The Journal of Transport Literature
www.journal-of-transport-literature.org
ITPS International Transport Planning Society

Modelo logit binomial com componentes principais para estimação de preferência por modo de transporte motorizado

Anabele Lindner; Cira Souza Pitombo+
Departamento de Engenharia de Transportes, Escola de Engenharia de São Carlos/USP

Article Info

Palavras-chave:
Escolha modal
Análise em Componentes Principais
Logit Binomial

Submitted 6 May 2015;
received in revised form 9 Jul 2015;
accepted 12 Jul 2015.

Resumo

Este trabalho apresenta um método sequencial, envolvendo aplicação de Análise em Componentes Principais (ACP) e logit binomial para previsão de escolha por modo de transporte motorizado. A aplicação da ACP permite reduzir o banco de dados multicolinear a componentes não correlacionadas entre si. Tais components são utlizadas, posteriormente, como variáveis explicativas em modelos logísticos binomiais. Os dados utilizados para o desenvolvimento deste trabalho são provenientes da entrevista domiciliar da Pesquisa Origem-Destino de 2007, realizada na Região Metropolitana de São Paulo. Os modelos obtidos apresentaram bom poder preditivo e valores coerentes e significativos de parâmetros calibrados. Na etapa de validação, foram obtidas taxas de acertos variando entre 69% a 92%. Finalmente, o método proposto é razoável, sendo uma boa alternativa para o caso de dados multicolineares em métodos de regressão.

+ Corresponding author. Escola de Engenharia de São Carlos/USP, Departamento de Transportes. Avenida Trabalhador Sancarlense, Parque Arnold Schimidt. 13566590 - São Carlos, SP - Brasil.
E-mail address: cira@sc.usp.br.

Introdução

Este trabalho visa apresentar um método sequencial, envolvendo aplicação de Análise em Componentes Principais (ACP) e logit binomial para previsão de escolha por modo de transporte motorizado. A aplicação da ACP permite reduzir o banco de dados multicolinear a componentes não correlacionadas entre si. Tais componentes, extraídas pela ACP, são utilizadas como variáveis explicativas no modelo logit binomial. Desta forma, trata-se de uma abordagem exploratória-confirmatória que permite investigar preferências por modo de transporte motorizado. O método proposto é eficiente, sobretudo para o caso de banco de dados com multicolinearidade.

Técnicas de regressão múltipla são versáteis e poderosas. São aplicadas em uma infinidade de casos, onde se deseja encontrar uma relação entre uma única variável dependente e diversas variáveis independentes, com estimação de parâmetros a partir de diferentes critérios. Um estimador de um parâmetro ou um vetor de parâmetros desconhecidos é uma variável aleatória cujo valor pode ser calculado a partir de uma amostra. O vetor de parâmetros pode ser estimado por vários métodos, tais como Método dos Mínimos Quadrados (Regressão Linear Múltipla) e o Método da Máxima Verossimilhança (Regressão Logística – logit), que são os mais utilizados.

Multicolinearidade ocorre quando duas ou mais variáveis explicativas são muito correlacionadas entre si. Utilizando-se apenas modelos de regressão, torna-se difícil distinguir suas influências separadamente. Outra suposição do modelo de regressão é que nenhuma relação linear exata pode existir entre quaisquer covariáveis ou combinações lineares destas. Quando se viola esta hipótese têm-se o problema de multicolinearidade perfeita. Por outro lado, se as variáveis não estão correlacionadas entre si, denomina-se, este caso, ausência de multicolinearidade, sendo chamada de ortogonal a regressão com estas variáveis. O caso intermediário, muito comum em problemas reais, ocorre quando a correlação entre duas ou mais variáveis é alta, sendo esta situação chamada de alto grau de multicolinearidade.

