MODELAGEM DA DEMANDA POR TRANSPORTE PÚBLICO NO ÂMBITO DE PONTOS DE PARADA

Samuel de França Marques
Cira Souza Pitombo
Universidade de São Paulo

RESUMO
A modelagem de embarques e desembarques por ponto de parada é um importante instrumento para o planejamento operacional do sistema de transporte público, além de contribuir para o desenvolvimento urbano orientado ao transporte sustentável. A variável de interesse, nesse caso, apresenta duas peculiaridades que afetam o desempenho das estimativas empreendidas: demonstra dependência espacial e são dados de contagens. Dessa forma, no intuito de incluir, paulatinamente, tais características ao processo de modelagem da demanda, o presente trabalho propõe a aplicação das regressões linear, de Poisson, Binomial Negativa, Geograficamente Ponderada e de Poisson Geograficamente Ponderada (GWPR) ao volume de desembarques ao longo de uma linha de ônibus da cidade de São Paulo. Os resultados, obtidos a partir de métricas de aderência, confirmaram a suposição de que incluir a assimetria e autocorrelação espacial dos dados, isoladamente e em conjunto, ao processo de modelagem da demanda por transportes, contribui para uma melhoria gradual das estimativas, com destaque para a ferramenta de estimativa local GWPR.

ABSTRACT
Boarding and alighting modeling at the bus stop level is an important tool for the operational planning of the public transport system, in addition to contributing to transit-oriented development. The interest variable, in this case, presents two particularities that influence the performance of proposed estimates: it demonstrates spatial dependence and it is a count data. Thus, in order to gradually include such characteristics in the demand modeling process, the present study proposes applying linear, Poisson, Negative Binomial, Geographically Weighted and Geographically Weighted Poisson (GWPR) regressions to the alighting data along a bus line from the city of São Paulo. The results, from goodness-of-fit measures, confirmed the assumption that adding data asymmetry and spatial autocorrelation, isolated and together, to the transportation demand modeling process, contributes for a gradual improvement in the estimates, highlighting the local estimation tool GWPR.

1. INTRODUÇÃO E *BACKGROUND*

O planejamento urbano alinhado ao de transportes é um dos pilares do desenvolvimento sustentável de cidades. As associações entre uso do solo e a mobilidade urbana subsidiam a elaboração de políticas públicas sustentáveis, primordiais para o estímulo à utilização do Transporte Público (TP), importante instrumento de inclusão social e acessibilidade. Nesse contexto, a modelagem de transportes é uma das ferramentas que, ao quantificar e explicar os efeitos de práticas urbanas sobre o deslocamento de pessoas e bens, dão suporte a essas políticas nas mais diversas escalas geográficas.

Condicionada, geralmente, pela disponibilidade de dados, a modelagem de viagens urbanas abrange diferentes abordagens, que se diferenciam pela unidade de agregação utilizada. No que se refere ao Transporte Público, podem ser encontrados estudos no âmbito de sistemas (Cervero e Dai, 2014; Hensher e Golob, 2008; Hensher *et al.*, 2014; Joonho *et al.*, 2019; Taylor *et al.*, 2009), zonas, bairros ou distritos (Chiou *et al.*, 2015; Kalaanidhi e Gunasekaran, 2013; Ma *et al.*, 2018; Siddiqui *et al.*, 2015), linhas de ônibus (Kyte *et al.*, 1985; Peng *et al.*, 1997), estações de trem, de metrô e pontos de parada (Gan *et al.*, 2019; Pulugurtha e Agurla, 2012; Zhu *et al.*, 2019) e individual ou do domicílio (Ewing *et al.*, 2014; Siddiqui *et al.*, 2015), englobando do nível mais agregado ao mais desagregado. De forma simplificada, o elemento de análise adotado influencia fortemente nos fatores intervenientes, ou variáveis explicativas, que poderão ser considerados no estudo.

A modelagem tradicional por zonas de tráfego assume um valor médio das variáveis

explicativas em cada unidade, o que impede a captura de variações no âmbito local e pode levar à falácia ecológica. Por sua vez, considerando o ponto de parada como unidade de análise, é possível obter, por meio de modelos, estimativas do volume de embarques e desembarques, de maneira rápida e econômica, subsidiando o planejamento da rede de TP (Cervero, 2006), sendo que tal modelagem é realizada em função das variáveis socioeconômicas, de uso do solo e do sistema de transporte do entorno desses pontos.

Os dados de viagens, contudo, que consistem na variável de interesse desses modelos, exibem duas características de fundamental importância para o desempenho das estimativas, são elas: referem-se a contagens, ou seja, podem assumir somente valores inteiros não negativos e possuir assimetria; e apresentam autocorrelação espacial, isto é, valores de demanda próximos entre si no espaço tendem a demonstrar um comportamento semelhante. Dessa forma, os modelos de demanda foram sendo aprimorados, ao longo dos anos, no intuito de contabilizar, no processo de modelagem, tais peculiaridades. No âmbito das unidades de interesse do planejamento urbano sustentável (pontos de parada e estações), é possível encontrar trabalhos que modelam a demanda por ponto de parada ou estação a partir da regressão linear clássica (Cervero, 2006; Gutiérrez *et al.*, 2011; Ryan e Frank, 2009). Esse modelo tradicional é apropriado para variáveis contínuas e seus resíduos não podem ser dependentes entre si, caso em que os pressupostos da regressão linear são violados (Yan e Su, 2009) e a inferência estatística fica comprometida, ou seja, o estimador deixa de ser o de menor variância. Artifícios como transformação nas variáveis e funções de decaimento foram adotados por alguns autores a fim de contornar tais problemas, embora a real natureza dos dados não tenha sido considerada. A expansão, na década de 1980, do modelo linear para outras distribuições de probabilidade, introduziu as regressões de Poisson e Binomial Negativa, que, diferentemente da distribuição normal, modelam dados de contagem. Tais modelos, que também já foram utilizados no tratamento de viagens por TP no âmbito de pontos de parada e estações (Choi *et al.*, 2012; Chu, 2004; Pulugurtha e Agurla, 2012), podem demonstrar um desempenho superior ao linear tradicional. Apesar disso, tais abordagens ainda ignoram a autocorrelação espacial presente na variável resposta.

