INVESTIGAÇÃO DE TÉCNICAS PARA A MODELAGEM DA DEMANDA POR TRANSPORTE PÚBLICO

Samuel de França Marques
Universidade de São Paulo
samuelmarques@usp.br

Cira Souza Pitombo
Universidade de São Paulo
cirapitombo@usp.br

INVESTIGAÇÃO DE TÉCNICAS PARA A MODELAGEM DA DEMANDA POR TRANSPORTE PÚBLICO

S. F. Marques e C. S. Pitombo

RESUMO

No intuito de possibilitar o adequado ajuste da oferta frente a variações na demanda por transportes, a modelagem de dados de viagens evoluiu gradativamente até que conseguisse abranger todas as peculiaridades inerentes a essa variável. Entretanto, percebe-se ainda uma carência de trabalhos que mostrem os ganhos proporcionados quando se considera o comportamento assimétrico dos dados de viagens urbanas em seu processo de modelagem, tanto na calibração de modelos, quanto validação dos mesmos. Dessa forma, o objetivo desse trabalho é comparar a modelagem da demanda pelo método clássico de regressão linear com aqueles em que se considera a não normalidade dos dados de viagem: modelos com transformações logarítmicas e regressões de Poisson e Binomial Negativa. Resultados obtidos a partir da modelagem das viagens diárias produzidas por transporte público, por bairro de Palmas-TO (Brasil), confirmaram que as últimas técnicas realmente conseguem alcançar um desempenho superior ao da regressão linear tradicional.

1 INTRODUÇÃO E *BACKGROUND*

O sistema de transporte e o ordenamento territorial são temas cuja associação é constantemente observada nas pesquisas em prol do desenvolvimento sustentável de regiões urbanas. Não são poucos os trabalhos que buscam prever o efeito de diversas variáveis sobre a geração de viagens em determinado local (Chakour e Eluru, 2013, 2016; Choi *et al*., 2012; Chow *et al*., 2006; Chu, 2004; Ewing *et al*., 2014). Haja vista que o planejamento de transportes repousa justamente na modelagem da demanda, o surgimento de técnicas que conseguissem capturar, de forma exata e precisa, a influência de variáveis socioeconômicas, da forma urbana e do sistema de transporte sobre a demanda se tornou inevitável.

Entretanto, as variáveis representativas da demanda por transportes possuem características bastante peculiares que, inicialmente, por simplificação ou inexistência da tecnologia computacional necessária, foram negligenciadas em sua modelagem. Uma dessas particularidades é que normalmente as observações de viagens são disponibilizadas na forma de contagens, o que implica distribuições de probabilidade assimétricas, isto é, distribuições não normais para a variável resposta.

Tendo em vista que as técnicas de regressão mais tradicionais não abrangem, em seu processo de calibração, tais características intrínsecas à demanda por transportes, a evolução da modelagem dessa variável permitiu que tais implicações fossem paulatinamente agregadas ao procedimento de estimação dos parâmetros do modelo. Por

exemplo, Thompson (1997) empregou um modelo direto multiplicativo (com transformações logarítmicas em ambos os lados da equação) para estimar o potencial de atração e produção de viagens entre pares de setores censitários. Tal formulação teve como forma funcional uma regressão de Poisson, na qual a variável dependente foi o número de viagens (total da semana) por transporte público entre os setores. O autor não discutiu, de forma aprofundada/explícita, a aderência dos dados ao modelo resultante da regressão, porém ponderou que as variáveis explanatórias foram significativas e tiveram sinais coerentes com o esperado.

A regressão de Poisson também foi empregada por Choi *et al*. (2012) para modelar o número de embarques (média dos dias da semana) de estação a estação em três períodos diferentes do dia: pico da manhã, meio do dia (vale) e pico da noite. Foram utilizados dois modelos, um multiplicativo, outro de Poisson. Ambos consideraram as mesmas variáveis explanatórias, porém, de uma forma geral, o modelo multiplicativo (com transformação log-log) demonstrou um ajuste ao banco de dados ligeiramente melhor que o de Poisson em termos da estatística F, R^2 ajustado e quantidade de variáveis estatisticamente significativas.

Em modelagem da demanda por transportes ao nível de ponto de parada, Chu (2004) utilizou a regressão de Poisson para relacionar o número de embarques em pontos de parada a seis categorias de covariáveis. O modelo teve um bom desempenho visto que conseguiu capturar a influência das seis categorias de covariáveis na produção de viagens por transporte público com todas elas tendo os sinais esperados. Já na regressão linear tradicional, os sinais de várias variáveis independentes não corresponderam ao esperado e diversas variáveis não passaram no teste *t*.

