FUNDAÇÃO GETÚLIO VARGAS
ESCOLA DE ADMINISTRAÇÃO DE EMPRESAS DE SÃO PAULO

Programa Institucional de Bolsas de Iniciação Científica (PIBIC)

Simulação de dados de mobilidade como estratégia para produção de análises sobre o transporte público

Aluno: Rodolfo Oliveira Lorenzo

Orientador: Eduardo de Rezende Francisco

Campo de estudo: Administração Pública e Estatística

São Paulo – SP

2019

Simulação de dados de mobilidade como estratégia para produção de análises sobre o transporte público

Resumo

A produção de dados de mobilidade urbana é uma importante atividade para a compreensão das dinâmicas de acessibilidade espacial nas cidades. Porém, a produção de dados empíricos confiáveis e em volumes suficientes para a análise da mobilidade é muito custosa. O presente trabalho explora uma estratégia de simulação de dados de viagens a partir de ferramentas de Big Data como uma possibilidade de atender essa necessidade. A ferramenta do Google Maps é usada para montar um banco de viagens simuladas, com origens e destinos parcialmente aleatorizados dentro do município de São Paulo, tanto para transporte público como para transporte privado.

A estrutura dos dados é analisada a partir de estatísticas descritivas e de visualizações espacializadas. Além disso, para validar a responsividade dos dados à realidade física do município, as medidas de comparação entre os tipos de viagens são analisadas a partir de diferentes estratégias de modelagem espacial, utilizando dados socioeconômicos de controle e de infraestrutura de transporte público nos distritos de São Paulo. São elaborados modelos regressivos SAR e GWR para a análise das relações entre os dados simulados e os dados empíricos de São Paulo. Os resultados indicam que os dados são explicados razoavelmente bem pelos indicadores de infraestrutura de transporte, mas que existem vieses relevantes a serem considerados na análise.

Palavras-chaves

Transporte, Mobilidade Urbana, Big Data, Simulação

1. Introdução

Uma das grandes dificuldades em realizar estudos quantitativos de mobilidade e acessibilidade urbana é o custo, em tempo e em recursos, para se produzir dados confiáveis e tempestivos sobre os comportamentos de locomoção urbana. Uma referência metodológica nesse sentido são as pesquisas de Origem e Destino, importantes ferramentas para não só avaliar os fluxos de pessoas nas cidades, mas também a condição de realização desses fluxos e questões socioeconômicas subjacentes aos comportamentos de mobilidade.

Porém, estudos extensos sobre mobilidade que utilizem essa abordagem, apesar de necessários, são pouco frequentes. Devido a seu custo, as instituições responsáveis por essas pesquisas realizam em intervalos periódicos relativamente longos. Por isso, estratégias de pesquisa alternativas se fazem necessárias para a abordagem de questões de mobilidade (que não dependam do detalhamento fornecido pelas pesquisas OD) nos intervalos dessas pesquisas.

Uma das alternativas possíveis é a simulação de dados de viagens. Essas podem ser feitas a partir de suposições acerca do comportamento gerador de viagens e das condições de mobilidade específicas. Uma das desvantagens dessa abordagem é que é preciso assumir 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. Uma abordagem menos ambiciosa depende de se abrir mão da simulação dos fluxos de viagem, preocupando-se somente com as condições de mobilidade - como no caso desse artigo, com os tempos de locomoção. Novamente, isso pode ser feito por uma simulação em que as condições de locomoção, como vias, velocidades, impedimentos e congestionamentos são simulados no modelo.

Porém, abordagens que levem em consideração muitos fatores na previsão da mobilidade podem gerar modelos extremamente complexos. Estratégias de simulação baseadas em Big Data, baseadas no uso e na disponibilidade intensiva de dados, permitem o uso de modelos mais simples, que dependem de menos suposições de comportamento e mais do acesso de grandes quantidades de dados para produzir as previsões de seus algoritmos. O presente trabalho busca explorar uma alternativa simples para simular os tempos de mobilidade, de forma a comparar a mobilidade do transporte privado e do transporte público. Adotando a postura de simular unicamente os tempos de viagens, o seja, abrindo mão de simular os

comportamentos de mobilidade (que podem ser captados nas pesquisas OD e em formas mais completas de simulação), foi explorada a possibilidade de utilizar ferramentas de Big Data para realizar a simulação de viagens para o município de São Paulo, de forma a capturar informações em tempo real da qualidade da locomoção no município, comparando o modal privado com o modal público. A partir da ferramenta da API do Google Maps, foram simuladas cerca de 125 mil viagens a partir dos dois modais de transporte. 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 à variáveis socioeconômicas e de infraestrutura de transporte, com o intuito de validar a relação dos dados simulados às variáveis físicas do município de São Paulo.

Foram identificadas relações consistentes entre a disposição espacial da cidade de São Paulo e os dados simulados, assim como foram indicados certos vieses da estratégia de simulação. Futuras abordagens podem se beneficiar de informações contidas nesse trabalho, ainda que uma validação mais direta, a partir de comparações com medidas empíricas de tempos de viagem seja necessária para o uso dessa estratégia com maior segurança.

2. Teoria

A questão da mobilidade em São Paulo

A mobilidade na metrópole de São Paulo é resultado de seu processo de urbanização e reflete as vicissitudes do mesmo. Durante o século XX o desenvolvimento das cidades brasileiras seguiu um padrão semelhante de crescimento intenso e periferização precária, gerando uma ocupação segregada do espaço urbano (Maricato, 2003; Rolnik e Klink, 2011). São Paulo, como principal expoente do crescimento urbano do período, não fugiu a esse padrão. Um intenso processo migratório acompanhado de uma rápida industrialização, cujo processo de capitalização drenava os recursos disponíveis, levou a formação de periferias extremamente vulneráveis, com péssimas condições de habitabilidade, além de distanciadas de forma considerável do centro da cidade (Kowarick, 1979). Mesmo considerando que a condição das periferias do município não foi uniformemente constituída, e que houve intervenções do Estado relativas à oferta de infraestrutura e serviços desde dos anos 70, ainda nos anos 2000 os indicadores das periferias apresentavam números consideravelmente piores que os das áreas centrais da cidade, com exceção particular das regiões centrais ocupadas por favelas e cortiços (Torres e Oliveira, 2001; Torres et al., 2003). O crescimento da mancha urbana de São Paulo, em seu processo de conurbação, levou a lógica da periferização para os municípios vizinhos, seguindo tendência já apontada no fim dos anos 70 (Kowarick, 1979), quando os indicadores sociais das periferias do município de SP, que eram muito piores que as áreas centrais, passaram a se estender para a periferia estendida, nos municípios vizinhos. Outro fator importante na relação centro-periferia que deriva também dessa formação urbana é a concentração de trabalhos disponíveis no Município de São Paulo. Apesar de muitos dos municípios vizinhos terem desenvolvido importantes economias geradoras de empregos, inclusive industriais (como no grande ABC), o censo de 2000 mostra que dos quase um milhão de habitantes da RMSP que trabalhavam ou estudavam fora do município de origem, mais da metade se dirigia para o município de São Paulo (Aranha, 2005) – revelando uma tendência centro-periferia metropolitana que se reproduz na escala municipal.

Dentro dessa demanda intensa por mobilidade, as decisões tomadas em relação à questão agravaram o cenário. Por muito tempo foi priorizada a mobilidade viária em detrimento dos trens e do metrô, com grandes projetos de expansão viária e anéis de circulação, e dentro dessa foi incentivado o uso de transporte individual, em razão de

incentivos à indústria automobilística a nível nacional, em detrimento do coletivo (Júnior, 2011; Gakenheimer, 1999; Silveira e Cocco, 2013; Wilheim, 2013; Scaringella, 2001). Esse foram fatores que contribuíram para a geração de uma infraestrutura viária incapaz de acompanhar as taxas de motorização do Brasil, além de um sistema de transporte público dependente de um empresariado ligado aos ônibus (Silveira e Cocco, 2013). Essa situação levou ao agravamento das condições de mobilidade para os moradores das periferias paulistanas que, dependendo tanto de carros como de transportes público, sofrem com viagens longas, congestionamento e saturação dos meios coletivos. Essa condição não é particularidade do Brasil, ou de São Paulo: cidades que passaram por intensos processos de urbanização associados a motorização apresentam grandes dificuldades para manter sua infraestrutura de transportes em compasso com a demanda (Gakenheimer, 1999).

A produção de estatísticas oficiais e o Big Data

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).

Ao mesmo tempo 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, que gera situações como a de Gana, em que a adoção de uma metodologia mais nova de cálculo de PIB indicou um crescimento de 60% desse[1]. A dificuldade desses países em construir essa capacidade passa pela falta de recursos financeiros, a baixa capacitação técnica do serviço público (causa e conseqüencia de uma fuga de cérebros para o setor privado), intervenções políticas na produção de dados, entre outros fatores (Letouzé e Jutting, 2014). Além disso, nesse cenário de fragilidade institucional se posiciona o desafio da crescente produção de dados e das novas formas de análises estatísticas que acompanham o termo Big Data.

Uma das primeiras definições de Big Data está relacionada às características dos dados englobados pela definição. 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. Ainda nessa direção existem definições que incluem outras características aos dados usados: 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) (Gandomi e Haider, 2015).

1http://www.reuters.com/article/2010/11/05/ozatp-ghana-economy-idAFJOE6A40BG20101105 6

Mas existem outras definições de Big Data, que partem de outros pressupostos. 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. Outras definições partem ainda de critérios voltados à implementação de sistemas, com 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 autonomamente de diversas fontes. É possível, a partir dessa capacidade, tentar reproduzir os indicadores oficiais já existentes, ou outros, mais granulares e inteligentes. Letouzé e Jütting (2014) 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. Nesse sentido, pensando no movimento de Big Data como um importante vetor de mudança na sociedade moderna, Letouzé e Jütting (2014) consideram 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). Mas a produção desses dados através de pesquisas empíricas de validade estatística, como a Pesquisa OD (METRO, 2008), 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) - esses dados 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.

Mobilidade e Acessibilidade

Em relação à mobilidade, a compreensão das formas de usos de diferentes modais em cada região podem 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 a 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).

Mas entender a mobilidade urbana, apesar de sua importância, não engloba toda a experiência de acesso a cidade. A informação de como os indivíduos se locomovem na cidade não nos informa se eles conseguem acessar as oportunidades que a cidade pode oferecer. Um conceito mais amplo, capaz de refletir o acesso dos indivíduos à cidade é a acessibilidade (Litman, 2003). A mobilidade, de acordo com a definição de Litman, é um meio para que os indivíduos cheguem aos seus destinos. Assim, para o estudo da acessibilidade, o que interessa em relação a mobilidade é o custo - tempo, dinheiro, desconforto ou risco - que ela implica aos indivíduos, e esse custo é um dos componentes das medidas de acessibilidade; o outro componente é a qualidade e a quantidade de oportunidades e sua distribuição no tecido urbano (Paéz, Scott e Morency, 2012).

As medidas de acessibilidade podem ser elaboradas baseadas nos indivíduos, associando a eles o valor da medida, ou baseadas nos lugares, em que a acessibilidade é um atributo do lugar. Ao mesmo tempo, as medidas podem ser centradas no local da origem das viagens potenciais ou no local de destino das viagens. Também, os dois componentes das medidas, o custo de transporte e a distribuição de oportunidades, podem ser abordados de forma normativa ou positiva. A abordagem positiva consiste em considerar o que de fato acontece, tanto em termos da mobilidade como da distribuição de oportunidades. A abordagem normativa considera o que deveria acontecer (em termos de mobilidade, qual é custo que deveria ser aceitável para o indivíduo) e em geral não se utiliza na distribuição das oportunidades (Paéz, Scott e Morency, 2012). Em relação aos tipos de indicadores de acessibilidade, a literatura abordada aponta quatro grupos: os indicadores "gravitacionais", os indicadores cumulativos, os indicadores baseados em utilidade e indicadores de espaço-tempo. Os indicadores gravitacionais, os cumulativos e os de espaço tempo são instâncias particulares da seguinte fórmula (Paéz, Scott e Morency, 2012; Kwan, 1998):

$$A_{ik}^{p} = \sum_j g\left(W_{jk}\right) f\left(c_{ij}^{p}\right)$$

A medida de Acessibilidade A é dada para a origem i e as oportunidades k para o 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 para o 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. As medidas de utilidade são baseadas no termo "log-sum" de "modelos discretos de escolha aplicados à análise se escolha de destino" (Paéz, Scott e Morency, 2012).

Alguns problemas 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 referência de origem impede que os indicadores deem 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).

3. Métodos

A execução do projeto pode ser descrita em dois grandes blocos: a simulação das viagens e a análise do banco gerado por essas simulações. Essas seções serão descritas separadamente.

Simulação de viagens

A simulação foi feita em duas etapas: primeiro a geração banco de endereços e depois a simulação das viagens propriamente ditas. De antemão algumas considerações precisam ser feitas:

- A relativa alta complexidade de simulações locais que consigam captar o comportamento em tempo real – com informações de trânsito – do tráfego implicava 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 para manter o caráter gratuito das requisições de previsão das viagens

- Essa limitação implicou em 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, a opção por usar a computação em nuvem para a simulação implicou na tentativa de reduzir a computação necessária para evitar problemas relacionados ao desempenho.

