Uso de técnica de mineração de dados no auxílio à modelagem de distribuição de viagens intermunicipais

Cira Souza Pitombo,
Henrique Stramandinoli Guimarães

Universidade de São Paulo/Escola de Engenharia de São Carlos, São Carlos, Brasil

RESUMO

O presente artigo tem dois objetivos de caráter investigativo: (1) testar a adequabilidade do uso de técnica de Mineração de Dados (MD) para modelagem de distribuição de viagens; (2) analisar a influência de variáveis desagregadas na escolha dos destinos de viagens intermunicipais. Os dados utilizados são provenientes de Pesquisa Origem/Destino, realizada em 2012 em onze municípios do estado da Bahia, Brasil. A técnica de MD utilizada foi Árvore de Decisão (AD). Foram feitas duas análises distintas, através da AD, e uma análise tradicional. Foi comparada a acurácia de três modelos obtidos e o modelo de AD que considerava características demográficas, de viagens e socioeconômicas foi o de melhor poder preditivo (86% de acertos). Vale ressaltar ainda que, apesar dos bons resultados observados, o presente trabalho atesta que, para este estudo de caso, a técnica é adequada. Outra conclusão relevante seria a pouca importância de variáveis individuais socioeconômicas quando comparadas às variáveis demográficas e de viagens na explicação do fenômeno.

1. INTRODUÇÃO

Diversos autores corroboram a afirmação de que comportamento relativo a viagens, sobretudo considerando escolhas discretas (modo de transporte, destinos, período do dia, motivo de viagem, etc.) é fortemente relacionado a características individuais, das viagens, do meio urbano e suas facilidades (Kitamura et al., 1997; Ortúzar e Willumsen, 2011).

O modelo mais tradicional na previsão de demanda por transporte é o modelo Quatro Etapas. O objetivo do modelo Quatro Etapas é estimar a demanda atual e futura por transporte. O modelo é divido em quatro fases distintas, mas interligadas ou sequenciais: I) Geração de viagens; II) Distribuição de viagens; III) Divisão modal; IV) Alocação do tráfego.

Este trabalho tem enfoque na segunda etapa do modelo sequencial tradicional, distribuição de viagens. Geralmente, os modelos de distribuição de viagens tradicionais (Fator de crescimento e Gravitacional) consideram variáveis agregadas em suas análises (Novaes, 1986), como características demográficas e de viagens. Suspostamente, tais modelos desconsideram que as escolhas dos destinos são feitas individualmente, podendo considerar características individuais e domiciliares, além das macro características (população, empregos, custo médio de viagem, etc.), usualmente investigadas. Através deste trabalho, será possível investigar, além das variáveis agregadas, aquelas variáveis relativas ao domicílio e/ou indivíduo que possivelmente influenciam as escolhas dos destinos.

Além disso, ao longo dos anos, técnicas de Mineração de Dados (MD) vêm sendo aplicadas à modelagem de demanda por transportes, alternativamente à modelagem tradicional. Escolhas relativas a viagens podem ser definidas como problemas de

reconhecimento de padrões, definidos por variáveis explicativas que determinam escolhas entre alternativas (Xie et al., 2007; Pitombo e Costa, 2014).

A técnica utilizada para auxílio às análises é a ferramenta de mineração de dados (MD) conhecida como Árvore de Decisão (AD). Um dos principais motivos que levam à escolha da AD é a sua capacidade de representar a natureza probabilística do objeto analisado, que no caso corresponde aos segmentos socioeconômicos e demográficos relacionados à distribuição de viagens.

Assim, o presente artigo tem dois objetivos de caráter investigativo: (1) testar a adequabilidade do uso de técnica de MD para distribuição de viagens; (2) analisar a influência de variáveis desagregadas na escolha dos destinos de viagens intermunicipais.

O método utilizado no trabalho segue as etapas sumariadas nas próximas seções. A seção 2 descreve a modelagem de demanda por transportes tradicional, bem como o emprego de técnicas de MD neste intuito. A seção 3 descreve o tratamento do banco de dados utilizados. Já a seção 4 descreve a aplicação da técnica de AD, resultados e discussões. Finalmente, a seção 5 descreve as principais conclusões obtidas.

2. Metodologia: Modelagens abordadas

2.1 Modelagem tradicional

Existem alguns modelos que analisam a distribuição de viagens, tais como: (1) Modelo de Fratar; e (2) Modelos gravitacionais.

No Modelo de Fratar a previsão do volume de viagens futuras entre um par de zonas é feita através da multiplicação do volume atual pelo produto dos fatores de crescimento previstos para as duas zonas com ajustamento para atratividade relativa das outras zonas.

$$Q_{ij}^{t} = Q_{ij}^{0}.F_i.F_j.L_i \quad (1)$$

Q_{ij}^{t}: número de viagens no ano t de i para j; Q_{ij}^{0}: número de viagens atuais de i para j; F_i: fator de crescimento da zona de origem i; F_j: fator de crescimento da zona de destino j; L_i: fator de ajuste das origens.