Tentativas de solucionar essa limitação culminaram no surgimento das regressões espaciais propriamente ditas que, ora consideram a autocorrelação a partir da inclusão, como covariável, da variável dependente defasada espacialmente, ou por meio dos resíduos do modelo, sendo que, em ambos os casos, a interação espacial é capturada através de uma matriz de pesos espaciais, normalmente baseada na distância entre os pontos do banco de dados (Fotheringham *et al.*, 2003). Tais técnicas também já foram utilizadas na modelagem de viagens por estação de TP (Gan *et al.*, 2019), embora, de acordo com Fotheringham *et al.*, (2003), esses modelos não reflitam, de forma local, a heterogeneidade espacial do banco de dados, já que neles a autocorrelação é expressa em termos de um único parâmetro apenas. A Regressão Geograficamente Ponderada, que gera um modelo diferente para cada coordenada geográfica, seria mais apropriada, nesse caso, para tratar a autocorrelação e heterogeneidade espacial dos parâmetros estimados (Brunsdon *et al.*, 1996). Nas aplicações de GWR à demanda por TP (Blainey e Mulley, 2013; Blainey e Preston, 2010; Cardozo *et al.*, 2012), os resultados sempre demonstram um desempenho melhor que o dos modelos globais.

A Regressão Geograficamente Ponderada, apesar de conseguir tratar, de maneira satisfatória, a dependência espacial do banco de dados, ainda sofre com a limitação de que, assim como o modelo linear, também pressupõe normalidade da variável de interesse, o que, no caso das

viagens por TP, não se verifica. Dessa forma, desenvolveram-se, recentemente, modelos geograficamente ponderados dedicados a dados de contagem, designados como Regressão de Poisson Geograficamente Ponderada (GWPR, *Geographically Weighted Poisson Regression*) e Regressão Binomial Negativa Geograficamente Ponderada (GWNBR, *Geographically Weighted Negative Binomial Regression*). Por serem bastante recentes, encontrou-se apenas um trabalho que aplica a GWPR à demanda por TP no âmbito de estações de metrô (Liu *et al.*, 2018) e a GWNBR a viagens urbanas de trem (Zhu *et al.*, 2019), os quais apontam, novamente, um melhor desempenho dos modelos locais frente à versão global dos mesmos, regressão de Poisson e Binomial Negativa, respectivamente.

Com base nos trabalhos citados anteriormente, as seguintes lacunas podem ser ressaltadas: 1) Aplicação de modelos espaciais no contexto de pontos de parada: as abordagens encontradas até o momento se restringem ao tratamento da assimetria demonstrada pelos dados de viagens por ponto de parada, ignorando a autocorrelação espacial potencialmente presente nos modelos, bem como ambas as características simultaneamente. 2) Abordagem da demanda por TP no âmbito de pontos de parada: apesar de as abordagens por estações de trem e metrô também representarem uma contribuição ao planejamento urbano sustentável, o espaçamento médio entre tais elementos não permite um detalhamento tão refinado das covariáveis, como no caso dos pontos de parada. O TP por ônibus, em contrapartida, é um sistema muito mais popular que o de transporte sobre trilhos, presente apenas nas maiores cidades. 3) Dos trabalhos cujo elemento de agregação é o ponto de parada (Cervero, 2006; Chu, 2004; Pulugurtha e Agurla, 2012; Ryan e Frank, 2009), os autores aplicam apenas um tipo de modelo, ora o linear tradicional, ora o apropriado para dados de contagens, não realizando comparação entre a versão clássica e a aprimorada, o que impede a visualização dos ganhos proporcionados pela utilização da regressão mais adequada. Mesmo nos outros estudos, que abordam a demanda por estações e nos quais aplicam-se mais de um tipo de modelo (Blainey e Mulley, 2013; Blainey e Preston, 2010; Cardozo *et al.*, 2012; Choi *et al.*, 2012; Gan *et al.*, 2019; Liu *et al.*, 2018; Zhu *et al.*, 2019), as regressões se concentram no tratamento de apenas uma das características comentadas anteriormente, ora a assimetria, ora a autocorrelação espacial, ou os autores não realizam comparação com o modelo linear tradicional. Dessa forma, consegue-se observar as melhorias proporcionadas apenas por meio da inclusão de uma ou outra particularidade na modelagem de viagens, mas nunca de ambas.

Por conseguinte, o objetivo do presente trabalho é modelar a demanda por transporte público, no âmbito de pontos de parada, a partir de modelos geograficamente ponderados para dados de contagem. Além disso, pretende-se estabelecer uma estrutura sequencial de modelos, desde o clássico linear até a Regressão de Poisson Geograficamente Ponderada, passando pelas regressões globais de Poisson e Binomial Negativa, e pela Regressão Geograficamente Ponderada tradicional. A intenção dessa proposta é permitir a comparação entre tais técnicas e a visualização dos ganhos gradativos alcançados por meio do tratamento da assimetria e da autocorrelação espacial isoladamente e, posteriormente, em conjunto. Tal análise será realizada a partir de um estudo de caso real, direcionado à linha 6045-10 do município de São Paulo.

2. MATERIAIS E MÉTODO

O banco de dados a ser utilizado no presente trabalho se baseia nos resultados de uma pesquisa de embarque e desembarque realizada em 8 linhas de ônibus da cidade de São Paulo – SP. Para cada sentido da linha (ida e volta), foi disponibilizada, pela São Paulo Transporte

S.A. (SPTrans), uma planilha contendo o número de embarques e desembarques por ponto de parada, codificados por um identificador, em 6 faixas horárias diferentes, que cobrem as 24 horas de uma terça-feira do ano de 2017. De posse dos identificadores dos pontos de parada e de suas respectivas coordenadas geográficas, também providenciadas pela SPTrans, tornou-se possível proceder à espacialização desse banco de dados.