Essas limitações definiram o processo de definição de endereços – se procurou sortear endereços em regiões mais densamente povoadas para evitar possíveis perdas. Ao mesmo tempo, para reduzir o esforço computacional, foi montada uma base de coordenadas offline, que foi usada pelo programa para sortear os endereços das viagens. Essa primeira etapa foi realizada nos seguintes passos:

I. Usando o software Qgis, foi gerada, a partir do shapefile do município de São Paulo censo de 2010, uma grade com quadrículas de 500m

II. A partir da informação da população por setor censitário, foram calculadas as populações de cada quadriculas e foram retiradas as quadriculas com população zero.

III. Dividindo esses setores em quintis de densidade populacional, foram sorteados aleatoriamente pontos de coordenadas dentro de cada quadricula, de acordo com o quintil: 5 pontos para o quintil mais populoso, e 1 ponto para o quintil menos populoso.

IV. A base de coordenadas resultante foi usada para o sorteio dos endereços.

A base de endereços resultante desse processo encontra-se representada na figura I do anexo.

A simulação propriamente dita das viagens foi feita a partir de um programa em Python, rodado no serviço de computação em nuvem da Google. A estrutura do programa seguiu a seguinte configuração:

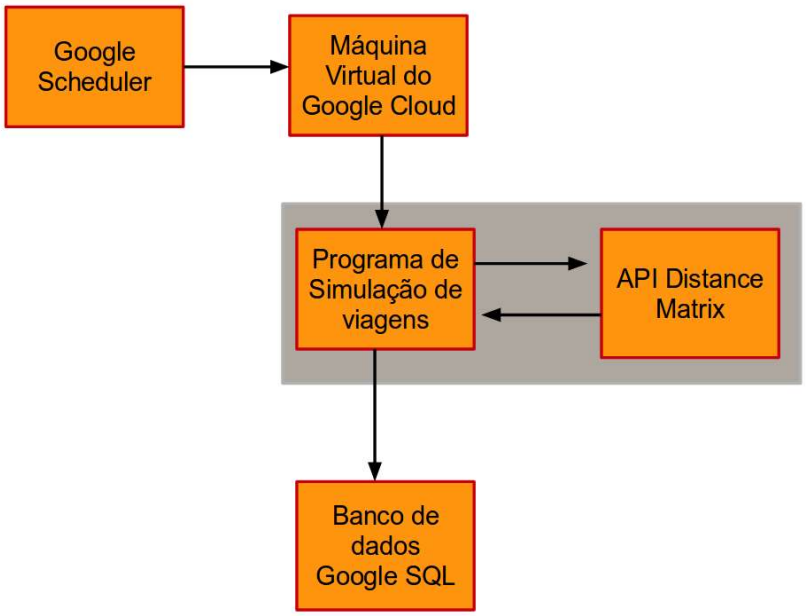


Figura 1:Estrutura do programa usado para o cálculo dos tempos de viagem

Foi usada uma ferramenta de agendamento (Google Scheduler) de ativação ligada a uma máquina virtual no ambiente em nuvem da Google. Por sua vez a máquina virtual rodou um script de inicialização que continha o programa usado para realizar a simulação. O

agendamento foi feito para os 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. Houve alguns problemas nesse agendamento, que serão abordados mais à frente. O programa iniciado pela máquina virtual seguiu as seguintes etapas:

I. Abrir uma conexão com o banco de dados SQL da nuvem da Google

II. definir uma função para inserção dos dados das viagens no banco de dados

III. Carregar o banco de coordenadas (previamente carregado à máquina virtual)

IV. Sortear dez coordenadas de origem e outras dez coordenadas de destino

V. Chamar a API Distance Matrix com as dez origens e os dez destinos, para viagens de transporte público

VI. Processar os resultados devolvidos pela API e armazenava em um vetor auxiliar

VII. Chamar novamente a API Distance Matrix com as mesmas dez origens e os dez destinos, para viagens de transporte privado

VIII. Processar os resultados devolvidos pela API e anexar ao vetor auxiliar

IX. Submeter o vetor auxiliar a função que insere os dados no Banco de Dados hospedado na nuvem.

Cada chamada da API Distance Matrix retornava uma lista com cem viagens (matriz de 10 origens e dez destinos) com dados de ano, hora, dia da semana, coordenadas da origem e do destino da viagem, endereços da origem e do destino da viagem, duração, distância e tarifa da viagem (para o transporte público). Assim, cada chamada da função anexava 200 viagens pareadas (100 de transporte público e 100 de transporte privado) ao Banco de Dados hospedado na nuvem. O período de simulação foi entre os dias 11 de fevereiro de 2019 a 5 de junho de 2019. Da metade do mês de março até a metade do mês de abril, como indicado no Figura 2, houve insuficiência do servidor em nuvem do Google para o qual o programa não estava preparado. Da metade de abril até a segunda semana de maio o programa ficou suspenso para correção; pela semana seguinte houve um aumento da intensidade da coleta em

quatro vezes para compensar o período anterior (chamadas nos dez e cinco minutos antes das horas cheias, nas horas cheias e cinco minutos após as horas cheias). Nas duas últimas semanas de maio e na primeira semana de junho o programa seguiu o mesmo padrão inicial. 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 de ônibus (98,47% de aproveitamento).

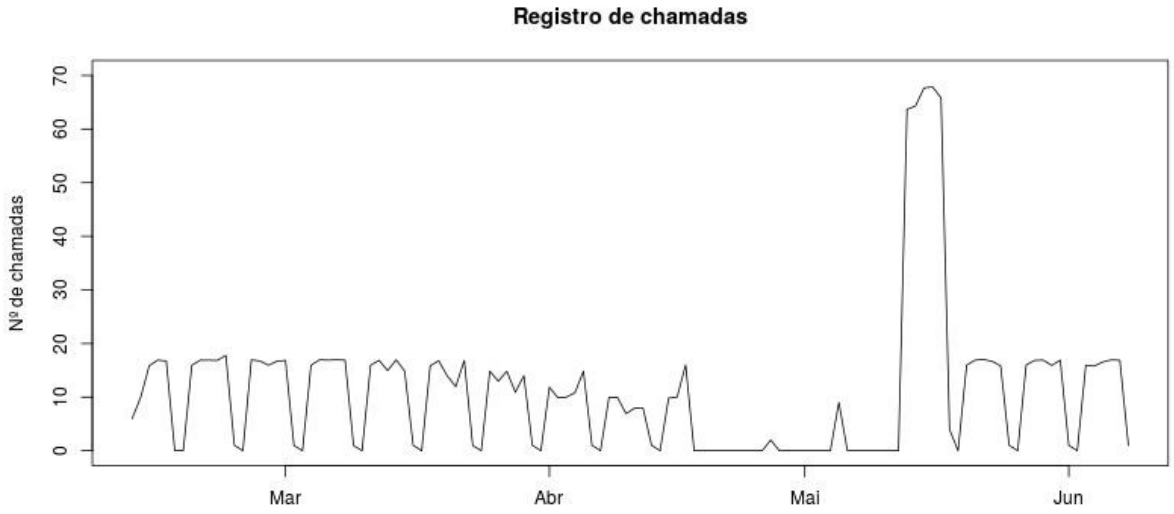


Figura 2: Registro das chamadas feitas à API do Google

Análise do banco de dados

A análise de dados seguiu três etapas. As duas primeiras consistiram em análises exploratórias dos dados e das medidas elaboradas para a análise, sendo a primeira etapa uma análise do agregado de todas as viagens e a segunda etapa uma análise exploratória da distribuição espacial das medidas no município. A terceira etapa de análise foi a modelagem das medidas elaboradas a partir de variáveis socioeconômicas e de variáveis de infraestrutura de transporte público nos distritos do município de São Paulo. Foram montados modelos de regressão simples (OLS), modelos de regressão de auto correlação espacial (SAR) e modelos de regressão espacialmente ponderados (GWR). Os resultados e características dos modelos foram comparados e discutidos.

Para a primeira análise foram realizadas estatísticas descritivas das medidas consideradas de interesse para entender a estrutura geral dos dados simulados. Após a coleta e antes da análise os dados armazenados foram processados. As entradas do banco de dados após o processamento de dados apresentam a seguinte estrutura:

Tabela 1:Estrutura do banco de dados

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

A variável tarifa não apresentou resultado consistentes para todas as chamadas de viagens de transporte público e por isso foi descartada. A partir do pareamento das viagens por modal (público e privado) foram criadas duas medidas, que foram alvo de análise do trabalho: (1) a diferença entre o tempo de viagem do modal público e do modal privado, dada por:

$$D_t = \frac{T_{público}}{T_{privado}}$$

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, calculada como:

$$R_t = \frac{T_{público}}{T_{privado}}$$

Posteriormente no trabalho as medidas serão referidas como D_t e R_t. Além de explorar as distribuições de D_t e R_t foram feitas análises para checar sua normalidade e suas possíveis dependências em função do horário e do dia da semana em que as viagens foram simuladas, como também as correlações das medidas com a distância das viagens e entre elas próprias.

Em seguida os dados de viagens, devido a sua natureza eminentemente geográfica, foram analisados a partir de abordagens espaciais. Para visualizar a distribuição das medidas em função da origem e do destino das viagens, foram elaboradas superfícies do município de São Paulo (considerando para cada superfície um distrito de São Paulo como origem das viagens) calculando para cada ponto da superfície a média das medidas analisadas dos vinte pontos mais próximos ao ponto da superfície, ponderados por uma gaussiana modificada centrada no ponto a ser calculado e pela distância euclidiana até os pontos de destino de viagens

próximos, similarmente ao processo de cálculo dos indicadores de gravidade (Kwan, 1998). O processo também foi repetido para as áreas de ponderação para observar as distribuições de forma mais granular. Para esse cálculo foi utilizado o algoritmo de regressão espacialmente ponderada (GWR), com um modelo simples de regressão como média da medida de análise.

Além da visualização dos dados a partir do GWR, foi feita para D_t e R_t análises dos I's de Moran e de Mapas de Indicadores Locais de Associação Espacial (LISA), tanto para o nível dos distritos como para as 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 melhor 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. Isso forneceu para cada medida um conjunto de dados georreferenciados no distrito de origem das viagens. Os dados que compuseram a base para a modelagem foram semelhantes aos usados no artigo anexo, mas modificados. Foram compilados dados socioeconômicos e de infraestrutura de transportes agregados por distrito e ponderados pela área dos distritos. As variáveis usadas estão descritas na Tabela 2. A partir desses dados foram calculados modelos de regressão linear, cujas variáveis foram reduzidas através de um processo stepwise, seguido da retirada de variáveis ainda insignificantes e de variáveis colineares. Esses modelos foram comparados aos modelos de regressão espacial.

O primeiro modelo de regressão espacial foi o modelo de auto regressão espacial (Autoregressive Spatial Model -SAR). A técnica consiste em utilizar as médias das variáveis dependentes dos vizinhos de cada observação como componentes da regressão a ser calculada. Isso é feito por meio da inclusão na expressão da regressão linear de um componente espacial:

$$Y_n = X_n\beta + \lambda W_n Y_n + \varepsilon_n$$

em que W_n é a matriz de vizinhanças. O efeito da inclusão do termo de "lag" espacial é a modelagem da correlação espacial do termo dependente de forma que ela não afete as outras variáveis do modelo.

No presente trabalho foram calculados modelos SAR usando o software GEODA para as duas medidas, incluindo todas as variáveis da tabela 1; as variáveis não significativas foram sendo retiradas uma a uma, até que um modelo significativo fosse encontrado. As variáveis foram então novamente incluídas, uma por vez, e mantidas quando sua significância fosse considerável ($p < 0,05$). Por último a colinearidade dos dados foi analisada, usando as correlações entre as variáveis restantes para retira-las do modelo. Caso necessário, mais variáveis foram retiradas para manter o critério de significância de $p < 0,05$.

O segundo modelo de regressão espacial usado foi o modelo de regressão geograficamente ponderada (Geographically Weighted Regression – GWR). Esse modelo, ao invés de capturar a variação espacial em um termo da regressão, permite que os coeficientes das variáveis da regressão variem no espaço, calculando um modelo para cada unidade espacial analisada (os distritos, no caso). Isso é feito a partir de uma "janela móvel": o algoritmo do modelo percorre cada ponto de análise e calcula um modelo de regressão utilizando as outras observações que estão dentro de uma janela definida por uma dada distância (que pode ser um valor fixo, ou dado por um fator variável, como o número de vizinhos mais próximo); a contribuição de cada observação é também ponderada por uma função de distância que favorece mais as observações mais próximas. Isso é feito para todas as unidades de análise, o que resulta em um conjunto de modelos regressivos com coeficientes variando no espaço.

