XXIII SEMEAD
Seminários em Administração

novembro de 2020
ISSN 2177-3866

Estratégia de Big Data Analytics de simulação de dados de mobilidade como abordagem para produção de indicadores sobre o transporte público

RODOLFO OLIVEIRA LORENZO
ESCOLA DE ADMINISTRAÇÃO DE EMPRESAS DE SÃO PAULO (FGV-EAESP)

EDUARDO DE REZENDE FRANCISCO
ESCOLA DE ADMINISTRAÇÃO DE EMPRESAS DE SÃO PAULO (FGV-EAESP)

Agradecimento à orgão de fomento:
Agradecimento ao CNPQ e ao GV Pesquisa por terem possibilitado a realização dessa pesquisa

Estratégia de *Big Data Analytics* de simulação de dados de mobilidade como abordagem para produção de indicadores sobre o transporte público

RESUMO

A transformação digital e as tecnologias de *Big Data Analytics* trazem inúmeras oportunidades para questões de interesse público. Pesquisas importantes de mobilidade urbana, como a Pesquisa Origem-Destino de São Paulo (Metrô, 2008), poderiam se beneficiar da possibilidade de obter dados a partir dessas novas estratégias. O presente trabalho analisa os tempos de viagem de transporte público e privado motorizado no município de São Paulo, simulando 257.400 viagens através da *Application Programming Interface* de roteamento do Google Maps. Foram analisadas duas medidas comparativas – diferença e razão dos tempos do transporte público pelos tempos do transporte privado motorizado – a partir da modelagem considerando regressão espacial por distritos com variáveis estruturais de transporte e socioeconômicas como controles. Os resultados indicam menores tempos para o transporte privado, mas desigualmente distribuídos pelo espaço. A presença de infraestrutura de transporte mostrou-se importante, mas variáveis socioeconômicas, como densidade populacional e raça, também se mostraram relevantes para explicar essa diferença. As regressões melhoraram o poder de explicação de 48,95% e 30,5% para 76,06% e 66,96%, respectivos as duas variáveis. Essa abordagem oferece uma perspectiva inovadora e tempestiva de *Big Data Analytics* para a gestão de informação sobre mobilidade urbana e transporte público no contexto municipal.

Palavras-chave: *Mobilidade urbana, Simulação de dados, Big Data Analytics.*

1. INTRODUÇÃO

A mobilidade urbana é um elemento explicativo essencial para entender questões urbanas, desde a dinâmica de valorização imobiliária (Heilmann, 2018; Fesselmeyer and Liu, 2018), da segregação urbana (Garcia-López e Moreno-monroy, 2018; Wang e Mu, 2018; Tabuchi, 2019) e da própria estrutura urbana (Behrens et al., 2017). Um dos desafios em realizar estudos empíricos quantitativos de mobilidade é o custo, em tempo e recursos, para a produção de dados confiáveis e tempestivos. Uma referência metodológica nesse sentido são as pesquisas domiciliares de Origem e Destino, usadas não só para avaliar os fluxos de pessoas e recursos nas cidades, mas também as condições de realização desses fluxos e as condições socioeconômicas subjacentes aos comportamentos de mobilidade. Estudos extensos sobre mobilidade que utilizam essa abordagem são pouco frequentes devido a seu custo; assim os intervalos entre edições de pesquisas do gênero são relativamente longos. Estratégias de pesquisa alternativas, com tempestividade e confiabilidade satisfatória, são importantes. A simulação de dados, potencializada pelo advento do *Big Data Analytics*, é uma dessas. Ela pode ser feita a partir de suposições acerca do comportamento gerador de viagens e das condições de mobilidade específicas (Tribby e Zandbergen, 2012). Uma das desvantagens dessa abordagem é que é preciso supor uma certa racionalidade dos agentes em movimento e certos comportamentos podem introduzir vieses importantes e ignorar tipos específicos de mobilidade (Kwan, 1998) nas simulações de origem e destino. Porém, abordagens que levem em consideração muitos fatores na previsão da mobilidade podem gerar modelos excessivamente complexos. O presente trabalho busca explorar uma alternativa simples e robusta, através de ferramentas de *Big Data Analytics*, para simular os tempos de mobilidade, de forma a comparar a mobilidade do transporte privado e do transporte público.

1.1. Problema da Pesquisa

O trabalho buscou comparar o modal de transporte privado com o modal de transporte público, explicitando o custo temporal relativo do transporte público. A partir do uso da interface de programação (ou doravante API, tradicional acrônimo para *Application Programming Interface*) do Google Maps, foram simuladas 128.700 viagens para cada um dos dois modais de transporte – para as viagens de transporte público foi registrada uma perda de 1,53% das viagens por falta de rotas disponíveis pela API. A comparação das viagens foi analisada a partir de visualizações e modelagens espaciais para verificar a estrutura dos dados simulados – procurando caracterizar possíveis vieses na simulação – e as dependências dos dados em relação a variáveis socioeconômicas e de infraestrutura de transporte, com o intuito de validar a relação dos dados simulados às variáveis físico-sociais do município de São Paulo.

1.2. Justificativas

Diante de novas estratégias de pesquisa, ainda é preciso ter em mente que a realização de qualquer pesquisa requer capacidade estatística instalada (Dargent et al., 2018), o que não é o caso para muitos dos países ou regiões que necessitam com urgência de dados para auxiliar na produção de suas políticas (Letouzé e Jütting, 2014). Há ao mesmo tempo uma discussão sobre o quanto as estatísticas existentes são capazes de atender as necessidades de quem necessita desses dados – devido a sua tempestividade ou precisão – Letouzé e Jütting (2014) discutem uma "desilusão estatística": há um descontentamento com a capacidade das burocracias estatais em produzir estatísticas confiáveis e relevantes, ao mesmo tempo que, em países pobres e em desenvolvimento, essa desilusão está associada a baixa capacidade estatística existente. À fragilidade institucional se soma o desafio da crescente produção de dados e das novas formas de análises estatísticas que acompanham essa nova geração de produção de dados, chamada de *Big Data*; constitui-se uma dupla camada de desafios para países emergentes. Nesse sentido, a aproximação de estratégias de *Big Data* (recentemente ressignificada com a adoção do termo completo *Big Data Analytics*, dotando-o de competências analíticas) aos já tradicionais estudos de mobilidade representa um passo na construção de dados mais acurados paras as intervenções urbanas.

2. REVISÃO DE LITERATURA

2.1. *Big Data*

Em que consiste o movimento de *Big Data*? Uma das primeiras definições de *Big Data* está relacionada às características dos dados envolvidos. O aumento da produção, capacidade de armazenamento e processamento de dados gerou uma grande potencialidade de aplicações analíticas. A princípios são três as dimensões definidoras dos dados envolvidos: Volume, Velocidade e Variedade (McAfee et al., 2012; Gandomi e Haider, 2015). Ainda nessa direção existem definições que destacam outras características dos dados usados: Veracidade (em relação a dados como o estado socioemocional de usuários de redes sociais); Variabilidade e Complexidade (variabilidade em relação aos ritmos do fluxo de dados e complexidade em relação ao uso de diversas fontes para os dados); e Valor (em relação ao baixo valor de um dado singular em comparação com o valor do agregado de dados) (Gandomi e Haider, 2015).