Posteriormente, as 16 linhas unidirecionais passaram por uma análise exploratória de dependência espacial por meio do cálculo do índice de Moran (Moran, 1948), com matriz de pesos baseada na distância euclidiana entre os pontos, para o número de embarques e desembarques por ponto de parada nas faixas horárias PM (05h às 08h59), EP (09h às 15h59), PT (16h às 19h59), PN (20h às 23h59) e ao total de passageiros que sobem e que descem das 05h às 23h59. Buscou-se encontrar uma linha de ônibus, dentro as 8 contempladas pela pesquisa sobe/desce, cujo volume de embarques e desembarques demonstrasse uma forte e significativa dependência espacial, ou seja, altos números do índice de Moran associados a baixos pseudo valores *p*. Nesse âmbito, a linha 6045-10-2 com seus 49 pontos de parada se destacou com relação ao volume de desembarques no total de viagens das 05h às 23h59. Dessa forma, estabeleceu-se como variável dependente o número de Desembarques na linha 6045-10-2, referente ao conjunto de viagens realizadas das 05h às 23h59.

Conforme mencionado na seção anterior, a modelagem da demanda por TP no âmbito de pontos de parada abrange, basicamente, três grupos de variáveis explicativas: socioeconômicas, de uso do solo e do sistema de transporte. Com base nisso, no caso do presente trabalho, foi possível coletar potenciais preditores tanto relacionados ao ponto de parada quanto referentes à sua área de influência, composta por um *buffer* de raio igual a 400m centrado nos pontos de parada (Zhao *et al.*, 2003). Relativos ao ponto, os seguintes dados foram coletados: 1) Distância euclidiana, em metros, até o terminal de ônibus mais próximo, conforme *shapefile* de terminais disponível no sítio eletrônico do GeoSampa; 2) Distância euclidiana até a estação de metrô ou de trem mais próxima (selecionou-se a menor entre as duas distâncias), também de acordo com malha digital do GeoSampa. Uma variável alternativa considerando-se a menor entre as distâncias euclidianas até o terminal, estação de metrô ou de trem mais próxima também foi incluída; e 3) Número e frequência média (em viagens por hora) das linhas de ônibus, em 2017, que passavam pelos pontos, exclusive a de interesse. Da área de influência, por sua vez, obtiveram-se os potenciais preditores a seguir: 4) População: recorte e interpolação areal da malha digital da Pesquisa O/D de 2017 (Metrô, 2019), dada em zonas de tráfego; 5) Área, em hectare, de 16 categorias de uso do solo predominante: recorte de *shapefile* disponível no GeoSampa, que detalha o uso do solo em São Paulo, no âmbito de quadras, em 2016. Esses dados também foram utilizados, em conjunto, para o cálculo do índice de entropia (Song *et al.*, 2013) no entorno dos pontos de parada, que reflete a mistura de usos do solo presente na região; 6) Renda familiar média e número médio de automóveis por domicílio, de acordo com a Pesquisa O/D de 2017. Nesse caso, calculou-se a média dos domicílios amostrados pela pesquisa que foram cobertos pelo *buffer*, sendo que, às áreas que não continham nenhum domicílio, foram atribuídos os resultados do recorte e interpolação areal dos dados agregados por zona; e 7) Número de vias e interseções presentes em cada *buffer*, conforme *shapefile* de linhas do *Open Street Map*. O levantamento dos potenciais preditores foi realizado em ambiente GIS.

Para evitar problemas de multicolinearidade e redundância de parâmetros, bem como identificar as variáveis com maior potencial de explicar Desembarques, calculou-se o

coeficiente de correlação linear de Pearson (ρ) entre todas as variáveis do banco de dados. Quando um par de potenciais preditores apresentava um valor de ρ igual ou maior que 0,60, a variável com menor correlação com Desembarques era descartada. Esse patamar foi admitido como aceitável no intuito de combater, também, o viés de variável omitida, uma vez que nem todos os pares de covariáveis com alta correlação representam uma relação de causa e efeito.

Após completado o banco de dados de Desembarques com seus potenciais preditores, seguiu-se para a etapa de modelagem. Nesse estágio, para cada tipo de modelo, buscou-se encontrar a combinação de variáveis explicativas que otimizasse as estimativas a partir da minimização do Critério de Informação de Akaike (AIC, *Akaike Information Criterion*) (Sakamoto *et al.*, 1986). Tal análise foi realizada em ambiente R (R Core Team, 2020), interface de programação aberta e livre. Inicialmente, calibrou-se o modelo linear tradicional (*reglin*), cuja estrutura está mostrada na Equação 1.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon \quad (1)$$

Em que a variável resposta *y* é composta pela combinação linear de x_k variáveis explicativas somadas a um erro aleatório ε. Os parâmetros β a serem estimados são números que refletem a contribuição de cada covariável à explicação da variância de *y*. No R, a regressão linear tradicional foi gerada e otimizada por meio do pacote "olsrr" (Hebbali, 2020). Em seguida, a não normalidade dos dados de contagem foi tratada por meio das regressões de Poisson (*regpoisson*) e Binomial Negativa (*regbin*), representadas pela Equação 2.

$$ln\ (\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \quad (2)$$

Na qual μ é o valor esperado da variável resposta. A regressão Binomial Negativa se diferencia da de Poisson por conseguir modelar o fenômeno da sobredispersão, que ocorre quando a variância da variável dependente supera a sua média (Hilbe, 2014). A escolha do melhor modelo linear generalizado teve o suporte dos pacotes "glmulti" (Calcagno, 2019) e "MASS" (Venables e Ripley, 2002). Posteriormente, o tratamento isolado da autocorrelação e heterogeneidade espacial se deu a partir do modelo GWR tradicional (Equação 3).

$$y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i) x_{ik} + \varepsilon_i \quad (3)$$

Em que (u_i,v_i) representam as coordenadas do *i*-ésimo ponto no espaço e $\beta_k(u_i,v_i)$ refere-se à realização da função contínua $\beta_k(u,v)$ no ponto *i* (Fotheringham *et al.*, 2003). No caso da GWR, a interação espacial entre o ponto em que o modelo será estimado e demais pontos do banco de dados é dada por um ponderador que varia em função da distância entre esses pontos e de um raio máximo (*bandwidth*) fora do qual admite-se dependência espacial nula. Por fim, o modelo espacial local que também considera a não normalidade dos dados de contagem está estruturado na Equação 4 (da Silva e Rodrigues, 2014; Nakaya *et al.*, 2005).