Já os modelos gravitacionais consideram que o número de pessoas que se movimentam entre quaisquer pares de cidades é proporcional ao tamanho delas e inversamente proporcional à distância entre elas. Uma das vantagens deste modelo é que se considera o efeito da distância espacial ou facilidade de iteração entre as regiões definidas pela função de impedância. Além disso, a taxa de geração de viagens em uma zona é proporcional à sua "massa", nesse caso podendo ser representada por variáveis como população, emprego, etc. (Novaes, 1986). Deve-se observar que a distância pode ser substituída por outros fatores, como tempo e custo da viagem, ou por diversos fatores compostos, denominados de impedância (resistência ao deslocamento). A forma básica do modelo gravitacional é dada na Equação 2.

$$V_{ij} = k_i.A_i.B_j.C_{ij}^{\theta} \quad (2)$$

k_i: constante de proporcionalidade; A_i: Viagens produzidas em i (Podendo ser População ou qualquer outra variável socioeconômica que caracterize a zona de origem i); B_j: Viagens atraídas a j (Podendo ser Empregos ou qualquer outra variável socioeconômica que caracterize a zona de destino j); C_{ij}: Custo da viagem de i para j (Podendo ser qualquer outra impedância como distância ou tempo de viagem, por exemplo).

2.2 Modelagem através de técnicas de Mineração de dados

As pesquisas e desenvolvimento de MD emergiram a partir dos anos 90 com o principal objetivo de procurar por informações úteis a partir de um grande conjunto de dados. Podem-se citar algumas técnicas de MD: Redes Neurais Artificiais. Árvores de Decisão, Algoritmos Genéticos, Regras de Associação. Recentemente, verifica-se o aumento de trabalhos que descrevem a aplicação de tais técnicas na análise do comportamento relacionado a viagens.

Com o objetivo de explorar o potencial de técnicas de MD para análise de viagens, Keuleers e Wets (2001) utilizaram regras de associação em um amplo conjunto de dados com o propósito de descobrir associações significativas entre padrões de atividades diárias e atributos individuais, domiciliares e relacionados ao meio. Rocha et al. (2015) utilizaram a técnica de Redes Neurais Artificiais para previsão de produção de viagens por Zonas de Tráfego na Região Metropolitana de São Paulo, Brasil.

Na literatura observa-se também a aplicação de CHAID – (*Chi-squared Automatic Interaction Detection*), um algoritmo de Árvore de Decisão, para análises de taxas de geração de viagens (Strambi e van de Bilt, 1998). Xie et al. (2007) compararam o uso de Árvores de Decisão e Redes Neurais para modelagem de escolha modal de viagens com motivo trabalho.

Pitombo et al. (2011) analisaram a influência de variáveis de uso do solo e socioeconômicas na escolha de padrões de viagens urbanas, através de uso de Árvore de Decisão. Enquanto, Pitombo et al. (2013) estudaram a influência de deslocamentos de trabalhadores através da mesma técnica de MD.

Considerando, especificamente, aplicação de técnicas de MD para modelagem de distribuição de viagens, são observados bem recentemente alguns trabalhos que comprovam aplicação e adequabilidade de técnicas de Redes Neurais Artificiais (Rasouli e Nikras, 2013, Gonçalves et. al, 2015). A aplicação de Árvore de Decisão, específica e exclusivamente para a etapa de distribuição de viagens, ainda é incipiente (Pitombo e Guimarães, 2015). Através deste trabalho, é explorado o potencial de tal técnica para obtenção de matrizes O/D.

3. DADOS

Os dados utilizados neste trabalho são provenientes de Pesquisa Origem/Destino, realizada em 2012 em onze municípios do estado da Bahia (Brasil). Os municípios que fazem parte da pesquisa são: Alagoinhas, Catú, Pojuca, Mata de São João, Dias D'Ávila, Camaçari, Simões Filho, Salvador, Candeias, Santo Amaro e Conceição da Feira. A amostra entrevistada é formada por usuários dos ônibus nos terminais rodoviários, moradores dos municípios pesquisados nos polos geradores de tráfego e com motoristas e passageiros de automóveis nas rodovias intermunicipais ao longo dos trechos em estudo. A pesquisa envolveu um levantamento do perfil socioeconômico de cada entrevistado, bem como informações inerentes à viagem intermunicipal.

A Figura 1 traz a localização dos trechos considerados. Os trechos têm 228 km de extensão de linha férrea e 249 km de distância rodoviária. A região compreende importantes polos de atividades no estado, observando-se muitas viagens intermunicipais pendulares com motivo trabalho entre tais cidades. Na região de abrangência do trabalho estão localizados centros urbanos de importância regional turística e histórica. A Tabela 1, em seguida, traz as principais informações censitárias dos municípios.

