**Título**

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

# 

# **1. INTRODUÇÃO**

A experiência da mobilidade urbana é parte integrante da vivência em cidade. O acesso às oportunidades de renda, consumo e lazer que uma cidade oferece são mediadas pela disponibilidade que os indivíduos têm de se locomover pela cidade; e essa locomoção implica em custos (tempo, dinheiro, informação, ...) que podem ser traduzidos em gastos para os indivíduos (Páez, Scott e Morency, 2012). Assim, frente a uma distribuição desigual de oportunidades no espaço urbano – como é o caso de cidades que viram um crescimento desordenado (Gakenheimer, 1999) – um sistema de mobilidade inclusivo tem um potencial equalizador. Mas o inverso também é verdade: em um contexto de oportunidades mal distribuídas, um sistema de mobilidade urbano espacialmente desigual pode reforçar ainda mais desigualdades pré-existentes, como é o caso de São Paulo (Torres et al., 2003, Torres e Oliveira, 2001).

Assim a geração de dados sobre a mobilidade urbana é essencial não só para o planejamento e gestão eficientes dos sistemas de transporte urbanos, mas também para a compreensão da distribuição social das oportunidades no tecido urbano. Referência nesse sentido são as pesquisas domiciliares de Origem e Destino, como a realizada pelo Metrô na Região Metropolitana de São Paulo (Metrô, 2017). Através dessas pesquisas é possível inferir os fluxos realizados de locomoção no contexto urbano, capturando as razões e meios dos deslocamentos e identificando os pólos de atração e geração dos fluxos. O levantamento de informações socioeconômicas dos indivíduos, famílias e domicílios também ajuda a identificar os padrões de mobilidade de diferentes camadas sociais da cidade. Porém, o custo elevado dessas pesquisas impede a sua realização freqüente, o que causa a sua obsolescência muito antes que seja possível realizar uma nova pesquisa. Nesse sentido, existem tentativas de se utilizar meios alternativos para a identificação de fluxos urbanos utilizando dados de bilhetagem (Arbex e Cunha, 2017; Guerra, Babosa e Oliveira, 2014), de localização de aparelhos celulares usando o sinal de telefonia (Kwan, 2016), ou a localização a partir de aplicativos de redes sociais (Noulas et al. 2012). Dentre essas estratégias, o uso de dados gerados pela interação de usuários em rede, em grandes volumes – mais especificamente em um contexto de Big Data (Lee e Kang, 2015) – apresenta uma oportunidade de gerar dados relevantes para a análise da mobilidade urbana a um baixo custo e com disponibilidade no momento da necessidade.

O trabalho apresenta uma estratégia de simulação de dados a partir de ferramentas de Big Data de tempos de viagens para transporte público e privado. O intuito é comparar os dados de tempos de deslocamento dessa estratégia com os dados registrados na pesquisa OD 2017 realizada pelo Metrô (Metro, 2017) no contexto do município de São Paulo, de forma a testar se o uso dessa ferramenta introduz vieses consideráveis na análise, relativos aos dados de mobilidade apresentados pela própria pesquisa OD. Para isso foram simuladas 257.400 viagens a partir da *Application Programming Interface* (API)de roteamento de transporte da *Google*, sendo esse um conjunto de viagens pareadas de transporte público e de transporte individual motorizado. Para comparar esses dados ao conjunto de viagens da Pesquisa OD, foi criado uma estrutura de comparação em que o tempo de cada viagem da OD foi comparada à média de tempos de viagens simuladas com trajetos semelhantes, considerando o modal usado nas viagens e considerando o horário de partida das viagens. Foram testadas possíveis dependências da diferença entre essas medidas, incluindo diferenças espaciais. O resultado indicou vieses consistentes de leve superestimação dos valores de viagem, mas com outras dependências envolvidas, como de distância e espaciais.

**1.1. Problema da Pesquisa**

O uso de uma estratégia de simulação de viagens rápida e barata através de aplicativos de roteamento baseado em *Big Data* pode ser muito útil para a analise da mobilidade urbana, mas é preciso verificar se esse tipo de análise pode gerar algum tipo de viés. O presente trabalho busca usar dados da Pesquisa OD como referência de tempos de viagens reais para comparar os dados simulados a fim de identificar se há variação nos dados e se é possível identificar variáveis que explique essa variação.

**1.2. Justificativas**

Uma estratégia ágil e barata de simulação de viagens pode ser usada para elaborar melhores indicadores de acessibilidade, de forma tempestiva e sob demanda. Uma boa identificação das limitações dessa estratégia pode permitir o uso de ferramentas que gerem informação em tempo rela de mobilidade, que pode ser usada para análises do comportamento urbano e da distribuição de mobilidade pelo espaço.

**2. REVISÃO DE LITERATURA**

**2.1. Mobilidade Urbana e Acessibilidade**

Dois conceitos relevantes para a análise do transporte nas cidades são a mobilidade e a acessibilidade. A Política Nacional de Mobilidade Urbana de 2012 define ambos os conceitos: a mobilidade urbana é definida como *“condição em que se realizam os deslocamentos de pessoas e cargas no espaço urbano”* e acessibilidade, como*“facilidade disponibilizada às pessoas que possibilite a todos autonomia nos deslocamentos desejados [...]”* (Brasil, 2012)*.* Outras definições de acessibilidade detalham melhor as nuances do conceito.Vasconcellos (2012b apud Caccia, 2015) distingue Microacessibilidade (que corresponde à definição da PNMU) e Macroaccessibilidade, que está relacionada à distribuição dos sistemas de transporte em relação às oportunidades disponíveis de acesso à cidade, definição que encontra eco em outros trabalhos (Kwan, 1998; Páez, Scott e Morency, 2012). A Macroacessibilidade interessa enquanto considera os dois lados do acesso à cidade: a mobilidade como forma de acessar as oportunidades e a distribuição dos usos urbanos de forma a reduzir a necessidade de locomoção.

Litman (2003) faz uma distinção entre as abordagens que se baseiam nos conceitos de tráfego, mobilidade e acessibilidade, e pondera as diferenças que esses enfoques implicam nos conceitos usados para gerir o transporte urbano. Abordagens focadas no tráfego focam em analisar e gerir as movimentações urbanas pensando no fluxo veicular; assim os usos do espaço urbano e as medidas elaboradas para conhecer o problema serão focados em aumentar a fluidez e velocidade dos veículos, que acaba por priorizar os carros e seus motoristas. Pensar em termos de mobilidade considera o fluxo de pessoas e cargas como prioridade; ainda que exista o foco em garantir o fluxo veicular, onde é possível que o transporte de massa, transportes compartilhados ou transportes ativos atendam demanda considerável de deslocamentos há a preocupação em incentivá-los através de disponibilidade de espaço urbano, infraestrutura e regulação. De acordo com o autor, a acessibilidade considera a habilidade dos indivíduos de acessar oportunidades no tecido urbano, e pensar o transporte urbano sob essa perspectiva implica considerar os diferentes modos de acessibilidade como importantes na medida em que eles colaboram para que um maior numero de pessoas acesse suas necessidades; o uso do solo implica não só dedicar espaço à mobilidade, mas a fomentar os diferentes usos de forma a reduzir a necessidade de mobilidade dos indivíduos. Assim, partindo dessa definição específica de acessibilidade, pensar a mobilidade urbana também passa por pensar a distribuição de oportunidades pelo território. Hagerstrand (1970) acrescenta e Kwan (1998) discute uma preocupação ao pensar a acessibilidade: não basta pensar na distribuição das oportunidades e de meios de locomoção. Utilizando o conceito de ‘espaço-tempo” de locomoção ao pensar indicadores de acessibilidade, esses trabalhos ressaltam que considerar padrões de locomoção específicos ao calcular a acessibilidade pode ser enganoso (como pensar em trajetos casa-trabalho ao verificar o acesso à cidade), levando a análise a desconsiderar padrões alternativos de locomoção. Considerando o conceito central de análise como sendo a acessibilidade, a mobilidade urbana é o meio pela qual a essa se realiza; assim, diferenças na mobilidade urbana (sejam relativas a renda, à ocupação profissional ou à responsabilidade de cuidado de filhos) podem implicar em diferentes acessibilidades para indivíduos vivendo em um mesmo tecido urbano. Assim, o estudo da mobilidade urbana é uma importante ferramenta para conhecer e gerir os sistemas urbanos. Pensando só em termos de eficiência: para se ter uma idéia, as perdas de eficiência da lentidão e congestionamentos na cidade de São Paulo já foram avaliados por um estudo da Fundação Dom Cabral de 2012 em 40 bilhões de reais anuais (Silveira e Cocco, 2013). O caso de São Paulo ilustra que o processo de crescimento e urbanização acelerado da cidade pode produziu um tecido urbano com enormes desigualdades na oferta de infraestrutura física, de serviços, de oportunidades e mesmo de mobilidade (Torres et al., 2003, Torres e Oliveira, 2001; Morandi, 2020), fatores que colaboram para a necessidade de mobilidade para que as pessoas cheguem às oportunidades. Os efeitos sociais das distancias necessárias a se percorrer para a empregabilidade são estudados para minorias sociais nos EUA, sob o conceito de *"Spatial Mismatch"* (Preston e Lafferty, 1999). Além disso, o foco dos investimentos em mobilidade tem um histórico de favorecimento dos modos privados individuais de locomoção, que geram um cenário em que há uma ocupação desequilibrado do espaço urbano dedicado para o deslocamento de carros (Caccia, 2015). Claramente, isso não é uma particularidade de São Paulo; o efeito de processos de intensa motorização que desafiam as capacidades fiscais e regulatórias municipais são um tema comum: Gakenheimer (1999) apresenta o problema da motorização acelerada como uma questão das grandes cidades de países em desenvolvimento, sintetizando fatores que colaboram para o agravamento dos congestionamentos em cidades de diversos países.

Uma forma bastante consolidada de estudo de mobilidade urbana é a elaboração de matrizes origem-destino de viagens com informações de demanda e modais usados. Os levantamentos mais completos são realizados por pesquisas domiciliares amostrais de origem-destino, capazes de levantar dados de motivação das viagens, itinerário completo, além de informações socioeconômicos acerca dos indivíduos viajantes, suas famílias e seus domicílios. Porém seu custo elevado motiva estratégias de estimação dessas matrizes de origem-destino a partir de outras informações disponíveis. Arbex e Cunha (2017) utilizar as informações de bilhetagem eletrônica do município de São Paulo para estimar os fluxos de passageiros entre as estações do sistema metroferroviário da cidade. A partir dessa matriz de fluxos os autores calculam os padrões de carregamento das várias rotas do sistema metroferroviário durante um dia comum de viagens assumindo um comportamento de escolha de trajetos dos usuários que considera tempo perdido em nas várias fases do uso do sistema. Além de usar os dados de bilhetagem nas estações de trem e metrô, são usadas as informações de embarque no sistema de ônibus para complementar as informações de destino das viagens de trem; ainda sim, existe uma quantidade expressiva de viagens que não são capturadas precisamente pelas estimações de origem e destino, ou por não seguirem as previsões necessários para a estimação (como a escolha do melhor caminho dentro do sistema metroferroviário), como por não haverem dados o suficiente para estimar a viagem (como um bilhete com uma única entrada no sistema de transporte no dia). Guerra, Babosa e Oliveira (2014) fazem um proposta metodológica nesse sentido para o município de Maceió/AL. Os dados de bilhetagem de ônibus são usados para estimar as origem e destinos de passageiros utilizando algumas suposições sobre o comportamentos dos usuários, uma vez que, assim como no caso de São Paulo, os cartões são validados somente na entrada no sistema: para um mesmo cartão são considerados os destinos da viagens os pontos de origem da viagem registrada logo após a primeira viagem. Para reduzir a incerteza das suposições sobre os resultados, as matrizes são calculadas para zonas de tráfego. O método é capaz de gerar informações sobre a origem e destino, mas existem incertezas que são difíceis de serem avaliadas, como o próprio trabalho ressalta. Outras formas de analisar mobilidade, a partir de outras fontes de dados é possível. Noulas et al. (2012) utilizam dados de localização da rede social *Foursquare* para comparar os padrões de deslocamento urbano em cidades como Singapura, Houston, São Francisco e outras. O trabalho analisa as distancias de deslocamento dos usuários da rede social entre lugares de interesse na cidade comparando modelos de atração gravitacional baseados na distância euclidiana e modelos de "ranqueamento" dos lugares de interesse a partir da distância. Essa estratégias buscam observar ou estimar os fluxos e deslocamentos que existem no espaço urbano; uma outra preocupação relacionada, mais afinada com o presente trabalho, é olha no sentido da (macro)acessibilidade; ou seja, o quanto as condições de mobilidade permitem o acesso à cidade.

**2.2. Indicadores de acessibilidade**

Páez, Scott e Morency (2012) e Kwan (1998) definem acessibilidade como o potencial de acessar oportunidades espacialmente distribuídas. Páez, Scott e Morency (2012) discute diferentes suposições para a implementação de indicadores de acessibilidade. Um indicador de acessibilidade pode analisar as duas interfaces da acessibilidade: a distribuição das oportunidades no espaço, medindo a acessibilidade a partir dos **destinos** pretendidos ou de interesse para os indivíduos em locomoção; e o custo de locomoção dos indivíduos no espaço, caso em que ponto de análise é na **origem** das viagens. Ao analisar cada uma das interfaces do problema, o indicador pode apresentar diferentes suposições para o que está sendo medido em termos de seu caráter **normativo**, considerando como deve ser distribuídas as oportunidades no espaço ou qual deve ser o custo/comportamento das viagem, ou  **positivo**, considerando como estão distribuídas de fato as oportunidades no espaço e quais os custos/comportamentos de viagens observados na cidade. Uma abordagem normativa em uma interface não implica necessariamente que a outra interface também será medida em termos normativos.