$$ln\ (\mu_i) = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i) x_{ik} \quad (4)$$

Assim como no modelo global, duas distribuições de probabilidade para a variável resposta são permitidas: Poisson e Binomial Negativa. Na análise do modelo generalizado global, a regressão de Poisson apresentou um melhor desempenho que a Binomial Negativa. Dessa forma, por brevidade, utilizou-se, nesse estágio da modelagem, apenas a Regressão de Poisson Geograficamente Ponderada. No âmbito da GWR e GWPR, a otimização do modelo pode ser realizada, inicialmente, por meio da seleção da função de ponderação (*kernel*) e respectivo *bandwidth* que minimizam o AIC da regressão. Com base nisso, analisaram-se o

kernel gaussiano e bi-quadrado, considerando-se apenas raios fixos, sendo que o segundo foi o que demonstrou os menores valores de AIC e, consequentemente, compôs todos os modelos geograficamente ponderados. Esse procedimento, bem como a otimização do modelo propriamente dito, foi efetuado conforme códigos disponíveis nos pacotes "sp" (Bivand *et al*., 2013; Pebesma e Bivand, 2005) e "GWModel" (Gollini *et al*., 2015; Lu *et al*., 2014) do R. A comparação entre os modelos supracitados foi realizada por meio de várias métricas de aderência, a saber: Erro Médio Absoluto (MAE), Raiz Quadrada do Erro Quadrado Médio (RMSE) (Hollander e Liu, 2008) e porcentagem de erro, que devem ser próximos de 0 para refletir um bom desempenho da técnica; razão entre desvio padrão dos valores reais e previstos (*SD_ratio*) (Pielke Sr, 2013) e coeficiente de correlação entre valores observados e estimados (R). Quanto mais próximos de 1 *SD_ratio* e R forem, melhor é o ajuste do modelo. Os resultados e discussão acerca destes estão descritos na Seção 3.

3. RESULTADOS E DISCUSSÃO

A Tabela 1 consolida as medidas descritivas dos dados utilizados no presente trabalho. Por brevidade, são mostradas apenas as variáveis explicativas que foram mantidas nos modelos finais.

Tabela 1: Medidas descritivas das variáveis utilizadas

Sentido	Variável\Descritivo	Média	Desvio Padrão	Mín.	Máx.	25%	50%	75%
Volta (49)	*Desembarques*	118,92	132,57	0,00	746,00	27,50	76,00	180,50
	População	3.522,00	2.144,84	490,93	8.510,98	1.587,71	3.211,40	4.993,73
	Frequência média (viagens/h)	4,27	0,79	2,30	5,82	3,77	4,00	5,26
	Distância metrô/trem (metros)	1.943,43	1.605,61	129,33	4.882,16	546,97	1.307,78	3.337,53
	Renda familiar média (R$)	4.215,28	2.078,05	1.976,34	9.669,07	2.479,38	3.551,73	5.691,72
	Número de interseções	110,29	64,42	35,00	343,00	64,00	93,00	141,00

Chama-se a atenção para o padrão de Desembarques, que varia entre 0 e 746, apresentando uma grande amplitude e desvio padrão. No conjunto de viagens realizadas das 05h às 23h59, em uma terça-feira de 2017, um total de 5.827 passageiros desembarcaram pela linha 6045-10-2. Observa-se, ainda, que essa variável demonstra a assimetria positiva especulada na Seção 3: sua mediana é menor que a média e, conforme esperado, o número nulo de usuários desembarcando ocorre uma única vez no conjunto de viagens realizadas das 05:00 às 23:59, no primeiro ponto de parada.

A amplitude de variação das variáveis relacionadas ao sistema de transporte (frequência média, número de interseções e a variável de proximidade intra e intermodal) revelam que a linha 6045-10-2 abrange tanto regiões com baixa cobertura da rede de transporte quanto áreas ricamente abastecidas pelo sistema. As variáveis de população e renda também demonstram um elevado desvio padrão, apontando para a presença, ao longo da linha analisada, de áreas com variados níveis de renda e de uso do solo para fins residenciais. É possível que as regiões de menor população sejam dedicadas, em parte, a áreas de comércio e serviços, ou seja, geradoras de empregos. Uma vez que as áreas centrais estão, geralmente, associadas a altas densidades viária e de linhas de ônibus, se essas regiões coincidem com as de menor população, a afirmação anterior se torna ainda mais plausível. A Tabela 2 consolida os modelos calibrados para Desembarques. Ressalta-se que foram mantidas, nos modelos finais, apenas aquelas variáveis explicativas cujos parâmetros resultaram estatisticamente significativos ($p < 0,05$).

Tabela 2: Modelos globais e locais para Desembarques ao longo da linha 6045-10-2 (N = 49)

Modelo\Parâmetro		Intercepto	População	Frequência	Distância metro/trem	Renda	Interseções
reglin		228,42797	0,03072	-64,70247	0,03009		
regpoisson		5,50300	0,00009	-0,58200	0,00030	-0,00006	0,00685
regbin		4,51100	0,00022	-0,63260	0,00043		0,00824
GWR	Mín	4,72971	0,00885	-92,95596	-0,02319		
	25%	68,31837	0,01019	-87,96254	0,00343		
	50%	160,59604	0,03986	-79,72072	0,02319		
	75%	275,46044	0,05917	-11,13289	0,03818		
	Máx	304,02187	0,06148	-1,40942	0,04853		
GWPR	Mín	-0,23606	-0,00030	-0,99380	0,00019	-0,00014	-0,00049
	25%	1,45965	0,00013	-0,38375	0,00038	-0,00010	0,00451
	50%	2,37237	0,00026	-0,04351	0,00052	-0,00007	0,00706
	75%	3,55485	0,00031	0,16659	0,00124	-0,00005	0,00868
	Máx	7,80060	0,00084	0,46932	0,00269	0,00017	0,01861

Nota: *reglin*, *regpoisson*, *regbin*, GWR e GWPR se referem, respectivamente, à regressão linear, de Poisson, binomial negativa, regressão geograficamente ponderada e regressão de Poisson geograficamente ponderada.