Geralmente, a multicolinearidade não aumenta o poder preditivo de modelos de regressão, sendo uma tarefa usualmente difícil a seleção de variáveis explicativas multicolineares (Camminatiello e Lucadamo, 2010). Um número alto de variáveis explicativas e correlacionadas pode tornar os modelos de regressão mais redundantes do que realmente bons. Para evitar os problemas provocados pela multicolinearidade o método mais simples é a eliminação, do modelo completo, das variáveis com os coeficientes estatisticamente não significativos para encontrar o melhor subconjunto de variáveis independentes. Outra alternativa, proposta por Hoerl e Kennard (1970), é o método de regressão denominado de "Ridge", que tem o objetivo de melhorar a precisão dos parâmetros estimados, sem o termo constante, por padronizar as variáveis independentes. Neter et al. (1989), no entanto, colocam como principal limitação do modelo anterior, a impossibilidade de fazer inferências sobre seus parâmetros.

Outra forma de obter um modelo adequado, quando algumas variáveis independentes são muito correlacionadas, é a partir da técnica de Análise em Componentes Principais, que tem a vantagem de não descartar nenhuma variável explicativa. Alguns autores propuseram métodos diferentes, com perda de pouca informação a respeito da variância dos dados e redução significativa de variáveis independentes (Wold, 1985; Frank et al., 1993; Aguilera et al., 2006).

Camminatiello e Lucadamo (2010) propuseram o modelo de regressão logística multinomial para dados multicolineares, desenvolvendo, a partir de dados simulados, uma extensão do modelo Principal Components Logistc Regression (PCLR). Aguilera et al. (2006) também desenvolveram método de extração de componentes e posterior uso em regressão logística com dados simulados. Chen et al. (2010) aplicaram metodologia similar para estudo de comportamento relativo a viagens. Na literatura de modelagem de transportes, por exemplo, o uso de tal método para aprimorar estimativas de demanda, resolvendo o problema de multicolinearidade, não é tão trivial.

Este trabalho visa melhorar estimativas, relacionadas a escolhas discretas de transportes (modo de transporte) através do uso de componentes principais em modelos logísticos tradicionais. Na Seção 1, serão apresentados conceitos básicos relativos às técnicas utilizadas; na Seção 2, são descritos materiais e método; na Seção 3 são apresentados os resultados obtidos e discussões. Finalmente, as conclusões são apresentadas e, em seguida, as referências bibliográficas utilizadas.

1. Técnicas abordadas

1.1 Análise em Componentes Principais (ACP)

A análise em componentes principais (ACP) é uma técnica de análise fatorial, que analisa inter-relações em uma estrutura multivariada. A ACP permite a transformação linear ótima de um conjunto de variáveis originais intercorrelacionadas em um novo conjunto de variáveis não correlacionadas. Esse procedimento se dá por uma rotação espacial dos dados, de forma que sejam salientadas as feições através da mudança no sistema de coordenadas e redução da massa de dados, com menor perda possível da informação (Jolliffe, 2002).

O entendimento da ACP está atrelado à aplicação da álgebra linear básica, em que são avaliados os autovetores e autovalores de uma matriz de covariância ou uma matriz de correlações. Dada uma matriz (m x n), em que m é o número de observações e n é o número de variáveis, a matriz de correlações é dada por:

$$R = \begin{bmatrix} z_{11} & z_{12} & z_{13} & \cdots & z_{1n} \\ z_{21} & z_{22} & z_{23} & \cdots & z_{2n} \\ z_{31} & z_{32} & z_{33} & \cdots & z_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ z_{m1} & z_{m2} & z_{m3} & & z_{mn} \end{bmatrix}$$

$$z_{mn} = \frac{x_{mn} - \bar{x}_n}{\sigma_n} \quad (1)$$

Onde, x1 são os m valores observados para a variável 1 e xn são os m valores observados para a variável n.

A determinação das componentes principais é dada pelos autovetores (v). Os autovetores são calculados em função dos autovalores da matriz R. Sendo I a matriz identidade, os autovalores (λ) da matriz S são os escalares que satisfazem a equação característica:

$$det[R - \lambda I] = 0 \quad (2)$$

Para cada autovalor, é possível calcular o respectivo autovetor v, conforme Equação 3.

$$(R - \lambda I) \times v = 0 \quad (3)$$

No caso geral, a matriz de autovalores é uma matriz diagonal, onde o número de autovalores é igual à dimensão da matriz R (n). O resultado para os autovetores é equivalente à matriz quadrada R (n x n).