Kwan (1998) faz uma análise comparativa entre indicadores integrais, que se dividem em indicadores Gravitacionais e Cumulativos, e indicadores de espaço-tempo. Os indicadores gravitacionais, os cumulativos e os de espaço tempo são instâncias particulares da seguinte fórmula, que relaciona a distribuição de oportunidades com o custo de viagem (Paéz, Scott e Morency, 2012; Kwan, 1998):

O indicador A é dado para a origem i e as oportunidades k relacionadas ao indivíduo p. A medida é dada em função do número de oportunidades W no local j – que é o destino - dado dentro de uma função de atratividade g. As oportunidades são multiplicadas por uma função de impedância f, que é um kernel em volta da origem i dado em função do custo de viagem c do local i para o j, em relação ao indivíduo p. Para os indicadores gravitacionais, a função g é uma função de atratividade do local j, que é dada em função das oportunidades k presentes. A função de impedância costuma ser uma função que varia de algum valor positivo na origem a 0 no infinito – por exemplo, uma exponencial negativa, ou uma potência invertida, ou uma gaussiana modificada (Kwan, 1998). Já para os indicadores cumulativos, a função f é uma inequação simples, em que seu valor é 1, se c está dentro de certo limite pré-definido, ou 0 se c está fora – o valor do indicador se refere ao número de oportunidades que estão dentro do raio de custo definido. Para os indicadores de espaço tempo, o custo c pode ser usado como uma região dentro de uma rede correspondente à área de caminho potencial (PPA) (Hägerstraand, 1970; Kwan, 1998), que reflete a área que o indivíduo é capaz de acessar dados os seus constrangimentos diários. Enquanto as duas primeiras medidas são baseadas em lugares, essa última é feita em relação aos indivíduos. Um problema dos indicadores relativos à lugares, como os cumulativos e de gravidade, é que eles ignoram as especificidades da mobilidade de indivíduos nas áreas analisadas. Por exemplo, casos específicos em que as mulheres consistentemente mostram padrões diferentes de acessibilidade, mesmo morando nas mesmas regiões, ou mesmo nas mesmas casas, que homens (Kwan, 1998; Paéz, Scott e Morency, 2012). Ao mesmo tempo, o uso de uma única referência de origem outra de destino impede que os indicadores dêem conta de comportamentos de mobilidade diferentes do padrão casa-trabalho. E como a implementação costuma ser feita a partir de dados agregados em métodos zonais, existem problemas de escolha de limites – o problema da unidade de área modificável (MAUP) - e possíveis falácias ecológicas (Kwan, 1998). Os indicadores de espaço-tempo, apesar de contornar alguns desses problemas, já que são baseados nos indivíduos e consideram os diferentes tipos de comportamento, apresentam uma implementação computacionalmente muito mais complexa e custosa, além de entregarem resultados que são menos capazes de caracterizar os lugares (Kwan, 1998). Existem várias formas de empregar os indicadores de acessibilidade.

Lessa, Lobo e Cardoso (2019) utilizam dados de presença de infraestrutura de transporte público no município de BH para elaborar um índice de acessibilidade para os Campos (bairros) da cidade, e comparar com a mobilidade efetivada medida pela pesquisa Origem-Destino realizada pela BHTRANS, agregando o numero de viagens pelo *Campo* de origem. A comparação indicou que a mobilidade está concentrada em bairros com menores índices de acessibilidade, revelando um desencontro da oferta de infraestrutura e da demanda de mobilidade. Ainda sim, a análise da agregação pela origem tenha escondido informações sobre os destino das viagens, que possivelmente explicariam a distribuição da acessibilidade.

Wang e Mu (2018), Hughes e MacKenzie (2016) e Insardi e Lorenzo (2019) utilizam dados de espera de aplicativos de corrida como medida de acessibilidade, e analisam a relação dessa medida com variáveis socioeconômicas espacializadas no contexto de Atlanta, Seattle e São Paulo respectivamente. Tribby e Zandbergen (2012) elaboram um modelo de predição de tempos de viagem a partir da simulação em redes virtuais do transporte público e da rede de ruas de Albuquerque. A partir desses tempos de viagem, os autores analisam o ganho de tempo de forma espacializada, avaliando a redução das viagens como um aumento da acessibilidade dos indivíduos favorecidos. Morandi (2020) utiliza os tempos de viagem entre origem e destinos distribuídos em uma porção da Região Metropolitana de São Paulo como componente de sua função distancia, relacionada à acessibilidade, para calcular um índice de segregação racial espacializado.

**2.3.*Big Data* e Estatísticas Oficiais**

As primeiras definições de Big Data estão relacionadas às características dos dados gerados dentro da tendência de crescimento constante da atividade digital. O aumento da produção, capacidade de armazenamento e processamento de dados gerou a potencialidade de aplicações analíticas que, se não apresentam necessariamente métodos inovadores em termos estatísticos, contam com importantes inovações computacionais. São usados três grandes conceitos definidores em relação aos dados envolvidos: Volume, Velocidade e Variedade (McAfee et al., 2012; Gandomi e Haider, 2015). De acordo com essa definição, o que caracteriza Big Data não é só o volume dos dados envolvidos, mas também a velocidade de produção de dados, com aplicações para a análise de dados produzidos em tempo real, e a variedade de formatos, com o uso de dados estruturados e não estruturados – como as interações em uma rede social. Gandomi e Haider (2015) comentam ainda de outras dimensões relevantes para esses dados: Veracidade (em relação a dados como o estado socioemocional de usuários de redes sociais, que mesmo tendo valor apresentam um grau de incerteza quanto ao seu conteúdo); Variabilidade e Complexidade (variabilidade em relação aos ritmos do fluxo de dados e complexidade em relação ao uso de diversas fontes para os dados, o que exige trabalho para agregá-los); e Valor (Em relação ao baixo valor de um dado singular em comparação com o valor que o grande agregado possui). Outro aspecto abordado por Gandomi e Haider (2015) são as diferentes técnicas usadas para analisar os diferentes tipos de dados - que incluem texto, áudio, vídeo, interações em redes sociais e modelos preditivos a partir desses dados - que ele denomina *Big Data Analytics* De acordo com os autores, essa intersecção entre *Analytics* e *Big Data* se desenvolveu rapidamente em alguns setores econômicos a partir do momento em que os enormes fluxos de dados passaram a conseguir ser aproveitados para gerar valor econômico.

Outras definições de *Big Data* partem de outros pressupostos. Pääkkönen e Pakkala (2015) analisam as arquiteturas de dados de casos de uso de *Big Data* (*Facebook*, *LinkedIn*, *Twitter*, *Netflix*, *Blockmon*, Network Measurement, FIU-Miner) e classificam as diferentes tecnologias usadas. Por um método indutivo através dos casos, os autores propõe um modelo analítico base para arquiteturas em *Big Data*, ajudando a caracterizar e descrever o modelo abstrato de estruturas de processamento envolvidas no *Big Data Analytics*. Já Letouzé e Jütting (2014) definem o movimento a partir de características sociológicas. Os três conceitos definidores de Big Data seriam a natureza dos dados (não o volume), que são gerados como rastros de atividade humana dentro da rede (como o comportamento em redes sociais) – “Crumbs” ou migalhas; as técnicas e a intenção envolvida na geração de “insights” a partir desses dados, que envolvem capacidades avançadas de armazenamento e computação e métodos e ferramentas quantitativos e computacionais avançados - “Capacities”; esses dados e essa técnicas são utilizados por comunidades específicas relacionadas ao desenvolvimento dessas aplicações, tanto dentro da comunidade de softwares abertos como dentro dos setor privado e de inteligência - “Communities” - os três C’s.

Outra discussão proposta por Letouzé e Jutting (2014) é sobre como os ecossistemas de estatísticas e indicadores oficiais (que incluem, além das instituições públicas direcionadas para a produção de estatísticas, todo o ambiente acadêmico e institucional próximos delas) podem se posicionar quanto ao movimento do *Big Data Analytics*. Há uma tensão entre as estatísticas oficiais e o Big Data na medida em que esse movimento provoca 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 autonomamente de diversas fontes. Existem propostas de a partir dessa capacidade tentar reproduzir os indicadores oficiais já existentes, ou outros, mais granulares e inteligentes. Os autores argumentam, porém, que a responsabilidade das agências oficiais, ao produzir os dados oficiais, não é só de 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. E de forma mais geral, a capacidade de produção de estatísticas oficiais confiáveis e periódicas é um fator essencial para a capacidade de um país tomar decisões racionais em relação ao futuro, baseada em evidências capazes de indicar algo da realidade (Dargent et al., 2018). Em termos do Estado, essa capacidade atende necessidades tanto para o desenvolvimento de novas políticas públicas como para o monitoramento e avaliação das existentes – em relação ao Brasil, a contabilidade populacional e a previsão de sua evolução são dados importantes para o repasse de recursos federais para os municípios. Para a sociedade civil e para o mercado, a produção de dados confiáveis permite que se realizem pesquisas relevantes aos diversos atores sociais e planejamento futuro em relação a evolução dos indicadores derivados desses dados. Ainda, para vários países e órgãos multilaterais, a participação em programas de ajuda financeira, ou mesmo parcerias dentro do setor privado, exigem a presença de indicadores sociais e econômicos confiáveis. De fato, tanto a necessidade interna do Estado como e demanda de atores externos ao Estado, ou externos ao país, são identificados como fatores de economia política que explicam o desenvolvimento dessa capacidade dentro do Estado (Dargent et al., 2018). E ainda antes da crítica derivada do movimento de *Big Data*, Letouzé e Jütting (2014) discutem uma “desilusão estatística”: há um descontentamento com a capacidade das burocracias estatais em produzir estatísticas confiáveis e relevantes – desde modelos tradicionais que não conseguem acompanhar períodos voláteis até medidas que são consideradas insuficientes para o que se propõe, como o PIB para medir bem estar. Ainda, em países pobres e em desenvolvimento essa desilusão está associada a baixa capacidade estatística existente. A dificuldade desses países em construir essa capacidade passa pela falta de recursos financeiros, a baixa capacitação técnica do serviço público (causa e conseqüência de uma fuga de cérebros para o setor privado), intervenções políticas na produção de dados, entre outros fatores - dificuldades que se somam aos novos desafios (Letouzé e Jutting, 2014).

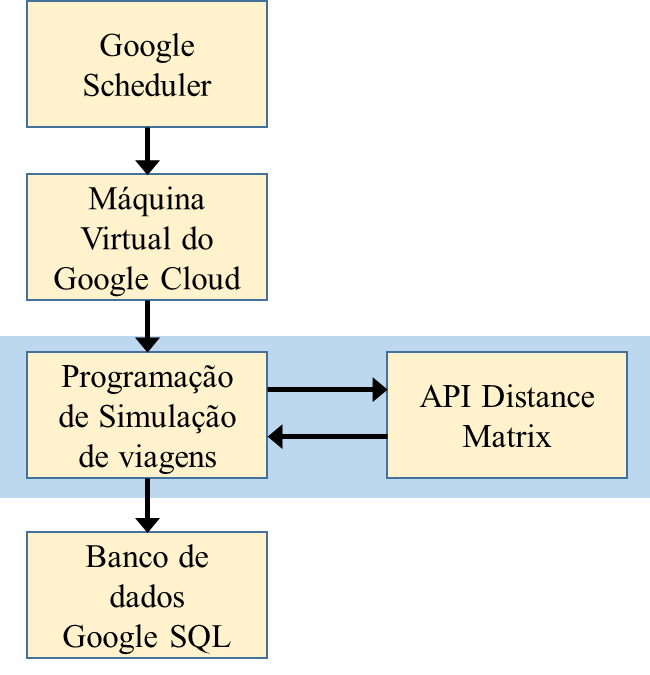
Como o movimento de *Big Data* se apresenta como um vetor de mudança na sociedade moderna, é interessante que haja movimentos de integração entre as estatísticas oficiais e essas novas técnicas de análise. De particular interesse para o presente trabalho, a produção de dados georreferenciados relativos 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 enorme volume de dados georreferenciados e, dependendo do uso de aplicativos, informações sobre os meios de transporte. Essa produção massiva de dados permite inclusive o uso desses dados para análises em tempo real, como os serviços de mapas para calcular rotas de transporte. Também pelo lado das estatísticas oficiais a produção de dados georreferenciados para entender os problemas urbanos, inclusive de mobilidade, é corrente e importante para embasar a adoção de políticas públicas específicas para cada localidade. A compreensão da dimensão geográfica dos problemas e da distribuição da infraestrutura presente e dos serviços ajudam a diagnosticar ineficiências e priorizar esforços, além de fornecer uma visão sistêmica dos indicadores sociais. Essa visão pode ajudar a escolher combinações de formas diferentes de intervenção pública (Torres et al., 2003, Torres e Oliveira, 2001). Como a produção desses dados através de pesquisas empíricas de validade estatística, como a Pesquisa OD (METRO, 2017), tende a ser bem custosa, o acesso a dados derivados dos novos aplicativos sociais que usam a localização podem permitir o acesso a informações de mobilidade de maneira muito mais barata, ainda que contendo algum grau de viés (Kwan, 2016). E enquanto esses dados podem fornecer informações valiosas sobre os padrões de mobilidade e acessibilidade das cidades (Noulas et al., 2012; Wang e Mu, 2018) a disponibilidade de dados e técnicas utilizando Big Data deve ser vista com cautela. Kwan (2016) alerta para vieses decorrentes do uso intensivo de algoritmos em qualquer tipo de análise. Mesmo que esses vieses não sejam particularidades de estratégias de *Big Data*, o uso intensivo de algoritmos no processamento e análise de dados 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.