Para o cálculo do modelo foi utilizado o pacote *GWmodel* do software R. Foram feitos modelos para as duas medidas utilizando todas as variáveis selecionadas. A princípio, antes da calibragem, foi usado o critério da janela de seleção com os 20 vizinhos mais próximos; e a função distância selecionada para a ponderação foi uma gaussiana. Usando uma função de seleção stepwise (*model.selection.gwr()*) foram retiradas algumas variáveis do modelo e foi calibrada uma nova quantidade de vizinhos para a janela de seleção utilizando a função *bw.gwr().* Calculado o modelo (usando a função *gwr.basic()*) a partir das variáveis até essa etapa e usando o critério da janela calibrado, foram retiradas as variáveis colineares, estabelecendo como critério VIFs menores que dez para as variáveis dos modelos de todos os distritos. A retirada das variáveis se deu na ordem das variáveis com a maior média (em

relação a todos os modelos) dos VIFs. Uma nova chamada de um processo stepwise foi feito e a seguir as variáveis foram sendo retirada a partir da significância global apresentada pelo modelo, até atingir o critério de significância desejado ($p< 0,05$)

Os modelos foram comparados e a discussão das variáveis capturadas foi realizada a partir dessas comparações.

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) 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, divididos pela área dos distritos	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., divididos pela área dos distritos	2010/2016
% de não brancos (pretos, pardos e Indígenas)	Dados do IBGE. Censo 2010	2010
Proporção de domicilios 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

4. Resultados

As figuras 3 e 4 representam a distribuição de frequências dos valores normalizados das diferenças de tempos e dos "Tempos Relativos". Os dois histogramas indicam uma distribuição razoavelmente próxima da distribuição T. Particularmente no caso de R_t a distribuição parece ser ligeiramente deslocada para a esquerda, mas foi assumido para fins de análise a normalidade dos dados.

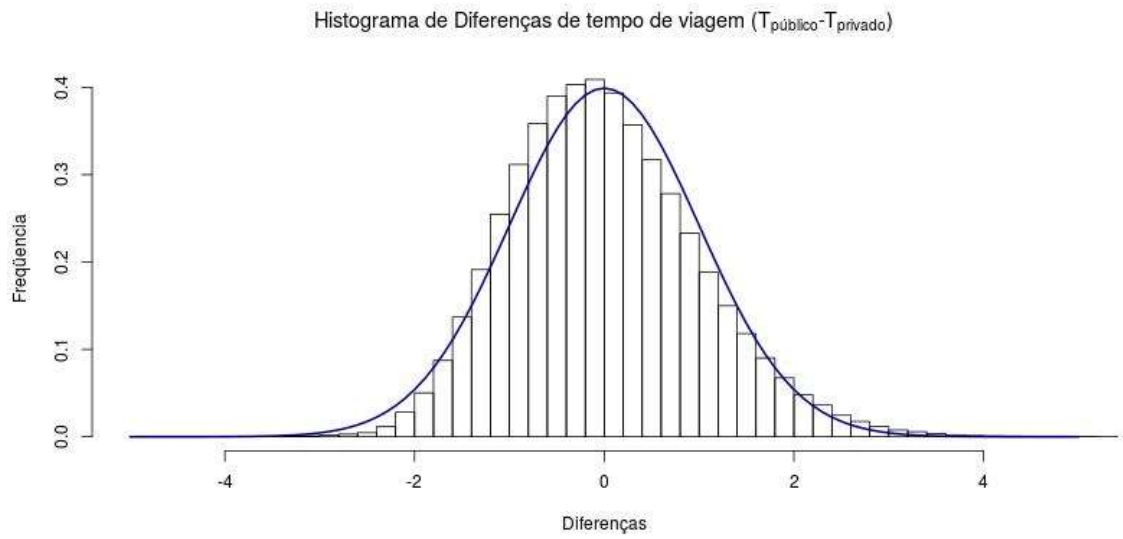


Figura 3: Histograma de D_t

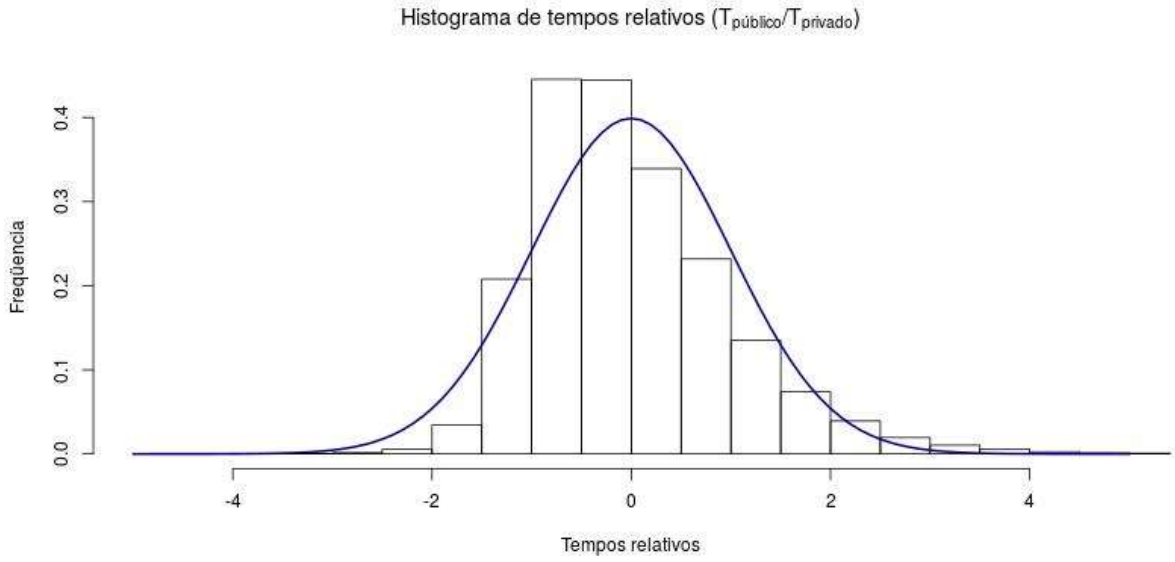


Figura 4: Histograma de R_t

Nos dois casos importante notar que as distribuições indicam uma diferença efetiva entre o tempo de transporte público e de transporte privado; para a D_t, utilizando um teste T de diferença de médias obtemos com 95% de confiança que a média da diferença está entre 3680s e 3697s; um teste T para o R_t indica com 95% de confiança que a média dessa medida está entre 2,396 e 2,404. Uma vez que as comparações são feitas entre viagens pareadas, o que essas medidas indicam, como esperado, é que 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.

Essa suspeita pôde ser averiguada ao analisar a relação dessas medidas com a distância das viagens. 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. D_t parece, de forma geral, crescer junto com as médias de distâncias de viagens; a correlação entre essas medidas, mesmo não sendo alta, é considerável: aproximadamente 0,564. A relação também é visível na comparação de distribuições na Figura 5. 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 algumas informações novas. A correlação é 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, a Figura 6 indica visualmente que os valores de distâncias maiores parecem tender 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. O comportamento que tende para a média é uma indicação dessa relação. 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 à trechos pedestres aumentam contribuem relativamente mais para o a razão dos tempos.

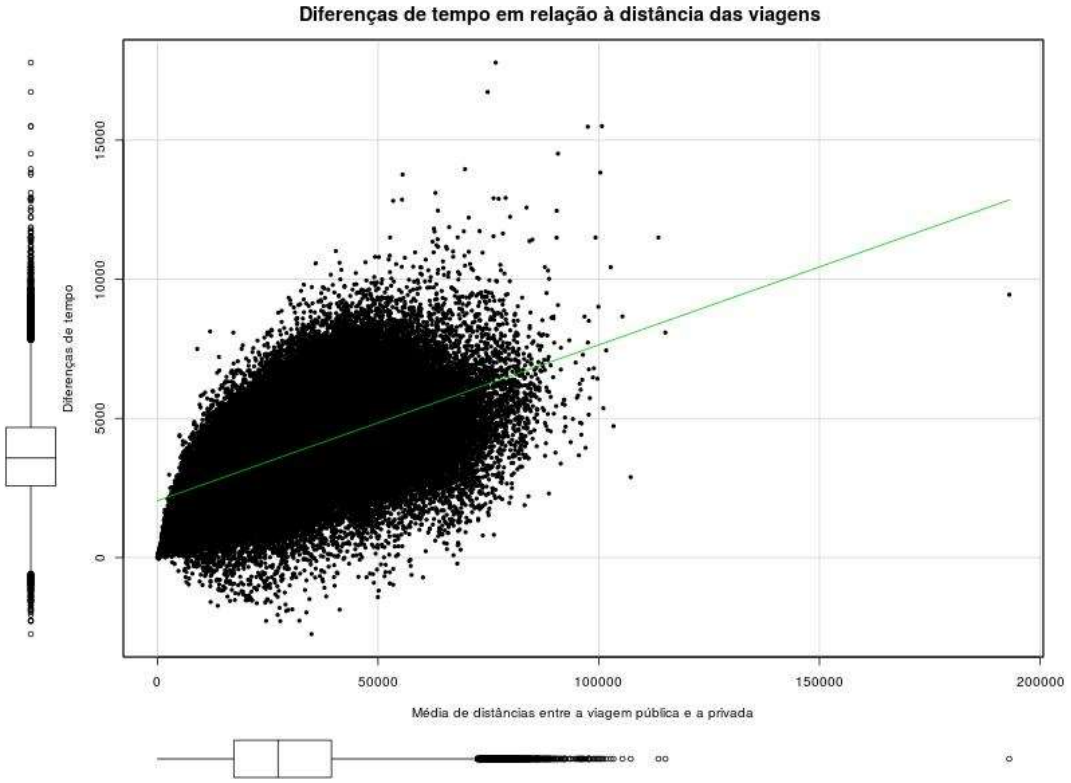


Figura 5: Gráfico de dispersão de D_t pela média de distâncias das viagens

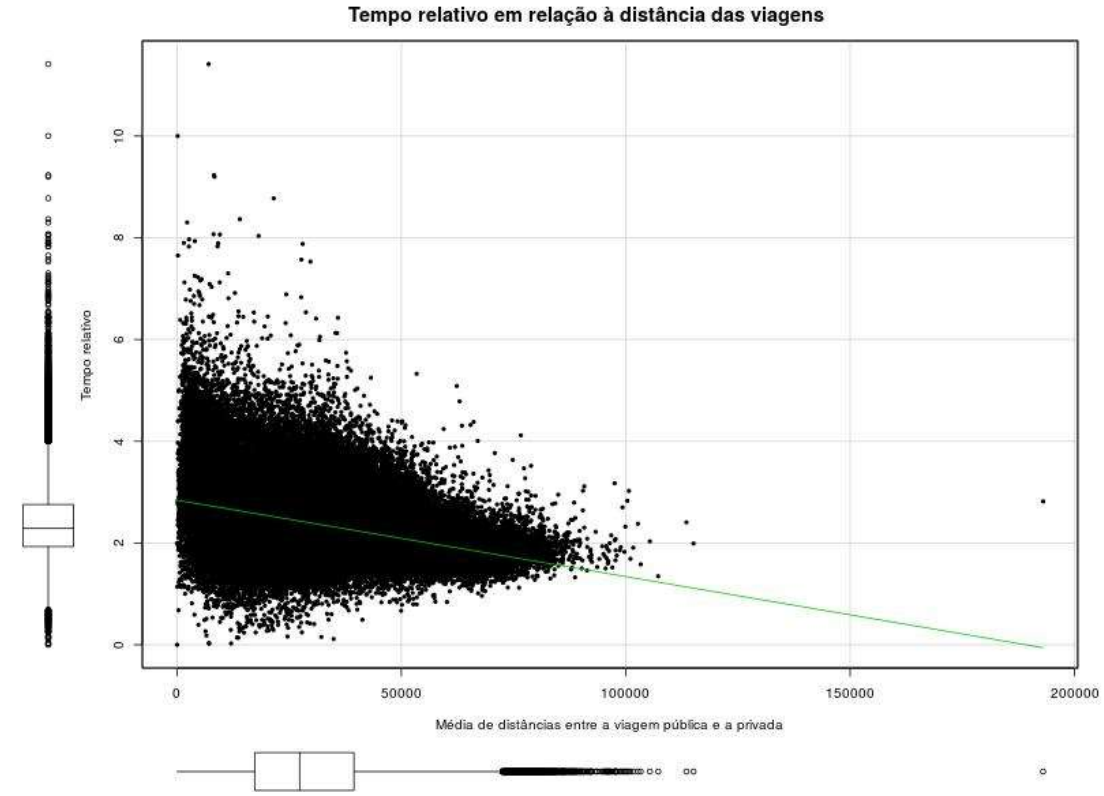


Figura 6: Gráfico de dispersão de R_t pela média de distâncias das viagens

Explorando outras possíveis dependências dos dados em relação aos dias da semana e aos horários, foram elaborados gráficos para explicitar a distribuição das viagens de acordo com essas variáveis. Enquanto a comparação dos boxplots dos dias não parece indicar nenhuma diferença significativa, tanto para a diferença de média como para o tempo relativo, as medianas dos boxplots das horas parecem apresentar alguma variação regular; mas a dispersão apresentada pelos dados inviabiliza qualquer afirmação sobre um padrão regular.