Um novo conjunto de variáveis (componentes W) pode ser obtido pela multiplicação dos coeficientes dos autovetores com o vetor de valores originais. Para isso, uma matriz quadrada A é composta usando os autovetores como colunas da matriz. As novas variáveis são combinações lineares das variáveis originais, calculadas conforme a Equação 4.

$$W = X \times A \quad (4)$$

onde, A é a matriz composta pelos autovetores e X é o vetor de observações original.

A componente principal é o arranjo que melhor representa a distribuição dos dados e a componente secundária é perpendicular a componente principal, e assim por diante. A importância de cada componente é avaliada por meio de sua contribuição, dada em porcentagem e calculada pela variância dividida pela variância total (autovalores). A fim de obter, de forma preliminar, o número de componentes a extrair e de reduzir a massa de dados, a matriz inicial não rotacionada (A) é pertinente. Contudo, na maioria dos casos, esta solução não fornecerá a interpretação mais adequada das variáveis. Em geral, a rotação é desejável, porque simplifica a estrutura, melhorando a interpretação e reduzindo ambiguidades que, frequentemente, acompanham componentes não rotacionadas (Hair et al., 2010). As rotações são caracterizadas em dois tipos: ortogonais e oblíquas. Neste trabalho a rotação ortogonal (Varimax) se mostrou conveniente.

1.2. Regressão Logística Múltipla (RLM)

Para prever escolhas discretas de modo de transporte, por exemplo, um determinado modo pode ser contrastado com uma opção alternativa, de maneira a ser transformado em uma probabilidade entre "0" e "1". Um dos modelos de escolha discreta mais utilizados na demanda por transportes é o logístico (Ortúzar e Willumsen, 2011). A Regressão Logística se diferencia da Regressão Linear, pois seus cálculos envolvem variáveis dependentes qualitativas, ao invés de quantitativas.

A regressão logística é caracterizada pela quantidade de valores a serem discretizados. Caso haja apenas a duas opções, a regressão logística é dita binomial, caso haja mais opções, a regressão é generalizada e nomeada multinomial. Estas regressões são também denominadas de Modelos Logit Binomial e Logit Multinomial. A Regressão Logística permite o uso de um modelo (curva em S) para prever a probabilidade π de um evento categórico. A modelagem da curva S é dada por uma transformação logística da probabilidade π, conforme a Equação 5, da função logística g(x).

$$g(x) = \ln\left(\frac{\pi}{1-\pi}\right) \quad (5)$$

Através da Equação 5, deriva-se a equação de calibração (Equação 6).

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (6)$$

Sendo x0=1, β0, β1,..., βn os coeficientes da equação de calibração e n o número de variáveis independentes. Em posse da equação de calibração e respectivos coeficientes, é simples calcular o valor de probabilidade π (Equação 7). A qualidade do ajuste pode ser mensurado por medidas estatísticas apropriadas, tais como: testes de regressão de Cox & Snell e Nagelkerke (Hair et al., 2010).

$$\pi = \frac{1}{1+e^{-g(x)}} \quad (7)$$

2. Materiais e Método

2.1. Banco de dados e área de estudo

Os dados utilizados neste trabalho são referentes à Pesquisa Origem/Destino (O/D) realizada na Região Metropolitana de São Paulo (RMSP) em 2007 pela Companhia do Metropolitano de São Paulo. Na Pesquisa O/D da RMSP de 2007 foram levantadas informações de 30 mil domicílios, escolhidos aleatoriamente e distribuídos em 39 municípios, das 460 Zonas de Tráfego, na RMSP.

Os dados da Pesquisa O/D foram manipulados de forma que cada registro estava relacionado a informações por domicílio. Ao final foram utilizados 18.733 domicílios na amostra final. Originalmente, a Pesquisa O/D indica os modos realizados para cada viagem. Neste trabalho, os modos de transporte foram classificados como: Transporte Individual Motorizado, Transporte não Motorizado e Transporte Público. O Transporte Individual Motorizado inclui viagens em que o indivíduo dirige automóvel ou motocicleta. O Transporte não Motorizado refere-se ao uso de bicicleta e modo a pé. O Transporte Público inclui viagens por ônibus do município de São Paulo, ônibus de outros municípios, ônibus metropolitano, microônibus do município de São Paulo, microônibus de outros municípios, microônibus metropolitano, metrô e trem.