Por fim, Ewing *et al*. (2014) se valeram da *Multilevel Modelling* para modelar cinco variáveis de demanda: viagens de carro, a pé, por bicicleta, por transporte público e quantidade de milhas percorridas por veículo (VMT, do inglês *Vehicle Miles Travelled*) no âmbito domiciliar, mas considerando também a influência de características regionais sobre as residências localizadas em uma mesma região. Tendo em vista a ocorrência de sobredispersão nos dados de contagens, a regressão binomial Negativa foi aplicada para modelar o número de viagens por domicílio por cada um dos modos citados anteriormente, mas apenas para aquelas residências em que o modo era utilizado. Excepcionalmente para a VMT, uma variável contínua, os autores empregaram uma regressão linear do logaritmo natural da VMT, dada a sua distribuição não normal e bastante inclinada para a direita. Os autores exaltaram as inúmeras aplicações possíveis dos modelos obtidos, indicando a coerência dos sinais verificados nos coeficientes estimados e sua respectiva significância.

Tendo em vista a natureza discreta geralmente associada às variáveis de demanda por transportes, não é incomum encontrar estudos cujos autores aplicaram transformações ou utilizaram modelos para dados de contagem no processo de modelagem da demanda. Entretanto, percebe-se ainda uma carência de trabalhos que mostrem os ganhos proporcionados quando se considera o comportamento assimétrico dos dados de viagens urbanas em seu processo de modelagem tanto na calibração quanto validação, abordagem ausente nos estudos supracitados. Além disso, percebe-se que, no Brasil, o emprego de regressões apropriadas para variáveis discretas ainda se concentra apenas na modelagem de acidentes (Cunto *et al*., 2012; Gomes *et al*., 2015; Gomes *et al*., 2016). Dessa forma, o objetivo desse trabalho reside em comparar a modelagem da demanda por transportes pelo método clássico de regressão linear com aqueles em que se considera a não normalidade

dos dados de viagem, confrontando as medidas de desempenho obtidas para tais categorias de modelos. A aplicação do procedimento metodológico é realizada para a cidade de Palmas (Tocantins, Brasil).

2 TÉCNICAS PARA A MODELAGEM DA DEMANDA POR TRANSPORTES

Tradicionalmente, os modelos mais usuais para modelagem da demanda por viagens são baseados na regressão linear múltipla (OLS, do inglês *Ordinary Least Squares*). Apesar de a estrutura de um modelo linear ser relativamente simples, seu processo de calibração e os testes de hipóteses associados a ele exigem o cumprimento de rigorosas suposições por parte do banco de dados utilizado, tais como: normalidade, linearidade e homocedasticidade. Entretanto, no que tange à modelagem da demanda por transportes, as variáveis de interesse usualmente desrespeitam os critérios listados acima, pois normalmente tratam-se de dados discretos em que a variância não é constante, as observações são dependentes entre si e a distribuição de probabilidade da variável resposta possui cauda fortemente alongada para a direita, ou seja, assimétrica.

Para contornar a violação das suposições que regem a regressão linear tradicional, a utilização de transformações nas variáveis surge como uma alternativa promissora. De acordo com Myers *et al*. (2012), transformações podem ser aplicadas para três propósitos diferentes: estabilizar a variância da variável resposta, tornar a distribuição da variável dependente mais próxima da distribuição normal e melhorar o ajuste do modelo ao banco de dados considerado. Se a variância dos dados varia com a média dos valores observados, a utilização de transformações na variável resposta permitiria, dessa forma, uma redução de escala na magnitude dos dados que levaria à constância de sua variância (Bartlett, 1947). Dessa forma, a regressão OLS poderia ser satisfatoriamente aplicada à modelagem da variável de interesse transformada, que se supõe possuir distribuição de probabilidade próxima da normal.

Os primeiros esforços para contabilizar, no processo de modelagem de dados discretos, a real distribuição de probabilidade que rege o comportamento de tais observações remontam ao surgimento dos modelos lineares generalizados (GLM, do inglês *Generalized Linear Models*). A estrutura dessa nova classe de regressões pode ser representada da seguinte maneira:

$$\eta(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon \quad (1)$$

em que μ é o valor esperado da variável resposta ou dependente, X_k são as variáveis explicativas, β_{k+1} representam os parâmetros a serem calibrados, ε trata-se dos resíduos do modelo e η refere-se à uma função cujo objetivo é ligar o componente aleatório do modelo, variável resposta, ao componente sistemático, a estrutura linear de covariáveis (Cordeiro e Demétrio, 2008; Myers *et al*., 2012; Nelder e Baker, 2004).