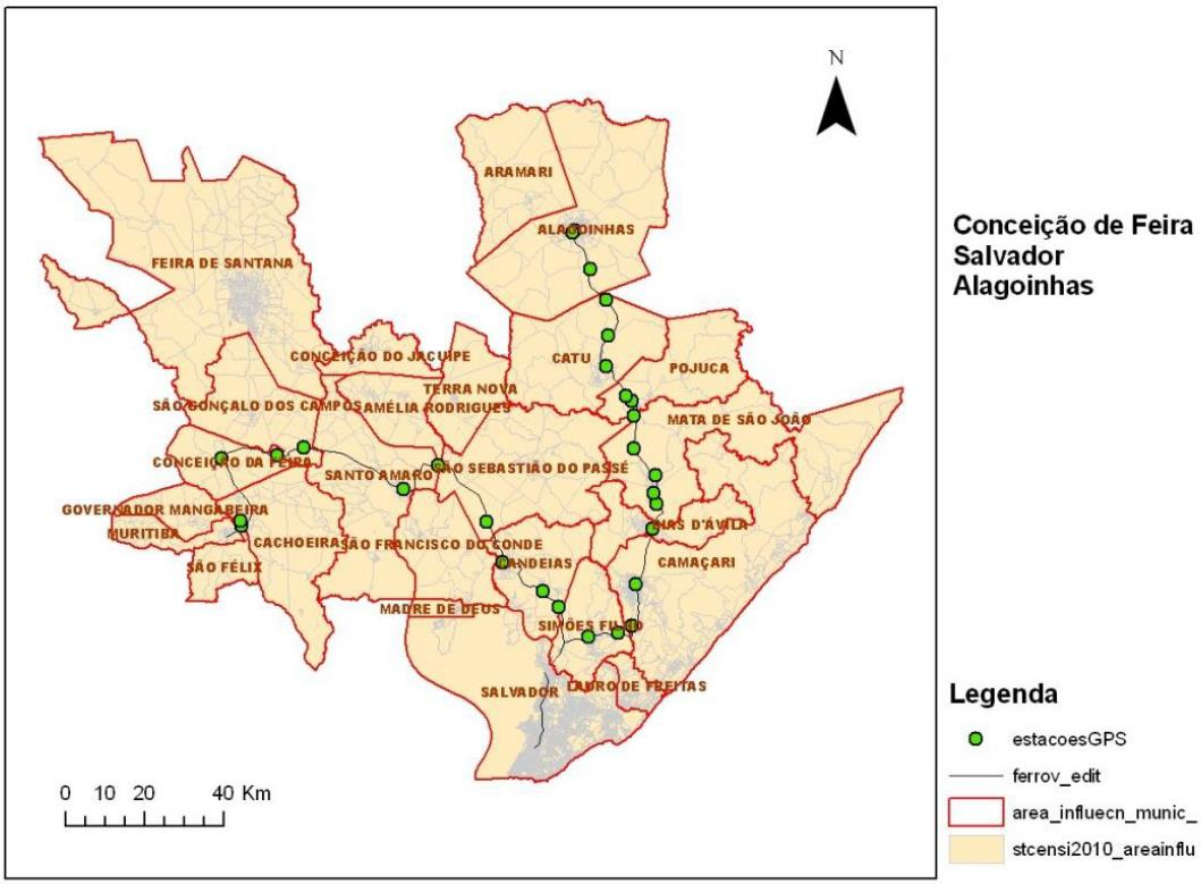


Figura 1 – Municípios que compõem a região estudada

Tabela 1 – Principais características dos municípios (IBGE, 2010).

Cidade	Pessoas ocupados que exerciam o trabalho principal em outro município	População acima de 15 anos	Pessoal ocupado	PIB per capita e preços concorrentes (U$)	Renda média per capita (R$)
Alagoinhas	6.128	108.688	24.967	11.370,00	400,00
Camaçari	15.000	175.144	84.458	55.063,52	377,50
Candeias	6.870	60.084	17.864	50.613,49	340,00
Catu	3.609	37.126	8.739	6.765,22	340,00
Conceição da Feira	1.200	19.579	1.521	5.113,57	266,25
Dias d'Ávila	7.734	46.581	17.233	32.732,93	344,33
Mata de São João	2.873	30.865	12.144	8.753,35	319,00
Pojuca	1.348	23.742	8.589	30.545,16	340,00
Salvador	60.219	2.015.074	910.402	13.728,08	510,00
Santo Amaro	3.453	40.954	6.055	6.598,49	280,00
Simões Filho	13.621	83.003	38.550	31.266,42	333,33

3.1 A amostra

A base de dados da pesquisa é composta por pessoas maiores de catorze anos (idade mínima para um indivíduo desacompanhado de um responsável viajar de transportes ferroviários ou rodoviários de passageiros entre as cidades intermunicipais) 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 (Centro de Estudos de Transportes e Meio Ambiente – 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 Pólos Geradores de Tráfego englobam os terminais rodoviários intermunicipais, centros comerciais, parques industriais, universidades e centro educacionais em que será assegurada a aleatoriedade de seleção das unidades amostrais. O plano amostral escolhido foi estratificado e os estratos foram constituídos por cada município que compõem a pesquisa. Selecionando uma amostra aleatória simples sem reposição de indivíduos em cada um dos estratos de forma independente. A amostra foi composta por onze estratos. 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. O tamanho amostral mínimo previsto foi de 3.029 entrevistas que foram alocadas de acordo com o critério da alocação proporcional dado por:

$$n_h = nW_h,\ h = 1, 2, 3, \dots, 11 \tag{3}$$

em que W_h é o peso do h-ésimo estrato, dado por:

$$W_h = \frac{N_h}{N},\ h = 1, 2, 3, \dots, 11 \tag{4}$$

em que N_h é número de indivíduos maiores de catorze anos no estrato h (município) e N é o tamanho da população, número total de pessoas maiores de catorze anos residentes nos municípios. Os tamanhos de amostras calculados pela expressão (4) são sempre arredondados para o inteiro imediatamente acima, quando fracionário. Para evitar problemas operacionais com amostras muito pequenas foi arbitrado um número mínimo de cem indivíduos para o tamanho amostral de cada estrato.

Para tratamento da amostra foram considerados dados de 3.300 indivíduos. 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 (Tabela 1) e variáveis de tempo e distância de viagem. Associadas a essas informações estão os padrões de deslocamentos dos indivíduos (Origem-Destino). Tais padrões codificados são exatamente a variável dependente categórica utilizada na aplicação da técnica de MD.