As variáveis originais, utilizadas para ACP foram: (1) coordenada geográfica x (UTM); (2) coordenada geográfica y (UTM); (3) Número de pessoas no domicílio; (4) Critério de Renda (variando de 1 a 8 – sendo 1 para maior renda e 8 para menor renda); (5) Quantidade de automóveis no domicílio; (6) Quantidade de motocicletas no domicílio; (7) Número total de viagens no domicílio; (8) Distância total de viagens realizadas no domicílio; (9) Distância de viagens realizadas por transporte individual motorizado no domicílio; (10) Distância de viagens realizadas por transporte não motorizado no domicílio; (11) Distância de viagens realizadas por transporte público no domicílio; (12) Quantidade de bicicletas no domicílio. A variável objeto do estudo (dicotômica) representa a preferência da maior parte das viagens motorizadas no domicílio: 0 – Transporte individual motorizado; 1 – Transporte Público.

2.2. Software

Este trabalho utilizou o software IBM SPSS Statistics 22 para a ACP e RLM, respectivamente.

2.3. Método

O método proposto para este trabalho é sequencial, formado por duas etapas principais: Aplicação da ACP e Calibração e Validação dos modelos logit binomial. A primeira etapa metodológica (ACP) envolve a transformação linear ótima do conjunto de dados das variáveis independentes em componentes. Para a sua aplicação, seguiu-se as seguintes etapas (Pitombo e Martins, 2014): (1) Cálculo da matriz de correlação das variáveis em estudo; (2) Extração e rotação das componentes, observando a variabilidade dos dados explicada; (3) Interpretação e nomenclatura para cada uma das componentes observando a contribuição (autovetores) de cada uma das variáveis.

A etapa de regressão logística utilizou as componentes, como variáveis independentes, e a variável preferência por modo motorizado como variável dependente. A calibração do modelo logit foi realizada com 70% da amostra, enquanto a validação foi realizada com 30% da amostra final. Vale ressaltar que foram calibradas equações logísticas de uma a n componentes principais, sendo n igual ao número de variáveis originais.

3. Resultados e discussões

3.1. Etapa 1: Aplicação da ACP

Como esperado, existem diversas variáveis correlacionadas entre si, fato que justifica a aplicação da ACP. Os valores de correlação mais altos encontrados foram entre (1) Quantidade de automóveis e critério de renda; (2) Total de viagens domiciliares e número de pessoas no domicílio; (3) Distância total de viagem e Distância de viagens realizadas por transporte público no domicílio.

Para extração das componentes mais significativas foram considerados os valores dos autovalores iguais ou superiores a um. Assim, foram extraídas cinco componentes com 68,7% da variabilidade dos dados explicada. Para finalidade de testes em modelos logísticos posteriormente, foram consideradas até 12 componentes, com 100% da variância explicada. A Tabela 1 descreve a variância explicada por cada uma das componentes.

Tabela 1: Proporção de variância explicada pelas componentes.

Componente	1	2	3	4	5	6	7	8	9	10	11	12
Autovalor	2,8	2,3	1,1	1,1	1,0	0,9	0,8	0,7	0,7	0,4	0,2	0,1
% Variância	23,1	19,1	9,3	9,0	8,2	7,6	6,9	6,1	5,4	3,1	1,9	0,4
% Acum.	23,1	42,1	51,5	60,4	68,7	76,2	83,1	89,2	94,6	97,7	99,6	100,0

Para se interpretar adequadamente e também dar nome às componentes, é necessário analisar os valores das cargas fatoriais (autovetores) das variáveis, ou seja, deve ser analisada a contribuição (negativa ou positiva) de cada uma das variáveis em relação a cada uma das doze componentes. Após o método de rotação Varimax, a Tabela 2 apresenta as variáveis e os valores dos autovetores em cada componente. Em seguida é apresentada a nomenclatura proposta.