As distribuições de probabilidade apropriadas para a modelagem da demanda por transportes são a distribuição de Poisson e a Binomial Negativa, que refletem o comportamento de dados de contagem (Cordeiro e Demétrio, 2008). De forma a permitir a comparação de resultados entre modelos obtidos a partir dessas duas distribuições, utiliza-se normalmente a mesma função de ligação, $\eta(\mu) = \ln \mu$, tanto para a regressão de Poisson quanto para a Binomial Negativa. De fato, uma das principais diferenças entre as duas distribuições reside em sua função de variância: enquanto a variância da distribuição de

Poisson equivale à média das observações (Equação (2)), a variância da Binomial Negativa se apresenta como uma função de segundo grau de sua média (Equação (3)), acrescentando à estrutura do termo quadrático o parâmetro *k*, também conhecido como parâmetro de dispersão (Hilbe, 2014).

$$V(\mu) = \mu \quad (2)$$

$$V(\mu) = \mu + k\mu^2 \quad (3)$$

A função de variância da distribuição binomial Negativa faz com que ela consiga modelar variáveis que apresentam o fenômeno da sobredispersão, que ocorre quando a variância dos dados supera a sua média. Dessa forma, quanto maior o valor do parâmetro de dispersão, mais conveniente se torna a utilização da Binomial Negativa em detrimento da regressão de Poisson. Porém, se *k* for igual a zero, então a dispersão do banco de dados pode ser devidamente controlada pela distribuição de Poisson.

A seção a seguir detalha os procedimentos realizados para a comparação do desempenho de GLM's que assumem para a variável resposta distribuições de probabilidade discretas com modelos nos quais se supõe a distribuição normal para os dados de viagens em sua forma bruta ou transformados a partir de seu logaritmo natural.

3 MATERIAIS E MÉTODO

O banco de dados utilizado para a calibração dos modelos de demanda refere-se às unidades geográficas básicas da cidade de Palmas, capital do estado do Tocantins, Brasil. Tais unidades são quadras de tamanho aproximadamente regular (700m x 700m) na área pertencente ao Plano Diretor original; e bairros de diferentes dimensões na região de expansão do perímetro urbano, extremo sul da cidade (Oliveira *et al*., 2014). Foram tabuladas informações socioeconômicas e demográficas bem como dados acerca do sistema de transporte da cidade. A Tabela 1 relaciona todas as variáveis contidas no banco de dados à fonte original de onde foram coletadas ou que contribuiu à sua obtenção.

Tabela 1 Variáveis contidas no banco de dados

Tipo de variáveis	Informações disponíveis	Fonte consultada
Demográficas e de uso do solo	População em 2010	Marques (2016)
	Densidade populacional bruta em 2010 (hab/ha)	
	Densidade de empregos (un/ha)	
	Índice de entropia	
Socioeconômicas	Renda per capita média em 2010	Mendes (2014)
Sistema de transporte	Viagens produzidas	Mendes (2014)
	Extensão do sistema viário (m)	Palmas (2016a)
	Quantidade de linhas de ônibus	SI Digital (2016)
	Acessibilidade à rede de transporte público	Palmas (2017)

O indicador de acessibilidade à rede de transporte público, considerado nesse trabalho, é uma medida da porcentagem de área em cada bairro/quadra que contém os pontos cuja distância ao ponto de parada mais próximo não excede 300m, valor equivalente a um nível de serviço regular de acordo com Ferraz e Torres (2004). O cálculo desse indicador teve

como base a malha digital das quadras/bairros de Palmas (Palmas, 2016b) e os pontos de parada do transporte público da cidade (Palmas, 2017).

O índice de entropia (I_{ent}), computado por Marques (2016) a partir da equação disponível em Song *et al*. (2013), busca refletir o quão bem diversificado o uso do solo se encontra em determinada unidade geográfica. Ele resulta em seu valor mínimo, 0, para a condição em que há uma única categoria de uso do solo, e no máximo, 1, para quando todas as categorias de uso do solo estão presentes em uma quadra/bairro ocupando a mesma porcentagem de área. Para o cálculo do índice de entropia correspondente a cada quadra/bairro de Palmas, foram consideradas seis categorias de uso do solo, a saber: residencial, comercial, serviços, industrial, institucional e outros (Marques, 2016).

A quantidade de viagens diárias produzidas por transporte público para cada quadra/bairro de Palmas foi levantada a partir do trabalho de Mendes (2014) por ocasião de uma pesquisa embarcada em 33 linhas de ônibus da capital realizada nos dias 4, 5 e 6 de junho de 2014. Por sua vez, o número de linhas cujos itinerários possuem trecho em comum com pelo menos uma das faces da quadra/bairro foi obtido por meio da sobreposição da malha digital das quadras/bairros de Palmas com a malha de linhas de ônibus (SI Digital, 2016).

Tendo em vista que a variável de demanda a ser modelada é a quantidade de viagens diárias produzidas por quadra/bairro de Palmas (VP), buscou-se encontrar, no banco de dados disponível, possíveis preditores lineares que melhor explicassem o comportamento da variável dependente definida. Para tanto, foram calculados os coeficientes de correlação linear de *Pearson* bem como sua respectiva probabilidade de significância.