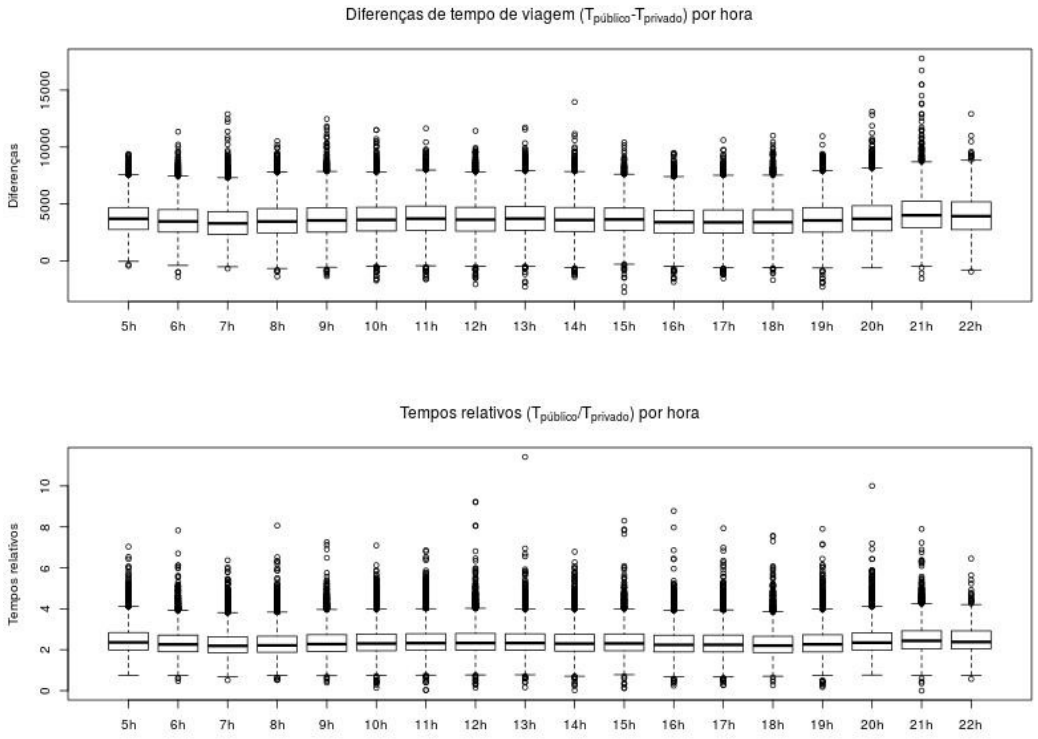


Figura 7: Boxplots de D_t e R_t em função das horas do dia

A análise exploratória espacial a partir das superfícies suavizadas foi feita tanto para os distritos como para as áreas de ponderação. O resultado foram superfícies suavizadas das medidas analisadas para todo o município; a qualidade dessas superfícies é diretamente relacionada a densidade de pontos nas proximidades das médias estimadas, o que implica que as superfícies calculadas para os distritos apresentam maior estabilidade das estimativas uma vez que foram calculadas a partir de um conjunto maior de pontos. A partir desses mapas foi possível identificar alguns padrões locais do comportamento das medidas que desaparecem na análise do conjunto global de dados. No exemplo de Tucuruvi é possível observar alguns

padrões particulares relativos à localização, como o acesso aos corredores de metro, visíveis nas faixas de valores baixos de D_t e R_t. Ao mesmo tempo a comparação das duas medidas permite entender melhor o comportamento das viagens. Em relação a zona leste é visível uma região mais clara no extremo leste (acompanhando os corredores de transporte público), mas há na zona que conecta ao centro um espaço de maiores valores de R_t que não se repetem claramente na superfície de D_t. Isso pode ser uma indicação de que, apesar da diferença de tempo dos modais se manter para viagens para essa região, o valor absoluto de tempo dos dois modais caiu em uma proporção semelhante.

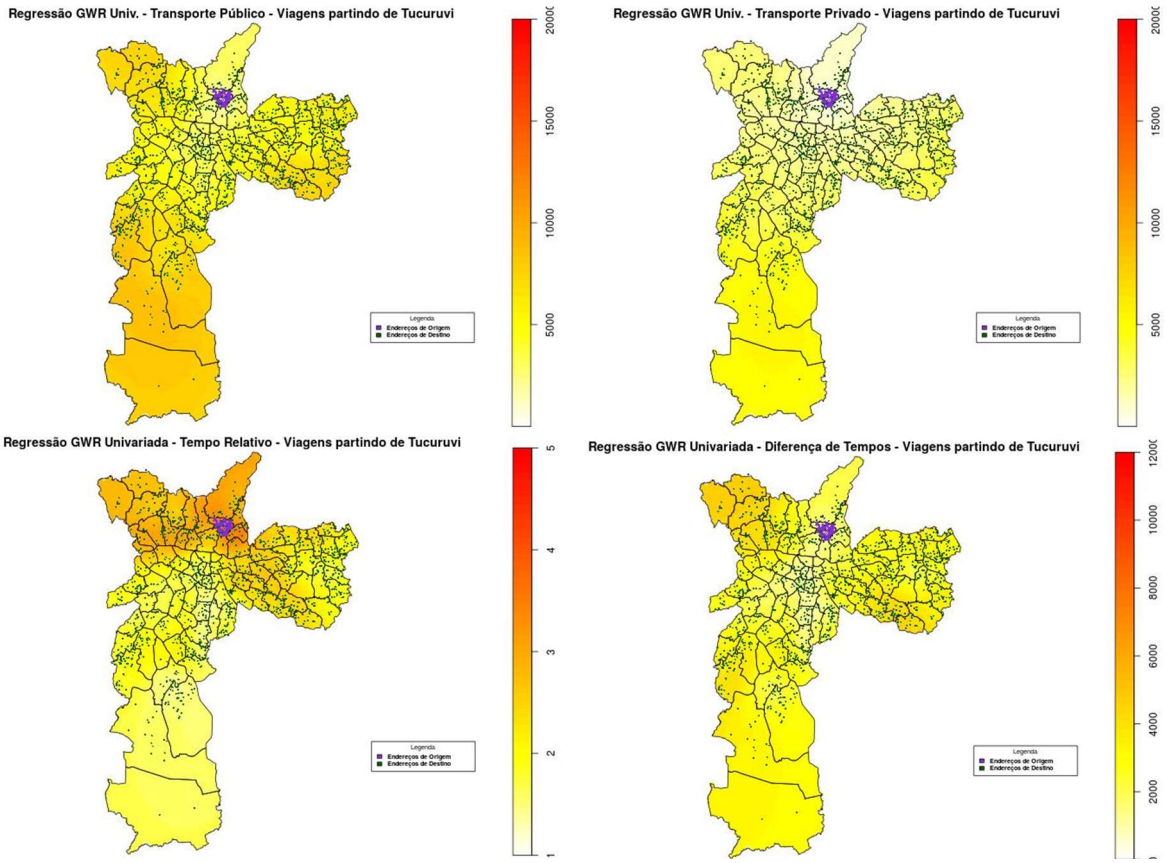


Figura 8: Superfícies suavizadas para T_{pub}, T_{priv}, D_t, e R_t com origem em Tucuruvi

Essa relação das medidas se dá pela própria forma como elas foram construídas: sendo ambas calculadas a partir dos tempos de viagem públicos e privados, quando conjugadas elas permitem extrair informações intuitivas dos valores absolutos dos tempo de viagens.

A próxima 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 partir dessa agregação foram calculados a partir do software Geoda os I's de Moran (Figura 9, 10, 11 e 12) e os mapas de associação espacial.

