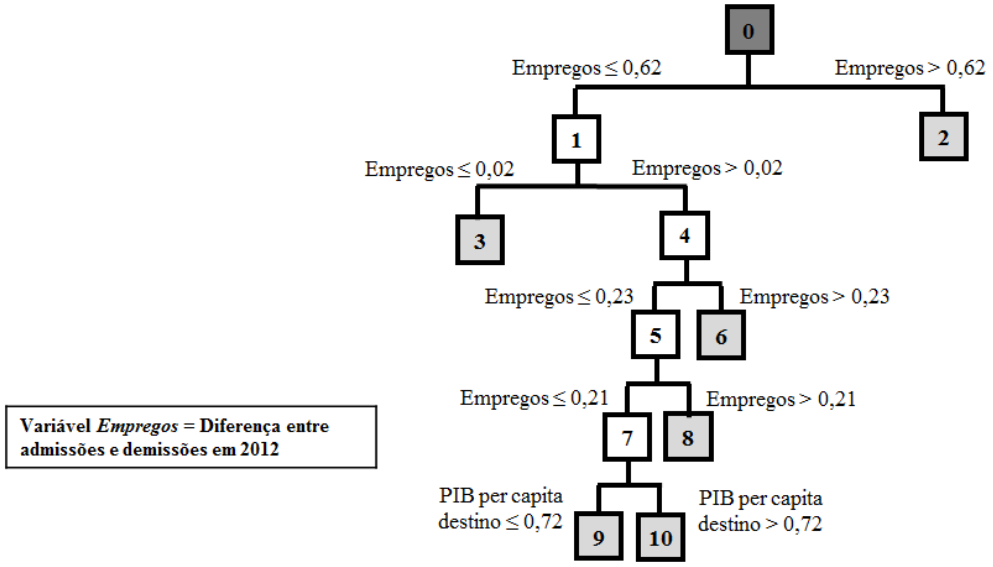

Figura 6. Mapa da Árvore de Decisão com as variáveis de corte normalizadas em cada divisão (Modelo 2: Algoritmo CART)

A Tabela 4 apresenta as informações dos nós-terminais, incluindo as três escolhas de destinos mais frequentes selecionadas pelo algoritmo CART. 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 Normalizado de Empregos* 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.

A primeira variável independente selecionada foi *Saldo Normalizado de Empregos,* 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*. A partir da amostra de validação do processo de Árvore de Decisão (AD), o algoritmo CART forneceu uma precisão de 85,9% de acertos.

Tabela 4: Municípios de destino mais frequentes e as condições de corte nos nós terminais (Modelo 2: 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	Empregos > 0,62	28,9
3	Empregos ≤ 0,02	18,4
6	Empregos ≤ 0,62; Empregos > 0,23	10,1
8	Empregos ≤ 0,23; Empregos > 0,21	14,6
9	Empregos > 0,02; Empregos ≤ 0,21; PIB per capita ≤ 0,72	21,7
10	Empregos > 0,02; Empregos ≤ 0,21; PIB per capita > 0,72	6,3

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*. 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.

4.3. Modelo 3: Algoritmo Genético

Para a obtenção do modelo de otimização, por meio da técnica do Algoritmo Genético, foi utilizado o *software* MathWorks MATLAB R2016B para calibrar os 130 parâmetros das funções utilidades dos onze municípios da análise. Foram obtidas oito replicações (ótimos locais), provenientes de testes. Todas as oito replicações tiveram percentuais de acertos altos, obtidos a partir da amostra de validação, entre aproximadamente 94% e 96%. Cada replicação teve uma duração aproximada de 96 horas e foram executadas em um computador com processador intel Core II5-3330, 30 GHz e memória 8GB.

Aplicando-se o procedimento auxiliar com o algoritmo CART, conforme descrito na Subseção 3.3, foi definido o resultado do modelo (replicação) referente ao teste 5 como o melhor conjunto de coeficientes, o qual caracterizou as funções utilidade, apresentadas na Tabela 5, de acordo com os seguintes valores de parâmetros.