O *software* IBM SPSS 22.0, utilizado para realização das análises têm uma limitação de 13 categorias para variável dependente. Desta forma, foram considerados na análise os 13 padrões mais frequentes na amostra de 3.300 indivíduos. Assim, a amostra final, após exclusão de padrões de deslocamentos menos frequentes, é composta por 2.144 indivíduos. A Tabela 2 apresenta as variáveis utilizadas na análise (variáveis independentes) e a Tabela 3 descreve a variável dependente e suas treze categorias selecionadas (padrões de deslocamentos mais frequentes).

4. APLICAÇÃO DA ÁRVORE DE DECISÃO

A técnica utilizada para a análise deste trabalho é Árvore de Decisão (AD), considerada uma forma simples de representação de relação ou de relações existentes em um conjunto de dados. Ela permite classificar uma base de dados em um número finito de classes,

com a qual é possível analisar um grande conjunto de dados, através de regras hierárquicas e da sua divisão em grupos, organizando os dados de maneira compacta e obtendo uma visão real da natureza do processo (Quilan, 1983).

Tabela 2 – Variáveis utilizadas na análise

Variável	Descrição
Categóricas Socioeconômicas e de viagens	
Sexo	(1) Masculino; (2) Feminino
Idade	(1) Até 19 anos; (2) 20 a 29 anos; (3) 30 a 39 anos; (4) 40 a 49 anos
	(5) 50 a 65 anos; (6) acima de 65 anos
Grau de Instrução	(1) sem instrução; (2) 1o grau; (3) 2o grau; (3) 3o grau
Renda	(1) 1 SM; (2) 1 a 3 SM; (3) 3 a 5 SM; (4) acima de 5 SM
Ocupação	(1) Comércio; (2) Indústria; (3) Serviços; (4) Agricultura; (5) Estudante
	(6) Aposentado; (7) Outros
Residência	(1) Própria; (2) Alugada; (3) Cedida
Carros no domicílio	(1) zero; (2)1; (3) 2; (3) 2 ou mais
Frequencia Semanal da viagem	(1) 1 dia; (2) 2 dias; (3) 3 dias; (4) 4 dias; (5) 5 dias; (6) 6 dias; (7) todos os dias
Motivo da viagem	(1) Trabalho; (2) Estudo; (3) Compras; (4) Lazer; (5) Saúde; (6) Visita
Tempo de viagem	(1) Até 30 min; (2) 30 a 60 min; (3) Acima de 60 min
Modo da viagem	(1) ônibus; (2) carro; (3) Van; (4) a pé; (5) Bicicleta ; (6) Motocicleta
Custo de viagem	(1) até R$ 5,00; (2) R$ 5,00 a R$ 10,00; (3) R$ 10,00 a R$ 20,00; (4) Acima de R$ 20,00
Forma de Pagamento	(1) Vale Transporte; (2) Dinheiro
Numéricas de viagens	
Distância	Distância em Km
Tempo de viagem	Tempo em minutos
Numéricas Sociodemográficas dos municípios	
Pessoas ocupadas que exerciam o trabalho principal em outro município	
População acima de 15 anos	
Pessoal ocupado	
PIB per capita e preços concorrentes	
Renda média per capita	

Tabela 3 – Caracterização da variável dependente

Código Origem	Município Origem	Código Destino	Município Destino	Trecho	Padrão de deslocamento	%	Dist (km)
06	Dias d'Ávila	09	Salvador	Dias d'Ávila - Salvador	0609	17,24	58,8
09	Salvador	11	Simões Filho	Salvador - Simões Filho	0911	15,78	30,4
09	Salvador	02	Camaçari	Salvador - Camaçari	0902	13,39	51,3
06	Dias d'Ávila	02	Camaçari	Dias d'Ávila -Camaçari	0602	7,14	21,0
09	Salvador	03	Candeias	Salvador-Candeias	0903	6,21	51,9
11	Simões Filho	09	Salvador	Simões Filho - Salvador	1109	5,76	30,4
09	Salvador	01	Alagoinhas	Salvador- Alagoinhas	0901	4,91	120,0
09	Salvador	10	Santo Amaro	Salvador - Santo Amaro	0910	4,22	107,0
06	Dias d'Ávila	01	Alagoinhas	Dias d'Ávila - Alagoinhas	0601	3,29	70,9
06	Dias d'Ávila	03	Candeias	Dias d'Ávila - Candeias	0603	2,43	37,4
06	Dias d'Ávila	10	Santo Amaro	Dias d'Ávila - Santo Amaro	0610	2,31	96,0
06	Dias d'Ávila	04	Catu	Dias d'Ávila - Catu	0604	2,27	39,0
09	Salvador	07	Mata de São João	Salvador - Mata de São João	0907	2,11	62,7

A hierarquia é denominada árvore e cada segmento é denominado nó. O segmento original contém o conjunto completo dos dados, referindo-se ao nó raiz da árvore. Este nó contém dados que podem ser subdivididos dentro de outros sub-nós, chamados de nós filhos. Quando os dados do nó não podem ser mais subdivididos dentro de um outro subconjunto ele é considerado um nó terminal ou folha.

Para geração do modelo de AD foi utilizado o *software* IBS SPSS 22.0. O algoritmo utilizado é uma variante do CART (do inglês, *Classification and Regression Tree*). O método de partição utilizado para a construção da árvore é descrito por Clark e Pregibon (1992). De

um modo geral, o algoritmo da árvore torna os subconjuntos resultantes cada vez mais homogêneos em relação à variável resposta, mediante sucessivas divisões binárias no conjunto de dados. A cada passo no crescimento da árvore, o particionamento dos dados se faz a partir da produção da minimização do desvio (Critério Gini) ou da entropia em todas as divisões permitidas nos nós da árvore (Breiman et al., 1984). Essa redução de entropia corresponde à diminuição da aleatoriedade ou dificuldade de previsão de uma variável resposta.