# **3. MÉTODO**

A comparação dos dados simulados com os dados da Pesquisa OD seguiu três etapas: a simulação e organização dos dados derivados da API do *Google*; a organização dos microdados da Pesquisa OD e o cálculo de medida comparativa com os dados simulados; e a análise propriamente da comparação entre as medidas de tempo simuladas e as medidas presentes na Pesquisa OD.

**3.1. Simulação de viagens**

A simulação foi feita em duas etapas. Primeiro foi gerado um banco de endereços parcialmente aleatorizado no município de São Paulo utilizando o software QGis; a partir dessa base de endereços foi feita a simulação das viagens propriamente ditas utilizando a API de roteamento do *Google*. A escolha dessa ferramenta levou em consideração a possibilidade de simular viagens considerando o efeito em tempo real do tráfego nas viagens, fator considerado um diferencial da estratégia adotada. Porém isso implicou em um número limitado de requisições de viagens, por questões de custo; essa limitação levou a uma série de escolhas para reduzir o número de viagens “perdidas” na simulação (quando as coordenadas usadas na API não correspondiam ou não podiam ser aproximadas a endereços válidos, como no caso de coordenadas dentro das represas de São Paulo). Para tentar obter diferentes padrões de mobilidade ao longo do tempo, foi decidido distribuir as viagens simuladas ao longo dos dias úteis (segunda-feira à sexta-feira, das 5h às 21h). Isso levou à opção por usar computação em nuvem para a simulação, para garantir a freqüência das chamadas à *API*. Foram adotadas, por isso, medidas para de reduzir o esforço computacional a fim de evitar problemas relacionados ao desempenho. Essas limitações definiram o processo de escolha dos endereços das viagem. Procurou-se sortear endereços em regiões mais densamente povoadas para evitar possíveis perdas, seguindo um modelo de densidade de probabilidade da população a partir dos setores censitários do censo. 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 dessa base de endereços 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 e destino. 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. 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: Elaborado pelo autor.

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). Durante o processo de chamada da API houve a indisponibilidade do servidor em nuvem que estava sedo utilizado, o que implicou na perda de um período de viagens. Foi aumentado o numero de chamadas nas semanas subseqüentes para compensar essa perda. O registro das chamadas está apresentado na Figura 2.

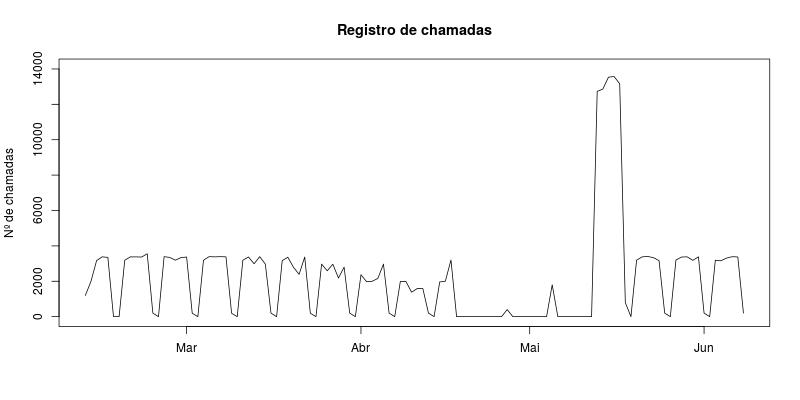


Figura 2. Registro da distribuição das chamadas da função no tempo. Elaborado pelo autor

**3.2. Organização dos microdados da Pesquisa Origem-Destino e Comparação**

Os microdados da Pesquisa OD apresentam entradas correspondentes a cada viagem realizada por indivíduo entrevistado (caso o indivíduo não tenha realizado nenhuma viagem, a entrada só corresponde aos dados socioeconômicos do entrevistado). Cada linha do banco registra anonimamente sobre o indivíduo sua identidade (dentro da base), a família na qual ele foi registrado na pesquisa, o domicilio em que ele reside, características físicas do domicílio, características socioeconômicas e ocupacionais do indivíduo e da família; sobre a viagem são registrados o propósito da viagem, o modal principal de deslocamento, o tempo da viagem e dos trechos à pé necessários para acessar os modais usados; as baldeações realizadas, as coordenadas, zonas e municípios da origem e do destino da viagem, a hora de saída e chegada, entre outras informações.

Para o presente trabalho foram filtradas informações relativas somente às viagens – modais principais, tempo de duração (da viagem e dos trechos a pé), horário de saída e coordenadas das origens e destinos. A partir dos modais principais, foram criados dois bancos de dados, um relativo ao transporte público (Metrô, Trem, Monotrilho, Ônibus/micro-ônibus/peruas municipais e metropolitanos) e transporte privado (Automóvel – passageiro ou motorista, táxi convencional e táxi não convencional). A opção dessa agregação foi feita em função da classificação modal que foi possível a partir dos dados da simulação das viagens.

O passo seguinte foi a espacialização dos bancos de dados, tanto o simulado como os bancos privado e público da OD. Utilizando o software aberto RStudio, foram utilizadas as coordenadas de origem e destino nos bancos para criar dois *Shapefiles* para cada banco, um a partir dos pontos de origem das viagens e outro a partir dos destinos, com os mesmos dados associados. Para a comparação do banco simulado com os bancos OD, foi criado um algoritmo, repetido para cada viagem dos Bancos OD.

Dada uma viagem OD, foram selecionadas todas as viagens simuladas com a origem dentro de um *buffer* de 1500m (escolhido por incluir 95% dos trajetos a pé para acessar os transportes); dessas viagens foi selecionado ainda o subconjunto de viagens simuladas em que os pontos de destino também se encontrassem dentro de um *buffer* de 1500m do ponto de destino da viagem OD. A partir desse conjunto de viagens simuladas foi calculada a média da duração simulada (considerando se a viagem comparada era de transporte público ou privado), que por sua vez foi subtraída da soma da duração da viagem OD, do tempo andando da origem e do tempo até o destino da viagem OD, convertidos em segundos. Mais explicitamente, a medida *D*  a ser analisada adiante foi calculada como segue em (1).

Esse algoritmo foi repetido para todas as viagens OD e a diferença calculada foi registrada para cada viagem. Uma variante desse algoritmo foi realizada também, adicionando mais uma restrição para a seleção das viagens: além dos *buffers* espaciais, foi utilizado um *buffer* temporal, filtrando as viagens simulada que começassem duas horas antes ou depois da viagem OD.

**3.3. Análise dos dados**

O intuito da análise foi verificar se existem e quais são as diferenças entre as medidas de tempo simuladas e as medidas de tempo da pesquisa OD. O processo de identificação desses 'vieses' passou por uma análise das dependências entre a diferença calculada e algumas medidas consideradas relevantes da OD, a modelagem desses dados para procurar explicar variações na medida e a verificação de dependências espaciais nos dados.

A confirmação da significância da diferença calculada foi feita a partir do teste T para diferença de médias entre as medidas de tempo da OD e calculadas a partir da simulação. Os teste foram realizados separadamente para o conjunto de viagens motorizadas privadas e para as viagens de transporte público e para os cenários com e sem o buffer de restrição de tempo. Em todos os cenários foram verificados *D* consistentes e significativos. O passo seguinte foi elaborar modelos lineares (OLS) para procurar explicar essa diferença. Foram usados como variáveis explicativas somente dados da Pesquisa OD sobre as viagens, descritas na Tabela 3. Para a variável Pontuação do Critério Brasil alguma entradas apresentavam valor nulo; para essas observações o valor do Critério Brasil foi substituído pela média do valor do critério Brasil nas zonas de domicílio do viajante. Para selecionar as variáveis significantes foi utilizado um processo stepwise através do pacote *stats* do software R, e foram retiradas as variáveis com significância inferior à p = 0,05.

Para a análise de dependências espaciais, as diferenças entre os tempos de viagem foram espacializadas dentro das zonas OD, calculando-se a média de *D* para as zonas de origem e de destino das viagens. A partir dessa espacialização foram calculados os Is de Moran Globais e os coeficientes de autocorrelação locais, que foram usados para a elaboração dos Mapas de Autocorrelação Espacial (*Lisa Maps*). Esse procedimento foi feito para os diferentes cenários de comparação e agregação, como indicado na tabela 1. Foi ainda explorada a técnica GWR (*Geographically Weighted Regression*) para visualização de possíveis variações na influencia das variáveis na medida *D*. Os mesmos modelos lineares encontrados na etapa anterior foram calculador localmente para pontos de uma grade regular criada sobre a área de análise. Para cada ponto foi calculado um modelo regressivo local a partir de um numero de pontos vizinhos encontrados pela função *gwr.bw,* do pacote *GWModel* do R; para o cálculo dessa quantidade de vizinhos foi usado somente 20% dos pontos de viagem, por razões computacionais. As variações espaciais dos coeficientes dos modelos analisados foram mapeadas e comparadas.

|  |  |  |
| --- | --- | --- |
| Transporte Público | Com restrição de tempo | Origem |
| Destino |
| Sem restrição de tempo | Origem |
| Destino |
| Transporte Privado | Com Restrição de tempo | Origem |
| Destino |
| Sem restrição de tempo | Origem |
| Destino |

Tabela 1. Combinação de fatores para os diferentes modelos de análise. Elaboração própria.

# **4. RESULTADOS E ANÁLISE**

**4.1. Diferença de médias**

A distribuição das medidas de *D* para os diferentes cenários estão apresentadas nas figuras 3 e 4. Visualmente é possível verificar que a restrição temporal alterou um pouco a distribuição de medidas, mas não alterou significativamente as regiões em que se encontram os valores mais altos e mais baixos no mapa. Para a distribuição de *D* nos dois modais, parece haver ainda alguma semelhança entre os mapas de origem e destino quanto a aglomeração de valores altos. *A* verificação da significância de *D* a partir dos testes T de diferença de médias confirmou a existência de uma diferença significativa , a um valor p de 0,05, para todos os cenários avaliados. Os resultados dos testes estão na tabela 2.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sem restrição de tempo | | Com restrição de tempo | |
|  | Transporte Público | Transporte Privado | Transporte Público | Transporte Privado |
| Média | -598.2338 | -583.9335 | -626.4945 | -579.2007 |
| Significância (Valor P) | < 2.2e-16 | < 2.2e-16 | < 2.2e-16 | < 2.2e-16 |
| Intervalo de Confiança | [-613.6560, -582.8116] | [-592.1072, -575.7599] | [-644.1967, -608.7922] | [-588.5032, -569.8982] |
| Graus de liberdade | 30494 | 35699 | 24494 | 27841 |

Tabela 2. Testes T de diferença de Médias para os diferentes modelos. Elaboração própria.

O uso do buffer de restrição temporal não pareceu mudar substancialmente o resultado, mesmo havendo uma perda significativa nos graus de liberdade dos dados (devido a viagens da OD para as quais nenhuma viagem simulada foi localizada dentro do buffer espacial e temporal). Isso parece ser corroborado pelo que se vê nos mapas de distribuição de *D*. A perda foi um pouco menos significativa para o transporte público (19,78%) que o privado (22,01%). De forma geral, esses resultados confirmam que, em média, os dados simulados subestimam as viagens OD em aproximadamente 10 minutos. A distribuição dessa diferença será melhor explorada, mas é possível supor que haja uma distribuição dessa diferença diferentes perfis de viagens e viajantes. Uma possibilidade é que as viagens calculadas pela API ofereçam sempre as rotas mais eficientes em termos de tempo, e que não sejam essas as rotas necessariamente as adotadas pelos indivíduos - ainda mais em relação ao transporte público. Ainda, isso pode ser uma deficiência da API ao contabilizar os tempos de espera envolvidos nas viagens, o que também afetaria mais as viagens de transporte público. Para explorar essas possibilidades foram realizados os modelos regressivos, apresentados a seguir.