Por outro lado, outras definições de *Big Data* que partem de pressupostos diferentes. Letouzé e Jütting (2014) definem o movimento a partir de características sociológicas. Os conceitos definidores de *Big Data* seriam três Cs – *digital* "Crumbs" (ou migalhas digitais, do inglês): a natureza dos dados, gerados como rastros de atividade humana na rede; "Capacities" (ou capacidades): as técnicas envolvidas na geração de "insights" a partir desses dados; e "Communities" (ou comunidades): as comunidades que dominam essas técnicas e desenvolvem

essas aplicações, que incluem tanto as comunidades de *softwares* abertos a grupos dentro dos setores privado e de inteligência. Outras definições partem ainda de critérios voltados à implementação de sistemas, como a classificação de arquiteturas de *Big Data* (Pääkkönen e Pakkala, 2015). A relação entre as estatísticas oficiais e o *Big Data* pode ser vista como representativa do conflito sobre a capacidade do Estado de fornecer dados ágeis e úteis. Por um lado, o *Big Data* é capaz de produzir informações a partir de dados produzidos em tempo real, coletados de diversas fontes. É possível, a partir dessa capacidade, tentar reproduzir os indicadores oficiais já existentes, ou outros, mais granulares e inteligentes. Em sociedades com a infraestrutura de comunicação e recursos humanos capacitados para isso, essas estratégias prometem reduzir o custo de produção de indicadores sem perda de qualidade. Porém Letouzé e Jütting (2014) argumentam que a responsabilidade das agências oficiais, ao produzir os dados oficiais, não é só gerar informações úteis: elas têm a função de produzir conhecimento sobre a sociedade. Além disso, elas são responsáveis por constituir um espaço deliberativo sobre o que merece ser medido na sociedade. Considerando o movimento de *Big Data* como um importante vetor de mudança na sociedade, Letouzé e Jütting (2014) consideram interessante movimentos de integração entre as comunidades responsáveis pelas estatísticas oficiais e essas novas técnicas de análise.

A produção de dados georreferenciados relativos à mobilidade é essencial para captar a distribuição da mobilidade no tecido urbano. Dentro dos meios de *Big Data*, os dados gerados pela utilização dos celulares – ainda mais no contexto em que volume da rede móvel supera o volume de rede fixa (Lee & Kang, 2015) – já fornece um grande volume de dados georreferenciados e sobre os meios de transporte. Essa produção massiva permite o uso desses dados para análises em tempo real, como fazem os serviços de roteamento de transporte. A compreensão da dimensão geográfica dos problemas, da distribuição da infraestrutura presente e dos serviços ajudam a diagnosticar ineficiências e priorizar esforços, permitindo uma visão sistêmica dos indicadores sociais e da prestação de serviços (Francisco, 2010). Essa visão pode ajudar a escolher combinações de diferentes formas de intervenção pública (Torres et al., 2003, Torres e Oliveira, 2001). Dados derivados dos novos aplicativos que usam a localização podem permitir o acesso a informações de mobilidade de maneira menos custosa, ainda que contendo algum grau de viés (Kwan, 2016) – dados que podem fornecer informações valiosas sobre os padrões de mobilidade e acessibilidade das cidades (Noulas, Scellato, Lambiotte, Pontil, Mascolo, 2012; Wang e Mu, 2018). Ao mesmo tempo a disponibilidade de dados e técnicas utilizando *Big Data* deve ser vista com cautela. Kwan (2016) alerta para vieses decorrentes do uso de algoritmos de *Big Data*. Mesmo que esses vieses não sejam particularidades dessas estratégias, o uso intensivo de algoritmos de análise tem o potencial de gerar interferência nos dados sem que seja possível ao pesquisador acompanhar os dados que serão usados, dado o seu volume. Por essa razão a importância da validação de estratégias de *Big Data* junto a estratégias tradicionais é importante para discernir os possíveis vieses introduzidos pelo processamento de dados.

2.2. Mobilidade Urbana

Em relação à mobilidade, a compreensão das formas de usos de diferentes modais em cada região pode ajudar a associar os padrões de mobilidade a certos grupos sociais, permitindo pensar em políticas voltadas para equilibrar os usos do espaço público para melhorar a mobilidade de quem mais precisa. Em São Paulo, estudos nessa direção identificam a dependência mais acentuada dos moradores periféricos de modais coletivos em relação aos individuais, mas também identificam uma expressiva periferia motorizada, que demanda espaço urbano para sua mobilidade (Requena, 2015). Há a associação entre os tempos médios

de viagem e a acessibilidade à rede de transportes rápidos (trem e metrô) nos distritos de São Paulo, e essas por sua vez têm associação com as rendas médias dos distritos, o que contribui para uma distribuição desigual da mobilidade (Morandi et al., 2013). A escolha da cidade de São Paulo como modelo da simulação foi feita em razão da extensa literatura que caracteriza a formação da cidade: o desenvolvimento do padrão centro – periferia (ou rico versus pobre), paradigmático do caso brasileiro (Kowarick, 19780; Maricato, 2003; Rolnik e Klink, 2011), a importância dos trajetos pendulares (Aranha, 2005) e a distribuição diferencial da infraestrutura e dos indicadores socioeconômicos pela cidade (Torres e Oliveira, 2001; Torres et al., 2003). Outro ponto importante é a centralidade que o transporte privado motorizado teve historicamente na mobilidade urbana (Júnior, 2011; Gakenheimer, 1999; Silveira e Cocco, 2013; Wilheim, 2013; Scaringella, 2001) que contribuem para o aumento do custo relativo do transporte público em relação ao transporte privado motorizado. Há que se considerar que a experiência de urbanização acelerada vivenciado por São Paulo também encontra eco em outros países em desenvolvimento de urbanização recente (Gakenheimer, 1999), que aumenta o valor comparativo do município.

3. MÉTODO

A estrutura analítica deste estudo está apresentada em dois grandes blocos: a simulação dos dados de mobilidade e a análise do banco de dados gerado pela simulação.

3.1. Simulação de viagens

A simulação foi feita em duas etapas: primeiro, a geração de um banco de endereços, e depois, a simulação das viagens propriamente ditas. Algumas considerações precisam ser feitas. A relativa alta complexidade de simulações locais que consigam captar o comportamento em tempo real – com informações de trânsito – da mobilidade implicou o uso de alguma ferramenta de previsão de tempos de viagem já estabelecida e acessível por meio remoto. A escolha feita (pela API Distance Matrix da Google) implicou em um número limitado de requisições de viagens, por questões de custo; essa limitação levou a escolhas para reduzir o número de viagens "perdidas" na simulação, incorridas quando as coordenadas usadas na API não correspondiam ou não podiam ser aproximadas a endereços válidos, como no caso de coordenadas nas represas de São Paulo. Ao mesmo tempo, foi feita a opção por usar a computação em nuvem para a simulação, o que levou ao esforço de reduzir a computação necessária para evitar problemas relacionados ao desempenho. Essas limitações definiram o processo de definição de endereços. Procurou-se sortear endereços em regiões mais densamente povoadas para evitar possíveis perdas, seguindo um modelo de densidade de probabilidade da população. Ao mesmo tempo, para reduzir o esforço computacional, foi montada uma base de coordenadas *offline*, que foi usada para sortear os endereços das viagens. O processo de geração da base foi feito nas seguintes etapas: 1) O mapa do município de São Paulo (em formato *shapefile*) foi dividido por uma grade com quadrículas de 500 metros de lado; 2) Foram calculadas as populações de cada quadrícula com dados do Censo Demográfico de 2010 e retiradas as quadrículas com população igual a zero; 3) As quadrículas foram divididas em quintis de densidade populacional, e sorteados aleatoriamente pontos geográficos dentro de cada quadrícula, de acordo com o quintil: 5 pontos para o quintil mais populoso, 4 para o 2º quintil, 3 para o 3º, 2 para o 2º e 1 ponto para o quintil menos populoso. 4) O conjunto de pontos resultante foi usado como base para o sorteio dos endereços de origem. Essa primeira etapa foi realizada utilizando bases cartográficas abertas do município de São Paulo e o software aberto QGIS. A simulação das viagens foi feita a partir de um programa desenvolvido em Python, executado no serviço de computação em nuvem da Google. A estrutura do programa seguiu a arquitetura apresentada

na Figura 1. Foi usada uma ferramenta de agendamento (Google Scheduler) de ativação ligada a uma máquina virtual no ambiente em nuvem da Google, para que a chamada ao serviço de viagem do Google Maps fosse realizada nos dias úteis da semana, a cada hora cheia, das cinco da manhã até às nove da noite. A intenção do espaçamento era obter amostragens de viagens em diferentes horários para comparar periodicidades diárias e horárias nas viagens. O programa seguiu as etapas descritas na Figura 1.