A AD assume a variável resposta como categórica seguindo uma distribuição multinomial e trata a árvore como modelo de probabilidade, empregando o desvio como critério de divisão. Este desvio é definido como recíproco da função verossimilhança elevada ao quadrado. Um dos principais motivos que levam à escolha da AD é a sua capacidade de representar a natureza probabilística do objeto analisado, que no caso corresponde aos segmentos socioeconômicos e demográficos relacionados à distribuição de viagens. A partir da utilização do modelo de árvore, reconhece-se que "indivíduos homogêneos" podem tomar diferentes decisões, e associá-los às probabilidades de diferentes respostas possíveis.

Neste trabalho foram feitas duas análises distintas, através do MD e uma análise tradicional. Assim, foram treinadas duas árvores variando-se as variáveis independentes consideradas: (1) Modelo desagregado contendo apenas variáveis socioeconômicas; (2) Modelo desagregado com variáveis socioeconômicas, variáveis agregados dos municípios e de viagens. O modelo 3 foi obtido através da abordagem tradicional (modelo gravitacional) com variáveis do município e de viagens.

4.1 Modelo 1: amostra desagregada com variáveis socioeconômicas

O primeiro modelo de AD a ser testado foi obtido considerando apenas as variáveis socioeconômicas como variáveis independentes, os padrões de deslocamentos (treze mais frequentes viagens intermunicipais da região) e a amostra desagregada de 2.144 indivíduos. As variáveis independentes socioeconômicas utilizadas foram: Sexo; Grau de instrução; Renda; Residência; Carros nos domicílio.

As análises foram realizadas através da AD apresentada pelo IBM SPSS 22.0, algoritmo CART. O critério adotado para a sua classificação foi o mínimo de 50 observações por nó terminal. 70% da amostra foi separada aleatoriamente para treinamento da AD enquanto 30% da amostra foi alocada para validação. A variável mais importante para segregação dos dados foi Grau de Instrução.

Da Figura 2, observam-se 19 nós terminais, ou seja, 19 classes de indivíduos com características distintas e seus respectivos padrões predominantes. Cada caixa da figura representa um Nó com os três padrões de deslocamentos mais frequentes. Cada Nó representa um grupo de indivíduos com características homogêneas e proporções de padrões de deslocamento.

A partir dos indivíduos correspondentes a 30% da amostra separada para teste do modelo de AD, foi possível obter o número de erros e acertos (Total de 25% de acertos). Fazendo ainda um teste qui-quadrado entre valores estimados e observados da amostra desagregada, a estatística qui-quadrado foi muito pequena e o teste corroborou com a hipótese nula que partia do pressuposto de que não há associação entre padrões estimado e observados.

Além disso, foi obtida a matriz O/D de valores estimados, fazendo uma agregação dos resultados. Posteriormente, é realizada uma comparação da matriz O/D de valores observados com as 3 matrizes obtidas com os três modelos descritos.

4.2 Modelo 2: variáveis socioeconômicas desagregadas, variáveis sociodemográficas agregadas e distâncias de viagens

Analogamente ao item anterior, o modelo de AD a ser testado foi obtido considerando as variáveis socioeconômicas desagregadas, além das variáveis sociodemográficas agregadas e distâncias de viagens como variáveis independentes. Os padrões de deslocamentos (treze mais frequentes viagens intermunicipais da região) foram as categorias de variável dependente, sendo a amostra desagregada formada por 2.144 indivíduos novamente. As variáveis independentes utilizadas foram as mesmas apresentadas anteriormente mais as seguintes variáveis agregadas: Distância (em km); Tempo (em minutos); População acima de 15 anos; População ocupada; PIB per capita (em reais); Renda média per capita (em reais); População ocupada que exercia o trabalho em outro município.

As análises foram realizadas através da AD apresentada pelo IBM SPSS 22.0, repetindo os mesmos procedimentos anteriores. O resultado gráfico pode ser observado na Figura 3, onde a variável mais importante para segregação dos dados foi População ocupada que exercia o trabalho em outro município. Observa-se que, desta vez, que as únicas variáveis selecionadas correspondem àquelas usualmente utilizadas no modelo gravitacional de distribuição de viagens. Repetindo o mesmo procedimento metodológico, 70% da amostra desagregada foi selecionado para treinamento da AD, enquanto 30% da amostra foi selecionado para teste da AD. O percentual de acertos aumentou consideravelmente com a inclusão de variáveis agregadas relativas aos municípios, bem como distâncias de viagens (85,7% de acertos). Adicionalmente, o teste qui-quadrado corroborou a hipótese alternativa de que há associação entre as duas variáveis qualitativas (padrões de deslocamentos observados x padrões de deslocamentos estimados).

4.3 Modelo 3: Modelo agregado tradicional com variáveis do município e de viagens

Para comparação dos modelos de MD com a abordagem tradicional, foi feita a calibração clássica de distribuição de viagens, através do modelo gravitacional. O modelo gravitacional foi calibrado a partir da escolha de variáveis tradicionais e amostra agregada contendo Número de viagens entre pares de Origem-Destino (variável dependente); População da Origem, População do destino e Distâncias entre origens e destinos (variáveis independentes). Para calibração do modelo foi utilizada a técnica de Regressão Linear Múltipla, após transformações das variáveis através dos seus logaritmos naturais. Finalmente, da equação 5 foi aquela ajustada.

$$V_{ij} = \frac{p_i^{0,172570} . p_j^{0,327892}}{d_{ij}^{0,15419}} \quad (5)$$

Para Vij = viagens de i para j; Pi = população de i; Pj = população de j; dij = distância de i para j. A partir dessa equação foram calculados os valores esperados para esse modelo gravitacional (modelo 3) e criada a matriz O/D estimada.