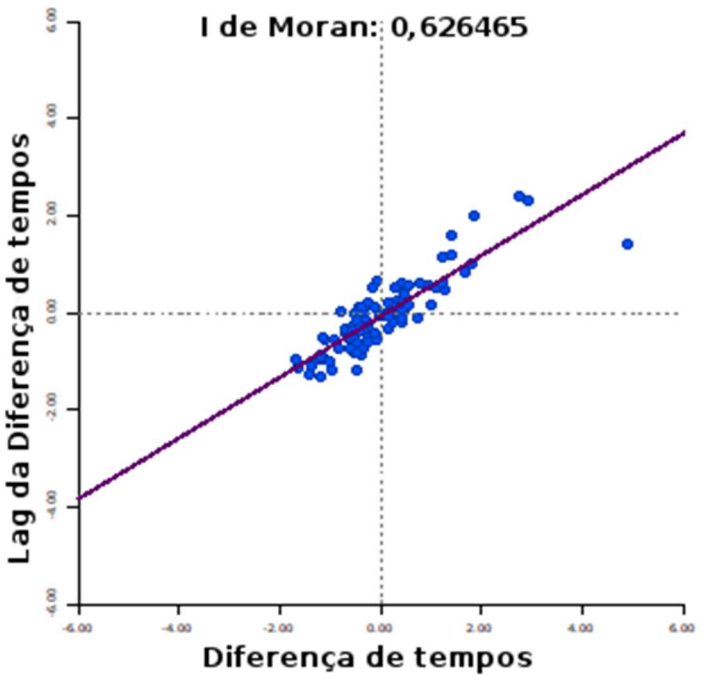


Figura 9: I's de Moran para D_t centrado nos distritos de origem

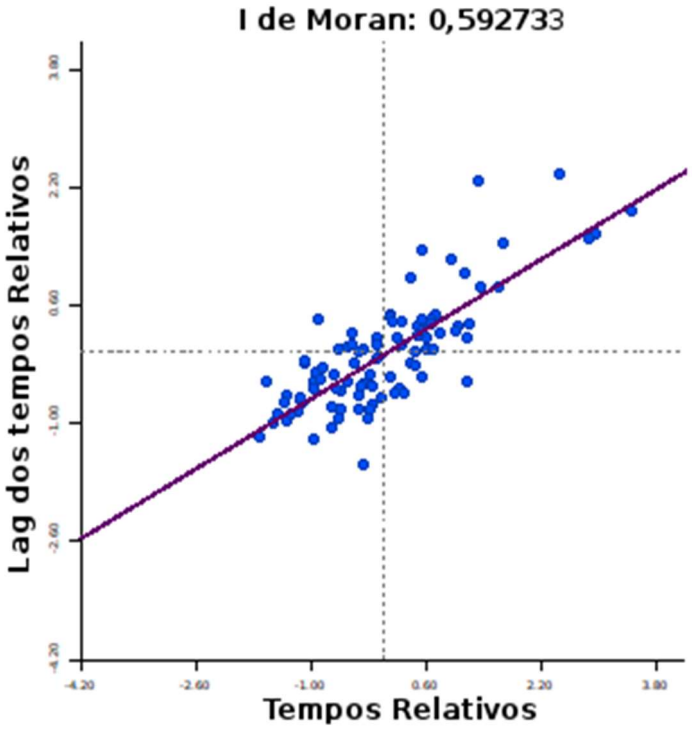


Figura 10: I's de Moran para R_t centrado nos distritos de origem

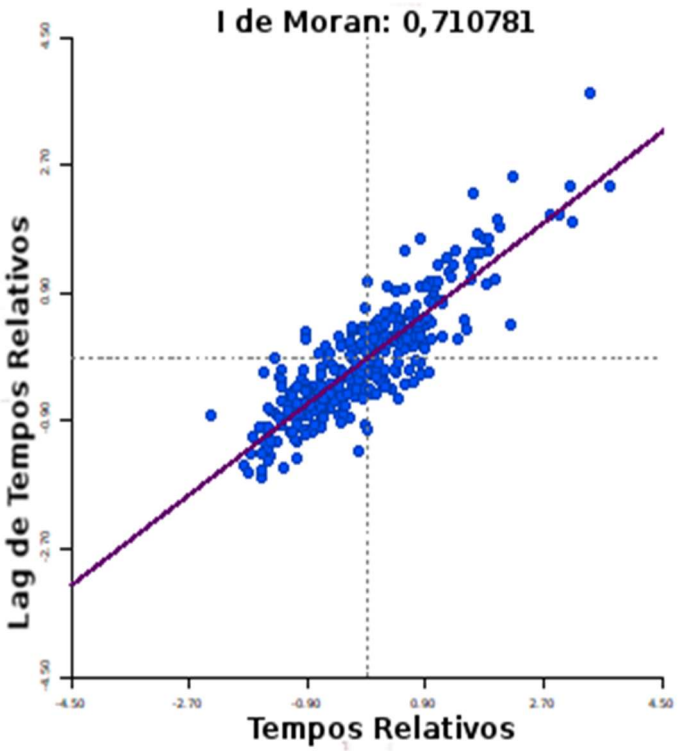


Figura 11: I's de Moran para R_t centrado nas áreas de ponderação

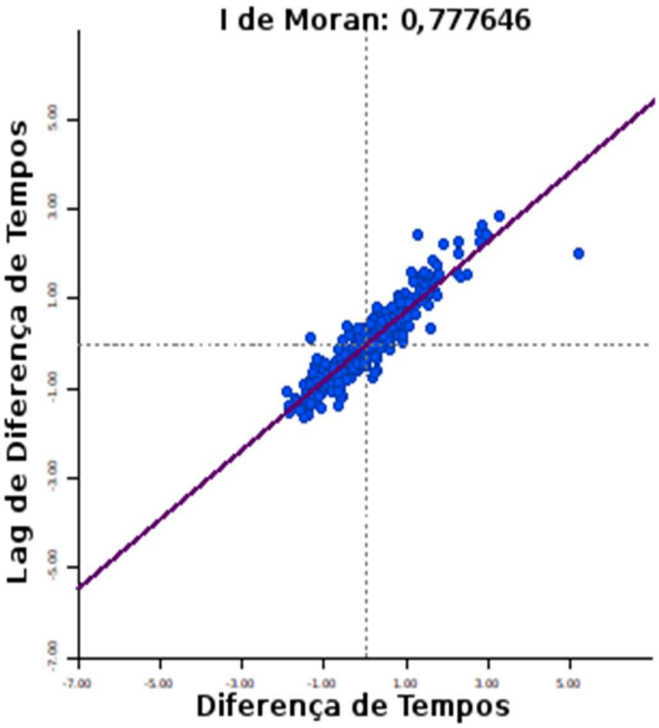


Figura 12: I's de Moran para D_t centrado nas áreas de ponderação

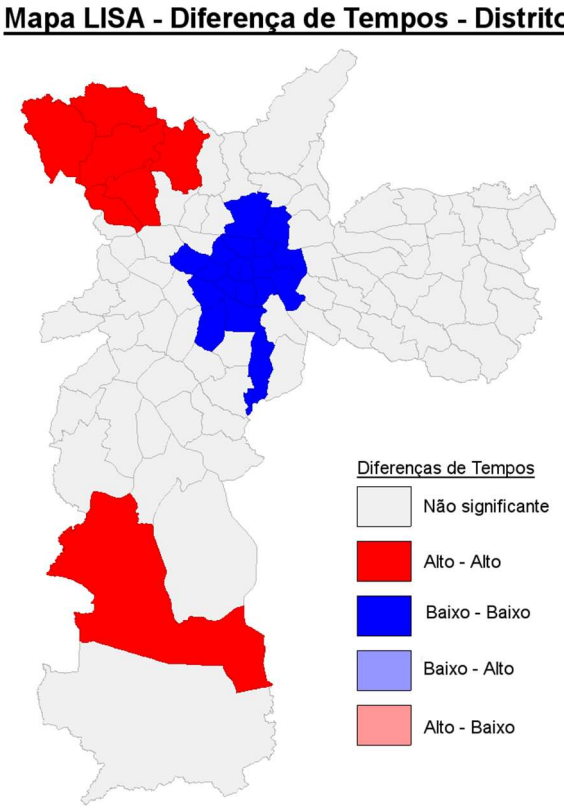


Figura 13: Mapa LISA para Dt centrado nos distritos

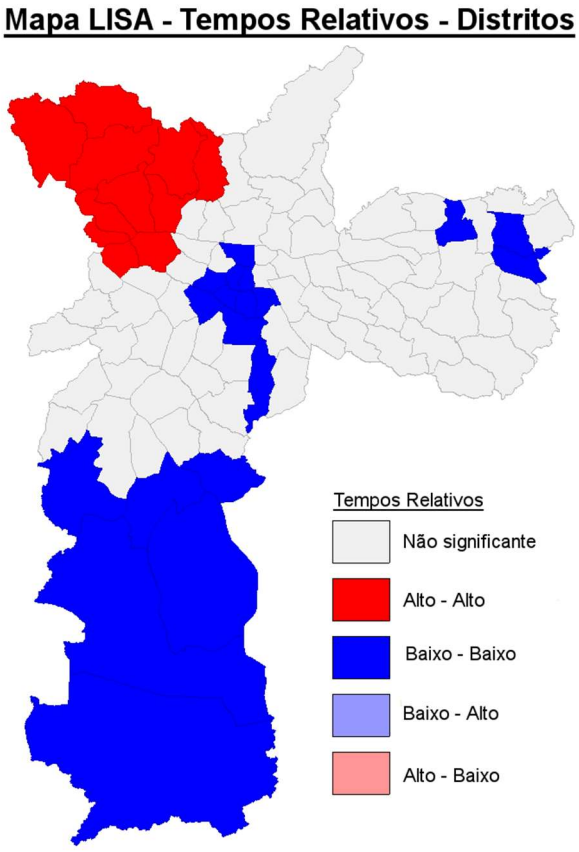


Figura 14: Mapa LISA para Rt centrado nos distritos

Os valores dos I's de Moran Globais já indicam a presença de associação espacial significativa entre as medidas dos distritos e das áreas de ponderação, com valores consistentemente acima de 0,5. 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.

Os gráficos dos I's de Moran Locais ressaltam a indicação de que há a associação espacial para as duas medidas nos dois níveis de análise. A diferença no I de Moran global entre as medidas pode ser interpretada pela diferente distribuição das observações nos quadrantes do gráfico dos I's de Moran locais. 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 presença dessas observações reduz o valor dos I's globais.

A análise dos mapas de associação espacial (LISA) feitos a partir da mesma análise derivada dos I's de Moran permite visualizar melhor as relações de clusterização espacial das zonas analisadas. No caso dos distritos 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 dos tempos relativos ser menor, há uma região comum de baixos valores para ambas as medidas.

Duas diferenças importantes são a presença de um cluster de baixos tempos relativos na zona leste e na zona sul – particularmente nesse segundo caso o mapa de diferenças indica justamente um cluster de altos valores. Ambas diferenças podem ser explicadas pelo comportamento já discutido das duas medidas: D_t tende a aumentar com o aumento das distancias das viagens enquanto R_t diminui. Ambas as regiões, por estarem relativamente longe do agregado das zonas dos municípios, apresentam uma proporção maior de viagens longas. A figura 15 mostra o a distribuição da média de distâncias das viagens pareadas (média da distância da viagem por transporte público e da viagem por transporte privado) por área de ponderação, ilustrando esse fenômeno. Parte disso se deu pela forma com que os dados foram simulados, privilegiando endereços em regiões mais populosas; mas parte também se dá pela organização das vizinhanças, com zonas centrais sendo mais próximas, na média, de todos os distritos.

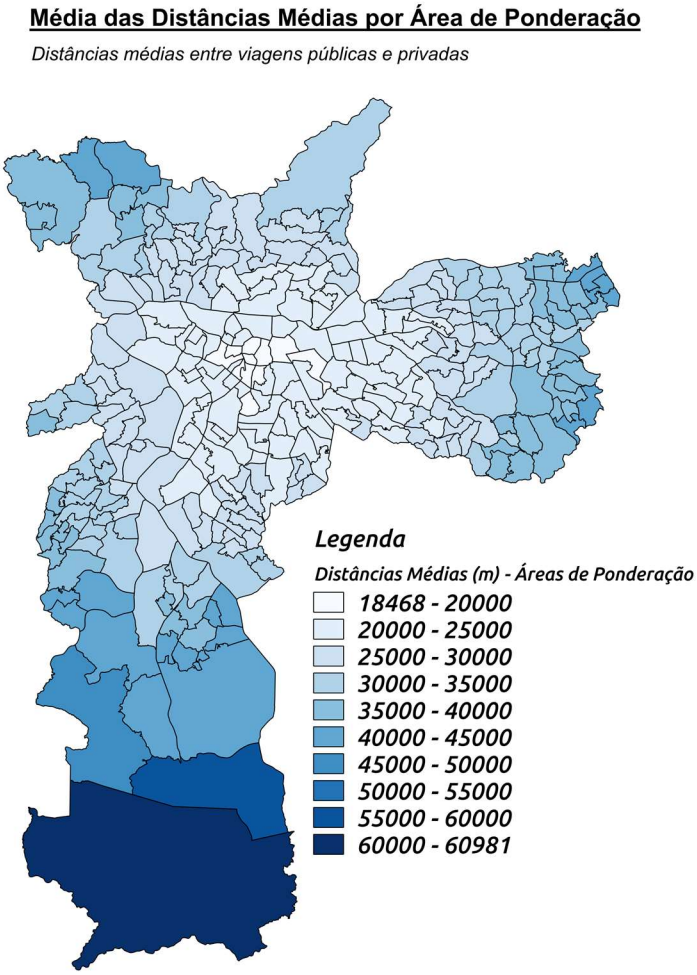


Figura 15: Distribuição das médias de distâncias de viagens pelas áreas de ponderação

Os mapas LISA para as áreas de ponderação (figuras 16 e 17) apresentam um grau maior de granularidade e, como indicam os I's de Moran globais e os gráficos de I's de Moran locais, apresentam mais núcleos de clusters que os distritos.

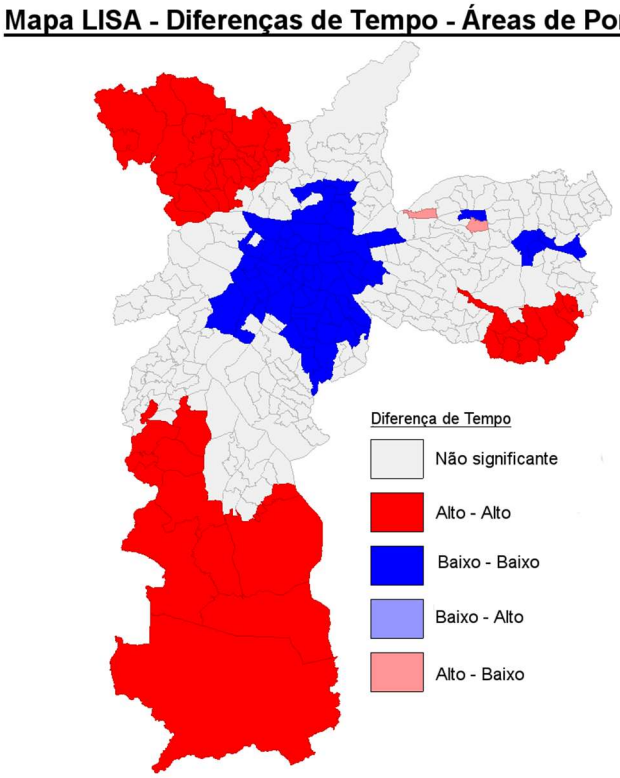


Figura 16: Mapa LISA para D_t centrado nas áreas de ponderação

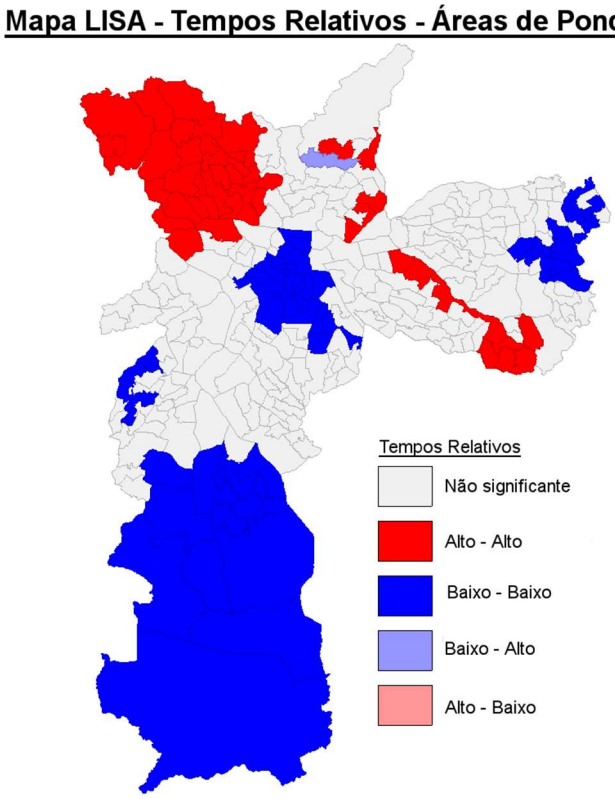


Figura 17: Mapa LISA para R_t centrado nas áreas de ponderação

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 baixo valores à sudoeste e um núcleo delgado de valores altos que se estende do sul da zona leste até a zona norte entre os núcleos de valores baixos no centro e na 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. Também aparece nos dois mapas um núcleo de baixos valores no extremo leste (já identificado no mapa de R_t dos distritos).

Uma explicação para a diferença do número de clusters entre os níveis de análise é que como os distritos apresentam uma área maior, há uma compensação no interior deles entre tendências diferentes das medidas. Dois exemplos podem ser dados: na comparação entre os mapas de D_t, o pequeno cluster de valores baixos à sudoeste presente no mapa de áreas de ponderação ocupa partes de dois distritos que não apresentam clusterização significativa. O núcleo do cluster de altos valores no sul da zona leste no mapa de áreas de ponderação engloba regiões de 4 diferentes distritos – o que indica que a dinâmica de clusterização, por se dar dentro dos distritos e pelas regiões fronteiriças deles, foi ocultada pelo nível de agregação distrital. Esse é um exemplo prático do MAUP (Problema da unidade de área modificável), e se as unidades menores não o resolvem, elas aumentam a granularidade dos dados e diminuem a dimensão espacial do erro.

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 (figuras 18 e 19) encoraja a idéia 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 medias apresenta uma linha de metro próxima, com a exceção do cluster delgado na figura KK. 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.