Determinadas as potenciais covariáveis a serem utilizadas para a modelagem da variável de demanda por transporte público, aplicou-se a técnica de regressão linear múltipla com seleção de variáveis pelo método *Stepwise* no intuito de se obter o modelo com melhor ajuste ao banco de dados. Após definidas as variáveis explanatórias que geravam o melhor modelo linear, utilizou-se essas mesmas covariáveis (ou sua correspondente transformada pelo logaritmo natural) para simular três novos tipos de modelos de regressão: o primeiro aplicando-se a transformação ln(VP+1), sugerida por Bartlett (1947) como apropriada para variáveis discretas com excesso de zeros, aos dados de viagens produzidas; o segundo considerando-se a distribuição de Poisson para a variável resposta e o terceiro com distribuição Binomial Negativa e parâmetro de dispersão estimado no processo de calibração. Além disso, empregou-se nos dois modelos de contagem a função de ligação logarítmica. Cabe ressaltar, ainda, que em cada categoria foram gerados vários modelos distintos, que se diferenciavam apenas pelas variáveis explicativas mantidas, a fim de se encontrar a combinação de covariáveis que proporcionava o melhor ajuste ao banco de dados. Nessa etapa, acrescentaram-se à lista de preditores as versões das variáveis população e sistema viário transformadas por seu respectivo logaritmo natural, haja vista a assimetria positiva também demonstrada por esses atributos.

Para validar os modelos gerados pelas simulações, foram reservadas 50 observações (30%) do banco de dados original e utilizaram-se somente 111 pontos na calibração das equações de regressão. Aos valores estimados pelo melhor modelo linear tradicional, logarítmico e generalizado foram aplicadas algumas das medidas de desempenho sugeridas por Hollander e Liu (2008) a fim de se avaliar o poder preditivo das equações obtidas. A seleção aleatória das duas sub-amostras bem como todas as análises foram realizadas por meio do *software* estatístico IBM SPSS 24.0. Os resultados alcançados estão apresentados

na seção 4.

4 RESULTADOS E DISCUSSÕES

A Tabela 2 consolida as principais medidas descritivas para as variáveis disponíveis no banco de dados.

Tabela 2 Medidas descritivas das variáveis contidas no banco de dados

Atributo\Medida descritiva	Média	Desvio padrão	Mín	25%	50%	75%	Máx
Viagens produzidas	580,96	1.496,35	0,00	0,00	95,00	595,00	12.378,00
População	1.087,27	1.838,58	0,00	13,00	491,00	1.506,50	16.888,00
Densidade populacional bruta	20,14	21,30	0,00	0,61	14,07	33,17	80,86
Densidade de empregos	14,53	49,63	0,00	0,48	3,83	9,94	589,82
Índice de entropia	0,25	0,17	0,00	0,06	0,28	0,39	0,56
Renda per capita média	1.496,71	1.053,76	0,00	765,61	1.339,75	1.910,18	5.472,29
Sistema viário	8.330,85	20.487,62	60,40	2.592,52	5.095,58	8.906,68	245.732,13
Quantidade de linhas	6,88	7,04	0,00	2,00	5,00	9,00	42,00
Acessibilidade	0,73	0,29	0,00	0,55	0,82	0,97	1,26

É importante notar que a maioria dessas variáveis, incluindo a de interesse, viagens produzidas, demostra uma assimetria bastante proeminente, resultando em um valor para a mediana consideravelmente inferior ao da média. Dessa forma, é possível afirmar que a distribuição de tais atributos possui cauda fortemente inclinada para a direita, o que contradiz a condição de normalidade suposta pela regressão linear tradicional. Cabe ressaltar ainda a existência de um excessivo número de quadras/bairros (61 observações, aproximadamente 38% do total disponível no banco de dados) que não produzem viagens por transporte público. A regressão linear de melhor desempenho obtida a partir das observações referentes a 111 quadras/bairros de Palmas é mostrada na Tabela 3.

Tabela 3 Resultados do modelo OLS

Parâmetros	OLS
Interseção	-
População	0,630***
Índice de Entropia	-859,318**
Sistema Viário	0,020***
R^2 ajustado	0,849
Estatística F	209,087***

Nota: *** e ** significativos ao nível de confiança de 100% e 99%, respectivamente (2 extremidades).

Nota-se que esse modelo em geral consegue explicar aproximadamente 85% da variância das viagens diárias produzidas por quadra/bairro de Palmas. Tanto as estatísticas *t* para cada variável independente quanto a estatística F para todo o modelo resultaram bastante altas, sendo estatisticamente significativas. Os sinais das covariáveis revelam que quanto maior a população e a extensão do sistema viário, maior o número de viagens produzidas.

Porém, quanto maior a diversidade no uso do solo, menor a demanda por transporte público. Isso pode ser explicado pelo fato de que quanto maior o número de categorias de uso do solo em determinada unidade de vizinhança, mais oportunidades são criadas para a utilização de modos não motorizados devido à eventual diminuição no comprimento das viagens. Cabe ressaltar, entretanto, que o índice de entropia é a covariável que menos contribuiu para a modelagem da variável de demanda considerada. O gráfico dos resíduos gerados pelo modelo OLS, bem como a adequação desses à distribuição normal são mostrados na Figura 1a e na Figura 1b, respectivamente.