4.4 Breve discussão acerca dos resultados

É possível observar pelos resultados obtidos nos três modelos que o modelo de AD (modelo 2), onde foram consideradas tanto as variáveis desagregadas quanto as variáveis agregadas e de distâncias, foi o de maior poder preditivo. Curioso comentar que o melhor modelo foi o desagregado com a AD, considerando variáveis clássicas da modelagem tradicional de distribuição de viagens, como população e distância. A Tabela 4, em seguida, mostra um resumo dos resultados obtidos para cada modelo. Até mesmo o modelo Gravitacional (modelo 3) teve um desempenho inferior ao modelo de AD com variáveis

agregadas e desagregadas (modelo 2).Vale ressaltar que o melhor modelo de AD, a despeito de melhor poder preditivo, ainda é um modelo não-paramétrico e exploratório. No entanto pode ser uma abordagem interessante de distribuição de viagens intermunicipais considerando comportamentos individuais (amostra desagregada).

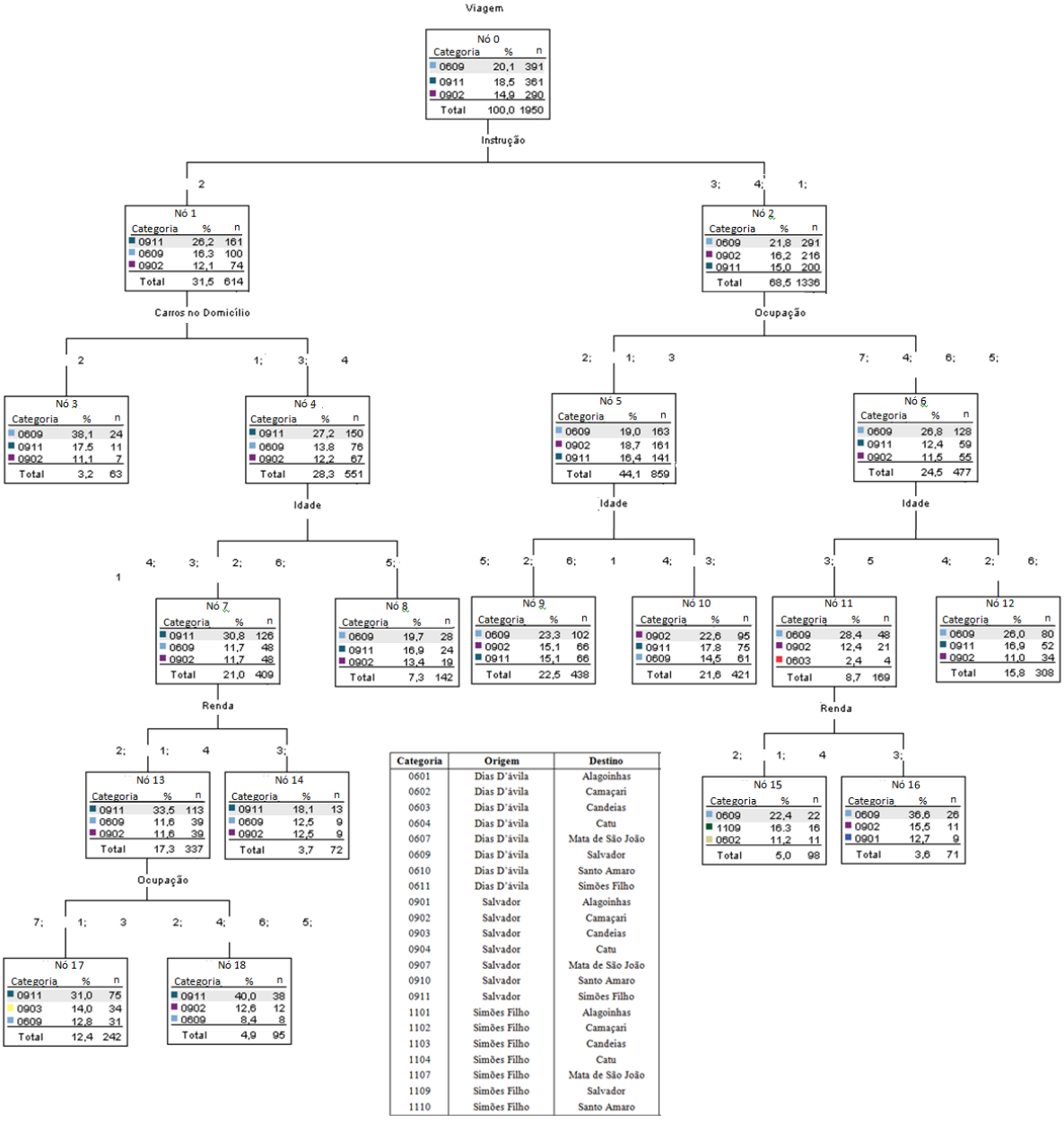
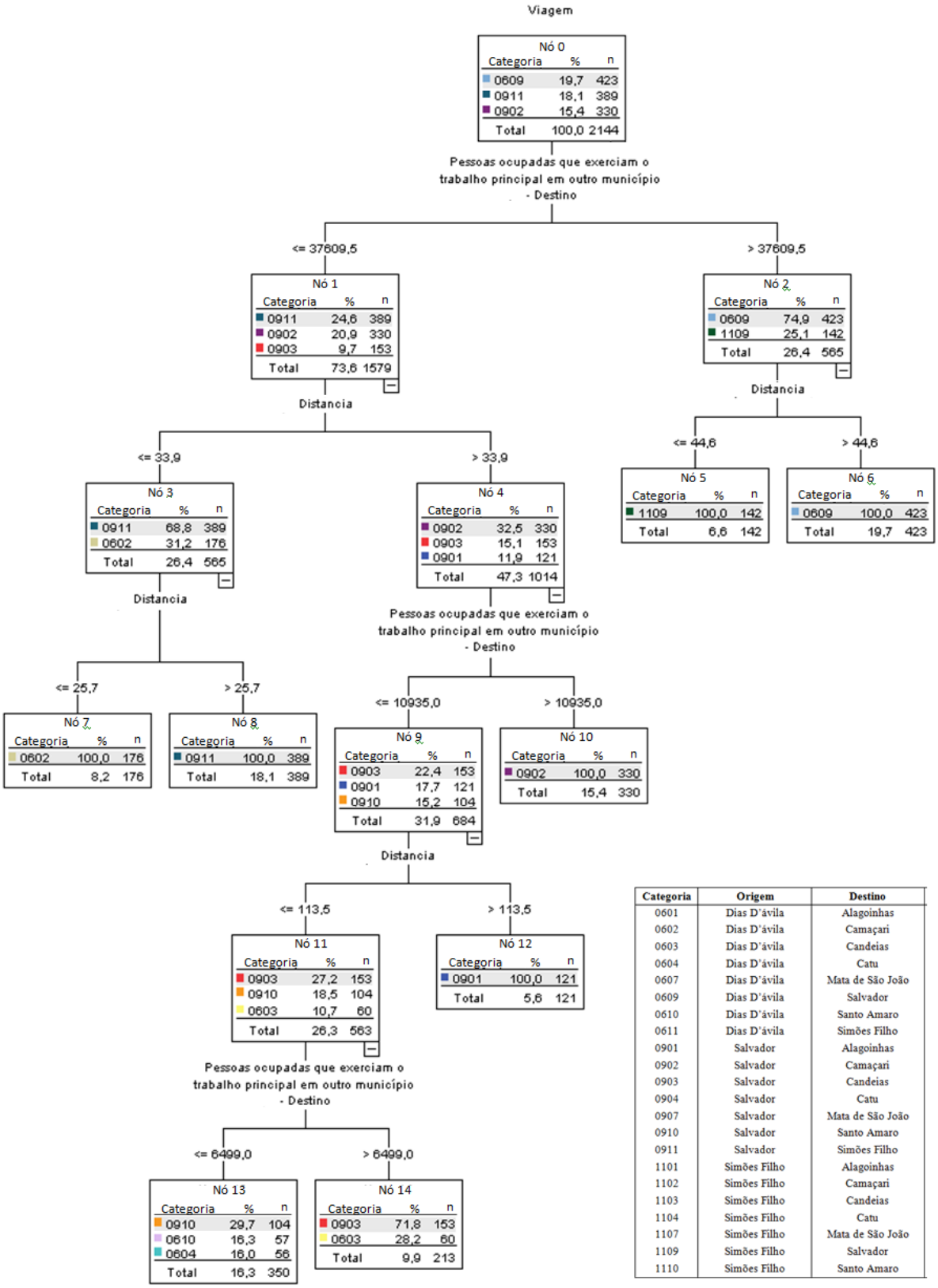