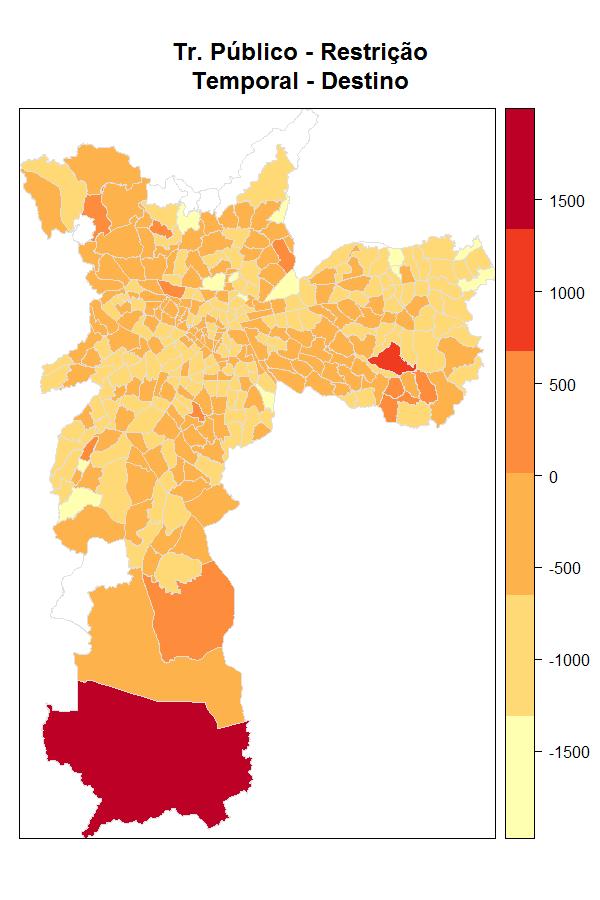
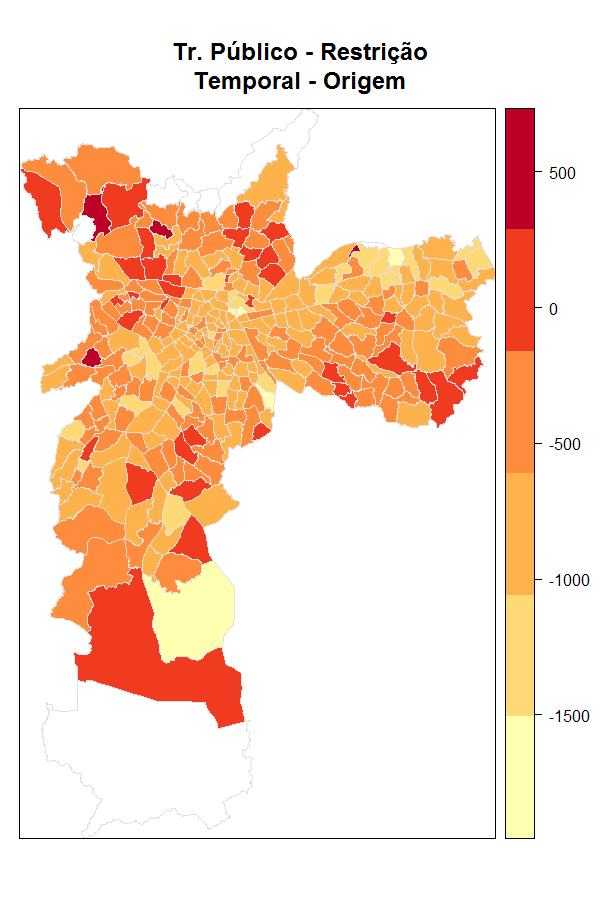
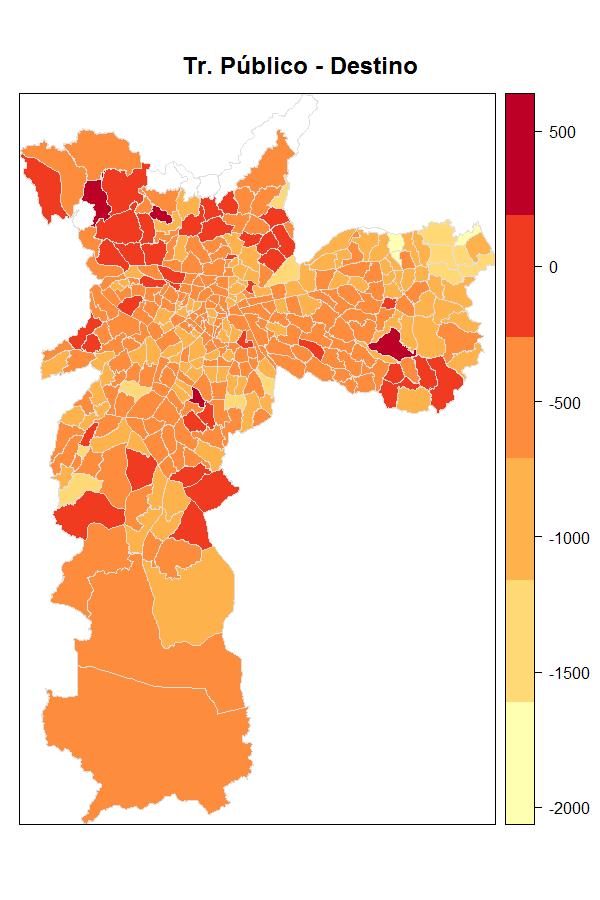
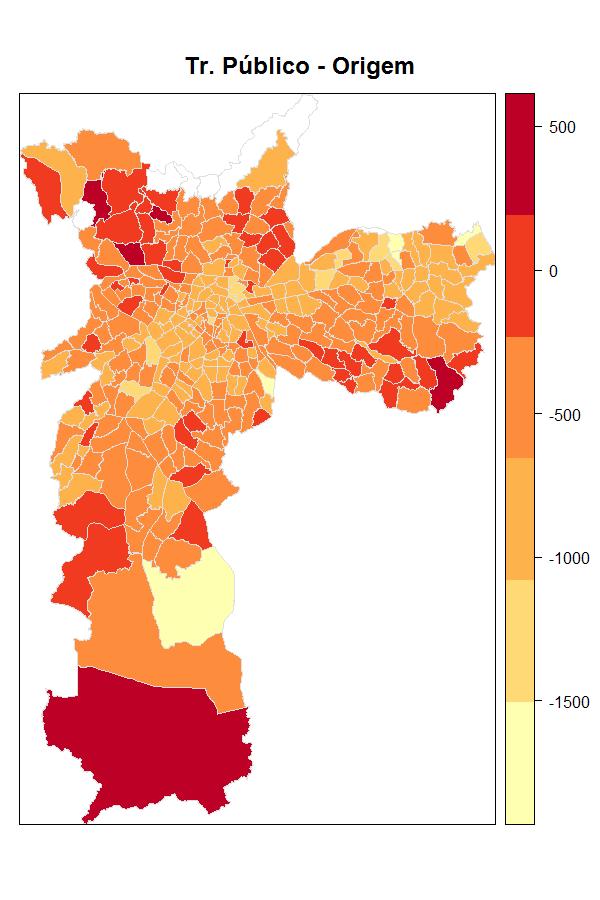
****

Figura 3. Mapas de distribuição da medida D para o transporte público nos casos sem restrição de tempo e com restrição de tempo e agregados pela origem e pelo destino. Elaboração própria.

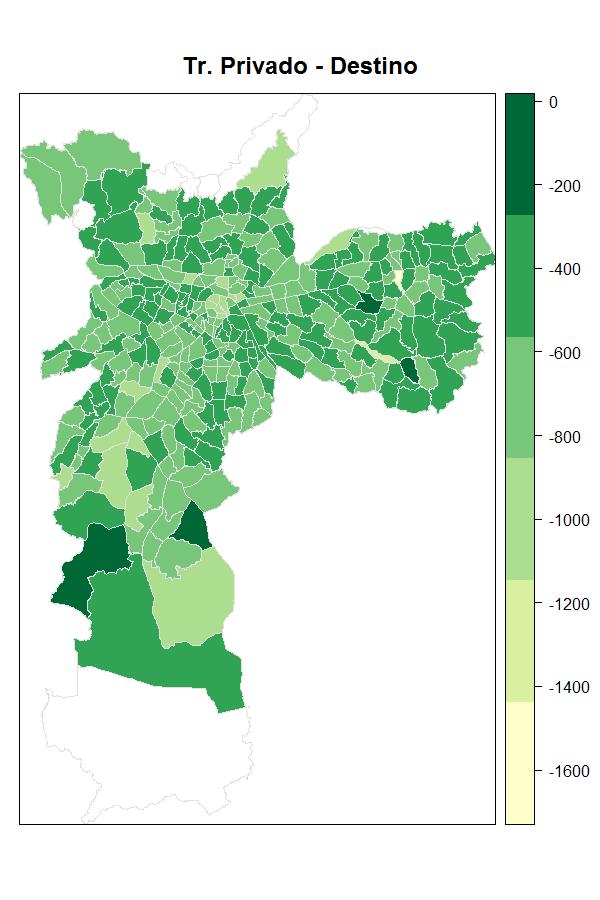
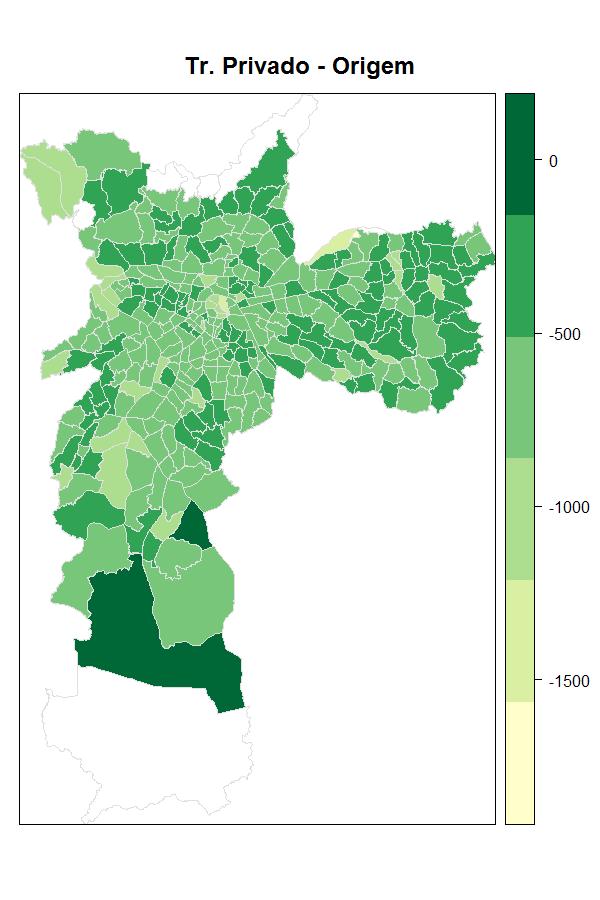
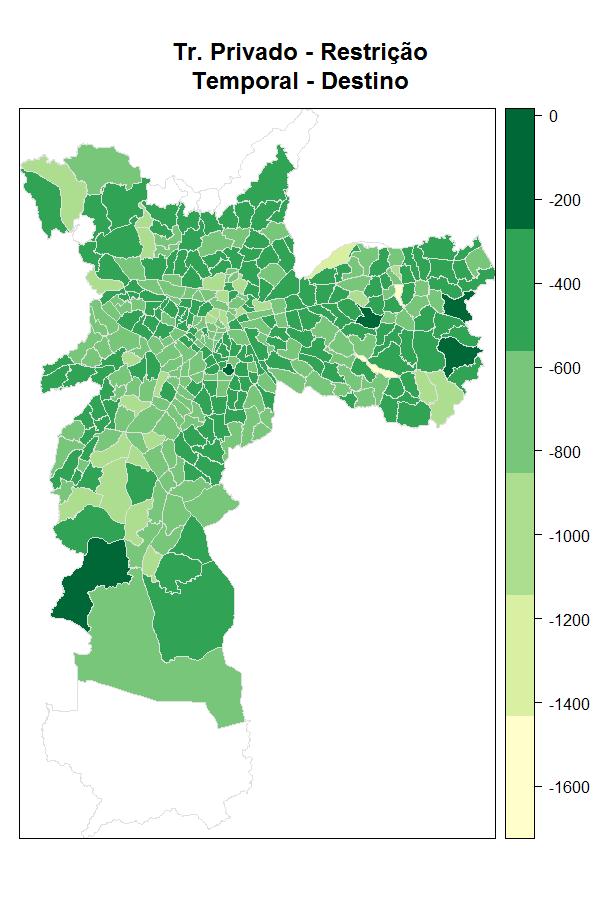
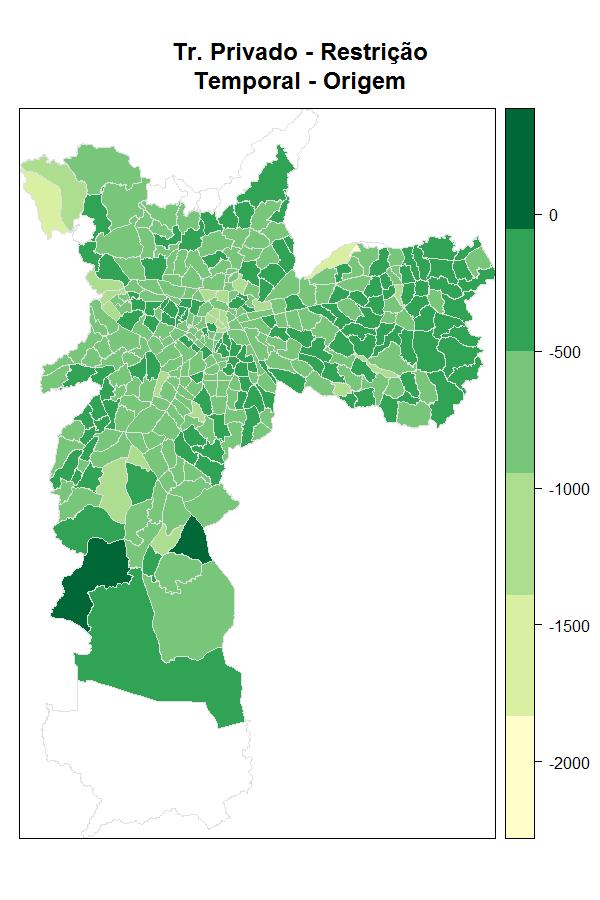
****

Figura 4. Mapas de distribuição da medida D para o transporte privado nos casos sem restrição de tempo e com restrição de tempo e agregados pela origem e pelo destino. Elaboração própria.