Tabela 5: Funções utilidades para cada município da replicação 5 do AG

Município	Funções utilidades do município
Alagoinhas	$U_1 = -3{,}52 + 0{,}14.\,\text{Idade} - 0{,}11.\,\text{Renda} - 0{,}16.\,\text{Instrução} - 0{,}06.\,\text{Carros no domicílio} + 0{,}26.\,\text{Frequencia semanal de viagem} - 0{,}67.\,\text{Tempo de viagem} + 0{,}63.\,\text{Custo da viagem} - 6{,}34.\,\text{Pop da cidade de origem} - 2{,}89.\,\text{PIB da cidade de orige} - 11{,}52.\,\text{PIB da cidade de destino} + 33{,}69.\,\text{Empregos} + 14{,}47.\,\text{Distância}$
Camaçari	$U_2 = -0{,}58 - 0{,}26.\,\text{Idade} - 0{,}24.\,\text{Renda} + 0{,}33.\,\text{Instrução} - 0{,}34.\,\text{Carros no domicílio} + 0{,}26.\,\text{Frequencia semanal de viagem} - 0{,}16.\,\text{Tempo de viagem} + 0{,}32.\,\text{Custo da viagem} - 2{,}88.\,\text{Pop da cidade de origem} - 3{,}41.\,\text{PIB da cidade de origem} + 9{,}35.\,\text{PIB da cidade de destino} - 44{,}87.\,\text{Empregos} + 1{,}73.\,\text{Distância}$

Tabela 5: Funções utilidades para cada município da replicação 5 do AG (continuação)