Abrir uma conexão com o banco de dados SQL da nuvem da Google
Carregar o banco de coordenadas
Sortear dez coordenadas de origem e dez coordenadas de destino
Chamar a API Distance Matrix com as dez origens e os dez destinos, para viagens de transporte público
Processar os resultados devolvidos pela API e armazenar em um vetor auxiliar
Chamar novamente a API Distance Matrix com as mesmas dez origens e os dez destinos, para viagens de transporte privado
Processar os resultados devolvidos pela API e anexar ao vetor auxiliar
Submeter o vetor auxiliar à função que insere os dados no Banco de Dados hospedado na nuvem.

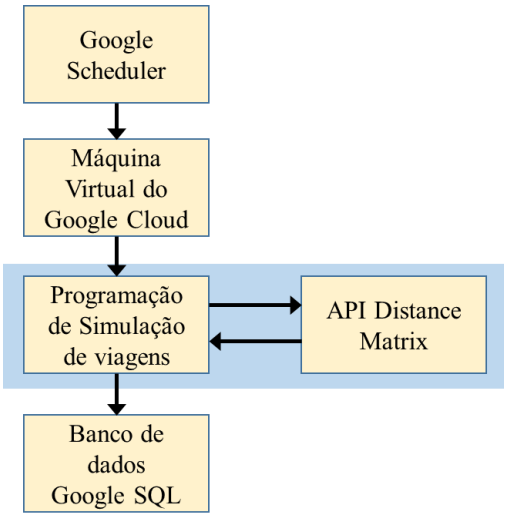

Figura 1. Etapas e Estrutura funcional do programa elaborado para simular os dados

Fonte: Os autores

Cada chamada da API Distance Matrix retornou uma lista com duzentas viagens com dados de horário e dia da semana, coordenadas da origem e do destino, endereços da origem e do destino da viagem, duração, distância e tarifa da viagem. Para cada par origem-destino houve registro de viagem de carro e de transporte público. O período de simulação foi entre os dias 11 de fevereiro de 2019 a 5 de junho de 2019. O total de viagens armazenadas no banco de dados nesse período foi de 257.400 viagens, sendo 253.450 viagens válidas – 128.700 (100% de aproveitamento) das viagens de carro e 126.725 das viagens transporte público (98,47% de aproveitamento).

3.2. Estrutura Analítica do Banco de Dados

A análise de dados seguiu três etapas: (i) análise exploratória do agregado de todas as viagens, (ii) análise exploratória da distribuição espacial das medidas obtidas no município, e, por fim, (iii) modelagem das medidas elaboradas a partir de variáveis socioeconômicas e variáveis de infraestrutura de transporte público nos distritos do município de São Paulo. Foram montados modelos de regressão simples (doravante OLS, ou *Ordinary Least Squares*) e modelos de regressão de autocorrelação espacial (doravante SAR, ou *Spatial Auto-Regressive models*). Para a primeira análise foram realizadas estatísticas descritivas das medidas consideradas de interesse para entender a estrutura geral dos dados simulados. A estrutura das entradas do banco de dados após o processamento de dados está apresentada na Tabela 1.

Tabela 1. Variáveis do banco de dados na nuvem

ID	Data	Hora	Dia	Latitude da origem	Longitude da origem	Endereço da origem
Latitude do destino	Longitude do destino	Endereço do destino	Duração (segundos)	Distância (metros)	Tarifa	Modal

Fonte: Os autores

A partir do pareamento das viagens por modal (público e privado) foram criadas duas medidas-alvo: (1) a diferença entre o tempo de viagem do modal público e do modal privado e (2) a razão entre o tempo de viagem do modal público pelo modal privado, ou "tempo relativo" que o transporte público demora mais que o transporte privado, conforme equações a seguir.

$$\textbf{(I)}\ D_t = T_{público} - T_{privado} \qquad \textbf{(II)}\ R_t = \frac{T_{público}}{T_{privado}}$$

As distribuições das medidas foram testadas para checar a sua normalidade, que foi considerada satisfatória. Em seguida, os dados de viagens devido a sua natureza eminentemente geográfica foram analisados a partir de abordagens espaciais. Para D_t e R_t foram feitas análises de autocorrelação espacial (utilizado para identificar padrões de dispersão, clusterização ou aleatoriedade) através do *I* de Moran e de mapas de Indicadores Locais de Associação Espacial (LISA), tanto para a granularidade de distritos como de áreas de ponderação. A partir dessas análises foi possível verificar a clusterização dessas medidas no município. Os resultados para as duas medidas foram comparados, assim como foram comparadas as diferenças entre os níveis de análise de distritos e de áreas de ponderação. Por último, foram realizadas as modelagens de regressões lineares simples e espaciais. Os modelos foram usados para descrever a distribuição e a relação entre as medidas elaboradas e o conjunto de variáveis que refletem condições socioeconômicas e de infraestrutura de transporte público nos distritos de São Paulo. Para isso as medidas D_t e R_t foram agrupadas em torno dos distritos de origem e para cada distrito foi considerada a média das medidas que partiam do distrito. Foram compilados dados socioeconômicos e de infraestrutura de transportes agregados por distrito e ponderados, caso fosse o caso, pela área dos distritos. As variáveis usadas estão descritas na Tabela 2. A partir desses dados foram calculados modelos de regressão linear, através de um processo *stepwise*, com nível de significância *a priori* de 5% e de análise de multicolinearidade. O modelo SAR foi então aplicado para as variáveis independentes finais do modelo (*X*), conforme equação (3) (W_n é a matriz de vizinhança e ρ é o coeficiente do termo espacial auto-regressivo). O software GeoDA foi utilizado para a aplicação do modelo SAR.

$$\textbf{(III)}\quad y_n = X_n\beta + \rho W_n y_n + \varepsilon_n$$

4. RESULTADOS E ANÁLISE

Os resultados indicaram que ambas as medidas D_t e R_t apresentam melhores condições de mobilidade privada que pública. Um teste *t* de diferença de médias com 95% de confiança indicou que a média da diferença de tempos está entre 3680s e 3697s; para R_t indicou com 95% de confiança que a média dessa medida está entre 2,396 e 2,404. Como esperado, as previsões de tempo de transporte público são consistentemente maiores que do transporte privado. Parte dessa diferença pode ser dada pelo fato de que nas previsões de transporte público são incluídos trechos pedestres, enquanto os trechos de transporte privado são completamente motorizados. Para cada uma das medidas foi feita uma comparação com a distribuição das médias entre as distâncias das viagens de transporte público e de transporte privado, conforme Figura 2. D_t cresce junto com as médias de distâncias de viagens; a correlação entre essas medidas é

considerável: aproximadamente 0,564. Esse dado indica que o tamanho das viagens (refletida nas médias de distância entre as viagens) tem alguma correlação com a diferença de tempos entre modais, ou seja, mesmo considerando a existência de trechos pedestres, a velocidade do transporte público é menor. A distribuição de R_t apresenta uma correlação de aproximadamente -0,359, não muito significativa, mas negativa. E apesar da correlação e da linha de regressão linear simples indicar uma relação negativa entre as distribuições, as distribuições na Figura 2 indicam visualmente que os valores de distâncias maiores tendem a um valor próximo à média da distribuição. A concepção dessa medida – a razão entre os tempos de viagem dos diferentes modais – seria, por princípio, menos variante em função da distância do que a diferença entre os tempos da viagem, uma vez que cada um dos tempos de viagem varia em função da distância. Uma possível interpretação para isso é que em viagens mais longas os trechos pedestres contam menos para a razão entre os meios de transporte, enquanto em viagens mais curtas os aumentos devidos a trechos pedestres contribuem relativamente mais para a razão dos tempos.