**4.2. Modelagem**

Para as regressões lineares foram usadas as variáveis da tabela 3. Mas antes foram analisadas as dependências entre as variáveis numéricas escolhidas para a regressão através painéis de correlação para cada um dos cenários (Figuras 5 e 6). Essa análise preliminar indicou pouca correlação entre a variável de interesse e as variáveis de renda familiar e de Pontuação do critério Brasil (que ainda apresentaram correlação alta entre si). A única variável numérica que apresentou uma correlação significativa com a diferença de tempos foi a duração da viagem na OD. Os diferentes modais foram representados por diferentes cores nos mapas de correlação. A medida *D* das viagens de transporte privado apresenta valores mais negativos e parece se correlacionar com rendas e pontuação do critério Brasil mais altas que as viagens de transporte público. Ainda, ao observar a correlação entre *D* e a duração de viagens da OD, o transporte privado apresenta uma correlação mais forte e caracteristicamente linear, enquanto o transporte público apresenta uma distribuição mais dispersa, apesar de ainda correlacionada. Isso pode indicar que as previsões de transporte público, dada a maior complexidade da realização desse tipo de viagem, seja menos precisa que as viagens de transporte privado. O mapa de correlações a partir do cenário de restrição do tempo de viagens não parece indicar diferenças consistentes ou relevantes quanto a distribuição das variáveis.

|  |  |  |
| --- | --- | --- |
| Variável | Descrição | Fonte |
| Duração da viagem na Pesquisa OD | Duração total da viagem na Pesquisa OD, somando o tempo da viagem com os tempos andando da origem e até o destino. | Microdados Pesquisa OD 2017 |
| Renda Familiar | Renda familiar do indivíduos que realizou a viagem. |
| Pontuação Critério Brasil | Pontuação no critério Brasil da família do viajante. |
| Viagem de Transporte Público | Variável dummy refletindo se a viagem foi feita usando transporte público - considerando o modo principal de viagem como Metrô, Trem, Monotrilho ou Ônibus, micro-ônibus e perua. (O transporte privado agrega motorista e passageiros de automóvel e usuários de taxi convencional e não convencional). |
| Grau de Instrução | Variável dummy considerando as categorias de não alfabetizado/Fundamental I incompleto, Fundamental I completo/Fundamental II incompleto, Fundamental II completo/Médio incompleto, Médio completo/Superior incompleto e Superior Completo. |
| Condição de Atividade | Variável dummy relativa à atividade exercida pelo viajante. Suas categorias são Trabalho regular, Faz bico, Em licença médica, Aposentado ou Pensionista, Sem trabalho, Nunca trabalhou, Dona de casa e Estudante. |
| Motivo da Viagem | Variável dummy relativa ao motivo do deslocamento. Foi feia uma agregação de motivos na origem e no destino, originalmente dados da Pesquisa OD. Viagem com origem ou destino no trabalho e origem ou destino na residência foram agregadas como viagens Casa-Trabalho; foi feita uma agregação análoga para motivos de viagem que incluíssem escola na origem ou no destino, agrupadas em Casa-Escola. Todas as outras viagens foram agrupadas em Outros Motivos |

Tabela 3. Variáveis usadas para modelagem. Elaboração própria.

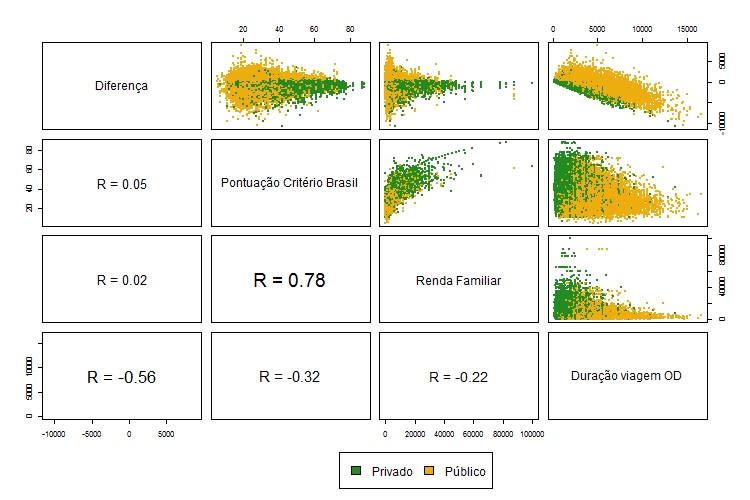


Figura 5. Mapa de correlações entre as variáveis numéricas para o cenário sem restrição de tempo. Elaboração própria.

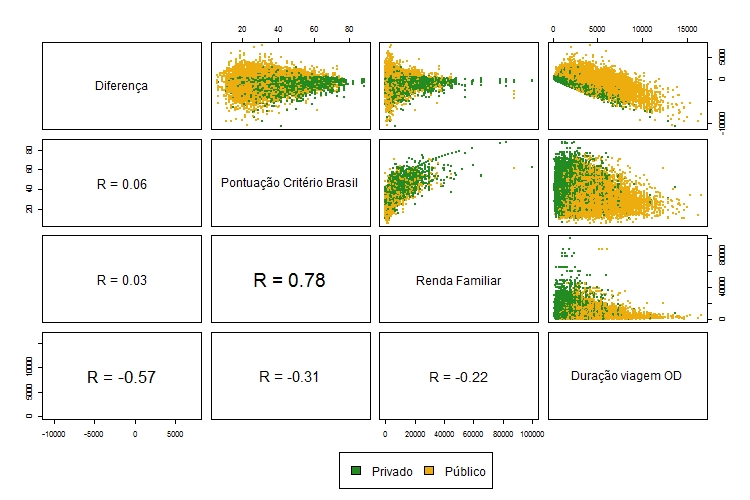


Figura 6. Mapa de correlações entre as variáveis numéricas para o cenário sem restrição de tempo. Elaboração própria.

Apesar das colinearidades indicadas a análise de correlações, foram feitas duas regressões lineares, uma para cada cenários de restrição temporal, com todas as variáveis da tabela 3. As não significâncias e as colinearidades foram retiradas através de um processo stepwise e manualmente. Em relação às variáveis categóricas, foi dotado como modelo base o indivíduo com trabalho regular e não alfabetizado ou com o Fundamental I incompleto, como motivo de viagem a ida ou volta do trabalho. Os resultados dos modelos de regressão linear estão apresentados na tabela 4.

# 

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variáveis** | | **Sem Restrição Temporal** | | | | **Com Restrição Temporal** | | | |
| Valor estimado | Erro P. | Valor T | Signif. | Valor estimado | Erro P. | Valor T | Signif. |
| Intercepto | | 247,470 | 10,2469 | 24,151 | 0,0000 | 212,105 | 18,5990 | 11,44 | 0,0000 |
| Duração OD | | -0,4727 | 0,0019 | -250,828 | 0,0000 | -0,4817 | 0,0023 | -209,211 | 0,0000 |
| Renda Familiar | | -0,0013 | 0,0005 | -2,479 | 0,0132 | -0,0027 | 0,0009 | -2,969 | 0,0030 |
| Pontuação Crit. Brasil | | -- | -- | -- | -- | 1,2571 | 0,5666 | 2,229 | 0,0265 |
| Transporte Público | | 1167,66 | 7,9036 | 147,737 | 0,0000 | 1130,95 | 9,6802 | 116,831 | 0,0000 |
| Grau de Instrução | Fund II incompl. | -- | -- | -- | -- | -- | -- | -- | -- |
| Médio incompl. | -- | -- | -- | -- | -- | -- | -- | -- |
| Sup. incompl. | 58,2627 | 8,4301 | 6,911 | 0,0000 | 67,3840 | 10,4461 | 6,451 | 0,0000 |
| Sup. completo | -33,6303 | 8,9701 | -3,749 | 0,0002 | -24,6232 | 11,7369 | -2,098 | 0,0359 |
| Condição de atividade | Faz bico | -43,1486 | 15,6972 | -2,749 | 0,0060 | -- | -- | -- | -- |
| Em licença médica | -- | -- | -- | -- | -- | -- | -- | -- |
| Aposentado/Pens. | -63,8789 | 10,8939 | -5,864 | 0,0000 | 54,5429 | 13,0241 | -4,188 | 0,0000 |
| Sem trabalho | -- | -- | -- | -- | -- | -- | -- | -- |
| Nunca trabalhou | -- | -- | -- | -- | -- | -- | -- | -- |
| Dona de casa | -- | -- | -- | -- | -- | -- | -- | -- |
| Estudante | -39,3245 | 12,5387 | -3,136 | 0,0017 | -31,1635 | 9,6802 | -2,028 | 0,0426 |
| Motivo da viagem | Escola | -67,9922 | 10,0754 | -6,748 | 0,0000 | -66,4131 | 12,5759 | -5,379 | 0,0000 |
| Outros | -68,3118 | 7,7863 | -8,773 | 0,0000 | -65,9121 | 9,3959 | -7,015 | 0,000 |
| R² | | 0,4999 | | | | 0,4838 | | | |
| R² ajustado | | 0,4998 | | | | 0,4837 | | | |
| Significância | | 0,0000 | | | | 0,0000 | | | |

Tabela 4. Modelos OLS para os cenários sem restrição de tempo e com restrição de tempo. Elaboração própria

O R² e o R² ajustado do modelo linear do caso sem restrição temporal indica que o modelo foi capaz de explicar quase metade da variação de *D*. A duração da viagem OD apresentou uma significância alta, com uma contribuição significativa: cada segundo a mais de tempo na viagem OD registrada colabora com quase menos meio segundo no valor de *D.* Isso pode indicar que a simulação subestima o tempo de viagens mais demoradas; ou que ele é capaz de achar caminhos mais eficientes em viagens longas do que as pessoas estão efetivamente viajando. Inversamente, para viagens curtas o modelo parece indicar que a simulação superestima o tempo gasto na viagem - o valor do intercepto 'equivale' à contribuição da duração OD para viagens de quase 9 minutos O componente de renda familiar desse modelo foi significante, mas com uma contribuição relativamente estreita: cada mil reais a mais de renda familiar está associado a uma redução de *D* de quase 1,3 segundos. Em relação às variáveis categóricas, o grau de instrução só foi significativo para as escolaridades Superior incompleto e completo, com contribuição de 58s e -33 segundos respectivamente, em comparação com o grau de instrução Analfabeto/Fundamental I incompleto. As condições de atividade significantes foram Faz bico, Aposentado ou Pensionista e Estudante; todas com uma tendência de subestimação modesta do tempo de viagem em comparação com quem tem trabalho regular. Isso pode indicar que o comportamento de diferentes perfis apresente uma relação diferente com a mobilidade. Mais significantes nesse sentido são as variáveis de motivos de viagem, na qual tanto o motivo Escola como Outros apresentam subestimação de seus tempos de viagem na simulação em relação ao motivo Trabalho-Casa. Essas duas variáveis indicam que as viagens de estudantes e viagens não ligadas ao trabalho não sejam bem representadas pela simulação (o transporte escolar já foi retirado da amostra, portanto os estudantes que estão no conjunto de dados estão se locomovendo de transporte público ou motorizado individual). É possível que as opções de caminhos oferecidos pela simulação não seja compatível com as opções disponíveis para os indivíduos - seja por uma questão de renda (o que não parece ser a indicação do modelo) ou de informação quanto aos possíveis caminhos.

O modelo considerando a restrição no tempo de viagens apresentou alguma variação nas variáveis. Entraram no modelo a variável relativa à pontuação no critério Brasil, com um coeficiente positivo (foi checada a colinearidade com a renda e não foi significativa para o modelo) e perdeu significância a categoria de condição de atividade Faz bico. Houve uma ligeira redução no R² do modelo e algumas variações nas variáveis, sendo o mais significante o aumento da significância da renda familiar, cujo coeficiente quase dobrou. Uma questão que pode ser discutida é a convivência entre as variáveis da pontuação do critério Brasil e da renda familiar no modelo. Uma possível interpretação é que a renda familiar pode estar mascarando desigualdades entre domicílios com números de trabalhadores diferentes: um exemplo - um domicílio com uma família nuclear sem filhos, com um só trabalhador que recebe um valor X pode ser equivalente nos dados a um domicílio com muitos membros que trabalham por rendas mais baixas que podem somar X. Já a pontuação do critério Brasil, por considerar a posse de eletrodomésticos, a posse de carros e motos, entre ou tos fatores que exigem poder de compra, pode conseguir fazer melhor a distinção de renda nos dados. Um possível ajuste para a renda familiar que evitaria essa questão é calcular a renda familiar per capita, que manteria a informação sobre a renda da família, mas consideraria as condições de distribuição da renda no domicílio.