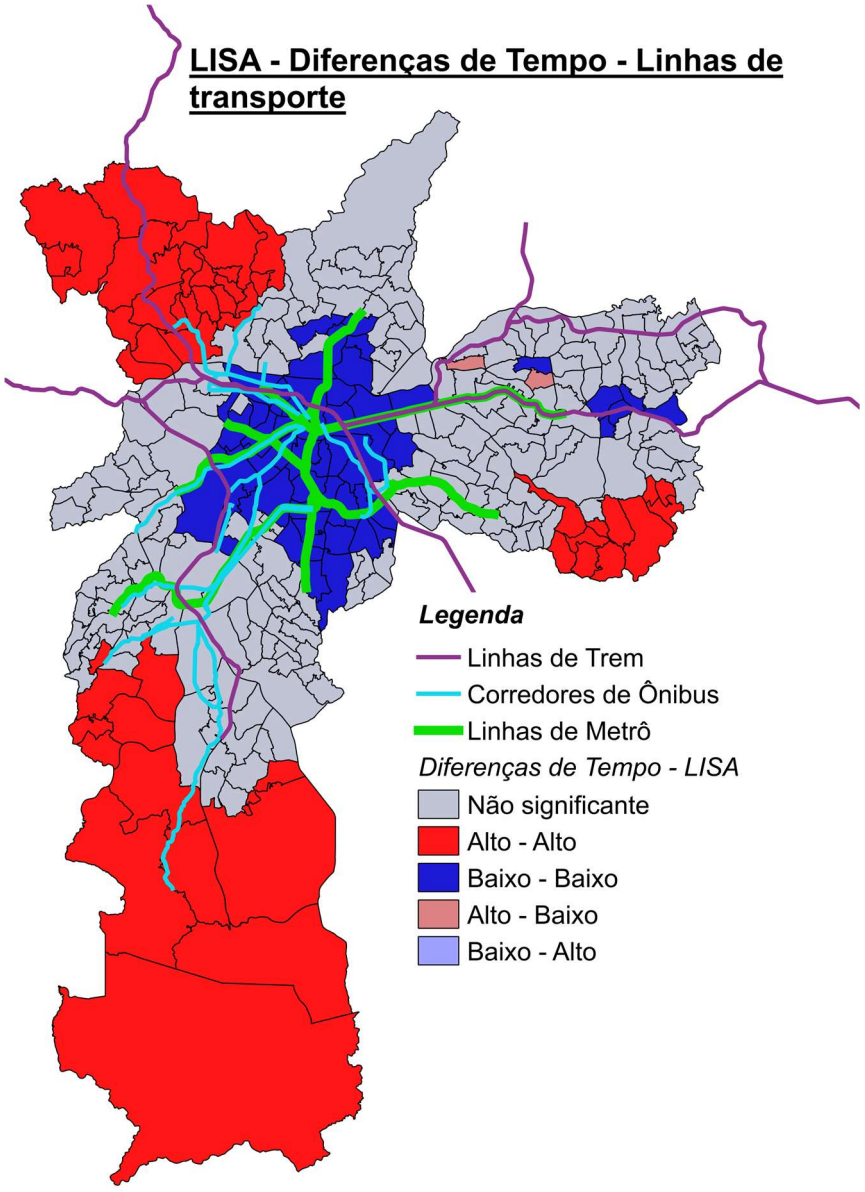


Figura 18: Mapa LISA para D_t centrado nas áreas de ponderação com as redes de transporte

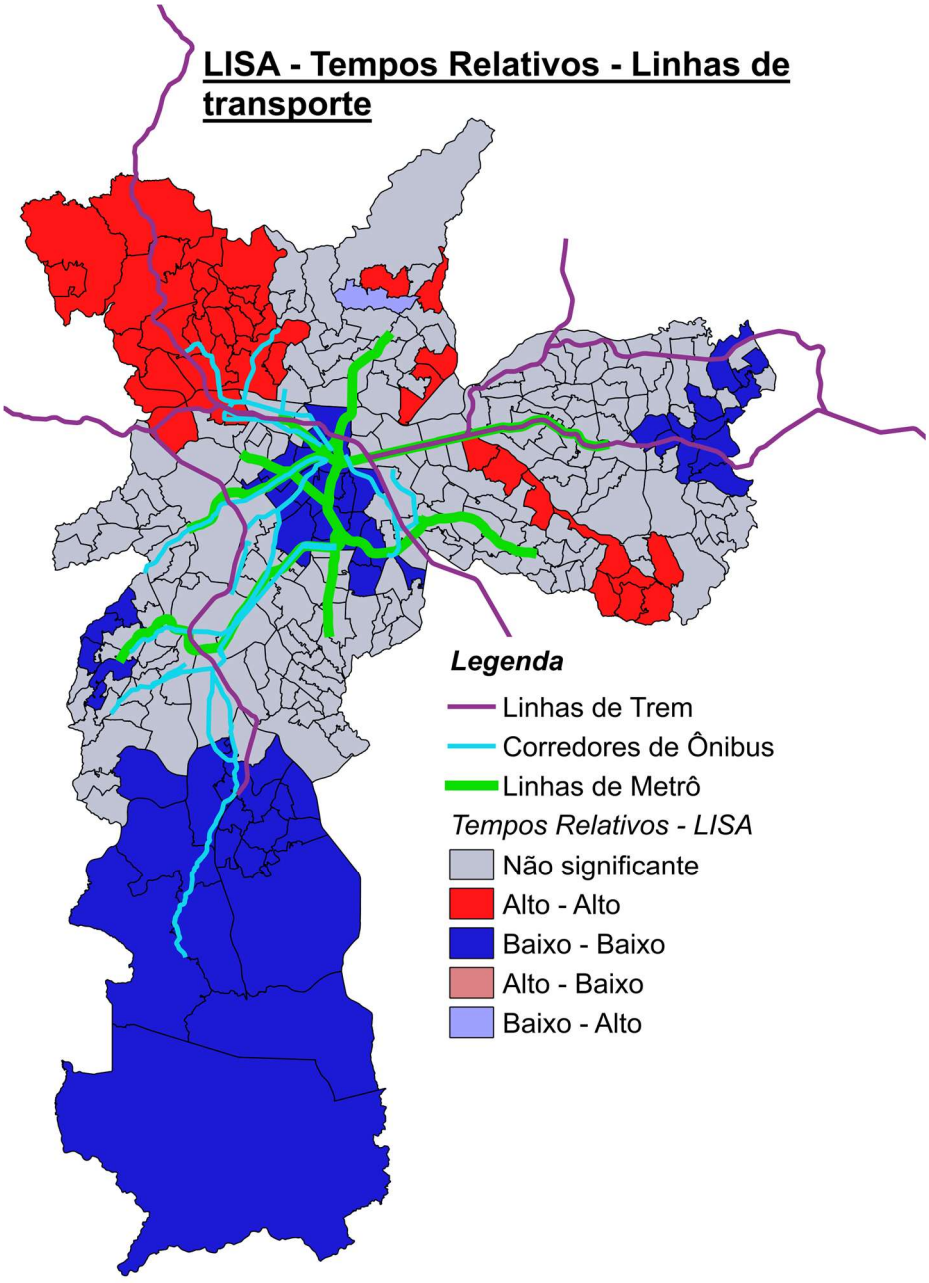


Figura 19: Mapa LISA para R_t centrado nas áreas de ponderação com as redes de transporte

Modelagem dos dados

A primeira modelagem dos dados feita foi uma OLS simples, a partir das variáveis da Tabela 2 para as duas medidas. Os resultados das duas regressões estão reproduzidos no anexo (Anexos II e III). 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. Além disso os resíduos também são espacialmente dependentes (figuras K e z), indicando a dificuldade dos modelos de lidar com a auto correlação espacial, mostrada na seção anterior.

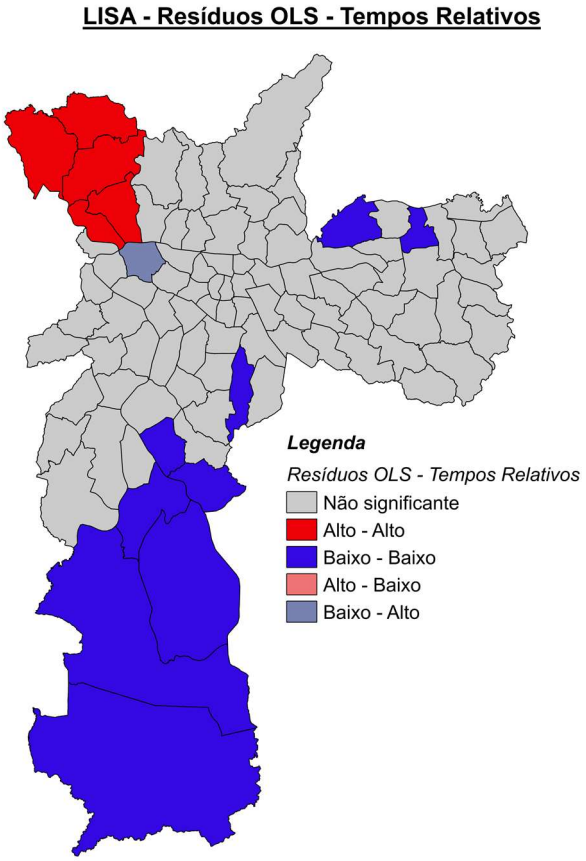


Figura 20: Mapa LISA para resíduos do modelo OLS de R_t

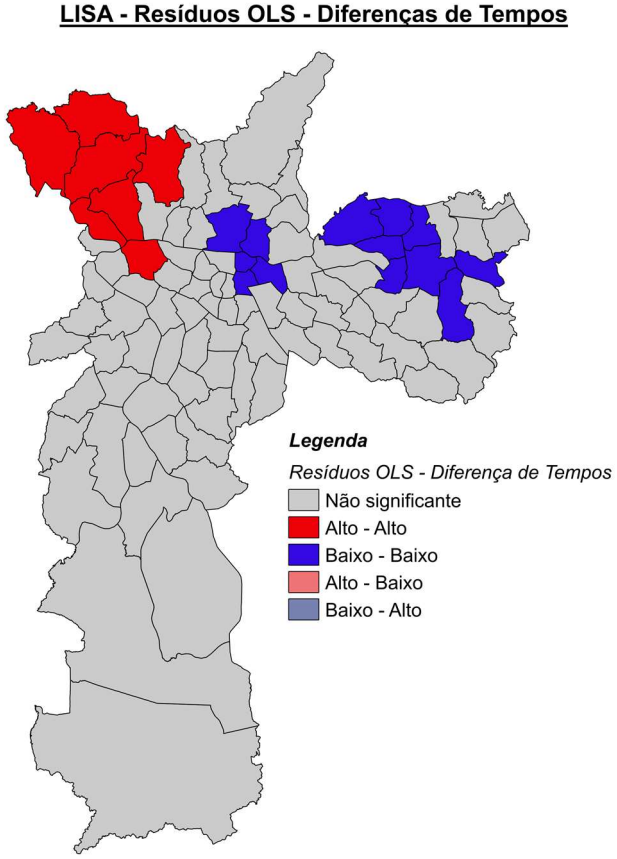


Figura 21: Mapa LISA para resíduos do modelo OLS de D_t

As variaveis selecionadas para as modelagens SAR e GWR para as duas medidas estão representadas na Tabela 3

Tabela 3: Variáveis selecionadas nos modelos SAR e GWR

D_t - SAR	R_t - SAR	D_t - GWR	R_t - GWR
DUMCPTM DENLINBUS PNBRAN2010 DENPOP2018	DUMCPTM DUMMETRO PDOMM2010	DUMCPTM DUMMETRO PNBRAN2010 DENPOP2018 DENPTBUS	DUMCPTM DUMMETRO PDOMM2010 DENPTBUS RENDP2010 DENEMPREG

Os modelos SAR foram calculados para as duas medidas e seus resultados estão no anexo (Anexo IV e V). Para ambas as medidas o R^2 aumentou seu valor – de 0.48947 para 0.760621 em D_t e de 0.304930 para 0.669616 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 (figuras 22 e 23)