Figura 2 - Árvore de Decisão de treinamento com variáveis socioeconômicas desagregadas – Modelo 1



Categoria	Origem	Destino
0601	Dias D'ávila	Alagoinhas
0602	Dias D'ávila	Camaçari
0603	Dias D'ávila	Candeias
0604	Dias D'ávila	Catu
0607	Dias D'ávila	Mata de São João
0609	Dias D'ávila	Salvador
0610	Dias D'ávila	Santo Amaro
0611	Dias D'ávila	Simões Filho
0901	Salvador	Alagoinhas
0902	Salvador	Camaçari
0903	Salvador	Candeias
0904	Salvador	Catu
0907	Salvador	Mata de São João
0910	Salvador	Santo Amaro
0911	Salvador	Simões Filho
1101	Simões Filho	Alagoinhas
1102	Simões Filho	Camaçari
1103	Simões Filho	Candeias
1104	Simões Filho	Catu
1107	Simões Filho	Mata de São João
1109	Simões Filho	Salvador
1110	Simões Filho	Santo Amaro

Figura 3 - Árvore de Decisão de treinamento variáveis socioeconômicas desagregadas, variáveis sociodemográficas agregadas e distâncias de viagens – Modelo 2

Tabela 4 – Resumo dos resultados obtidos nos testes para cada modelo

Modelo	Média dos erros	Variância dos erros	Person entre observados e estimados
Modelo de AD: variáveis desagregadas (modelo 1)	91,09	28888,56	0,82
Modelo de AD: variáveis desagregadas e agregadas	27,82	3159,11	0,95
Modelo Gravitacional	37,06	3936,82	0,82

5. CONCLUSÕES

A partir do presente trabalho foi possível: (1) testar a adequabilidade do uso de técnica de Mineração de Dados para modelagem de distribuição de viagens intermunicipais; e (2) analisar a influência de variáveis desagregadas na escolha dos destinos de viagens intermunicipais.