**4.2. Diagnóstico de dependência espacial**

Para a análise da dependência espacial, a medida *D* foi agregada nas Zonas OD, regiões utilizadas pela própria Pesquisa OD para agregação de seus dados. Foram utilizadas dois padrões de agregação, um em torno das zonas de origem das viagens OD e outro em torno das zonas de destino. Nos dois casos foi calculado o valor da média de *D* na zona. As agregações também consideraram as categorias de restrição temporal e tipo de transporte. O processo de agregação apresentou algumas zonas em que não foram encontradas viagens da OD que tenham sido pareadas com viagens simuladas - essas zonas foram registradas como sem dados, estão indicadas nos mapas e não foram contabilizadas para o cálculo das autocorrelações. O casos específicos da restrição temporal observaram um maior numero de zonas sem dados, dado o critério mais estrito para o pareamento das viagens. Foram calculados os Is de Moran, os diagramas de espalhamento de Moran e os mapas de autocorrelação espacial (LISA) para cada cenário. Os mapas LISA apresentam seis categorias: as zonas classificadas como Alto-Alto apresentam valores relativamente altos de *D* e são rodeadas por zonas que também apresentam valores relativamente altos; analogamente, zonas classificadas como Baixo-Baixo são zonas com valores relativamente baixos de *D* e cercadas por valores baixos de *D*. No caso Baixo-Alto, a zona apresenta um valor relativamente baixo e é cercada por valores relativamente altos e a lógica inversa para o caso Alto-Baixo. Não significante é referente as zonas em que a autocorrelação não foi considerada significante, e sem dados são zonas sem nenhuma medida *D* registrada*.*

**4.2.1 Transporte Público**

As agregações de viagens de transporte público apresentam alguns padrões claros de clusterização espacial de *D.* A figura 7 apresenta os diagrama de espalhamento para os níveis de agregação tanto da origem como do destino, e percebe-se pela distribuição dos pontos que há uma tendência visível de organização. Os mapas na figura 8 permitem posicionar espacialmente esses resultados.

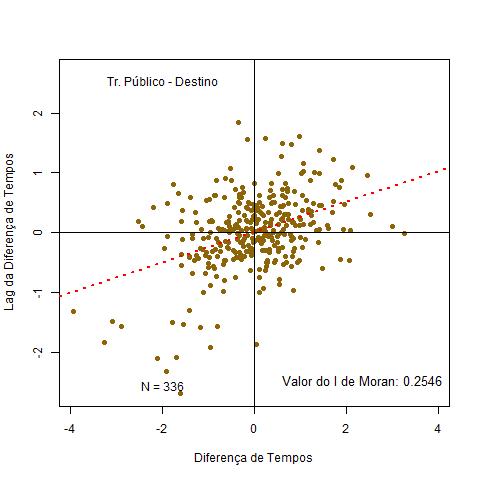
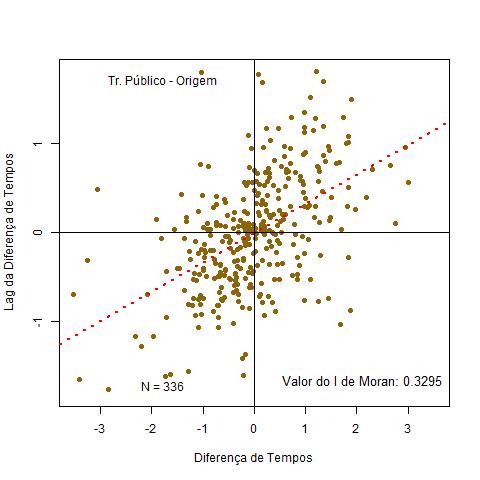


Figura 7. Diagramas de dispersão dos coeficientes de autocorrelação espacial para *D* do transporte público agregado nas zonas de origem e de destino. Elaboração própria.

É visível nos mapas LISA que há aproximadamente quatro zonas de clusterização dos dados. Apesar de não serem exatamente sobrepostas, os dois mapas apresentam zonas de altos valores de  *D* (que indicam que a simulação superestima os tempos de viagem) para as viagens que partem e chegam na região noroeste, na região norte do centro expandido e na região sul da zona leste; Há ainda um pequeno cluster ao sul do centro expandido. Há também um cluster de baixos valores de *D* (que indica que a simulação subestima os tempos de viagem) na região norte da zona leste. Há um pequeno núcleo de valores baixos no centro da cidade, mas que só aparece na agregação em torno da origem. O fato de haver essa sobreposição pode indicar que as medidas altas e baixas de diferença são devidas à viagens feitas localmente, que impactam tanto a origem como o destino das viagens. Uma outra possibilidade para explicar essa sobreposição é que a agregação de viagens em momentos diferentes do dia pode ter um efeito de aumentar a distância entre a previsão e a mobilidade verificada na OD, ainda mais se na região houver grande variação nos tempos de viagem durante o dia. Como a medida de tempo simulada comparada ao tempo da viagem OD é uma média de todas as viagens próximas, pode haver uma compensação das medidas simuladas capaz de impactar na medida da diferença. Nas regiões que são alvo de lentidões localizadas no tempo, essa média de viagens simuladas pode considerar viagens em tempos não compatíveis com o tempo da viagem OD. Apesar da perda de viagens, o cenário com restrição de tempo para o pareamento das viagens ajuda a verificar a robustez das informações encontradas, no sentido que ele evita esse efeito de comparar viagens muito diferentes. Os gráficos na Figura 9 são referentes aos diagramas de espalhamento do cenário de restrição de tempo para transporte público, assim como os mapas LISA na Figura 10. É visível pelos diagramas que há uma redução da clusterização. Na agregação pela origem, a redução parece pequena: o índice de Moran Global caiu de 0,3295 para 0,3024. Já para a agregação nas zonas de destino a redução foi maior, de 0,2546 para 0,1654. Os mapas LISA na Figura 10 mostram a distribuição espacial das autocorrelações espaciais.

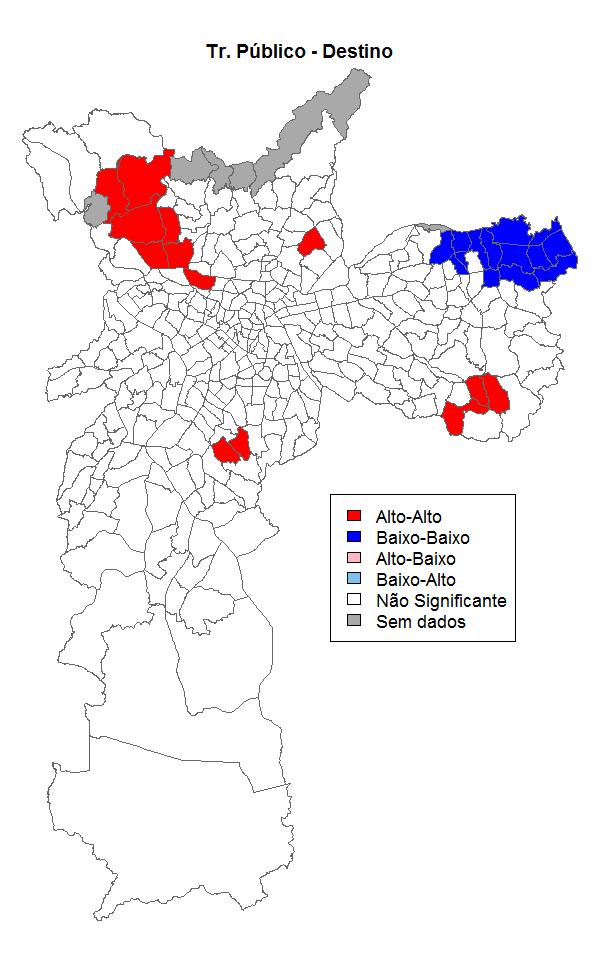
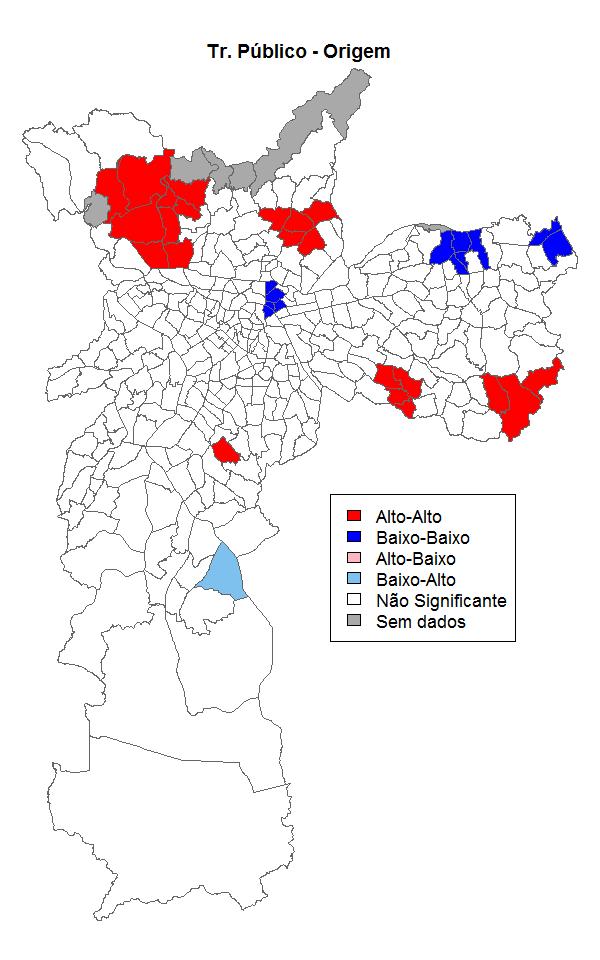


Figura 8. Mapas LISA para *D* do transporte público agregado nas zonas de origem e de destino. Elaboração própria.

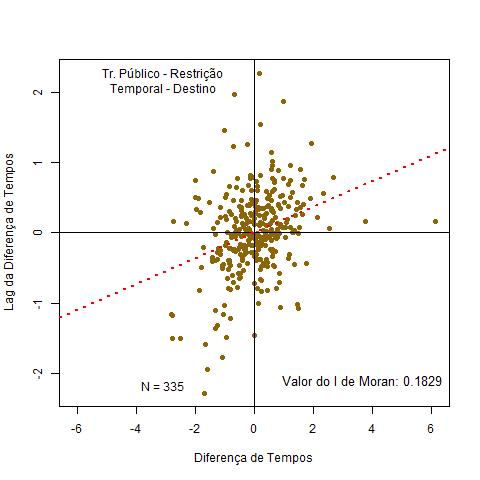
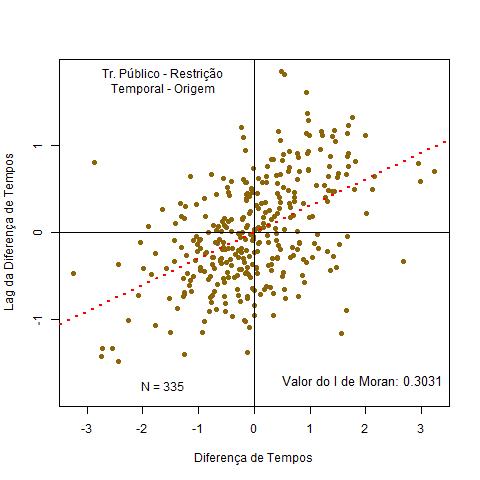


Figura 9. Diagramas de dispersão dos coeficientes de autocorrelação espacial para D do transporte público no cenário de restrição temporal agregado nas zonas de origem e de destino. Elaboração própria.