(1) Componente 1: Distância de viagens realizadas por transporte público no domicílio; (2) Componente 2: Distâncias de viagens realizadas por modo individual motorizado no domicílio; (3) Renda; (4) Distâncias de viagens realizadas por modo

não motorizado no domicílio; (5) Quantidade de bicicletas no domicílio; (6) Número de pessoas no domicílio; (7) Quantidade de motocicletas no domicílio; (8) Latitude; (9) Longitude; (10) Total de viagens domiciliares; (11) Quantidade de automóveis no domicílio; (12) Distâncias totais.

Tabela 2: Matriz de autovetores das componentes

Variável	Componente											
	1	2	3	4	5	6	7	8	9	10	11	12
Latidute	0,1	0,0	0,1	0,0	0,0	0,0	0,0	**1,0**	0,1	0,0	0,0	0,0
Longitude	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,1	**1,0**	0,0	0,0	0,0
N Pessoas	0,2	0,0	0,0	0,1	0,1	**0,9**	0,0	0,0	0,0	0,3	0,0	0,0
Renda	0,1	-0,2	**0,9**	0,1	-0,1	0,0	0,0	0,1	0,0	-0,1	-0,3	0,0
Nmotocicletas	0,0	0,1	0,0	0,0	0,1	0,0	**1,0**	0,0	0,0	0,0	0,0	0,0
Nautomóveis	-0,1	0,3	**-0,5**	-0,1	0,1	0,1	0,1	0,0	0,0	0,1	**0,8**	0,0
Nbicicletas	0,0	0,1	-0,1	0,0	**1,0**	0,1	0,1	0,0	0,0	0,1	0,1	0,0
Total de viagens	0,2	0,2	-0,1	0,1	0,1	**0,3**	0,0	0,0	0,0	**0,9**	0,1	0,0
Distância total	**0,8**	**0,5**	0,0	0,2	0,1	0,1	0,0	0,0	0,0	0,2	0,0	**0,2**
Distância Transp Público	**1,0**	-0,2	0,1	0,0	0,0	0,1	0,0	0,1	0,0	0,0	-0,1	**-0,1**
Distância individual motorizado	0,0	**1,0**	-0,2	0,0	0,1	0,0	0,1	0,0	0,0	0,1	0,2	0,0
Distância não motorizado	0,1	0,0	0,1	**1,0**	0,0	0,1	0,0	0,0	0,0	0,1	0,0	0,0

3.2. Etapa 2: Modelos logit binomial

Com 70% da amostra final de domicílios, foram gerados doze modelos logísticos com a inclusão de componentes por modelo. Os valores de parâmetros calibrados para cada variável independente foram considerados significativos. Além disso, a magnitude dos valores calibrados, bem como o sinal, em todos os casos, são coerentes. O poder preditivo do modelo (pseudo R2) melhora bastante quando são utilizadas 5 componentes apenas, para o caso de se considerar as componentes com autovalores maiores ou iguais a uma unidade. Quanto maior a variância explicada utilizada para calibração dos modelos, maior incremento da sua acurácia. Desta forma, para o caso do uso de doze componentes (100% da variância explicada), o pseudo R2 chega a um valor equivalente a 0,87. Uma das grandes vantagens do método é a garantia de que as componentes utilizadas (novas variáveis explicativas) não são correlacionadas entre si. A Tabela 3 apresenta principais resultados dos modelos logísticos.