O método utilizado envolvia a aplicação da técnica de Mineração de Dados em um banco desagregado, proveniente de uma Pesquisa Origem/Destino, realizada em onze cidades da Bahia. O intuito seria testar o poder preditivo de técnicas de Árvore de Decisão, considerando variáveis individuais e variáveis tradicionalmente utilizadas na modelagem de distribuição de viagens.

O modelo de Árvore de Decisão, que envolvia apenas variáveis socioeconômicas individuais (Modelo 1), teve um poder preditivo muito baixo (25% de acertos). Na etapa posterior, foram incorporadas ao Modelo 1, variáveis agregadas tradicionais. A adição de tal informação aumentou consideravelmente o percentual de acertos (86%), bem como diminuiu significativamente a média dos resíduos.

Finalmente, foi feita uma análise comparativa das matrizes Origem-Destino obtidas para as onze cidades estudadas, considerando valores observados e valores estimados pelos modelos 1 e 2 e pelo modelo gravitacional calibrado para variáveis de população e distância entre cidades (modelo 3).

Observou-se que o modelo gravitacional teve acurácia semelhante ao modelo de Árvore de Decisão que incorporava variáveis demográficas e distâncias de viagens. No entanto, este último ainda mostrou-se superior. Vale ressaltar ainda que, apesar dos bons resultados observados, o presente trabalho atesta que, para este estudo de caso, a técnica é adequada. Pode-se afirmar então que pode ser uma técnica alternativa às abordagens tradicionais, com vantagens relacionadas à sua fácil aplicabilidade, sem restrições de tipo de variáveis de entrada e suposições matemáticas rígidas. A principal desvantagem seria a sua característica não paramétrica que não permite testar, por exemplo, a significância dos coeficientes estimados das variáveis explicativas. Outra conclusão relevante seria a pouca importância de variáveis desagregadas quando comparadas às variáveis demográficas e de viagens na explicação do fenômeno.

AGRADECIMENTOS

O presente trabalho foi realizado com apoio do CNPq. Os autores também agradecem ao CETRAMA (Centro de Estudos de Transportes e Meio Ambiente – UFBA).

REFERÊNCIAS

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

Clark, L. A. e Pregibon, D., Tree-Based Models. *In Statistical Methods in S* (eds. J. M. Chambers and T. J. Hastie). AT&T Bell Laboratories and Wadsworth & Brooks/Cole (1992).

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

Instituto Brasileiro de Geografia e Estatística – IBGE, Censo Demográfico Brasileiro (2010). Disponível em < http://www.ibge.gov.br>. Acesso em 20 de agosto de 2014.

Keuleers, B. e Wets, G., Using association rules to identify patterns in activity diary data. IN: 80th Annual Meeting of Transportation Research Board, Washington, D.C. Compendium of Papers CD-ROM (2001).

Kitamura, R.; Mokhtarian, P.L. e Laidet, L., A micro-analysis of land use and travel in five neighborhoods in the San Francisco Bay Area. *Transportation* 24, 125–158 (1997).

Novaes, A. G., *Modelos em Planejamento Urbano, Regional e de Transportes*. São Paulo: Edgard Blücher. 290p (1986).

Ortúzar, J. D. e Willumsen, L. G., *Modelling Transport*. Londres: Wiley. 4ª ed. 586p (2011).

Pitombo, C. S. e Costa, A. S. G., Decision Tree application for modal choice. In: Panam 2014, 2014, Santander. Panam 2014 (2014).

Pitombo, C. S.; Kawamoto, E. e Sousa, A. J., An exploratory analysis of relationships between socioeconomic, land use, activity participation variables and travel patterns. *Transport Policy* (Oxford), v. 18, p. 347-357 (2011).

Pitombo, C. S.; Kawamoto, E. e Sousa, A. J. . 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 (2013).

Pitombo, C. S.; Guimaraes, H. S. Uma análise desagregada na Modelagem de distribuição de viagens intermunicipais. In: XXIX Anpet - Congresso de pesquisa e ensino em transportes, 2015, Ouro Preto (2015).

Quinlan, I.R., Learning Efficient Classification Procedures and their Application to Chess end-Games. *Machine Learning: An Artificial Intelligence Approach*, p. 463-482 (1983).

Rasouli, M. e Nikraz, H., Trip Distribution Modelling Using Neural Network. Australasian Transport Research Forum 2013 Proceedings 2 - 4 October 2013, Brisbane, Australia (2013).

Rocha, S. S. ; Pianucci, M. N. ; Pitombo, C. S. ; Cunha, A. L. B. N. . Uso de redes neurais para previsão de produção de viagens: uma análise agregada. In: XXIX Anpet - Congresso de Pesquisa e Ensino em Transportes, 2015, Ouro Preto. Congresso de Pesquisa e Ensino em Transportes (2015).

Strambi, O. e van de Bilt, K., Trip Generation Modeling Using CHAID, a Criterion-Based Segmentation Modeling Tool. *Transportation Research Record Journal of the Transportation Research Board*; 1645(1):24-31. DOI: 10.3141/1645-04 (1998).

Xie, C.; Jinyang, L. e Parkany, E., Work Travel Mode Choice Modeling with Data Mining: Decision Trees and Neural Networks. *Transportation Research Record: Journal of the Transportation Research Board*, volume 1854, pp 50-61. DOI: 10.3141/1854-06 (2007).