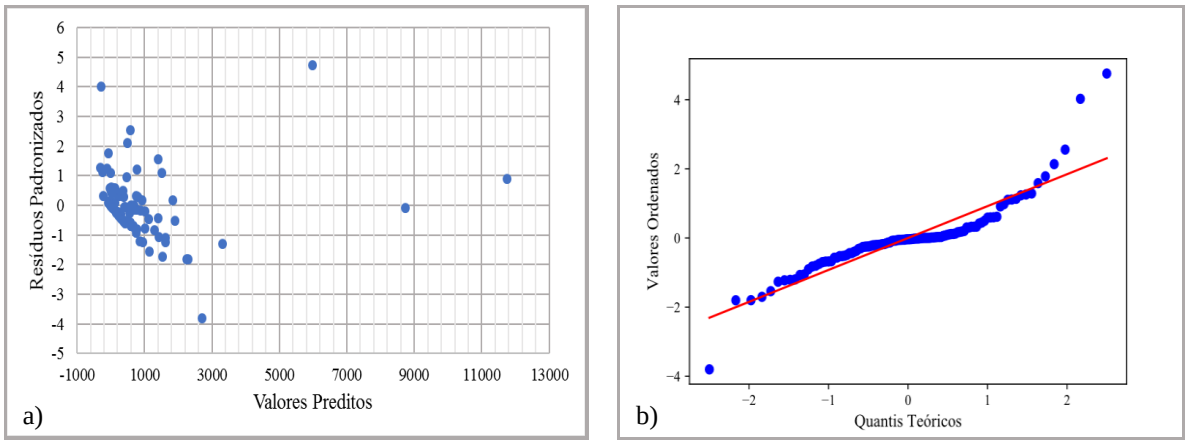

Fig. 1 Diagnóstico dos resíduos do modelo OLS

A Figura 1a revela que os resíduos não apresentam variância constante, uma vez que à medida em que os valores preditos aumentam, há também um aumento no valor absoluto das diferenças entre o número de viagens observado e estimado pelo modelo, fenômeno conhecido como heterocedasticidade. Sob essa condição, os erros padrões das estimativas dos coeficientes do modelo tornam-se enviesados e, consequentemente, os testes de hipóteses associados a esses parâmetros (como o teste *t*, por exemplo) podem gerar resultados enganosos (Baltagi, 2011). A Figura 1b, por sua vez, mostra que não há um ajuste razoável dos resíduos à curva de distribuição normal (reta em vermelho), situação comprovada pelos testes de Kolmogorov-Smirnov e Shapiro-Wilk realizados. Além disso, a curva formada pelos resíduos revela que eles são dependentes entre si. Dessa forma, tanto as suposições de normalidade e homocedasticidade quanto de independência entre os termos de erro foram violadas.

Em seguida, foram calibrados os demais modelos propostos neste trabalho. A Tabela 4 compara os resultados encontrados para o modelo OLS e logarítmicos com aqueles obtidos por meio da utilização de distribuições de contagem para a variável resposta. Apresentam-se apenas os modelos de melhor desempenho em cada categoria, ou seja, aqueles nos quais todos os parâmetros estimados foram significativos ($p < 0,01$).

Tabela 4 Comparação entre os modelos OLS, logarítmicos e de dados de contagem

Parâmetros	OLS	LN I	LN II	Poisson	Binomial Negativa
Interseção				5,082	
População (Pop)	0,630			2,08E-04	
Índice de Entropia	-859,318		10,077	2,982	5,705

Parâmetros	OLS	LN I	LN II	Poisson	Binomial Negativa
Sistema Viário (SV)	0,020				
ln(Pop + 1)*		0,738			
ln(SV)			0,177		0,519
Parâmetro de dispersão					5,044
Medidas de desempenho					
Verossimilhança de log	-885,726	-262,947	-256,523	-41.751,458	-652,424
AIC	1.779,451	529,895	519,046	83.508,917	1.310,848
BIC	1.790,290	535,314	527,175	83.517,045	1.318,977

Nota: * foi necessário acrescentar uma unidade à variável População a fim de se evitar problemas com zeros.

Percebe-se que, no segundo modelo com transformações e nas regressões de Poisson e Binomial Negativa, o índice de entropia passa a exercer um efeito positivo sobre a produção de viagens por quadra/bairro, em contraposição ao modelo OLS. Além disso, o parâmetro de dispersão estimado mostra a melhor adequação do banco de dados à distribuição Binomial Negativa do que às distribuições normal e Poisson. As medidas de desempenho corroboram essa conclusão: para o modelo Binomial negativo obteve-se a maior verossimilhança de log e os menores valores de AIC e BIC, em comparação com a regressão OLS e de Poisson.