Tabela 3: Síntese dos resultados dos modelos logísticos

	Variância Explicada	Modelos	Pseudo R^2	
			Cox & Snell	Nagelkerke
1	23,1	1.4CP1+0.415	0,17	0,24
2	42,1	2.2CP1-2.4CP2+0.47	0,42	0,56
3	51,46	3.2CP1-3.2CP2+1.7CP3+0,7	0,55	0,73
4	60,44	3.2CP1-3.2CP2+1.7CP3+0,17CP4+0,7	0,55	0,73
5	68,65	3.2CP1-3.2CP2+1.7CP3+0,17CP4-0.21CP5+0,7	0,55	0,74
6	76,20	3.2CP1-3.2CP2+1.7CP3+0,18CP4-0.22CP5+0.42CP6+0,8	0,56	0,75
7	83,13	3.5CP1-3.4CP2+1.8CP3+0,19CP4-0.23CP5+0.44CP6-0.48CP7+0,8	0,56	0,76
8	89,21	3.6CP1-3.5CP2+1.8CP3+0,19CP4-0.23CP5+0.44CP6-0.48CP7+0.21CP8+0,8	0,57	0,76
9	94,61	3.6CP1-3.5CP2+1.8CP3+0,19CP4-0.24P5+0.44CP6-0.48CP7+0.21CP8+0.06CP9+0,8	0,57	0,76
10	97,72	3.7CP1-3.6CP2+1.9CP3+0,26CP4-0.25P5+0.46CP6-0.49CP7+0.24CP8+0.06CP9-0.62CP10+0,8	0,58	0,77
11	99,62	4.9CP1-4.8CP2+2.4CP3+0,3CP4-0.39P5+0.63CP6-0.66CP7+0.36CP8+0.08CP9-0.91CP10-2.02CP11+1,0	0,65	0,87
12	100,00	4.2CP1-4.8CP2+2.4CP3+0,3CP4-0.39P5+0.67CP6-0.67CP7+0.39CP8+0.09CP9-0.97CP10-2.06CP11-0,57CP12+1,1	0,65	0,87

Modelos	% de variância	Taxa de acertos
1	23.06%	69.4%
2	42.13%	79.0%
3	51.46%	87.0%
4	60.44%	87.0%
5	68.65%	87.0%
6	76.20%	87.3%
7	83.13%	87.3%
8	89.21%	87.2%
9	94.61%	87.4%
10	97.72%	87.5%
11	99.62%	92.5%
12	100.00%	92.3%

Tabela 4: Taxa de acertos da validação

Com 30% da amostra de domicílios restante, foi feita a validação dos doze modelos calibrados anteriormente. Após a obtenção de probabilidades e categorias modais estimadas, foram calculadas as taxas de acertos para todos os modelos. A Tabela 4 indica os resultados de taxa de acertos para a validação. O modelo que utiliza apenas a primeira componente

(aproximadamente 23% da variância explicada) acerta 69,4%. Já o modelo que utiliza toda a variância explicada, através de doze componentes, tem 92,3% de acertos. A validação mostrou alta taxa de acertos para todos os modelos.

Conclusões

Este trabalho apresentou um método sequencial, envolvendo aplicação de Análise em Componentes Principais (ACP) e logit binomial para previsão de escolha por modo de transporte motorizado. Na primeira etapa foi realizada a extração das componentes a partir das variáveis originais socioeconômicas domiciliares, variáveis de viagens e locacionais. Em seguida, foi utilizado o método de rotação varimax para facilitar a interpretação das componentes. A partir das componentes extraídas na primeira etapa, são calibrados e validados n modelos (sendo n igual ao número total de variáveis originais e 100% da variância total explicada). Os modelos calibrados tiveram bom poder preditivo, com alta taxa de acertos, além disso, os parâmetros obtidos foram significativos e coerentes.

As componentes relativas às distâncias percorridas por cada modo motorizado foram aquelas mais significativas dos modelos. O parâmetro positivo para a componente relativa à Renda (CP3), por exemplo, faz sentido, pois, em escalara likert, a renda decresce de 1 para 8. Assim, o aumento numérico da componente Renda significa diminuição do poder aquisitivo, portanto maior probabilidade para escolhas do transporte coletivo. Já componente 11, que representa a quantidade de automóveis no domicílio, possui valor negativo no modelo calibrado, com relação inversa à escolha do transporte coletivo.