Município	Funções utilidades do município
Candeias	$U_3 = -4{,}91 - 0{,}02.\,\text{Idade} - 0{,}42.\,\text{Renda} - 0{,}05.\,\text{Instrução} - 0{,}76.\,\text{Carros no domicílio} + 0{,}20.\,\text{Frequencia semanal de viagem} - 0{,}63.\,\text{Tempo de viagem} + 0{,}29.\,\text{Custo da viagem} - 3{,}11.\,\text{Pop da cidade de origem} - 5{,}19.\,\text{PIB da cidade de origem} + 17{,}55.\,\text{PIB da cidade de destino} - 9{,}00.\,\text{Empregos} + 3{,}68.\,\text{Distância}$
Catu	$U_4 = +2{,}09 - 0{,}54.\,\text{Idade} - 0{,}13.\,\text{Renda} + 0{,}43.\,\text{Instrução} + 0{,}02.\,\text{Carros no domicílio} + 0{,}34.\,\text{Frequencia semanal de viagem} - 0{,}49.\,\text{Tempo de viagem} + 0{,}69.\,\text{Custo da viagem} - 3{,}48.\,\text{Pop da cidade de origem} - 0{,}77.\,\text{PIB da cidade de origem} - 21{,}89.\,\text{PIB da cidade de destino} - 4{,}50.\,\text{Empregos} + 7{,}55.\,\text{Distância}$
Conceição da Feira	$U_5 = -3{,}70 + 0{,}32.\,\text{Idade} + 0{,}04.\,\text{Renda} - 0{,}30.\,\text{Instrução} - 0{,}05.\,\text{Carros no domicílio} + 0{,}34.\,\text{Frequencia semanal de viagem} - 0{,}60.\,\text{Tempo de viagem} + 0{,}72.\,\text{Custo da viagem} - 3{,}14.\,\text{Pop da cidade de origem} + 3{,}81.\,\text{PIB da cidade de origem} - 21{,}67.\,\text{PIB da cidade de destino} - 11{,}86.\,\text{Empregos} + 19{,}17.\,\text{Distância}$
Dias D'Ávila	$U_6 = 2{,}79 + 0{,}36.\,\text{Idade} + 0{,}01.\,\text{Renda} + 0{,}55.\,\text{Instrução} - 0{,}53.\,\text{Carros no domicílio} + 0{,}17.\,\text{Frequencia semanal de viagem} + 0{,}01.\,\text{Tempo de viagem} + 0{,}21.\,\text{Custo da viagem} - 2{,}90.\,\text{Pop da cidade de origem} - 1{,}36.\,\text{PIB da cidade de origem} - 2{,}94.\,\text{PIB da cidade de destino} - 28{,}17.\,\text{Empregos} + 6{,}26.\,\text{Distância}$
Mata de São João	$U_7 = -1{,}15 + 0{,}37.\,\text{Idade} - 0{,}03.\,\text{Renda} + 0{,}25.\,\text{Instrução} - 0{,}05.\,\text{Carros no domicílio} + 0{,}26.\,\text{Frequencia semanal de viagem} + 0{,}21.\,\text{Tempo de viagem} + 0{,}45.\,\text{Custo da viagem} + 1{,}15.\,\text{Pop da cidade de origem} + 1{,}20.\,\text{PIB da cidade de origem} - 23{,}56.\,\text{PIB da cidade de destino} + 10{,}83.\,\text{Empregos} - 1{,}23.\,\text{Distância}$
Pojuca	$U_8 = -3{,}21 + 0{,}06.\,\text{Idade} - 0{,}12.\,\text{Renda} - 0{,}03.\,\text{Instrução} - 0{,}43.\,\text{Carros no domicílio} + 0{,}10.\,\text{Frequencia semanal de viagem} - 0{,}11.\,\text{Tempo de viagem} + 0{,}42.\,\text{Custo da viagem} - 5{,}56.\,\text{Pop da cidade de origem} - 4{,}80.\,\text{PIB da cidade de origem} + 7{,}15.\,\text{PIB da cidade de destino} + 4{,}93.\,\text{Empregos} + 11{,}77.\,\text{Distância}$
Salvador	$U_9 = -12{,}11 + 0{,}44.\,\text{Idade} - 0{,}73.\,\text{Renda} - 0{,}28.\,\text{Instrução} - 0{,}14.\,\text{Carros no domicílio} + 0{,}17.\,\text{Frequencia semanal de viagem} - 1{,}16.\,\text{Tempo de viagem} + 0{,}69.\,\text{Custo da viagem} - 7{,}92.\,\text{Pop da cidade de origem} - 3{,}76.\,\text{PIB da cidade de origem} - 17{,}25.\,\text{PIB da cidade de destino} + 56{,}85.\,\text{Empregos} + 11{,}06.\,\text{Distância}$
Santo Amaro	$U_{10} = 0{,}48 + 0{,}38.\,\text{Idade} + 0{,}25.\,\text{Renda} - 0{,}04.\,\text{Instrução} + 0{,}18.\,\text{Carros no domicílio} + 0{,}36.\,\text{Frequencia semanal de viagem} - 0{,}14.\,\text{Tempo de viagem} + 0{,}81.\,\text{Custo da viagem} + 1{,}75.\,\text{Pop da cidade de origem} + 5{,}11.\,\text{PIB da cidade de origem} - 26{,}65.\,\text{PIB da cidade de destino} - 19{,}48.\,\text{Empregos} + 5{,}01.\,\text{Distância}$
Simões Filho	$U_{11} = 0$

Como no algoritmo CART, variáveis agregadas como *Empregos* e *PIB* tiveram maior influência (maiores valores de coeficientes associados) nas escolhas das cidades de destino. Também é possível verificar a influência das variáveis desagregadas através da magnitude e sinais dos parâmetros estimados. Uma das desvantagens do AG é a ausência de teste de hipótese para testar a significância estatística dos parâmetros calibrados.

4.4. Comparações das Abordagens

4.4.1. Distribuição das distâncias de viagens

A distribuição das distâncias das viagens estimadas e observadas, ilustrada na Figura 7, foi considerada como uma medida de desempenho para avaliar e comparar os modelos previamente estimados. O intuito deste critério é minimizar as diferenças entre as distribuições da função de impedâncias entre as viagens observadas e estimadas.

A Figura 7 indica uma forte associação entre a distribuição das distâncias de viagens dos valores observados e os estimados pelos algoritmos de AM. O Algoritmo Genético apresentou o

melhor desempenho para qualquer faixa de distância destacada e algoritmo CART foi melhor para viagens menores que 30 km, enquanto o Modelo Gravitacional melhor estima as viagens com distância menores que 15 km e distâncias entre 75 e 90 km.