Tabela 2. Dados socioeconômicos e de infraestrutura de transporte público usados para modelagem

Dado do distrito	Fonte	Data
Densidade Populacional	Dados originais de população do censo demográfico de IBGE, com reajuste anual calculado pela Fundação SEADE. Retirado do portal de indicadores dos municípios paulistas (IMP) Fundação SEADE, divididos pela área dos distritos	2010/2018
Densidade de Domicílios Particulares Permanentes	Dados originais do censo demográfico de IBGE, com reajuste anual calculado pela Fundação SEADE. Retirado do portal de indicadores dos municípios paulistas (IMP) da Fundação SEADE, divididos pela área dos distritos	2010/2018
Renda per Capita – Censo Demográfico (Em reais correntes)	Dados originais do censo demográfico de IBGE. Retirado do portal de indicadores dos municípios paulistas (IMP) Fundação SEADE	2010
Densidade de Empregos (Comércio, Serviços, Indústria de Transformação, Construção Civil)	Portal Infocidade do município de São Paulo. Fonte original dos dados: Ministério do Trabalho e Emprego. Relação Anual de Informações Sociais – Rais,	2010/2016
Densidade de Estabelecimentos (Comércio, Serviços, Indústria de Transformação, Construção Civil)	Portal Infocidade do município de São Paulo. Fonte original dos dados: Ministério do Trabalho e Emprego. Relação Anual de Informações Sociais – Rais.	2010/2016
% de não brancos (pretos, pardos e indígenas)	Dados do IBGE. Censo 2010. Para referência sobre a agregação ver Fernandes (2017).	2010
Proporção de domicílios com carro e moto	Amostra Censo IBGE. A proporção de motorização por distrito (de carros e motos) a partir dos domicílios ponderados da amostra.	2010
Densidade de pontos de ônibus	Quantidade de pontos divididos pela área dos distritos. Portal Geosampa	2018
Densidade de quilometragem linhas de ônibus	Quilometragem de linhas de ônibus dividida pela área dos distritos Portal Geosampa	2018
Densidade de linhas de ônibus	Quantidade de linhas de ônibus dividida pela área dos distritos Portal Geosampa	2018
Acesso a Estações de Metrô	Variável dummy para distritos com estações de metrô a no máximo 200 metros de seus limites. Portal Geosampa	2018
Acesso a Estações da CPTM	Variável dummy para distritos com estações de CPTM a no máximo 200 metros de seus limites. Portal Geosampa	2018

Fonte: Confeccionado pelos autores, a partir das fontes descritas para cada variável

A primeira etapa da análise espacial foi a identificação de *clusters* das medidas nos distritos e nas áreas de ponderação de São Paulo. Para isso as medidas foram agregadas à divisão

geográfica da origem das viagens – para cada zona de origem foi calculada a média das medidas relativas à zona. A Figura 3 indica a distribuição das duas medidas nos distritos. A partir dessa agregação foram calculados a partir do software R e do pacote `spdep` o indicador de autocorrelação espacial *I* de Moran (Figura 2) e os mapas de autocorrelação espacial local (Figuras 4, 5 e 6).

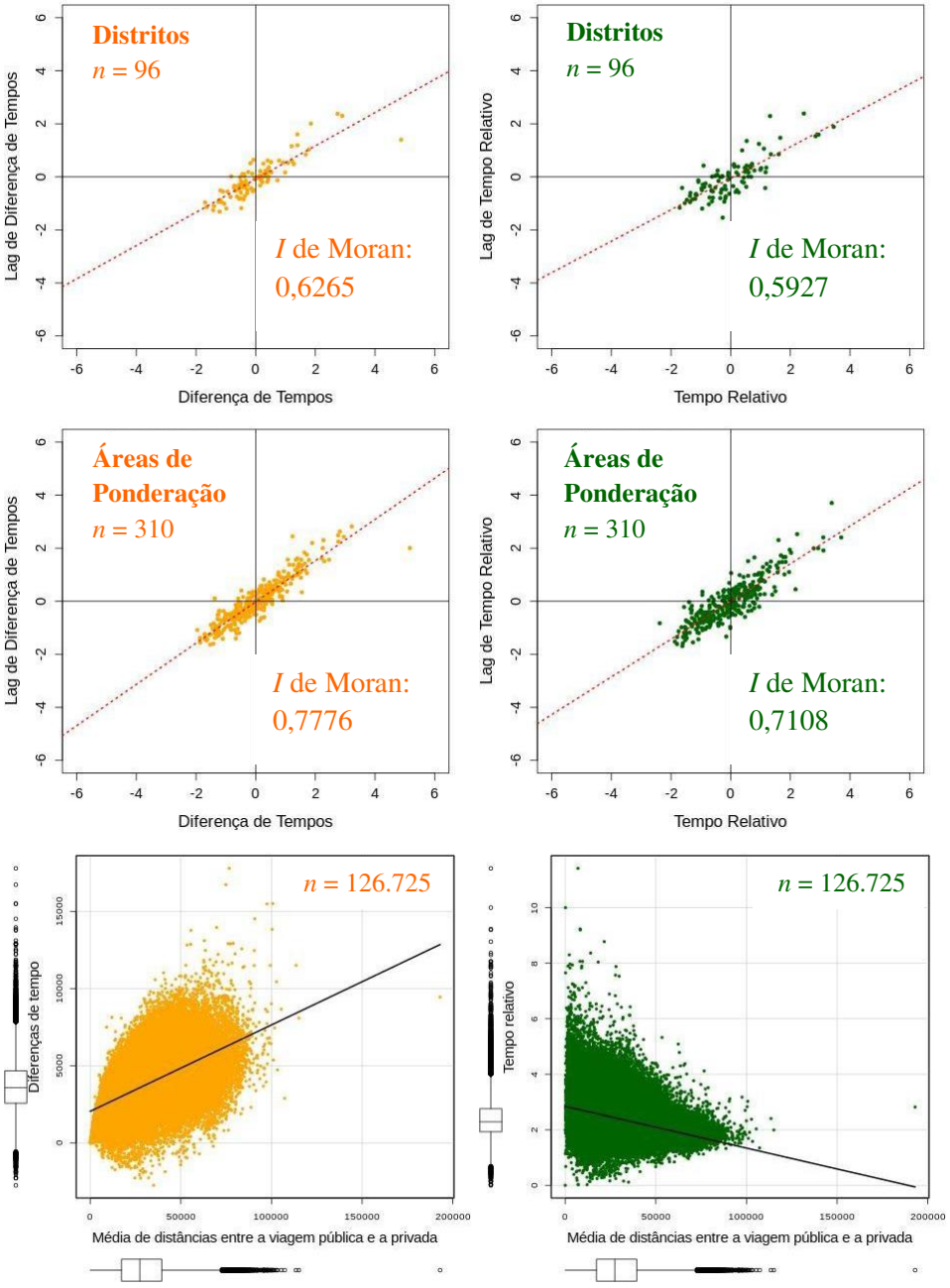

Figura 2. Diagramas de Dispersão das Distribuições de: (acima e meio) *I* de Moran da Diferença de Tempos e *I* de Moran dos Tempos Relativo, (abaixo) Distribuição de D_t pelas médias de distâncias de viagem e Distribuição de R_t pelas médias de distâncias de viagem
Fonte: Os autores, a partir da ferramenta R