Há que se ressaltar, também, o bom desempenho dos dois modelos com transformações logarítmicas, tanto em termos da otimização da função de verossimilhança quanto minimização dos critérios AIC e BIC. De fato, quando o segundo modelo logarítmico é calibrado pelo método dos mínimos quadrados, as variáveis Índice de Entropia e ln(SV) conseguem explicar 76,0% da variância de ln(VP+1) com estatísticas F e *t* altamente significativas. Nesse caso, o primeiro modelo também se destaca: contendo apenas a covariável ln(Pop+1), seu R^2 ajustado alcança 0,733, se aproximando do desempenho dos modelos OLS e daquele comentado anteriormente. As estatísticas F e *t* para essa regressão são ainda maiores que as de LN II.

Apesar de o parâmetro estimado para a covariável População ou ln(Pop+1) não ser significativo em algumas regressões indicar uma possível anomalia no banco de dados, ressalta-se que a porcentagem de área residencial utilizada no cálculo do índice de entropia possui alta e significativa correlação com a população residente em cada quadra/bairro. Dessa forma, é possível afirmar que, mesmo mantendo-se somente o índice de entropia como variável explicativa do modelo, as variações no número de habitantes em cada quadra/bairro ainda serão capturadas no cálculo desse índice. Portanto, indiretamente, o efeito de qualquer aumento ou diminuição na população certamente impactará a produção de viagens. É importante reiterar também que todas as variáveis explicativas mantidas nos modelos não demonstraram estar fortemente correlacionadas entre si, impedindo, dessa forma, a ocorrência de multicolinearidade. Apesar de a maioria das correlações ter sido significativa, o maior valor de R entre duas dessas covariáveis é 0,530, referente às variáveis População e Sistema Viário.

No que concerne à amostra de validação, a Tabela 5 sintetiza os resultados das métricas de erros aplicadas bem como mostra o valor do coeficiente de correlação linear de *Pearson* entre os valores reais e preditos e sua probabilidade de significância.

Tabela 5 Resumo das medidas de desempenho aplicadas à amostra de validação

Modelos\Métricas		Erro Quadrado	Erro Médio	Erro Absoluto Médio	Raiz do Erro Quadrado Médio	R
OLS	VP=0.630Pop-859.318I_{ent}+0.020SV	19.743.358	148,03	375,79	628,38	0,697**
LN I	VP=[(Pop+1)$^{0.738}$] -1	26.128.983	-200,31	293,82	722,90	0,626**
LN II	VP=[$SV^{0.177}exp$(10.077I_{ent})] -1	31.831.611	-189,43	345,01	797,89	0,108
Poisson	VP=exp(5.082+0.0002Pop+2.982I_{ent})	19.899.599	161,47	421,19	630,87	0,652**
Binomial Negativa	VP=$SV^{0.519}exp$(5.705I_{ent})	28.734.442	156,34	477,81	758,08	0,359*

Nota: ** e * significativos ao nível de confiança de 99% e 95%, respectivamente (2 extremidades).

De acordo com a Tabela 5, das quatro métricas de erros aplicadas, três delas resultaram menores para o modelo OLS que para os demais. Entretanto, no que concerne ao erro quadrado e à raiz do erro quadrado médio, o desempenho da regressão de Poisson se aproximou consideravelmente do referente ao modelo linear. A ocorrência de erros médios negativos nos dois modelos com transformações logarítmicas revela que, em média, os valores preditos foram inferiores aos reais, enquanto nas outras regressões observa-se o inverso, ou seja, superestimação das viagens produzidas. O coeficiente de correlação linear de *Pearson*, por sua vez, aponta para ajustes comparáveis da amostra de validação à regressão linear, ao primeiro modelo logarítmico e à regressão de Poisson. A Fig. 2 compara o histograma dos valores observados na amostra de validação com o número de viagens previsto pelo modelo OLS, o primeiro modelo logarítmico e pelas regressões para dados de contagem.