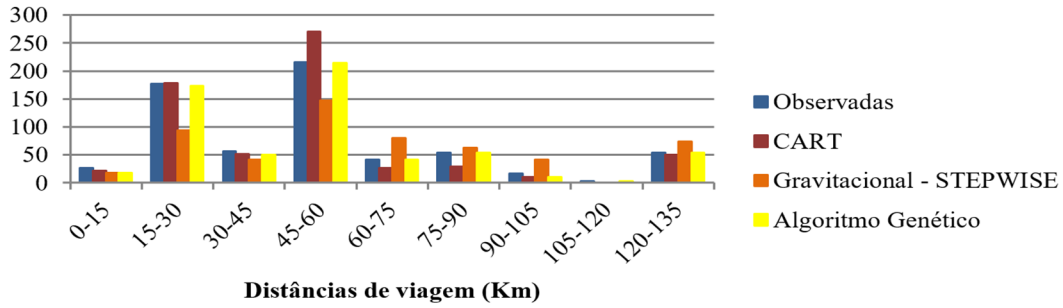

Figura 7. Histograma das distâncias de viagens

Além da interpretação visual, testes não paramétricos, como os testes estatísticos *Mann-Whitney* e *Kolmogorov-Smirnov*, foram realizados a fim de comparar as distribuições das distâncias de viagens observadas às distribuições das distâncias de viagens estimadas pelas diferentes abordagens. O teste da mediana foi realizado para testar se as amostras têm medidas centrais semelhantes. Com base nos resultados, pode-se afirmar que os Algoritmos Genéticos (AG), seguido pelo CART, forneceram as estimativas mais precisas. Ambos os algoritmos de AM têm a mesma distribuição de probabilidade dos valores observados, bem como, as mesmas medidas centrais. No entanto, para o modelo gravitacional, esta hipótese foi rejeitada.

4.4.2. Medidas de qualidade do ajuste

Para o critério da qualidade do ajuste, algumas medidas relativas ao erro foram avaliadas, como as mostradas nas Equações 6 a 9, 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*.

$$\frac{1}{N}\sum\left(x_i - y_i\right)^2 \tag{6}$$

$$\sqrt{\frac{1\sum\left(x_i - y_i\right)^2}{N}} \tag{7}$$

$$\frac{1}{N}\sum\left(x_i - y_i\right) \tag{8}$$

$$\frac{1}{N-1}.\sum\frac{\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sigma_x \sigma_y} \tag{9}$$

Em que, 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 6 apresenta os valores das medidas propostas. Observando as medidas estatísticas, pode-se notar que o procedimento de otimização do AG, seguido pelo CART, proporcionou melhores resultados do que o modelo gravitacional.

Tabela 6: 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.	80,61	8,98	4,52	0,73

4.4.3. Perspectiva qualitativa

O algoritmo CART fornece a medida de importância das variáveis explicativas selecionadas no modelo. Esta medida é expressa em termos de quantidade normalizada, em relação à variável com a maior dimensão de importância. A Tabela 7 apresenta a importância normalizada das variáveis selecionadas por meio do algoritmo CART, sendo que a variável *N° de carros no domicílio* foi a única variável desagregada selecionada. Esta variável foi selecionada como divisor substituto a partir do nó raiz. Sendo assim, é possível verificar a sua influência na escolha dos destinos, fixando-a como variável divisória a partir do nó raiz. A Figura 8 apresenta as diferenças nas escolhas de destinos quando a variável desagregada (*N° de carros no domicílio*) é selecionada como divisor substituto no nó raiz (nó 0). Percebe-se, por exemplo, que o número de carros influencia positivamente a escolha do município de Salvador, é praticamente indiferente à escolha do município de Camaçari e influencia negativamente a escolha do município de Simões Filho.