A regressão linear ótima para Desembarques conteve três variáveis explicativas: população, distância metrô/trem e frequência. O sinal negativo da frequência revela que regiões com uma cobertura densa da rede de TP apresentam um volume de passageiros desembarcando menor que áreas menos abastecidas pelo sistema. Tal conclusão mostra que a maioria das viagens na linha 6045-10-2 são atraídas para lugares com uma acessibilidade ao TP menor que nas regiões centrais. Esse destino pode se referir à moradia dos usuários de TP, indicando que, provavelmente, o sentido de volta da linha 6045-10 atende uma parcela considerável de viagens trabalho-casa. Cabe destacar que o R^2 ajustado desse modelo foi de 0,60 e seu intercepto resultou positivo e significativo, fazendo com que as chances de o modelo prever um número de Desembarques negativo sejam reduzidas.

As regressões de Poisson e Binomial Negativa acrescentaram, ao conjunto de preditores da *reglin*, as variáveis renda e número de interseções e somente número de interseções, respectivamente. Supondo que, na linha 6045-10-2, prevalecem as viagens de volta do trabalho para casa, é possível afirmar que o sinal do coeficiente de renda condiz com o esperado, ou seja, volumes maiores de Desembarques ocorrem em pontos de parada que atendem, em sua maioria, uma população de baixa renda. Por outro lado, haja vista que as altas densidades viárias são, geralmente, características de regiões centrais, isto é, atratoras de viagens, quanto maior o número de interseções no entorno do ponto de parada, maior será o número de passageiros desembarcando por ele. Na GWR para Desembarques, mantêm-se, novamente, as mesmas variáveis explicativas presentes no modelo linear tradicional e o sinal dos coeficientes se mantém constante, em sua maioria. Além disso, o R^2 ajustado dessa regressão aumenta para 0,74, o que representa uma melhoria de 23% com relação ao modelo global. Por sua vez, o coeficiente de determinação local varia de 0,34 a 0,88, sendo que 75% dos pontos de parada tiveram um valor de R^2 acima de 0,63. No caso dos parâmetros, os seguintes valores *t* foram obtidos: para população e frequência média, 65% e 70% dos coeficientes apresentaram uma estatística *t* superior, em módulo, a 1,75 e 1,92, respectivamente; distância metrô/trem, por outro lado, demonstrou valores *t* acima de 1,69 em 55% dos pontos de parada.

Seguindo o mesmo padrão que a GWR tradicional, a GWPR também manteve os mesmos preditores que apareceram na regressão de Poisson final. Observa-se, nesse caso, que apenas a variável de frequência média não demonstrou relativa constância no sinal de seu coeficiente. Cabe ressaltar, entretanto, que sinais positivos para esse preditor não podem ser considerados como estranhos: os pontos de parada cuja frequência média das linhas que passam por eles impacta positivamente o volume de Desembarques possivelmente servem como nós de integração intramodal. Além disso, a depender do horário do dia e considerando que as regiões com elevada frequência são também aquelas que detêm significativa parcela dos empregos ao longo da linha, o fluxo pode ser, de fato, do tipo casa-trabalho nesses pontos. No que tange à importância dos parâmetros, os valores *t* encontrados sugerem que população, frequência média, distância metrô/trem, renda familiar e número de interseções foram estatisticamente significativos em 90%, 65%, 100%, 75% e 80% dos pontos de parada, respectivamente. Um resultado interessante a ser comentado é que a maioria dos casos em que a frequência média não foi considerada significativa se refere a pontos em que seu coeficiente foi positivo. Mesmo assim, houve várias ocorrências de coeficientes positivos significativos. A Tabela 3 resume os resultados das métricas de aderência aplicadas aos modelos globais e locais de Desembarques.

Tabela 3: Métricas de aderência para modelos globais e locais de Desembarques

Sentido	Modelo\Métrica	MAE	RMSE	*SD_ratio*	R
Volta (*Desembarques*)	*reglin*	62,126	80,190	0,791	0,791**
	regpoisson	48,704	64,381	0,972	0,877**
	regbin	72,184	193,283	2,156	0,816**
	GWR	42,869	60,160	0,890	0,889**
	GWPR	26,595	41,616	0,955	0,948**

Nota: *reglin*, *regpoisson*, *regbin*, GWR, GWPR, MAE, RMSE, *SD_ratio* e R se referem, respectivamente, à regressão linear, regressão de Poisson, regressão binomial negativa, regressão geograficamente ponderada, regressão de Poisson geograficamente ponderada, erro médio absoluto, raiz quadrada do erro quadrado médio, razão entre desvio padrão dos valores reais e previstos e coeficiente de correlação linear de *Pearson*. ** A correlação é significativa ao nível de confiança de 99% (1 extremidade).

Nota-se que, de uma forma geral, as técnicas podem ser ranqueadas, do desempenho mais fraco ao mais forte, como se segue: 1) regressão Binomial Negativa; 2) regressão linear tradicional; 3) regressão de Poisson; 4) GWR; e 5) GWPR. O desempenho inferior da *regbin* frente ao modelo linear e de Poisson provavelmente provém da não ocorrência do fenômeno da sobredispersão na variável de interesse. Retomando a descrição contida na Seção 2, a *regpoisson* representa uma contribuição ao tratamento da real natureza das variáveis de interesse, o que inclui a assimetria de Desembarques; GWR, que se refere a um modelo para dados contínuos, lida com a autocorrelação e heterogeneidade espacial; GWPR consegue incorporar as duas vantagens das regressões anteriores. Com base nisso e utilizando os resultados de MAE para Desembarques, os ganhos, ou seja, reduções relativas no erro provenientes da incorporação da assimetria e da autocorrelação, de forma isolada e em conjunto, ao processo de modelagem do volume de Desembarques no âmbito de pontos de parada, podem ser ilustrados da seguinte forma: -19,71% na *regpoisson*; -24,98% na GWR; -48,10% na GWPR, usando, como referência, o erro médio absoluto obtido na regressão linear.