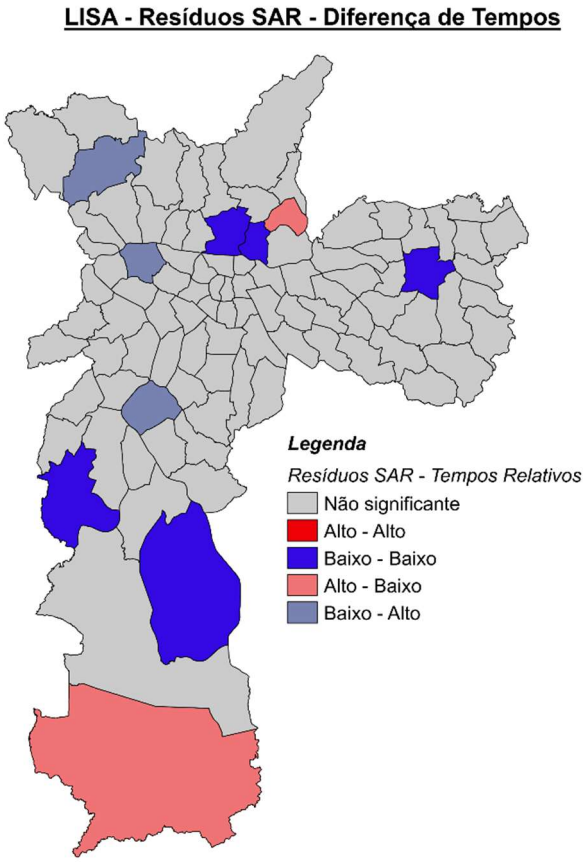


Figura 22: Mapa LISA para resíduos do modelo SAR de D_t

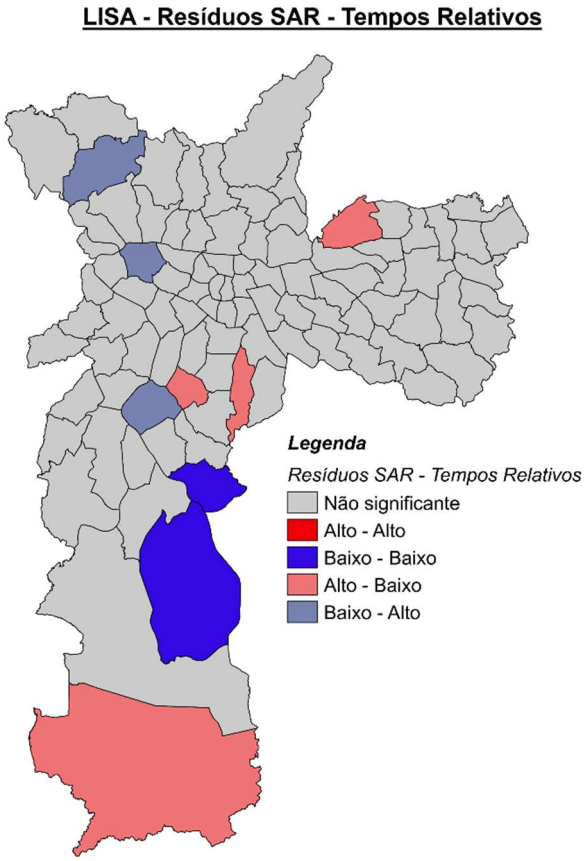


Figura 23: Mapa LISA para resíduos do modelo SAR de R_t

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

Variable	Coefficient	Std.Error	z-value	Probability
W_Tempo_dif	0.756164	0.0638933	11.8348	0.00000
CONSTANT	1013.64	265.464	3.81836	0.00013
DENLINBUS	-5.87178	2.79589	-2.10015	0.03572
PNBRAN2010	1183.8	348.151	3.40026	0.00067
DUMCPTM	-260.749	92.4829	-2.81943	0.00481
DENPOP2018	-0.0246901	0.00937455	-2.63374	0.00845

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	0.745058	0.0671267	11.0993	0.00000
CONSTANT	0.529982	0.172999	3.0635	0.00219
PDOMM2010	0.341544	0.115849	2.94817	0.00320
DUMMETRO	-0.146594	0.0333705	-4.39294	0.00001
DUMCPTM	-0.0638153	0.0304845	-2.09337	0.03632

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 (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. Dt, como mostrado, 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 (Figuras 5 e 6). 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 originadas nas 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 ao 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.

A modelagem GWR apresenta um conjunto de figuras para sua descrição. Cada modelo GWR apresenta um conjunto de regressões lineares para cada um dos distritos. As saídas do programa de cálculo da função estão no anexo (Anexo VI eVII). A análise das variáveis desse modelo é mais complexa que a do modelo SAR, uma vez que cada variável tem um valor diferente em cada distrito. A partir desse modelo é possível explorar a relação de variáveis não estacionárias com a variável dependente, variando a seus coeficientes e sua significância pela superfície de análise. Os R^2 dos dois modelos são maiores que os dos modelos OLS simples, mais menores que dos modelos SAR: 0,704 para D_t e 0,626 para R_t..

Para D_t as Figuras 24 a 34 mostram os valores dos coeficientes e a os valores T deles em cada distrito, além do R^2 em cada distrito. A significância é dada pelo valor de um teste T para cada distrito e está representada no mapa de TV (T-values). Uma aproximação de valores de T que tornam os coeficientes significativos ($p < 0,05$) para os nossos modelos, considerando o tamanho da amostra usada para o cálculo de cada modelo distrital, deve ser de 2,042 (valor para 30 graus de liberdade – as bandas de vizinhos incluem 31 (D_t) e 35 (R_t) vizinhos).

Uma das características do modelo é a possibilidade de variação dos coeficientes da regressão no espaço. Isso também implica que certas variáveis explicativas podem ser significantes em algumas regiões e não em outras. Na figura 25, vemos que a densidade populacional apresenta valores T maiores que o limite de significância (onde $p < 0,05$) principalmente na zona sul, onde a densidade está associada à redução da diferença de tempos. No resto do município a variável não apresenta significância considerável; a troca de sinais do coeficiente da variável na zona leste é também representativa da perda de significância da variável, indicando que a relação que existe na zona sul não está presente na zona leste.

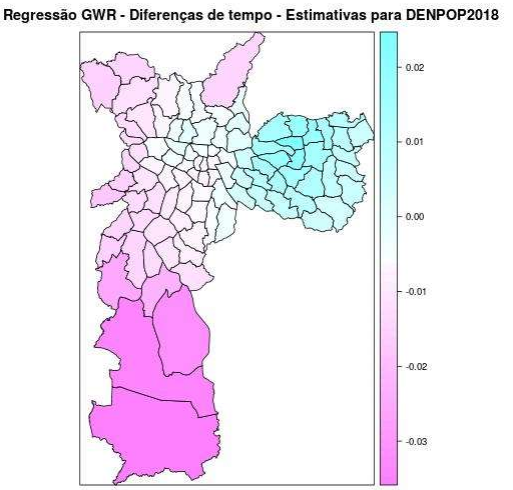


Figura 24: Coeficientes de densidade populacional do modelo GWR para D_t

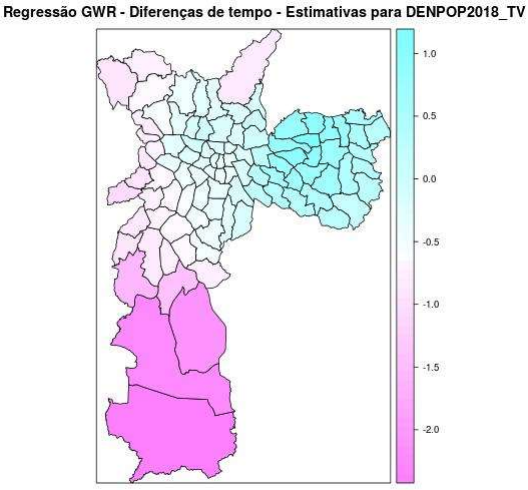


Figura 25: Valores T para densidade populacional do modelo GWR para D_t

A densidade de pontos de ônibus, apresentada na figura 26, indica que a variável é significativa em quase todo o município, mas não no centro de São Paulo. A variável está associada sempre a redução de D_t e tem o valor absoluto de seu coeficiente maior nas regiões mais afastadas das periferias e menor na área do centro expandido e nas regiões mais próximas das periferias.

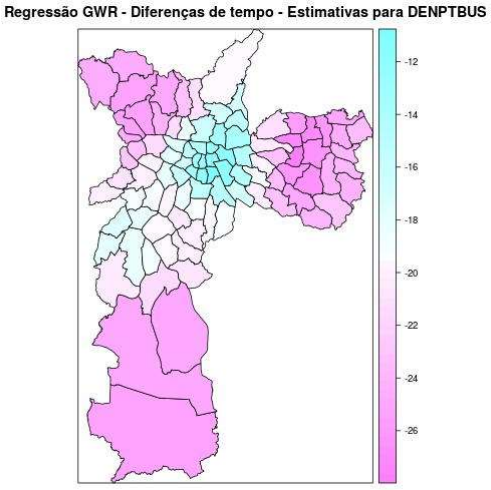


Figura 26: Coeficientes para densidade de pontos de ônibus do modelo GWR para D_t

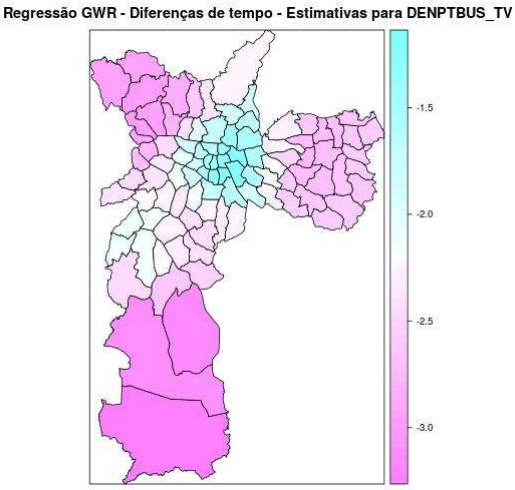


Figura 27: Valores T para densidade de pontos de ônibus do modelo GWR para D_t

O acesso à CPTM (Figuras 28 e 29) é significante principalmente na zona leste, com a zona sul apresentando Marsilac dentro da margem de significância. Onde é significante, a presença de estações da CPTM está associada a redução de D_t. Vale ressaltar que como a variável DUMCPTM é uma dummy de acesso, o fato de ela ser significante em distritos onde não há

acesso de CPTM (zona sul da zona leste) indica que a ausência de CPTM está associada a esse aumento de D_t na região.

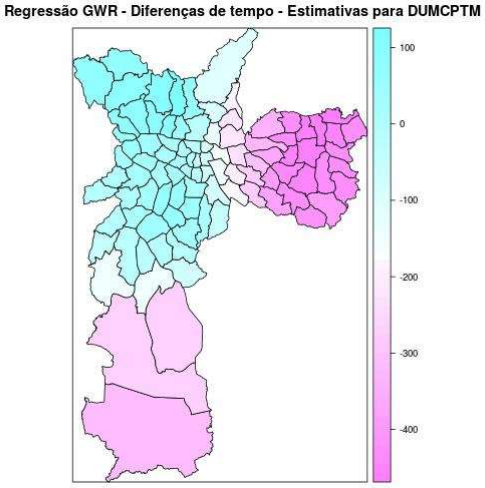


Figura 28: Coeficientes de acesso à CPTM do modelo GWR para D_t

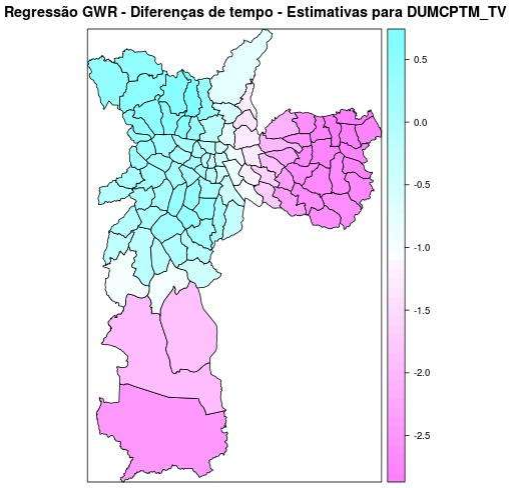


Figura 29: Valores T para acesso à CPTM do modelo GWR para D_t

A variável de presença do Metrô apresenta significância na zona norte e em boa parte do centro (Figuras 30 e 31). O sinal da contribuição é negativo, com a presença de estações próximas reduzindo D_t. O valor absoluto do coeficiente é maior na zona norte, com contribuições mais moderadas na região central. O mesmo comentário sobre a variável dummy DUMCPTM vale para a DUMMETRO, principalmente em sua região de maior significância, a zona norte: a ausência de acesso na região é associada ao aumento de D_t.

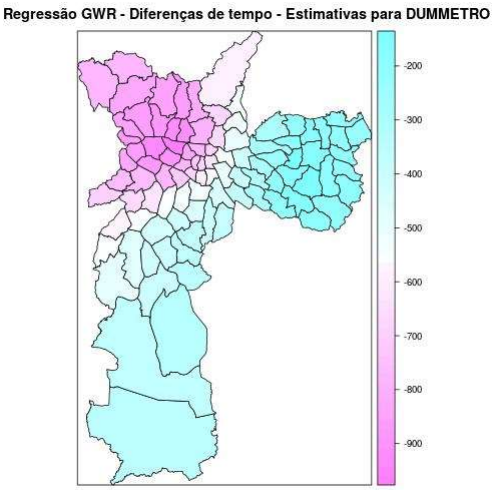


Figura 30: Coeficientes de acesso ao Metrô do modelo GWR para D_t

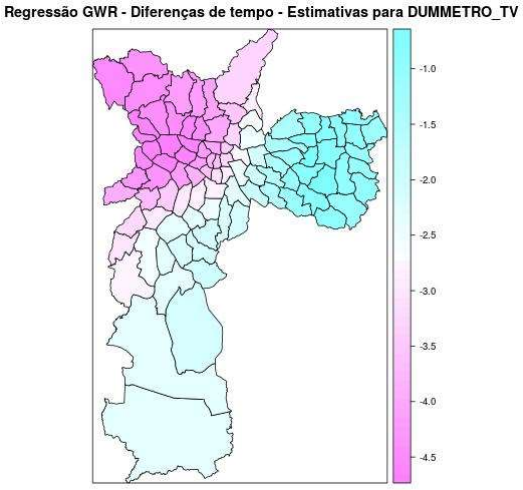


Figura 31: Valores T para acesso ao Metrô do modelo GWR para D_t

A variável de proporção da população não branca (Figuras 32 e 33) não apresenta significância no centro (onde também seus valores são baixos) e na parte mais próxima da zona leste, mas é significante para o resto do município. Seu coeficiente, onde é significante, é positivo – há uma associação positiva entre D_t e a proporção de não brancos nos distritos.