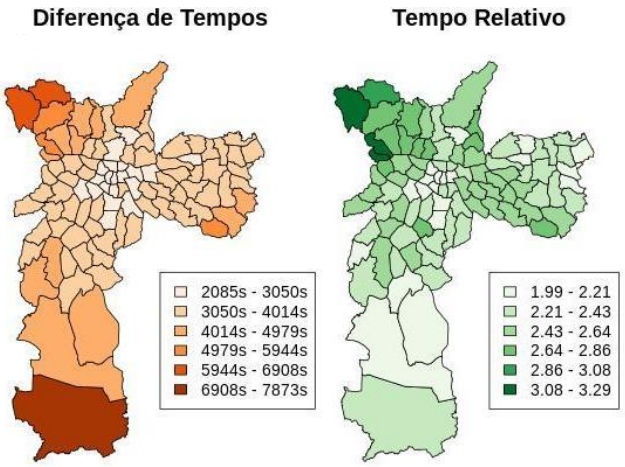

Figura 3. Mapas distribuição das medidas D_t e R_t.
Fonte: Os autores, a partir da ferramenta R
Nota: n = 96

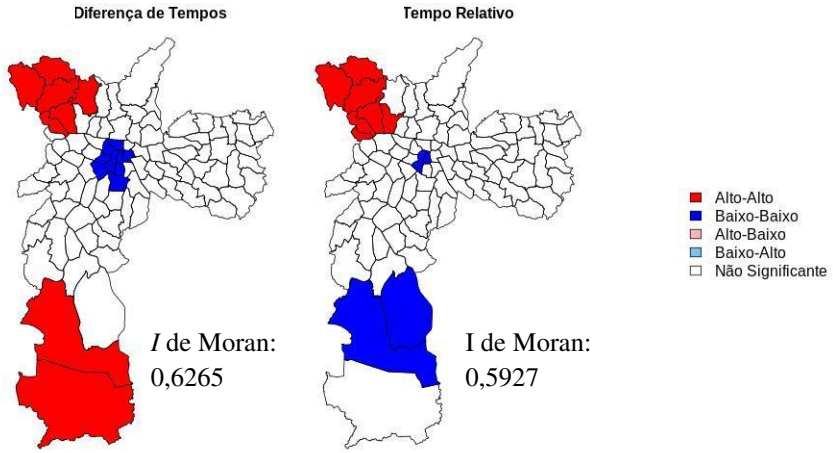

Figura 4. Mapas de autocorrelação espacial (LISA *maps*) nos distritos.
Fonte: Os autores, a partir da ferramenta R
Nota: n = 96

Os valores globais dos *I* de Moran e os gráficos de dispersão já indicam a presença de associação espacial forte entre as medidas dos distritos e das áreas de ponderação, com valores consistentemente acima de 0,5 (todas as medidas são significantes a 0,05). A classe Alto-Alto indica que as medidas no distrito são relativamente altas e as dos distritos vizinhos também - indicando um cluster de medidas altas. Analogamente, um cluster Baixo-Baixo indica um comportamento semelhante com valores baixos das medidas. Altos-Baixos e Baixos-Altos indicam um distrito cujas medidas diferem das de sua vizinhança, apontando possíveis outliers locais (e indicando menor autocorrelação espacial local). É notável que a associação é mais significativa nas áreas de ponderação (o que indica maior clusterização dos dados) que nos distritos, e mais fortes para a D_t que R_t. A diferença no *I* de Moran global entre as medidas pode ser interpretada pela diferente distribuição das observações nos quadrantes do gráfico de autocorrelação espacial. Os gráficos de tempo relativo apresentam mais pontos nos quadrantes 1 e 4, o que pode indicar mais observações que são outliers em relação a sua vizinhança. A análise dos mapas de autocorrelação espacial (LISA *maps*) permite visualizar melhor as relações de clusterização espacial das zonas analisadas. No caso dos distritos (Figura 4) as duas medidas apresentam dois núcleos de clusters em comum: uma região de altos tempos relativos

e alta diferenças de tempo na zona noroeste do município, e na região central (apesar do núcleo do cluster de R_t ser menor) há uma região comum de baixos valores para ambas as medidas. Uma diferença importante é a presença de um cluster de baixos tempos relativos na zona sul, onde o mapa de diferenças de tempo indica justamente um cluster de altos valores.

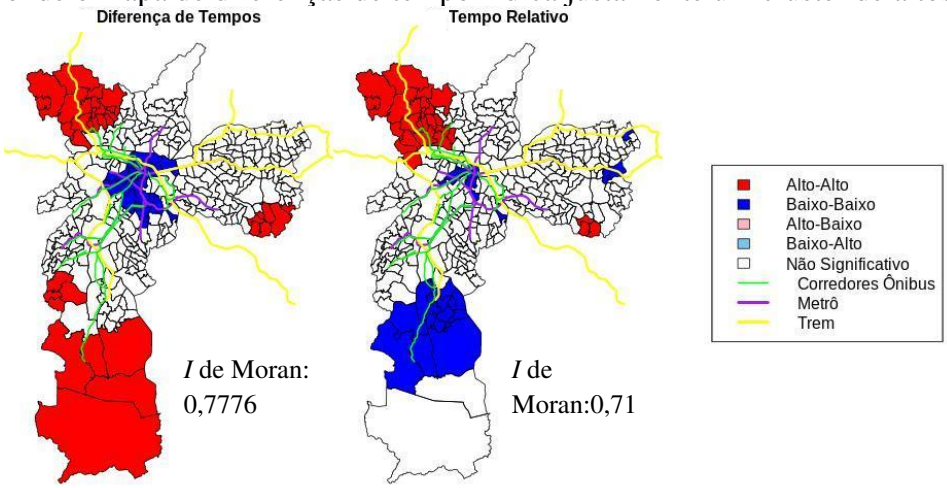

Figura 5. Mapas de autocorrelação espacial (LISA *maps*) nas áreas de ponderação e as redes de transporte
Fonte: Os autores, a partir da ferramenta R
Nota: n = 310 áreas de ponderação

Os mapas LISA para as áreas de ponderação (Figura 5) apresentam um grau maior de granularidade e, como indicam os *I* de Moran globais e os gráficos de autocorrelação espacial local, e apresentam mais núcleos de clusters que os distritos. Os mesmos núcleos são visíveis nos dois mapas; no mapa de R_t os clusters de baixo valor no centro e ao sul e o cluster de altos valores a noroeste estão presentes, enquanto no mapa de D_t os núcleos de altos valores a noroeste e ao sul e de baixos valores no centro também se repetem. Porém são novos no mapa de R_t os núcleos de baixos valores ao norte da zona leste e um núcleo de valores altos ao sul da zona leste. Enquanto o primeiro não aparece no mapa de D_t, o último encontra reflexo no mapa: há um núcleo de valores altos concentrado no sul da zona leste. Relativo a essa maior nuclearização, como os distritos apresentam uma área maior, há uma compensação no interior deles entre tendências diferentes das medidas D_t e R_t.
A partir do mapeamento desses núcleos é possível comparar o agrupamento das medidas com a presença de sistemas de transporte público de alta capacidade. A comparação visual (Figura 5) encoraja a ideia de que há uma identificação entre os agrupamentos baixos e a proximidade do Metrô e da CPTM. A partir do contraste visual, a presença do Metrô parece fortemente relacionada à redução das medidas de análise, e consequentemente dos tempos de viagem do transporte público – nenhum dos clusters de valores altos de ambas as medidas apresenta uma linha de metrô próxima. A presença da CPTM parece ter um efeito relevante, mas reduzido: na zona leste, onde ela atravessa, não há clusters de valores altos e há a clusterização de valores baixos próximos à extremidade da zona leste. Porém, na zona norte e na zona sul, a presença da CPTM parece não ser tão efetiva, já que ao norte ela cruza um cluster de altos valores para ambas as medidas, e ao sul ela atravessa uma região com altos valores para D_t. Os corredores de ônibus parecem ter pouco impacto na clusterização das medidas, além do fato de eles estarem associados às outras estruturas de transporte. Mas onde há somente os corredores de ônibus, não parece haver impacto significativo.