Tabela 7: 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%

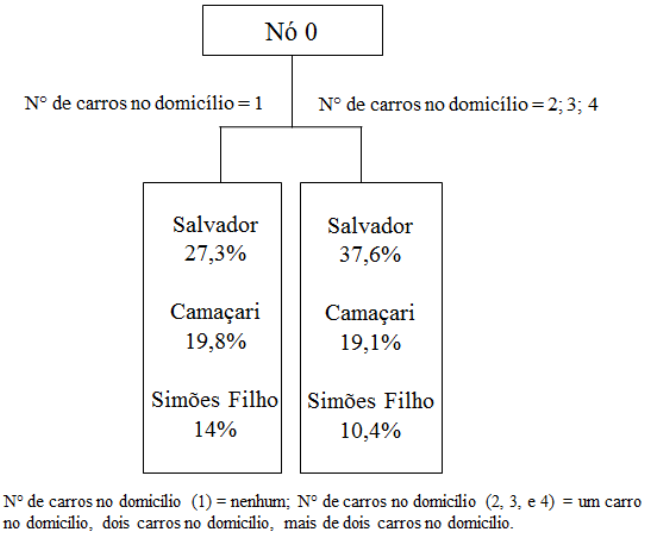

Figura 8. Variável *N° de carros no domicílio* utilizada como divisor substituto no nó raiz

Enquanto, através do CART, foi possível verificar a influência apenas da variável *N° de carros*

no domicílio, a partir da calibração do AG, foi possível avaliar os 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. A Tabela 8 traz informações referentes à influência de cada variável da escolha de cada cidade, levando-se em conta o modelo (5) escolhido.

Tabela 8: Influência de cada variável na escolha de destino intermunicipal

	Alagoinhas	Camaçari	Candeias	Catu	Feira	Dias D'Avila	João	Pojuca	Salvador	Santo Amaro	Filho
Idade	+	-	+	-	-	-	+	+	+	+	**REF**
Grau de Instrução	+	+	+	-	-	+	+	-	-	-	**REF**
Renda	+	-	+	+	+	+	+	-	-	+	**REF**
Nº 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**

REF = Função utilidade nula (referência)

5. 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 destinos. Além disso, os algoritmos de Aprendizagem de Máquinas, aqui utilizados, não partem de suposições matemáticas rigorosas, sendo possível utilizar qualquer tipo de variável (qualitativa/quantitativa) 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 se 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. Ressalta-se também o procedimento metodológico auxiliar para a escolha da melhor replicação (dentre os ótimos locais) do AG, com auxílio do CART. Além disso, é uma ferramenta viável inclusive para casos de obtenção de matrizes futuras a partir de matriz semente incompleta, pois é possível gerar um conjunto de dados sintéticos através dos diversos resultados obtidos através do AG.

Como a otimização proposta através do AG baseou-se no modelo logit multinomial, sugere-se para trabalhos futuros, a comparação entre o modelo logit multinomial, calibrado neste trabalho através de AG, e o modelo logit multinomial tradicional.

AGRADECIMENTOS

Esta pesquisa foi patrocinada pela Coordenação de Aperfeiçoamento do Pessoal de Ensino Superior (CAPES) e pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Os autores também agradecem ao grupo CETRAMA (Universidade Federal da Bahia) pela cessão dos dados.

REFERÊNCIAS

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

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

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.

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

De Souza, A. D. (2017); *Comparação de algoritmos de Aprendizagem de Máquinas para análise desagregada de viagens intermunicipais*. 84 f. Dissertação de Mestrado. Departamento de Engenharia de Transporte. Escola de Engenharia de São Carlos.

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

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), v. 2, p. 175-186.

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

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

Mozolin, M.; Thill, J.C.; Linn, U.E. (2015) 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

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

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, p. 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. *Journal of Transport Literature*, v. 7, p. 59-86. DOI: 10.1590/s2238-10312013000300004

Pitombo, C. S.; De Souza, A.D.; Lindner, A. (2017) Comparing decision tree algorithms to estimate intercity trip distribution. *Transportation Research Part C* , v. 77, p. 16-32. DOI: 10.1016/j.trc.2017.01.009

Pulugurta S, Arun A, Errampalli M (2013) Use of Artificial Intelligence for Mode Choice Analysis and Comparison with Traditional Multinomial Logit Model, *Procedia - Social and Behavioral Sciences,* v. 104, p. 583-592. DOI: 10.1016/ j. sbspro.2013.11.152

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

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

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

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