A fim de analisar, de forma desagregada, a taxa de erros encontrada para a regressão linear, de Poisson, GWR e GWPR, apresentam-se, na Figura 1, mapas contendo a variação do desvio, em porcentagem, entre o valor real e estimado por esses modelos para o número de

Desembarques por ponto de parada. Para uma melhor visualização dos resultados, preferiu-se utilizar o *buffer* que representa a área de influência do ponto de parada ao invés do ponto propriamente dito. Cabe lembrar ainda que o primeiro ponto de Desembarques possui um volume observado de passageiros igual a 0, motivo pelo qual ele foi omitido nos mapas.

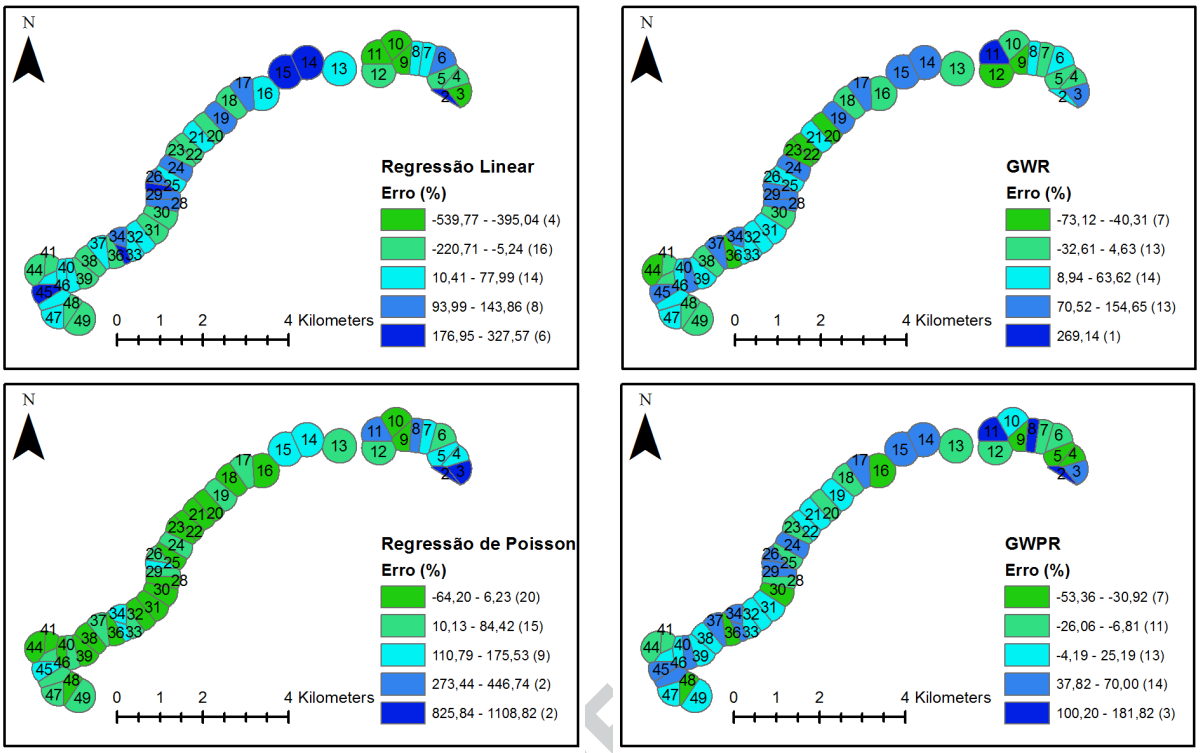

Figura 1: Taxas de erro provenientes de modelos globais e locais para Desembarques

A partir da numeração das áreas de influência, percebe-se que a linha se inicia na extremidade nordeste e finaliza seu itinerário no canto sudoeste. A Figura 1 corrobora, mais uma vez, a melhoria gradual no desempenho das estimativas a partir da regressão linear, passando pela regressão de Poisson, pela GWR e finalizando na GWPR. Constata-se, nos modelos locais, com destaque para a GWPR, valores extremos (mínimos e máximos) menores que para os demais. Além disso, a representatividade das categorias intermediárias de erro aumenta de acordo com a sequência: *reglin*, *regpoisson*, GWR e GWPR, indicando um melhor desempenho dos últimos modelos frente aos primeiros. Os melhores resultados dos modelos locais frente aos globais se refletem nas medidas de tendência central e dispersão do erro em porcentagem: *reglin*, *regpoisson*, GWR e GWPR apresentam, respectivamente, um erro médio absoluto de 113%, 108%, 55% e 36%, com desvios padrão de 124%, 200%, 51% e 34%, aproximadamente. Considerando-se a mediana do erro absoluto, os seguintes valores são observados: 72% para a *reglin*; 50% para a *regpoisson*; 42% para a GWR; e 28% para a GWPR.

Uma análise minuciosa dos percentis da distribuição dos erros em porcentagem, para os 48 pontos de parada expostos nos mapas da Figura 1, permite extrair o seguinte resultado: considerando-se a faixa de erros que varia de -30% a 30%, aproximadamente, a porcentagem do banco de dados situada nesse intervalo é de 24%, 32%, 41% e 49%, para a *reglin*, *regpoisson*, GWR e GWPR, respectivamente. Dessa forma, a GWPR, que, dentre os modelos comparados, é o único que contabiliza a autocorrelação e heterogeneidade espacial, além da

real natureza dos dados de contagem, exibe aproximadamente 24 pontos de parada com um volume estimado de Desembarques com erro na estreita faixa de -30% a 30% do valor real.

Basicamente, os modelos globais se diferenciam dos locais na medida em que, no segundo tipo de regressão, pontos de parada com valores iguais das variáveis explicativas dificilmente terão um valor previsto idêntico para Desembarques, já que, nesse caso, o resultado também depende do arranjo espacial dos pontos de parada. Além disso, a GWR e GWPR são considerados modelos locais por permitirem, entre outras conveniências, a discriminação da heterogeneidade espacial dos parâmetros do modelo. Nesse contexto, tais regressões contribuem para o conhecimento da forma com que a demanda por TP de cada região responderia, de forma local, a mudanças no uso do solo e no sistema de transporte, guiando o planejamento urbano associado ao de transportes. Conforme mostrado na Tabela 2, a amplitude de variação dos parâmetros na GWR e GWPR corroboram a existência dessa heterogeneidade espacial. Por fim, uma vez que ambos os métodos originam uma superfície contínua de valores estimados, abrangendo, também, os locais/pontos de parada não amostrados, é possível, principalmente no caso da GWPR, a obtenção de estimativas aceitáveis com um número reduzido de informações. A seção a seguir sintetiza as conclusões alcançadas no presente trabalho e pontua sugestões para trabalhos futuros.