A primeira modelagem dos dados foi uma regressão linear OLS, a partir das variáveis da Tabela 2 para as duas medidas. Os resultados das duas regressões estão reproduzidos na Tabela 4. Os modelos finais após a retirada de variáveis colineares e variáveis não significantes apresentam R^2 relevantes, mas não maiores que 0,5. Além disso, os testes de Jarque-Bera e Breusch-Pagan indicam a não normalidade dos erros e heterocedasticidade dos resíduos. Os resíduos são espacialmente dependentes (Figura 6), sugerindo a adoção de modelos espaciais, conforme discutido anteriormente.

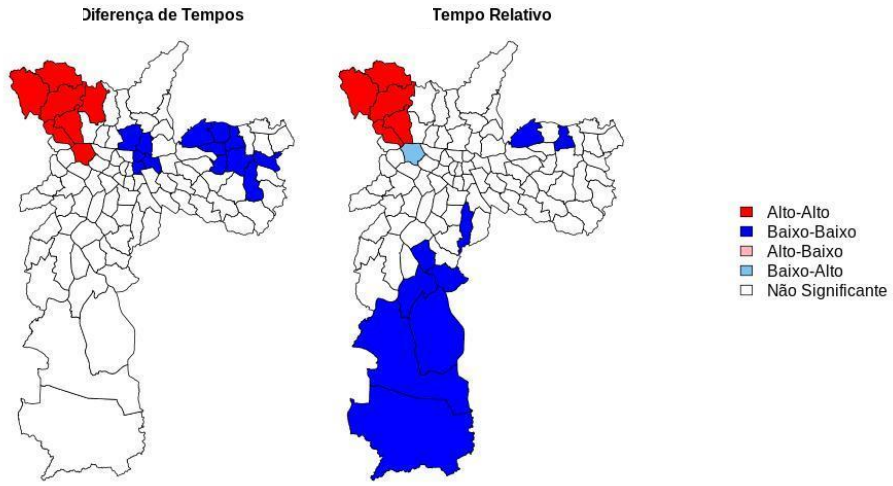

Figura 6. Mapas de autocorrelação espacial (LISA) dos resíduos da regressão OLS
Fonte: Os autores, a partir da ferramenta GeoDA e R

A forte dependência espacial dos resíduos do modelo OLS reforçou a adoção de modelos espaciais. A Tabela 3 descreve as variáveis independentes selecionadas para os modelos SAR de explicação de D_t e de R_t.

Tabela 3. Variáveis dos modelos SAR para D_t e R_t

D_t – SAR	R_t – SAR
Acesso à estação de CPTM	Acesso à estação de CPTM
Densidade de linhas de ônibus	Acesso à estação de Metrô
% de não brancos	Proporção de domicílios com moto
Densidade populacional	

Fonte: Os autores.

Os modelos SAR foram calculados para as duas medidas. Para ambas o R^2 aumentou seu valor – de 0,4895 para 0,7606 em D_t e de 0,3049 para 0,6696 em R_t – mas para D_t os resíduos parecem manter a sua heterocedasticidade. Para ambos os modelos há a indicação de que ainda existe dependência espacial que não é explicada pelas variáveis utilizadas nos modelos. Apesar disso, a distribuição espacial dos resíduos não indica padrões espaciais da variação dos modelos, conforme Figura 7.

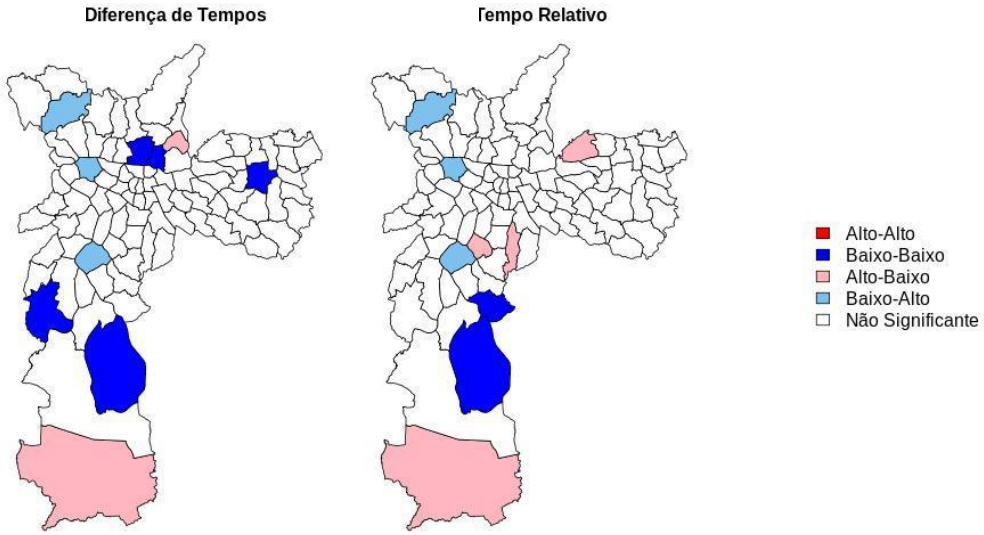

Figura 7. Mapas de autocorrelação espacial (LISA) dos resíduos da regressão SAR
Fonte: Os autores, a partir da ferramenta R e GeoDA

Os coeficientes selecionados no modelo SAR de D_t estão destacados abaixo:

Variable	Coefficient	Std.Error	z-value	Probability
W_Tempo_dif	0,7562	0,0639	11,8348	0,0000
CONSTANT	1013,64	265,464	3,8184	0,0001
Dens. Linhas ônibus	-5,8718	2,7959	-2,1001	0,0357
% de não brancos	1183,8	348,151	3,4003	0,0007
Acesso CPTM	-260,749	92,4829	-2,8195	0,0048
Densidade Pop	-0,0247	0,0094	-2,6337	0,0084

Os coeficientes das variáveis indicam que D_t aumenta em distritos com maior proporção de não brancos e que distritos com maior densidade populacional e maior densidade de linhas de ônibus tem menor diferença entre os tempos de viagem pública e privada. A proximidade de estações de trem também contribui para a redução dessa diferença. O modelo SAR para R_t apresenta outras variáveis:

Variable	Coefficient	Std.Error	z-value	Probability
W_Temp_rel	0,7451	0,0672	11,0993	0,0000
CONSTANT	0,5230	0,1730	3,0635	0,0022
Prop. Dom. Mot.	0,3416	0,1158	2,9482	0,0032
Acesso Metrô	-0,1466	0,0334	-4,3929	0,0000
Acesso CPTM	-0,0638	0,0305	-2,0934	0,0363