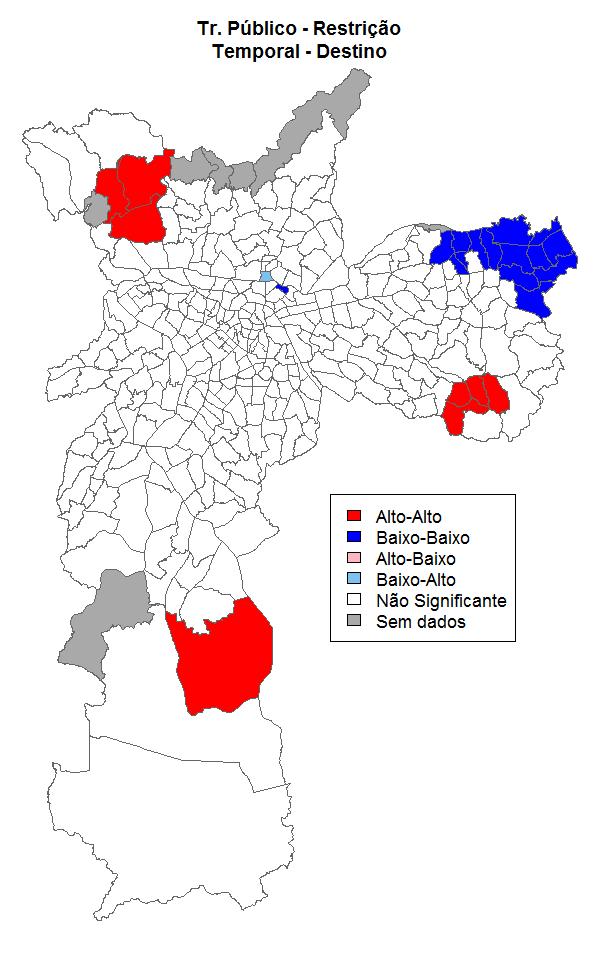
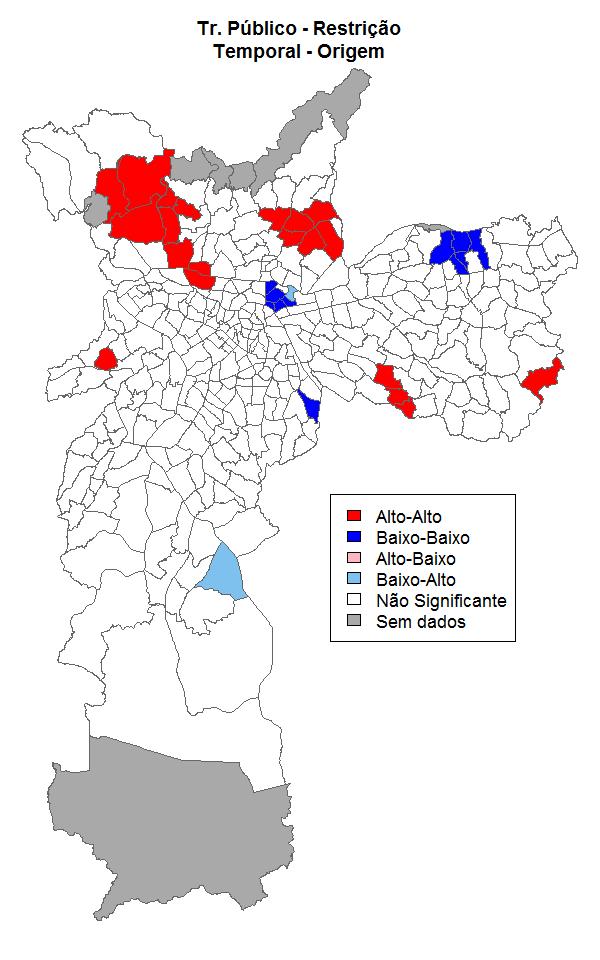


Figura 10. Mapas LISA para D do transporte público no cenário de restrição temporal agregado nas zonas de origem e de destino. Elaboração própria.

Os mapas LISA apresentam um cenário semelhante ao apresentado para o cenário sem restrição temporal. No caso da origem a localização dos clusters é quase a mesma, com variações no tamanho, que foi em geral reduzido. Aparecem novos clusters, relativamente pequenos, no sudoeste do município. Na agregação dos destinos, se repetem os maiores clusters do cenário sem restrição de tempo, que coincidem com clusters da agregação na origem.

O que esses resultados parecem indicar é que há uma variação consistente da medida *D* em certas regiões do município. Isso é um sinal de que há uma diferença espacialmente localizada entre as medidas realizadas na simulação e a mobilidade verificada através da medida OD. Isso não quer dizer ainda que o único fator responsável por essa distribuição é o espaço; outras variáveis que também variem no espaço e que influenciem D podem ser as responsáveis pelo padrão observado. Mas isso requer um esforço de identificação de potenciais variáveis em um contexto espacializado. De qualquer forma, essa consistência na variação pode levar a vieses em estudos que utilizem as medidas simuladas de mobilidade para o transporte público, caso não se tenham sido identificadas os padrões de variação dessa medida.

**4.2.2 Transporte Privado**

Os diagramas de espalhamento de Moran para o transporte privado sem restrição temporal estão apresentados na figura 11. Já de antemão se verifica que há uma clusterização muito menor de *D*, verificável através dos valores do Índice Global de Moran - 0,178 na Origem e 0,105 no Destino.

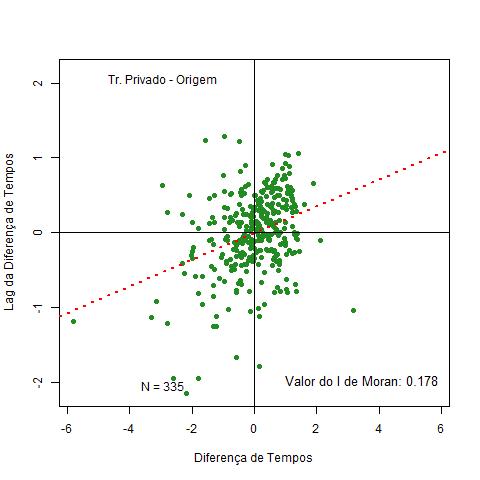
**

Figura 11. Diagramas de dispersão dos coeficientes de autocorrelação espacial para D do transporte privado agregado nas zonas de origem e de destino. Elaboração própria.

# Os mapas LISA na Figura 12 para o transporte privado apresentam isso com mais clareza. Na agregação em torno das zonas de origem há alguns pequenos clusters de valores baixos nas franjas do centro expandido do municípios, na zona noroeste e no norte da zona leste mais próxima do centro e há também um cluster no centro.Além disso há um núcleo de valores altos no norte da zona leste. Mas com exceção do núcleo da região central e do núcleo de valores baixos situado próximo da zona sul, nenhum dos núcleos da agregação na origem encontra eco na agregação no destino. Isso pode indicar que há uma distribuição mais equilibradas das médias de *D* nas zonas, o que pode indicar que há menor interferência do espaço nas previsões que são realizadas pela simulação nas viagens de carro no município.

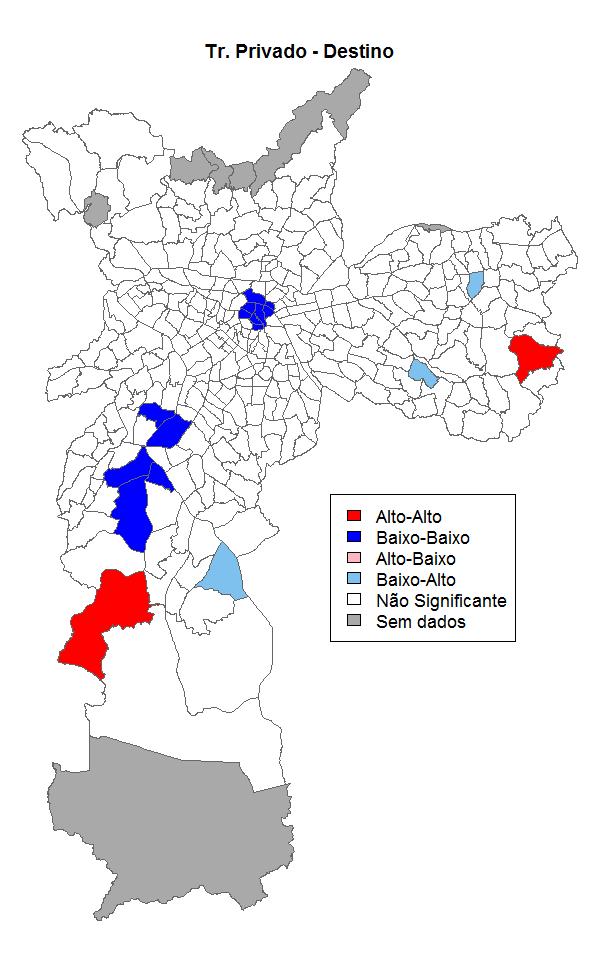
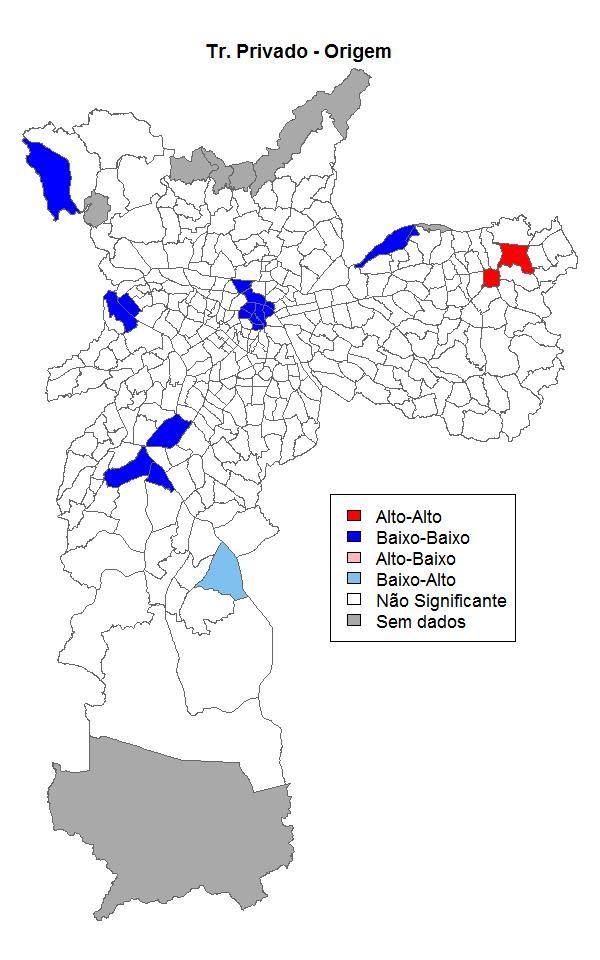


Figura 12. Mapas LISA para D do transporte privado agregado nas zonas de origem e de destino. Elaboração própria.

A comparação dos casos sem restrição temporal e com restrição temporal - apresentados nas Figuras 13 e 14 - parece indicar que os cluster de valores baixos da origem são robustos em relação a agregação temporal - particularmente os clusters no centro, na zona noroeste, na zona leste próxima e na zona sul. Já em relação à agregação nos destino, só parecem relevantes os cluster d valores baixos no centro e na zona sul. De qualquer forma, o nível de clusterização de *D* no contexto do transporte privado é consideravelmente menor do que no contexto do transporte público. Esse comportamento pode ser fruto da maior precisão do serviço de roteamento para viagens motorizadas, dada a menor dificuldade de simulação de rotas possíveis das viagens. Como existem maiores restrições tanto físicas como informacionais para viagens de transporte público, é possível que haja maior distância entre os caminhos executados e os caminhos escolhidos pela API de roteamento. Já nos carros, existe um menor número de variáveis que interferem no deslocamento, sendo que parte delas, como o trânsito ou fechamentos de ruas, são previstas com maior facilidade.

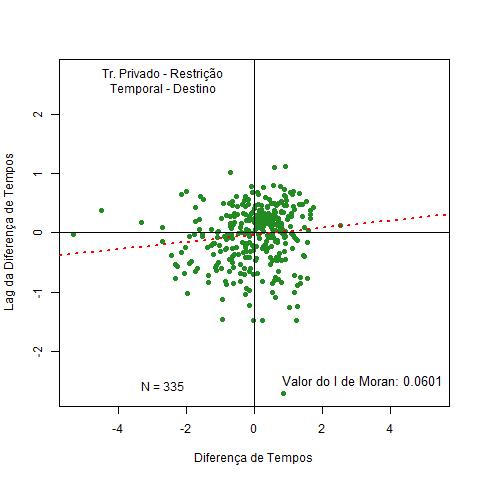
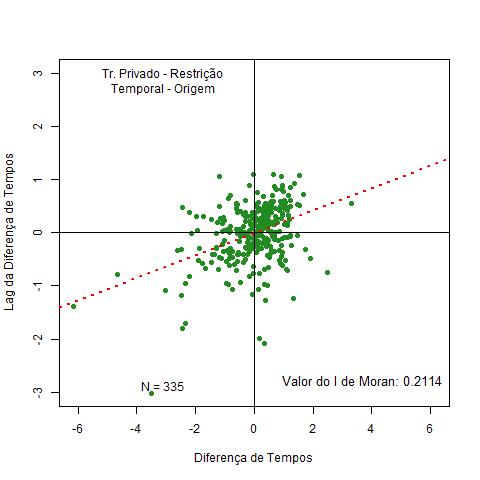


Figura 13. Diagramas de dispersão dos coeficientes de autocorrelação espacial para D do transporte privado no cenário de restrição temporal agregado nas zonas de origem e de destino. Elaboração própria.