4. CONCLUSÕES E RECOMENDAÇÕES FINAIS

Partindo da observação de que, na literatura científica, não foram encontrados trabalhos que avaliassem o impacto da incorporação da autocorrelação espacial, isolada e em conjunto com a assimetria de variáveis que representam a demanda por transporte público no âmbito de pontos de parada, o presente trabalho propôs a aplicação e comparação de modelos globais e locais para dados contínuos e discretos à variável de Desembarques ao longo de uma linha de ônibus da cidade de São Paulo. Conforme esperado, os resultados obtidos, a partir de métricas de aderência, mostraram que, de fato, há uma melhoria gradual nas estimativas à medida em que as duas peculiaridades da demanda por transportes são tratadas pela modelagem.

Nesse contexto, o presente trabalho buscou contribuir para a solidificação e avanço metodológico da modelagem de embarques e desembarques no âmbito de pontos de parada, utilizando uma interface de programação aberta, livre e de fácil acesso. Entretanto, cabe ressaltar que a seleção da melhor técnica dependerá fortemente da natureza do banco de dados disponível, cujas características devem ser exploradas em fase anterior à de modelagem, bem como do propósito da investigação. Uma contribuição prática empreendida pelo artigo diz respeito à ferramenta GWPR incluída na análise: por conseguir gerar estimativas em qualquer ponto do espaço, esse modelo pode ser aplicado, com sucesso, a variáveis de difícil aquisição / alto custeio, como, por exemplo, Embarques e Desembarques, a fim de se obter o número de passageiros embarcando e desembarcando em pontos de parada não amostrados em uma pesquisa sobe/desce. Aliado a esse fato, as relações dos preditores socioeconômicos e demográficos, de uso do solo e do sistema de transportes com a demanda por transporte público por ônibus, discriminadas no presente trabalho, também representam uma contribuição prática, na medida em que subsidiam a promoção de boas políticas no planejamento urbano orientado ao transporte sustentável. Por fim, no intuito de estimular a consolidação da modelagem apropriada da demanda por TP no âmbito de pontos de parada, alguns tópicos podem ser recomendados para trabalhos futuros, tais como: calcular métricas de aderência com base em uma amostra de validação à parte da de calibração utilizada na presente análise. Esse procedimento permitiria verificar se as técnicas de melhor desempenho

na calibração também se destacariam na validação. Além disso, tendo em vista que os dados de Embarques e Desembarques se referem à agregação de um total de 4 faixas horárias, seria interessante tratar cada um desses conjuntos de viagens, de forma desagregada.

Agradecimentos
Às agências de fomento FAPESP (Processo 2019/12054-4) e CNPq (Processo 304345/2019-9). Os autores também agradecem à SPTrans, pela pesquisa de Embarque e Desembarque utilizada nesse trabalho.

REFERÊNCIAS BIBLIOGRÁFICAS

Bivand, R. S.; Pebesma, E. e Gomez-Rubio, V. (2013) Applied spatial data analysis with R, 2.ed. Springer, NY. https://asdar-book.org/

Blainey, S., e Mulley, C. (2013) Using geographically weighted regression to forecast rail demand in the Sydney region. *Australasian Transport Research Forum*.

Blainey, S., e Preston, J. (2010) A geographically weighted regression based analysis of rail commuting around Cardiff, South Wales. *12th World Conference on Transport Research*. Lisbon, Portugal.

Brunsdon, C., Fotheringham, A. S., e Charlton, M. E. (1996) Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geographical Analysis*, *28*(4), 281–298. doi:10.1111/j.1538-4632.1996.tb00936.x

Calcagno, V. (2019). glmulti: Model Selection and Multimodel Inference Made Easy. R package version 1.0.7.1. https://CRAN.R-project.org/package=glmulti

Cardozo, O. D., García-Palomares, J. C., e Gutiérrez, J. (2012) Application of geographically weighted regression to the direct forecasting of transit ridership at station-level. *Applied Geography*, *34*(Supplement C), 548–558. doi:https://doi.org/10.1016/j.apgeog.2012.01.005

Cervero, R. (2006) Alternative Approaches to Modeling the Travel-Demand Impacts of Smart Growth. *Journal of the American Planning Association*, *72*(3), 285–295. doi:10.1080/01944360608976751

Cervero, R., e Dai, D. (2014) BRT TOD: Leveraging transit oriented development with bus rapid transit investments. *Transport Policy*, *36*, 127–138. doi:https://doi.org/10.1016/j.tranpol.2014.08.001

Chiou, Y. C., Jou, R. C., e Yang, C. H. (2015) Factors affecting public transportation usage rate: Geographically weighted regression. *Transportation Research Part A: Policy and Practice*, *78*, 161–177. doi:10.1016/j.tra.2015.05.016

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. doi:10.1007/s11116-011-9368-3

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

da Silva, A. R., e Rodrigues, T. C. V. (2014) Geographically Weighted Negative Binomial Regression—incorporating overdispersion. *Statistics and Computing*, *24*(5), 769–783. doi:10.1007/s11222-013-9401-9

Ewing, R., Tian, G., Goates, J. P., Zhang, 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. doi:10.1177/0042098014560991

Fotheringham, A. S., Brunsdon, C., e Charlton, M. (2003) *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons.

Gan, Z., Feng, T., Yang, M., Timmermans, H., e Luo, J. (2019) Analysis of Metro Station Ridership Considering Spatial Heterogeneity. *Chinese Geographical Science*, *29*(6), 1065–1077. doi:10.1007/s11769-019-1065-8

Gollini, I; Lu, B.; Charlton, M.; Brunsdon, C. e Harris, P. (2015) GWmodel: An R Package for Exploring Spatial Heterogeneity Using Geographically Weighted Models. *Journal of Statistical Software*, *63*(17), 1-50. http://www.jstatsoft.org/v63/i17/.