O modelo para R_t varia positivamente em distritos com maior proporção de domicílios que possuem motos, enquanto a proximidade de estações de metrô e de trem está relacionada a redução das razões de tempos de viagem pública pela privada. Uma possível explicação para a relação da proporção de domicílios com motos é que em distritos onde a locomoção pública é consideravelmente mais lenta que a privada, há maior proporção de motorização; alternativamente, como a motorização está correlacionada à renda e como há uma motorização

forte em zonas centrais (mais ricas), onde a malha viária é mais abundante, as viagens privadas nessas regiões tendem a ser mais eficientes que as públicas. A diferença das medidas pode também trazer informações sobre os modelos. As duas medidas apresentam particularidades em sua variação. D_t apresenta valores pequenos para viagens curtas e um crescimento linear em função da distância, enquanto R_t apresenta valores altos em viagens curtas, mas que tendem para um valor próximo da média do conjunto de dados (Figura 2). Essa característica dos dados se reflete nas médias dos distritos de origem, gerando para D_t um padrão centro e periferia claro, derivado da própria simulação: como os destinos das viagens oriundas das extremidades do município têm muito mais probabilidade de serem sorteados a uma distância maior do que se a origem fosse o centro (já que a extremidade é relativamente mais longe da maioria dos outros pontos da cidade do que o centro), as médias de diferenças de tempo dos distritos da periferia são maiores que as médias de distritos mais centrais. Essa dependência de D_t em relação à distância pode explicar em parte a associação de algumas das variáveis que foram selecionadas que variam também segundo um padrão centro-periferia (proporção de não brancos no distrito e densidade de linhas de ônibus, principalmente). Como R_t não é sensível à distância, a distribuição das médias distritais não segue o padrão centro-periferia. Ao mesmo tempo, apesar de ser muito sensível a pequenas distâncias (nas quais o valor de R_t é alto), ao agregar as viagens em torno das médias dos distritos de origem a estrutura da simulação compensa em parte esse desvio. Por construção da simulação, são os distritos mais densamente povoados os mais sujeitos a esse desvio, uma vez que a densidade de endereços na base foi mais densa nessas regiões – o que aumentaria a probabilidade de viagens próximas.

5. DISCUSSÃO FINAL E PRINCIPAIS CONCLUSÕES

O presente trabalho explorou uma abordagem de simulação de dados de viagens a partir de ferramentas de Big Data. Foi feita uma análise exploratória dos dados simulados e das relações dos dados com variáveis de infraestrutura de transporte e de variáveis socioeconômicas de controle. A intenção do trabalho era verificar possíveis vieses dessa estratégia, bem como avaliar o quão responsivo os dados simulados são aos dados empíricos que refletem a infraestrutura de mobilidade no município; essa responsividade foi pensada como uma primeira validação dos dados simulados. Uma primeira limitação do experimento foi a quantidade de dados que puderam ser simulados. À diferença das pesquisas empíricas em que o fluxo das viagens é um dos dados extraídos a partir da amostragem estatísticas das entrevistas, como essa simulação não incluiu nenhuma suposição sobre o comportamento dos viajantes, os fluxos de viagens foram arbitrários (apesar da distribuição de endereços ter seguido a densidade populacional – mesmo que por limitações da simulação – inserindo não intencionalmente um comportamento de viajantes priorizando viagens de áreas densas para áreas densas). Mesmo assim, as quantidades de viagens entre certas origens e certos destinos, quando as medidas são agrupadas nas médias distritais, podem influenciar sobremaneira as medidas agregadas. Essa mesma questão, mesmo existindo em bancos de origem e destino, é menos arbitrária, já que a dimensão dos fluxos também é uma medida estatisticamente válida em uma pesquisa OD bem realizada. Outro viés identificado na simulação foi o aumento das distâncias de viagens em distritos mais afastados. Por mais que esse seja um padrão real de viagens no município de São Paulo, o padrão identificado nos dados é dado puramente pela relação da distância entre os distritos e a distribuição espacial dos endereços de sorteio. Esse "desbalanceamento" da quantidade de viagens afeta as médias distritais das medidas, de forma que são ressaltadas as dependências espaciais não correlacionadas a presença e qualidade da infraestrutura de transportes. O balanceamento de viagens a partir de densidades de viagens registradas em pesquisas Origem e Destino pode ajudar a "calibrar" esse viés com a contribuição que existe na

constituição espacial da cidade – de fato as distâncias entre os distritos importa. Ao mesmo tempo, a adoção dos modelos espaciais contribui para isolar em parte a influência da distância nos dados usados. Os modelos calculados a partir da simulação e a comparação entre o modelo OLS e o SAR estão resumidos abaixo:

Tabela 4. Resumo dos modelos OLS e SAR para D_t e R_t

	OLS D_t	**OLS R_t**	**SAR D_t**	**SAR R_t**
R^2	0,4895	0,305	0,7606	0,6696
R^2 ajustado	0.4670	0,2744	-	-
Nº de variáveis	4	4	4	3
Variáveis	Acesso à estação de CPTM	Proporção de domicílios com carro	Acesso à estação de CPTM	Acesso à estação de CPTM
	Densidade de estabelecimentos	Acesso à estação de Metrô	Densidade de linhas de ônibus	Acesso à estação de Metrô
	% de não brancos	% de não brancos	% de não brancos	Proporção de domicílios com moto
	Densidade populacional	Renda per capita	Densidade populacional	
Fator espacial	Não há		Coeficiente espacial dependente da vizinhança – contribui para retirar a perturbação da autocorrelação espacial das outras variáveis	
Vantagens	Simplicidade do modelo e fácil interpretação dos resultados e de sua qualidade		Enquanto mantém a simplicidade, os modelos SAR conseguem captar bem a variância espacial da amostra, obtendo os maiores R^2 dentre os modelos usados	
Desvantagens	Não consegue lidar com as dependências espaciais dos dados. Mesmo selecionando variáveis semelhantes aos modelos espaciais, os R^2 são muito menores		Apesar conseguir lidar bem com a não-estacionariedade da variável dependente, os modelos não conseguem processar bem a não-estacionariedade das variáveis explicativas.	

Fonte: Os autores.

As modelagens espaciais para as duas medidas refletiram a distribuição de infraestrutura de transporte público. As variáveis de acesso a Metrô e CPTM foram captadas em quase todos os modelos, assim como quase todos os modelos capturaram alguma medida de densidade de infraestrutura de ônibus. Algumas variáveis socioeconômicas refletiram distribuições centro-periferia presentes nos dados – padrão que era mais presente em D_t que em R_t – como a porcentagem de não brancos nos distritos e a densidade populacional. Os modelos lineares simples, apesar de selecionarem variáveis semelhante (principalmente no caso de D_t), apresentaram pouca capacidade de explicar as variações nos dados. Apesar dos vieses, as medidas parecem manter relações consistentes com as variáveis relativas ao transporte público. Uma das possibilidades de melhora da qualidade dos modelos pode ser a inclusão de variáveis relacionadas a infraestrutura que afetem o desempenho do transporte privado. Os modelos utilizados neste trabalho pecam nesse sentido. Outro fato que pode melhorar a qualidade dos modelos é um estudo mais aprofundado das características da distribuição dos dados para a tomada de decisões de política pública.