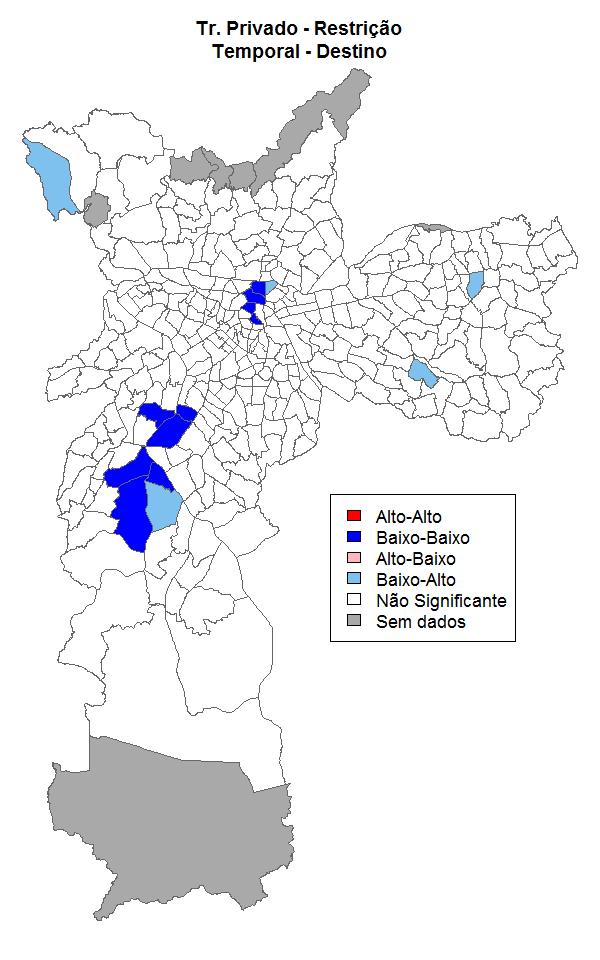
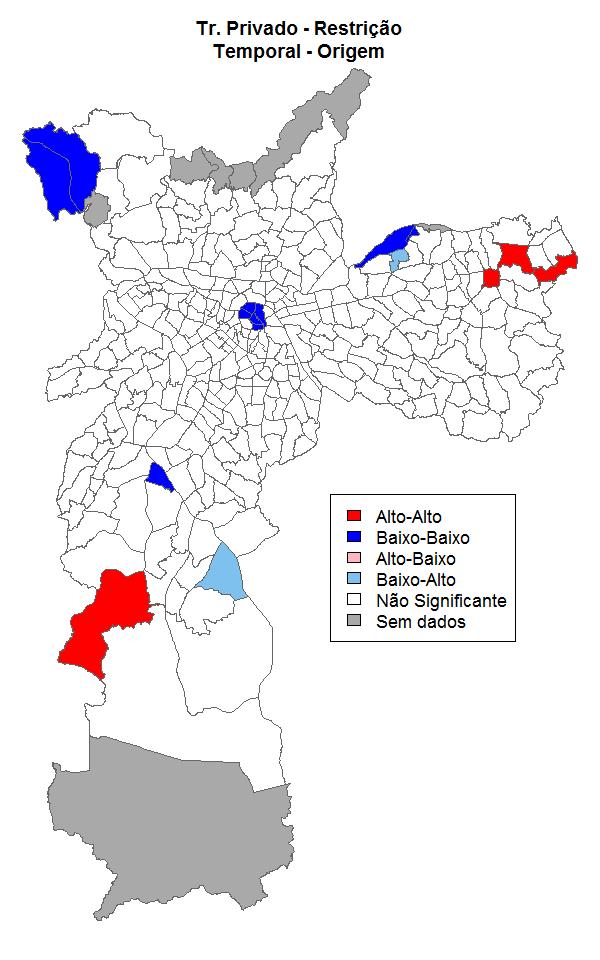


Figura 14. Mapas LISA para D do transporte privado no cenário de restrição temporal agregado nas zonas de origem e de destino. Elaboração própria.

**4.3. Modelagem espacial exploratória- GWR**

Para um estratégia exploratória da variação espacial das variáveis dos modelos encontrados, foi usada a ferramenta de regressão espacial GWR (*Geographically Weighted Regression*). Foram elaborados mapas com os valores previstos dos coeficientes a partir das amostras locais selecionadas pelo programa - foram usados os modelos OLS encontrados anteriormente e foram usados os pontos de origem e destino das viagens para a modelagem, gerando dois mapas de variação para cada variável de cada modelo. Os resultados para as variáveis de duração da viagem OD, do uso de transporte público e do motivo de viagem casa-escola estão apresentados nas imagens S, G, H respectivamente. Os resultado relativos as outras variáveis estão no Anexo I.

O comportamento da variável de duração da viagem OD (Figura 15) parece indicar que para todo o território analisado, viagens mais compridas colaboram para que a simulação subestime o tempo real, ou pelo menos o tempo registrado na OD, das viagens. Mas isso não é uniformemente distribuído. De forma razoavelmente previsível, as regiões do centro expandido e as regiões da zona leste que tem acesso ao metrô e trem parecem ter maiores valores absolutos do coeficiente relativo à duração da viagem OD, indicando que viagens com destino ou origem (mais claramente mostrado nos mapas) nessas regiões tem sua duração subestimada pelos dados simulados. De outra forma, a duração registrada na OD de viagens com algum pé na periferia contribui menos para a subestimação da viagem real do que a distância OD das viagem concentradas no centro expandido. Um explicação possível para isso é que as viagens realizadas pelos modais analisados nas periferias já apresentam comprimentos maiores e tempos, enquanto as viagens concentradas no centro expandido não costumam ser tão longas quanto essas. Como as diferenças entre as medidas foram calculadas em valor absoluto, ao considerar a influencia relativa do tempo OD em *D*, os menores valores de diferença de viagens centrais curtas impactam mais que os valores absolutamente maiores das viagens da periferia mas que são relativamente menores, dado o tamanho das viagens realizadas.

A variável de uso do transporte público (Figura 16), assim como a variável de duração da Viagem OD, apresenta o mesmo sinal em toda a região do estudo - o uso de transporte está associado a superestimação do tempo de viagem pela simulação. Mas há uma concentração de valores menores do coeficiente nas regiões em que há o acesso ao transporte metroferroviário (mas principalmente metroviário). Uma interpretação para isso é que nessas regiões a previsão de viagens de transporte público é mais precisa dada a maior pontualidade dos serviços do metrô e dada a menor quantidade de incertezas na viagem, como trânsito, tempo de espera no ponto, baldeações com maior tempo de espera. Esses fatores podem levar o próprio serviço de roteamento prever maiores tempos de viagem.

O último painel é relativo à dummy de motivo de viagem casa-escola (Figura 17). É possível observar regiões vermelhas representando valores negativo na zona sul e na zona norte, aparecendo consistentemente nos quatro cenários apresentados. O que isso pode indicar é que o perfil de viagem de pessoas que fazem o trajeto de e para essas regiões apresenta maior tempo de viagem do que o que as simulações estão calculando. Isso pode indicar uma falha da estratégia em capturar um perfil de mobilidade diferente do relacionado ao trabalho; mas o fato de haver uma redução considerável nos valores absolutos dessas regiões quando os dados são do cenário de restrição de tempo pode indicar que a significância dessa variável seja relacionada à incapacidade da simulação de considerar os horários de pico - que acabam subestimados ao se calcular a média dentro do buffer. Um conjunto de dados simulados maior e com viagens mais freqüentes permitiria uma restrição mais estrita dos tempos das viagens e pode ser capaz de reduzir uma a incerteza quanto à origem da variação da variável.

# **C:\Users\Administrador\Desktop\tcc\711_GWR_Origem_DuraçãoViagemOD.jpegC:\Users\Administrador\Desktop\tcc\712_GWR_Destino_DuraçãoViagemOD.jpegC:\Users\Administrador\Desktop\tcc\712_GWR_OrigemRestriçãoTemporal_DuraçãoViagemOD.jpegC:\Users\Administrador\Desktop\tcc\711_GWR_DestinocomRestriçãoTemporal_DuraçãoViagemOD.jpeg**

Figura 15. Mapas com as distribuições espaciais da variável Duração Viagem OD do modelo GWR para os cenários sem e com restrição temporal, considerando os pontos de origem e de destino das viagens. Elaboração própria.

# **C:\Users\Administrador\Desktop\tcc\711_GWR_Origem_TransportePúblico.jpegC:\Users\Administrador\Desktop\tcc\712_GWR_Destino_TransportePúblico.jpegC:\Users\Administrador\Desktop\tcc\712_GWR_OrigemRestriçãoTemporal_TransportePúblico.jpegC:\Users\Administrador\Desktop\tcc\711_GWR_DestinocomRestriçãoTemporal_TransportePúblico.jpeg**

Figura 16. Mapas com as distribuições espaciais da variável Transporte Público do modelo GWR para os cenários sem e com restrição temporal, considerando os pontos de origem e de destino das viagens. Elaboração própria.

# **C:\Users\Administrador\Desktop\tcc\711_GWR_Origem_MotivoEscola.jpegC:\Users\Administrador\Desktop\tcc\712_GWR_Destino_MotivoEscola.jpegC:\Users\Administrador\Desktop\tcc\712_GWR_OrigemRestriçãoTemporal_MotivoEscola.jpegC:\Users\Administrador\Desktop\tcc\711_GWR_DestinocomRestriçãoTemporal_MotivoEscola.jpeg**

Figura 17. Mapas com as distribuições espaciais da variável Motivo Escola do modelo GWR para os cenários sem e com restrição temporal, considerando os pontos de origem e de destino das viagens. Elaboração própria.

# **5. DISCUSSÃO FINAL E PRINCIPAIS CONCLUSÕES**

O presente trabalho buscou elaborar e validar uma estratégia de simulação de dados que pudesse ser usada para analisar a mobilidade e a acessibilidade urbana. Foi feita a comparação dos dados simulados com os dados da pesquisa domiciliar OD 2017 a fim de identificar os possíveis vieses e limitações dessa estratégia. Três coisas ficaram claras a partir desse exercício. A primeira delas é que há um viés considerável nas simulações no sentido de subestimar as viagens realizadas. Em parte, isso pode ser compreendido como a dificuldade das simulações de dar conta das incertezas de prever os itinerários de viagem. Nesse sentido, o fato dos dados simulados superestimarem as viagens de transporte público vai nessa direção - os fatores de incerteza em um trajeto de transporte publico são muito maiores que os fatores de transporte privado.

Um segundo ponto a ser considerados é que diferentes perfis de viajantes parecem ter sido identificados na análise, com um certo viés particular para cada um deles. Isso conversa com o argumentos de Kwan (1998) de que certos padrões de mobilidade baseados em diferentes perfis podem ser negligenciados pelo modelo de análise adotado. Mas diferentemente do caso analisado por Kwan (1998), as diferenças no caso particular são encontradas fazendo o controle da origem e do destino de todas as viagens realizadas pelos indivíduos, incluindo as viagens que não são tradicionalmente relacionadas ao trajeto trabalho-casa e casa-trabalho. Do que foi possível identificar no trabalho, é preciso considerar que essa estratégia de simulação pode não ser capaz de representar alguns perfis e alguns padrões de mobilidade.

E terceiro, há uma distribuição espacial dos vieses que dificulta o uso do modelo. Particularmente no caso do transporte público, há uma enorme variação nas incertezas do tempo da viagem dependendo de qual sistema de transporte está sendo acessado, e isso se soma às imprecisões que existem dado ao trânsito e outras alterações de tráfego. Da mesma forma, os padrões de mobilidade e os perfis de usuários podem ser consistentemente diferentes em diferentes regiões da cidade, o que implica em considerações de custo, modal e trajeto diferentes - considerações difíceis de serem previstas no modelo.

O presente trabalho buscou contribuir com uma proposta de estratégia de analise de mobilidade dos indivíduos na cidade. Apesar da identificação de vieses e limitações do modelo, há um ganho explicativo em usar funções de distância para o cálculo de indicadores de acessibilidade que levem em consideração o tempo de deslocamento dos indivíduos no espaço urbano ao invés de funções de distância que simplificam o custo de locomoção envolvido no acesso às oportunidades urbanos.

# **REFERÊNCIAS BIBLIOGRÁFICAS**

ARBEX, R. O., da CUNHA, C. B. Estimação da matriz origem-destino e da distribuição espacial da lotação em um sistema de transporte sobre trilhos a partir de dados de bilhetagem eletrônica. *Revista Transportes*, *25*(3), 166, 2017.

BRASIL. Lei nº 12.587, de 3 de janeiro de 2012, que institui as diretrizes da Política nacional de mobilidade Urbana. *Diário oficial da União*, *4,* 2012.

CACCIA, L. S. Mobilidade urbana: políticas públicas e apropriação do espaço em cidades brasileiras, 2015.

CIA. DO METROPOLITANO DE SÃO PAULO. Pesquisa Origem-Destino 2017. 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.

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.

GUERRA, A. L., BABOSA, H. M., de OLIVEIRA, L. K. Estimativa de matriz origem/destino utilizando dados do sistema de bilhetagem eletrônica: proposta metodológica. *Transportes*, *22*(3), 26-38, 2014.

HUGHES, R., & MACKENZIE, D. Transportation network company wait times in Greater Seattle, and relationship to socioeconomic indicators. *Journal of Transport Geography*, *56*, 36-44, 2016.

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.

INSARDI, A., & LORENZO, R. O. Measuring accessibility: A big data perspective on Uber service waiting times. *Revista de Administração de Empresas*, *59*(6), 402-414, 2019.

LEE, J. G., KANG, M. Geospatial big data: challenges and opportunities. Big Data Research, 2(2), 74-81, 2015.

LESSA, D. A., LOBO, C., & CARDOSO, L. Accessibility and urban mobility by bus in Belo Horizonte/Minas Gerais–Brazil. *Journal of Transport Geography*, *77*, 1-10, 2019.

LETOUZÉ, E., JÜTTING, J. Official statistics, big data and human development: towards a new conceptual and operational approach. Data Pop Alliance and PARIS21, 2014.

LITMAN, T. Measuring transportation: traffic, mobility and accessibility. *Institute of Transportation Engineers. ITE Journal*, *73*(10), 28, 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.

MORANDI, E. L. *Close but far: the real distance between blacks and whites in the largest Brazilian city* (Master Thesis), 2020.

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.

PRESTON, V., MCLAFFERTY, S. Spatial mismatch research in the 1990s: progress and potential. *Papers in regional science*, *78*(4), 387-402, 1999.

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.

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.

TORSTEN, H. What about people in Regional Science?. *Regional Science Association*, *24*(1), 6-21, 1970.

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.