A componente que representa número total de viagens domiciliares (CP10) também possui valor negativo no modelo calibrado, com relação inversa à escolha do transporte coletivo. Sabe-se, portanto, que o automóvel viabiliza maior número de viagens domiciliares. Desta forma, pode-se afirmar que os resultados encontrados com os modelos logísticos binomiais com componentes principais corroboram com diversos resultados encontrados na literatura de escolha modal (Ahern e Tapley, 2008; Bhat, 1997; Brownstone et al., 2000).

Finalmente, sabe-se que a multicolinearidade é um problema para o caso de modelos de regressão. Este trabalho apresenta uma forma de mitigar essa dificuldade, comum na modelagem de demanda por transportes, buscando melhorar estimativas relacionadas a escolhas discretas de transportes através do uso de componentes principais em modelos logísticos tradicionais.

Agradecimentos
Ao CNPq e CAPES pelo suporte financeiro fornecido à presente pesquisa.

Referências

Aguilera, A.M.; Escabias, M.; Valderrama, M.J. (2006) Using principal components for estimating logistic regression with high-dimensional multicollinear data. Computational Statistics & Data Analysis 50, pp. 1905 – 1924. DOI:10.1016/j.csda.2005.03.011

Ahern, A.; Tapley, N. (2008) The use of stated preference techniques to model modal choices on interurban trips in Ireland. Transportation Research Part A: Policy and Practice, vol 42, 1, 15-27. DOI: 10.1016/j.tra.2007.06.005.

Bhat, C. (1997). Work travel mode choice and number of non-work commute stops. Transpor-tation Research Part B: Methodological 31 (1), pp. 41–54. DOI: 10.1016/S0191-2615(96)00016-1

Brownstone, D.; Bunch, D.; Train, K. (2000) Joint mixed logit models of stated and revealed preferences for alternative-fuel vehicles. Transportation Research Part B: Methodological, vol 34, 5, pp 315-338. DOI: 10.1016/S0191-2615(99)00031-4.

Camminatiello, I.; Lucadamo, A. (2010) Estimating Multinomial Logit Model with Multicollinear Data. Asian Journal of Mathematics & Statistics, vol 3,2, 93-101. DOI: 10.3923/ajms.2010.93.101.

Chen, Z. J., Cheng, L., Deng, H. N., Zhang, J. K (2010) Analyzing Residential Travel Mode Choice Based on Principal Component Analysis. Proceedings of the 10th International Conference of Chinese Transportation Professionals. 2010.

Frank, I.E.; Friedman, J.H.; Wold, S.; Hastie, T.; Mallows, C. (1993) A statistical view of some chemometrics regression tools. Technometrics, 35, pp 109-148.

Hair Jr, J. F.; Black, W. C.; Babin, B. J.; Anderson, R. E. (2010) Multivariate Data Analysis. Prentice Hall. 7a ed. 785 p.

Hoerl, A.; Kennard, R.W. (1970) Biased Estimation for Nonorthogonal Problems. Technometrics, Vol. 12, No. 1 (Feb., 1970), pp. 55-67.

Jolliffe, I. T. (2002) Principal Component Analysis. 2a ed. Springer 2002. 518p.

Neter, J.; Wasserman, W.; Kutner, M.H. (1989). Applied Linear regression Models. 2nd Edition. Irwin Homewood IL.

Pitombo, C. S., Gomes, M. M. (2014). Study of Work-Travel Related Behavior Using Principal Component Analysis. Open Journal of Statistics,4(11), 889.

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

Wold, H. (1985) Partial Least Squares, Encyclopedia of statistical sciences, vol 6. Pp 581-591.

Abstract

This paper presents a sequential method involving application of Principal Component Analysis (PCA) and binomial logit for forecasting of motorized travel mode choice. The application of ACP reduces the multicolinear database into uncorrelated components. Such components are used later as explanatory variables in binomial logit models. The data used are from the Origin-Destination Survey carried out in 2007 in São Paulo Metropolitan Area. The gotten models showed good accuracy and consistent and significant calibrated parameters. In the validation step, the hit rates were obtained ranging from 69% to 92%. Finally, the proposed method is reasonable to be a good alternative for the case of multicollinear data used in regression methods.

Key words: travel mode choice; Principal Component Analysis; Logistc regression methods.