A possibilidade de abordagens como essa substituírem as pesquisas empíricas como a OD podem estar distantes. As medidas de demandas de viagens, com informações estatisticamente relevantes sobre escolhas do modal, objetivos da viagem, divisões socioeconômicas, entre

outros, são difíceis de se simular. Mas não quer dizer que os resultados da OD não podem ser enriquecidos com essa abordagem que, se alimentada com informações da própria pesquisa, pode fornecer estimativas úteis de tempos de viagens. No futuro, esse tipo de abordagem tem potencial para contribuir para a redução dos custos e do aumento da tempestividade da produção de informações de mobilidade; os dados simulados a partir de ferramentas de roteamento usadas no dia a dia podem oferecer informações relevantes sobre mobilidade urbana e podem ser importantes ferramentas para analisar a mobilidade em São Paulo e outras cidades.

Como destacam Letouzé e Jutting, as estruturas institucionais de produção de dados oficiais têm um papel importante na escolha e na discussão dos dados a serem produzidos; o uso e validação dessas novas estratégias de *Big Data Analytics* são importantes para aumentar a eficiência da produção de dados e validar e fomentar o debate público acerca de suas vantagens e limites. Para manter a capacidade dessas instituições de fornecer respostas às demandas crescentes de questões sociais, uma sinergia maior entre essas técnicas e as instituições que exercem a capacidade estatística oficial é muito bem-vinda. Mas como adverte Kwan (2016), é preciso conhecer as limitações e os vieses do uso intensivo de algoritmos para que se possa estimar a qualidade dos dados gerados. A praticidade e a tempestividade dessa estratégia devem passar por esforços de validação com dados empíricos para que possa oferecer garantias de sua significância para a análise de problemas reais da mobilidade urbana, em particular, e da sociedade, em geral.

REFERÊNCIAS BIBLIOGRÁFICAS

ARANHA, V. Mobilidade pendular na metrópole paulista. São Paulo em perspectiva, v. 19, n.4, p. 96-109, 2005.

BEHRENS, K., MION, G., MURATA, Y., & SUEDEKUM, J. Spatial frictions. Journal of Urban Economics, 97, 40–70, 2017. https://doi.org/10.1016/j.jue.2016.11.003

CIA. DO METROPOLITANO DE SÃO PAULO. Pesquisa Origem-Destino 2007. São Paulo: Secretaria de Transportes Metropolitanos, 2008.

DARGENT, E., LOTTA, G., MEJÍA, J. A., MONCADA, G. A quem importa saber?: a economia política da capacidade estatística na América Latina, 2018

FERNANDES, G. A. A. L. Is the brazilian tale of peaceful racial coexistence true? some evidence from school segregation and the huge racial gap in the largest brazilian city. World Development, Elsevier, v. 98, p. 179–194, 2017.

FESSELMEYER, E., & LIU, H. (2018). Regional Science and Urban Economics How much do users value a network expansion? Evidence from the public transit system in Singapore. Regional Science and Urban Economics, 71, 46–61, 2017.https://doi.org/10.1016/j.regsciurbeco.2018.04.010

FRANCISCO, E. R. Indicadores de renda baseados em consumo de energia elétrica: Abordagens domiciliar e regional na perspectiva da estatística espacial. 2010. 381 f. Tese (Doutorado em Administração de Empresas) - Escola de Administração de Empresas de São Paulo, Fundação Getulio Vargas, São Paulo, 2010.

GANDOMI, A., HAIDER, M. Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, v. 35, n. 2, p. 137-144, 2015.

GAKENHEIMER, R. Urban mobility in the developing world. Transportation Research Part A: Policy and Practice, 33(7-8), 671-689, 1999.

GARCIA-LÓPEZ, M. &MORENO-MONROY, A. I. (2018). Regional Science and Urban Economics Income segregation in monocentric and polycentric cities: Does urban form really matter?. Regional Science and Urban Economics, 71, 62–79, 2017. https://doi.org/10.1016/j.regsciurbeco.2018.05.003

HEILMANN, K. (2021). Regional Science and Urban Economics Transit access and neighborhood segregation. Evidence from the Dallas light rail system. Regional Science and Urban Economics, 73, 237–250, 2018. https://doi.org/10.1016/j.regsciurbeco.2018.10.007
JÚNIOR, J. A. O. Direito à mobilidade urbana: a construção de um direito social. Revista dos Transportes Públicos – ANTP – Ano, 33, 1o, 2011.
KOWARICK, L. *A espoliação urbana* (Vol. 44). Editora Paz e Terra, 1980.
KWAN, M. P. Space-time and integral measures of individual accessibility: a comparative analysis using a point-based framework. Geographical analysis, v. 30, n. 3, p. 191-216, 1998.
_____ Algorithmic geographies: Big data, algorithmic uncertainty, and the production of geographic knowledge. *Annals of the American Association of Geographers*, *106*(2), 274-282, 2016.
LEE, J. G., KANG, M. Geospatial big data: challenges and opportunities. Big Data Research, 2(2), 74-81, 2015.
LETOUZÉ, E., JÜTTING, J. Official statistics, big data and human development: towards a new conceptual and operational approach. Data Pop Alliance and PARIS21, 2014.
MARICATO, E. Metrópole, legislação e desigualdade. Estudos avançados, 17(48), 151-166, 2003.
MCAFEE, A., BRYNJOLFSSON, E., DAVENPORT, T. H., PATIL, D. J., BARTON, D. Big data: the management revolution. Harvard business review, 90(10), 60-68, 2012.
NOULAS, A., SCELLATO, S., LAMBIOTTE, R., PONTIL, M., MASCOLO, C. A tale of many cities: universal patterns in human urban mobility. PloS one, 7(5), e37027, 2012.
PÄÄKKÖNEN, P., PAKKALA, D. Reference architecture and classification of technologies, products and services for big data systems. Big Data Research, 2(4), 166-186, 2015.
REQUENA, C. *A mobilidade paulistana: viária e desigual* In: MARQUES, E. (Org) *A metrópole de São Paulo no século XXI,* 1 ed. , São Paulo: Editora Unesp, cap. 13, pp 395-421, 2015.
ROLNIK, R., & KLINK, J. Crescimento econômico e desenvolvimento urbano: por que nossas cidades continuam tão precárias? Novos estudos-CEBRAP, (89), 89-109, 2011.
SCARINGELLA, R. S. A crise da mobilidade urbana em São Paulo. São Paulo em perspectiva, 15(1), 55-59, 2001.
SILVEIRA, M. R., COCCO, R. G. Transporte público, mobilidade e planejamento urbano: contradições essenciais. Estudos avançados, São Paulo, v. 27, n. 79, p. 41-53, 2013. Disponível em <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-40142013000300004&lng=en&nrm=iso>. Acesso em 1 Junho de 2018.
TABUCHI, T. (2019). Do the rich and poor colocate in large cities?. Journal of Urban Economics, 113, 103186, 2018. https://doi.org/10.1016/j.jue.2019.103186
TORRES, H. D. G., MARQUES, E., FERREIRA, M. P., BITAR, S. Pobreza e espaço: padrões de segregação em São Paulo. Estudos avançados, 17(47), 97-128, 2003.
TORRES, H. D. G., & OLIVEIRA, G. C. D. Primary education and residential segregation in the Municipality of São Paulo: a study using geographic information systems. In International Seminar on Segregation in the City, pp. 26-28, Julho de 2001.
TRIBBY, C. P., ZANDBERGEN, P. A. High-resolution spatio-temporal modeling of public transit accessibility. Applied Geography, 34, 345-355, 2012.
WANG, M., & MU, L. Spatial disparities of Uber accessibility: An exploratory analysis in Atlanta, USA. Computers, Environment and Urban Systems, 67, 169-175, 2018.
WILHEIM, J. Mobilidade urbana: um desafio paulistano. Estudos avançados, 27(79), 7-26, 2013.