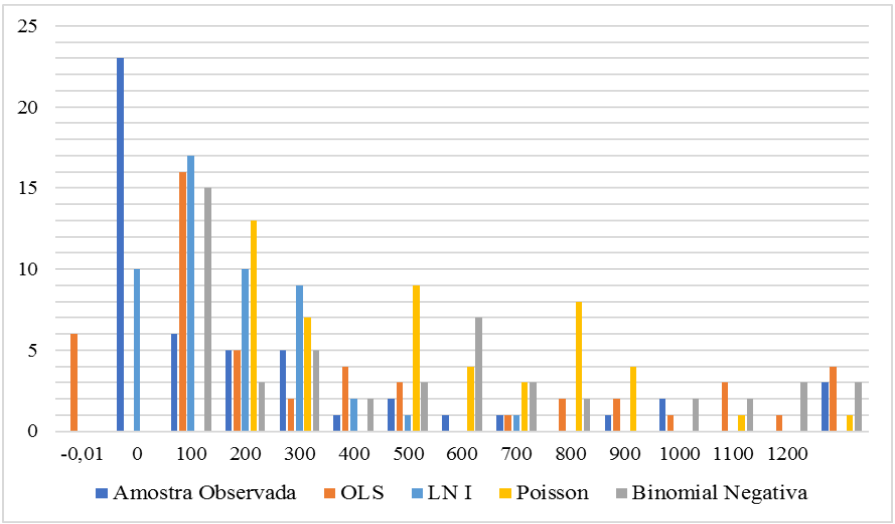

Fig. 2 Histograma de valores observados e estimados pelos diferentes modelos

A leitura do gráfico acima sugere que, para diferentes intervalos, a frequência de viagens preditas por cada modelo se aproxima razoavelmente do número real de viagens naquele intervalo, embora não seja possível apontar as regressões de melhor desempenho apenas pela análise do histograma. Contudo, o comportamento da amostra de valores observados mostra que os dados realmente possuem uma assimetria consistente com as distribuições de contagem, em contraposição à normalidade imposta pela regressão linear tradicional. Além disso, é importante observar que a regressão linear demonstra o inconveniente de

prever valores negativos para uma variável que só pode assumir valores maiores ou iguais a zero, como é a demanda. Devido a essa continuidade da distribuição normal, o modelo OLS, para a amostra de validação, gerou seis valores negativos. Ressalta-se ainda que o modelo com transformações logarítmicas foi o único a retornar resultados nulos quando o número real de viagens era zero. Entretanto, mesmo essa regressão conseguiu prever apenas 10 das 23 observações nulas.

4 CONCLUSÕES E CONSIDERAÇÕES FINAIS

O objetivo desse trabalho foi comparar o desempenho de um modelo de regressão linear tradicional com modelos obtidos a partir de transformações logarítmicas e aqueles apropriados para variáveis discretas, que compõem o conjunto dos modelos lineares generalizados, técnica que se tornou bastante proeminente a partir da década de 1970. Para tanto, modelou-se o número de viagens diárias produzidas por transporte público por quadra/bairro de Palmas-TO (Brasil) em função de variáveis demográficas, de uso do solo e do sistema viário. Os resultados apontaram um melhor ajuste dos modelos com transformações logarítmicas e da regressão Binomial Negativa em detrimento do modelo linear e de Poisson, porém, em termos de validação, destacaram-se principalmente as regressões de Poisson e OLS.

O melhor ajuste da amostra de calibração à regressão Binomial Negativa e aos modelos com transformações traz à tona algumas reflexões sobre a natureza do banco de dados utilizado. Conclui-se que a aplicação do logaritmo natural às contagens de viagens produzidas foi eficaz na estabilização de sua variância, bem como na correção da assimetria que essa variável apresentava. Além disso, é possível afirmar que o modelo linear generalizado com a distribuição Binomial Negativa para a variável resposta realmente consegue contabilizar parte da variabilidade da demanda que a distribuição de Poisson não considera, o fenômeno conhecido como sobredispersão. Nesse quesito, a distribuição normal é deficiente, pois além de ser simétrica, ainda representa dados contínuos, ou seja, pode assumir valores negativos.

A análise realizada com a amostra de validação, por sua vez, revelou que, novamente, a consideração da não normalidade inerente às informações de viagens conseguiu alcançar resultados satisfatórios: o modelo de Poisson, juntamente com o linear, foram os que apresentaram os menores erros e coeficientes de correlação mais próximos de 1. Entretanto, a análise de frequência dos valores reais e estimados mostrou que o excesso de observações nulas pode ter sido o fator que mais comprometeu o desempenho dos modelos nesse quesito.

O fato de a regressão Binomial Negativa não ter sido contemplada como a técnica mais precisa na validação pode indicar a existência de uma aleatoriedade substancial no banco de dados. Essa condição pode ser proveniente de uma eventual amostragem insuficiente quando da pesquisa embarcada que deu origem ao número de viagens produzidas por transporte público por quadra/bairro de Palmas, o que pode ter levado à quantidade substancial de quadras/bairros com nenhuma viagem produzida. Outra fonte possível de erros seria a disparidade entre o ano de coleta das informações utilizadas na modelagem. Ademais, o número excessivo de zeros observados na variável dependente poderia apontar para a necessidade de aplicação de outra classe de regressões, os modelos lineares generalizados inflacionados de zeros. Além disso, pondera-se que tais modelos ainda sofreriam com a carência da segunda característica intrínseca às variáveis de demanda: a

autocorrelação espacial, responsável pela interdependência entre os termos de erro. Dessa forma, pretende-se propor, em trabalhos futuros, a utilização de modelos espaciais locais, como a Regressão Geograficamente Ponderada, associados a distribuições de contagem e que considerem o excesso de observações nulas da variável resposta.

5 AGRADECIMENTOS

Às agências de fomento CAPES, CNPq e FAPESP.

6 REFERÊNCIAS

Baltagi, B. H. (2011) Econometrics, Springer Texts in Business and Economics, Springer.

Bartlett, M. S. (1947) The Use of Transformations. Biometrics, 3(1), 39–52.