Gutiérrez, J., Cardozo, O. D., e García-Palomares, J. C. (2011) Transit ridership forecasting at station level: an approach based on distance-decay weighted regression. *Journal of Transport Geography*, *19*(6), 1081–1092. doi:https://doi.org/10.1016/j.jtrangeo.2011.05.004

Hebbali, A. (2020). olsrr: Tools for Building OLS Regression Models. R package version 0.5.3. https://CRAN.R-project.org/package=olsrr

Hensher, D. A., e Golob, T. F. (2008) Bus rapid transit systems: a comparative assessment. *Transportation*, *35*(4), 501–518. doi:10.1007/s11116-008-9163-y

Hensher, D. A., Li, Z., e Mulley, C. (2014) Drivers of bus rapid transit systems – Influences on patronage and service frequency. *Research in Transportation Economics*, *48*, 159–165. doi:https://doi.org/10.1016/j.retrec.2014.09.038

Hilbe, J. M. (2014) *Modeling Count Data*. Cambridge University Press, Cambridge. doi:DOI:

10.1017/CBO9781139236065

Hollander, Y., e Liu, R. (2008) The principles of calibrating traffic microsimulation models. *Transportation*, *35*(3), 347–362. doi:10.1007/s11116-007-9156-2

Joonho, K., Daejin, K., e Ali, E. (2019) Determinants of Bus Rapid Transit Ridership: System-Level Analysis. *Journal of Urban Planning and Development*, *145*(2), 4019004. doi:10.1061/(ASCE)UP.1943-5444.0000506

Kalaanidhi, S., e Gunasekaran, K. (2013) Estimation of Bus Transport Ridership Accounting Accessibility. *Procedia - Social and Behavioral Sciences*, *104*, 885–893. doi:https://doi.org/10.1016/j.sbspro.2013.11.183

Kyte, M., Stoner, J., e Cryer, J. (1985) Development and application of time-series transit ridership models for Portland, Oregon. *Transportation Research Record*, *1036*, 9–18.

Liu, Y., Ji, Y., Shi, Z., e Gao, L. (2018) The Influence of the Built Environment on School Children's Metro Ridership: An Exploration Using Geographically Weighted Poisson Regression Models. *Sustainability*, *10*(12), 4684.

Lu, B.; Harris, P.; Charlton, M. e Brunsdon, C. (2014). The GWmodel R package: further topics for exploring spatial heterogeneity using geographically weighted models. *Geo-spatial Information Science*, *17*(2), 85-101. https://www.tandfonline.com/doi/abs/10.1080/10095020.2014.917453

Ma, X., Zhang, J., Ding, C., e Wang, Y. (2018) A geographically and temporally weighted regression model to explore the spatiotemporal influence of built environment on transit ridership. *Computers, Environment and Urban Systems*, *70*, 113–124. doi:10.1016/j.compenvurbsys.2018.03.001

Metrô (2019) Pesquisa de Origem e Destino de 2017 (Banco de dados). Companhia do Metropolitano De São Paulo, Secretaria Estadual dos Transportes Metropolitanos. Disponível em: http://www.metro.sp.gov.br/pesquisa-od/. Acesso em: abr. 2020

Moran, P. A. P. (1948) The interpretation of statistical maps. *Journal of the Royal Statistical Society. Series B (Methodological)*, *10*(2), 243–251.

Nakaya, T., Fotheringham, A. S., Brunsdon, C., e Charlton, M. (2005) Geographically weighted Poisson regression for disease association mapping. *Statistics in Medicine*, *24*(17), 2695–2717. doi:10.1002/sim.2129

Pebesma, E.J. e Bivand, R.S. (2005) Classes and methods for spatial data in R. *R News*, *5*(2), https://cran.r-project.org/doc/Rnews/.

Peng, Z.-R., Dueker, K. J., Strathman, J., e Hopper, J. (1997) A simultaneous route-level transit patronage model: demand, supply, and inter-route relationship. *Transportation*, *24*(2), 159–181. doi:10.1023/A:1017951902308

Pielke Sr, R. A. (2013) *Mesoscale meteorological modeling*. Academic press.

Pulugurtha, S. S., e Agurla, M. (2012) Assessment of models to estimate bus-stop level transit ridership using spatial modeling methods. *Journal of Public Transportation*, *15*(1), 33–52. Obtido de https://scholarcommons.usf.edu/cgi/viewcontent.cgi?article=1095&context=jpt

R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ryan, S., e Frank, L. (2009) Pedestrian Environments and Transit Ridership. *Journal of Public Transportation*, *12*(1), 39–57. doi:10.5038/2375-0901.12.1.3

Sakamoto, Y., Ishiguro, M., e Kitagawa, G. (1986) *Akaike information criterion statistics*. D. Reidel Publishing Company.

Siddiqui, S., Amirhossein, J., e Hossain, F. (2015) *Increasing Transit Ridership in Small Urban Areas : A case study of Streamline in Bozeman , MT*. doi:10.13140/RG.2.1.3488.5847

Song, Y., Merlin, L., e Rodriguez, D. (2013) Comparing measures of urban land use mix. *Computers, Environment and Urban Systems*, *42*, 1–13. doi:10.1016/j.compenvurbsys.2013.08.001

Taylor, B. D., Miller, D., Iseki, H., e Fink, C. (2009) Nature and/or nurture? Analyzing the determinants of transit ridership across US urbanized areas. *Transportation Research Part A: Policy and Practice*, *43*(1), 60–77. doi:https://doi.org/10.1016/j.tra.2008.06.007

Venables, W. N. e Ripley, B. D. (2002) *Modern Applied Statistics with S*. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Yan, X., e Su, X. G. (2009) *Linear regression analysis: theory and computing*. World Scientific.

Zhao, F., Chow, L.-F., Li, M.-T., Ubaka, I., e Gan, A. (2003) Forecasting Transit Walk Accessibility: Regression Model Alternative to Buffer Method. *Transportation Research Record*, *1835*(1), 34–41. doi:10.3141/1835-05

Zhu, Y., Chen, F., Wang, Z., e Deng, J. (2019) Spatio-temporal analysis of rail station ridership determinants in the built environment. *Transportation*, *46*(6), 2269–2289. doi:10.1007/s11116-018-9928-x