Chakour, V. e Eluru, N. (2013) Examining the Influence of Urban form and Land Use on Bus Ridership in Montreal, Procedia - Social and Behavioral Sciences, 104, n. Supplement C, 875–884.

Chakour, V. e Eluru, N. (2016) Examining the influence of stop level infrastructure and built environment on bus ridership in Montreal, Journal of Transport Geography, 51, n. Supplement C, 205–217.

Choi, J., Lee, Y. J., Kim, T. e Sohn, K. (2012) An analysis of Metro ridership at the station-to-station level in Seoul, Transportation, 39(3), 705–722.

Chow, L.-F., Zhao, F., Liu, X., Li, M-T e Ubaka, I. (2006) Transit Ridership Model Based on Geographically Weighted Regression, Transportation Research Record: Journal of the Transportation Research Board, 1972, 105–114.

Chu, X. (2004) Ridership models at the stop level, National Center for Transit Research, University of South Florida.

Cordeiro, G. M. e Demétrio, C. G. B. (2008) Modelos lineares generalizados e extensões, Piracicaba.

Cunto, F. J. C.; Castro Neto, M. M. e Barreira, D. S. (2012) Modelos de previsão de acidentes de trânsito em interseções semaforizadas de Fortaleza, Transportes, 20(2), 57–64.

Ewing, R., Tian, G., Zhang, JP G. M., Greenwald, M. J., Joyce, A., Kircher, J. e Greene, W. (2014) Varying influences of the built environment on household travel in 15 diverse regions of the United States, Urban Studies, 52(13), 2330–2348.

Ferraz, A. C. P. e Torres, I. G. E. (2004) Transporte público urbano, 2. ed., RiMa Editora.

Gomes, M. J. T. L.; Torres, C. A. e Cunto, F. J. C. (2016) Avaliação da dependência espacial na modelagem do desempenho da segurança viária em zonas de tráfego, Transportes, 24(4), 56–59.

Gomes, M. J. T. L.; Torres, C. A.; Oliveira Neto, F. M. e Cunto, F. J. C. (2015) Análise exploratória para a modelagem da frequência de acidentes de trânsito agregados ao nível de zonas de tráfego, Transportes, 23(4), 42–50.

Hilbe, J. M. (2014) Modeling Count Data, Cambridge University Press, Cambridge.

Hollander, Y. e Liu, R. (2008) The principles of calibrating traffic microsimulation models, Transportation, 35(3), 347–362.

Marques, S. F. (2016) A influência da forma urbana na viabilidade financeira do sistema de transporte público por ônibus em cidades médias: o caso de Palmas – TO, 166 f. Trabalho de Conclusão de Curso (Graduação em Engenharia Civil) – Fundação Universidade Federal do Tocantins, Palmas.

Mendes, F. C. (2014) Estudo da geração de viagens por transporte público em Palmas-TO, 62 f. Trabalho de Conclusão de Curso (Graduação em Engenharia Civil) – Fundação Universidade Federal do Tocantins, Palmas.

Myers, R. H., Montgomery, D. C., Vining, G. G. e Robinson, T. J. (2012) Generalized linear models: with applications in engineering and the sciences, John Wiley & Sons, v. 791.

Nelder, J. A. e Baker, R. J. (2004) Generalized Linear Models, in Encyclopedia of Statistical Sciences, John Wiley & Sons, Inc.

Oliveira, L. A., Cruz, S. N. e Pereira, A. P. B. (2014) Identificação da Estrutura Espacial Urbana: o caso de Palmas, in E. C. Kneib (org.), Projeto e Cidade: Centralidades e Mobilidade Urbana, Faculdade de Artes Visuais da Universidade Federal de Goiás, Goiânia, 169-196.

Palmas (2016) Arquivo shp do sistema viário da cidade de Palmas. GeoPalmas, Divisão de Georreferenciamento, Secretaria de Desenvolvimento Urbano Sustentável – SEMDUS, Prefeitura de Palmas.

Palmas (2016) Arquivo kml detalhado das quadras e bairros da cidade de Palmas. GeoPalmas, Divisão de Georreferenciamento, Secretaria de Desenvolvimento Urbano Sustentável – SEMDUS, Prefeitura de Palmas.

Palmas (2017) Arquivo kml dos pontos de parada da rede de transporte público da cidade de Palmas. Secretaria Municipal de Acessibilidade, Mobilidade, Trânsito e Transporte – SMAMTT, Prefeitura de Palmas.

SI Digital (2016) Mapa do itinerário das linhas de ônibus da cidade de Palmas. Meu Busão Palmas, SI Digital.

Song, Y., Merlin, L. e Rodriguez, D. (2013) Comparing measures of urban land use mix. Computers, Environment and Urban Systems, 42, 1–13.

Thompson, G. (1997) Achieving Suburban Transit Potential: Sacramento Revisited, Transportation Research Record, 1571, 151–160.