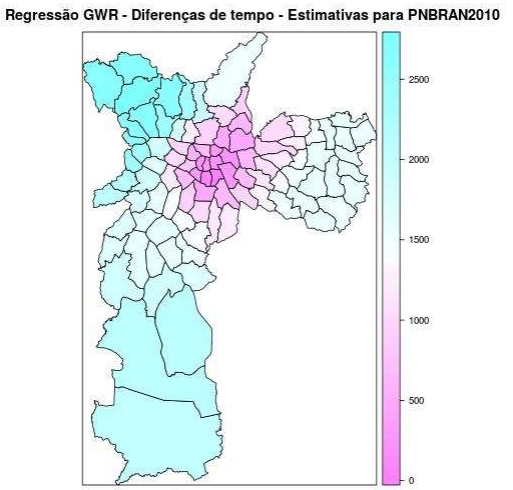


Figura 32: Coeficientes para a proporção de não brancos do modelo GWR para D_t

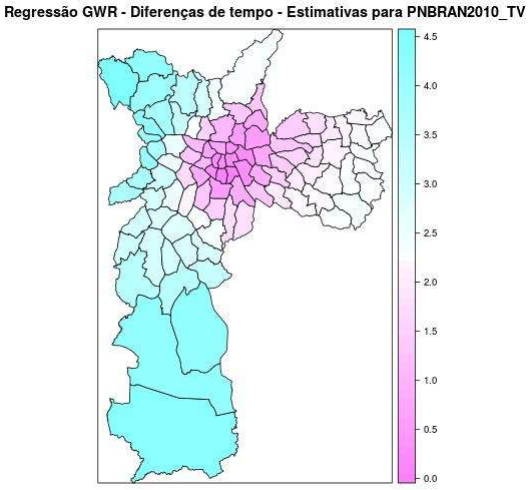


Figura 33: Valores T para proporção de não brancos do modelo GWR para D_t

A Figura 34 indica o desempenho em termos de R^2 para os modelos lineares de cada distrito. O modelo foi muito mais capaz de captar a variância dos dados nas zonas centrais e na zona norte, além do começo da zona sul. Para a zona leste e a zona sul o modelo é significativamente menos capaz. Isso indica que faltam variáveis relevantes para a análise, que poderiam ser significantes nessas zonas.

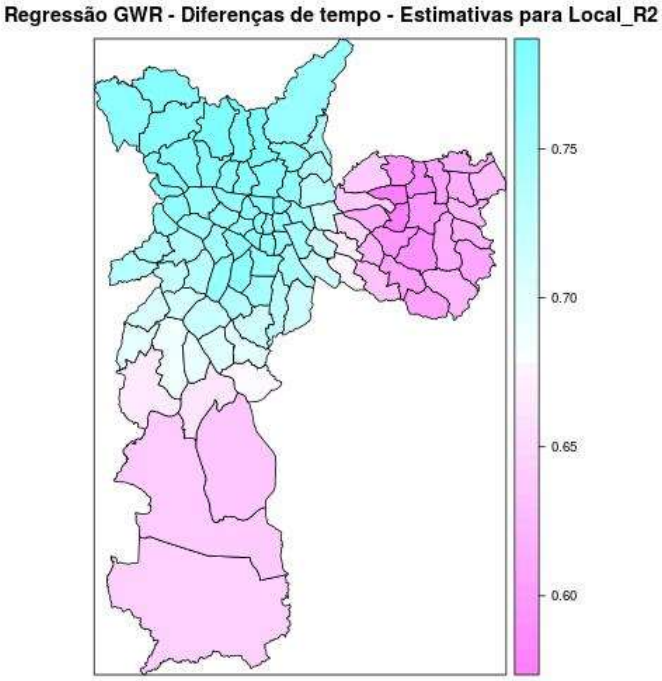


Figura 34: Valores de R^2 do modelo GWR para D_t

Para R_t as figuras 35 a 47 mostram os valores dos coeficientes e a significância deles em cada distrito, além do R^2 em cada distrito.

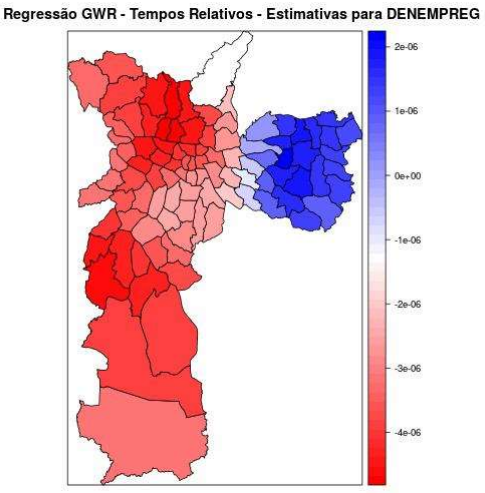


Figura 35: Coeficientes para a densidade de empregos do modelo GWR para R_t

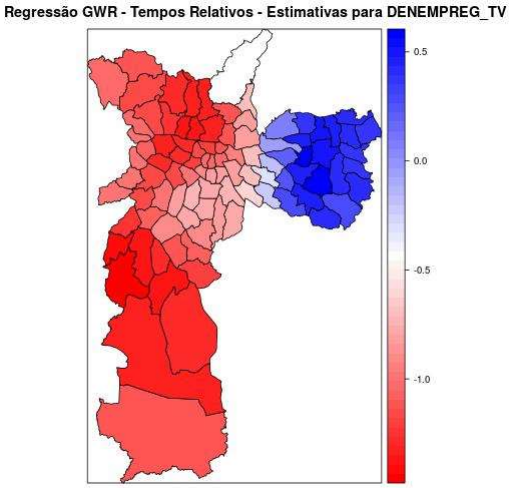


Figura 36: Valores T para a densidade de empregos do modelo GWR para R_t

A variável de densidade de empregos (Figuras 35 e 36) é pouco significante para quase todo o município de São Paulo. As regiões onde seus valores t estão próximos do limiar da confiança são parte da região sul e a região ao norte do centro, onde os valores de seu coeficiente indicam que a densidade de empregos está associada a menores R_t.

A densidade de pontos de ônibus (Figuras 37 e 38), assim como a densidade de empregos, apresenta pouca significância local nos distritos. A área onde ela mais se aproxima da significância é na zona Sudoeste, onde a maior densidade de pontos de ônibus está associada a maiores R_t.

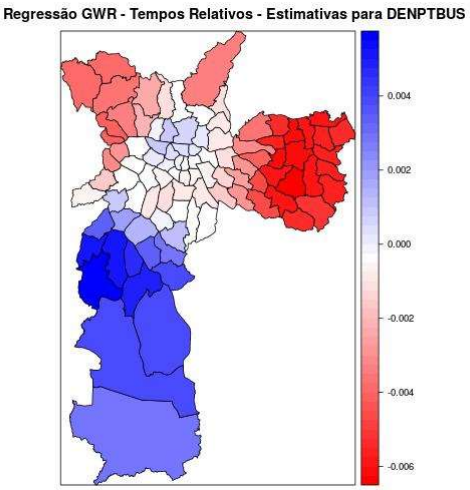


Figura 37: Coeficientes para a densidade de pontos de ônibus do modelo GWR para R_t

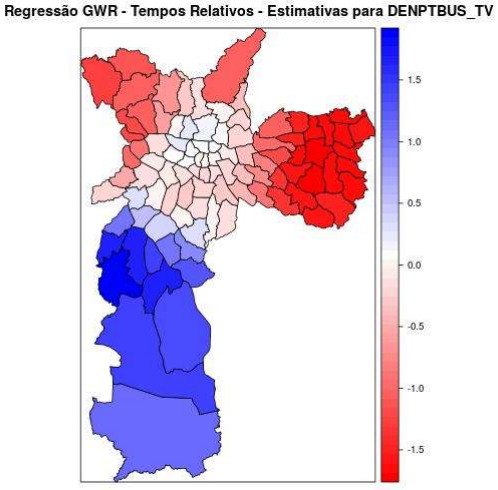


Figura 38: Valores T para a densidade de pontos de ônibus do modelo GWR para R_t

A variável de acesso à CPTM (Figuras 39 e 40) apresentou significância para a zona leste, onde ela está associada à redução dos R_t. Essa relação está de acordo com o modelo de D_t, com exceção da zona sul, que naquele modelo é significante. Essa diferença pode estar relacionada à já discutida dependência de D_t das distâncias de viagem, que afetam sobretudo a zona sul.

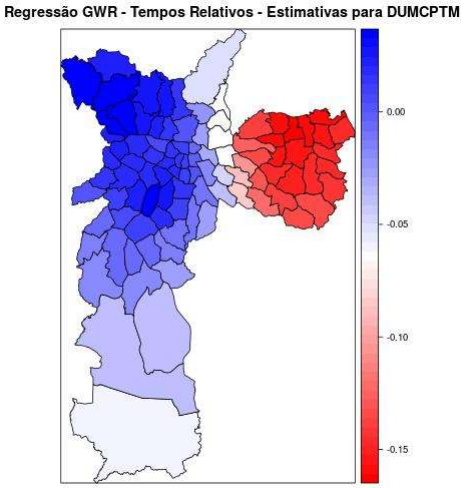


Figura 39: Coeficientes de acesso à CPTM do modelo GWR para R_t

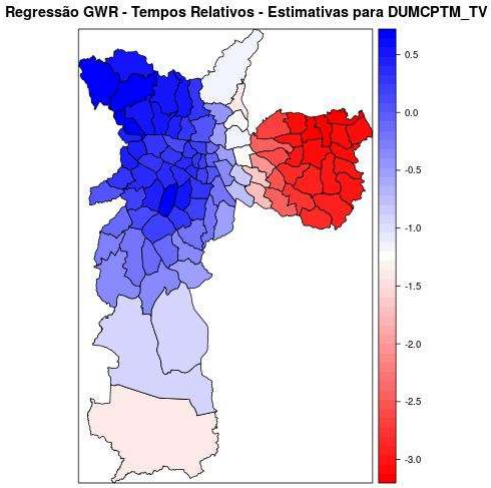


Figura 40: Valores T para o acesso à CPTM do modelo GWR para R_t

A variável de acesso ao Metrô (Figuras 41 e 42) é aparentemente significativa onde a CPTM não o é, e vice-versa. O acesso ao Metrô contribui, assim como a CPTM, para a redução dos R_t, mas os coeficientes da contribuição do metrô são significativamente maiores, em números absolutos, que os da CPTM. Para as duas medidas vale o mesmo comentário feito no modelo de D_t: como elas são dummies de acesso, sua significância também vale como a associação da ausência de acesso nos R_t.

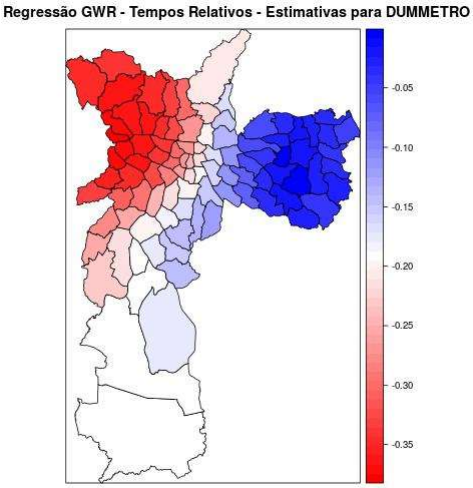


Figura 41: Coeficientes de acesso ao Metrô do modelo GWR para R_t

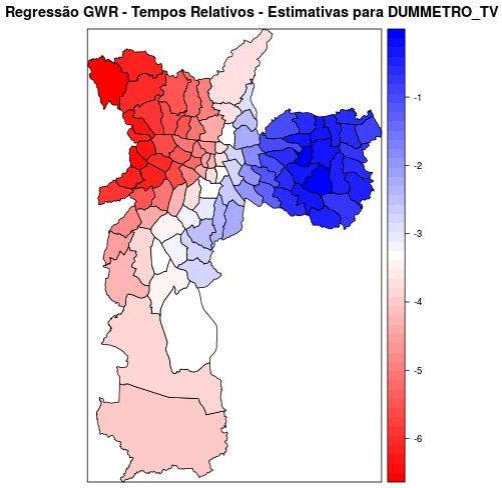


Figura 42: Valores T para o acesso ao Metrô do modelo GWR para R_t

A variável de taxa de motorização de motos (Figuras 43 e 44) dos domicílios por distrito apresenta significância nas zonas Leste e Sul e nas franjas do centro que fazem fronteira com essa região. A variável está associada com o aumento de R_t, com coeficientes bastante expressivos. Uma interpretação para o comportamento dessa variável é que regiões onde o R_t é alto apresentam um diferencial importante de tempo entre os meios de transporte, o que constituiria um incentivo para a posse de motos.

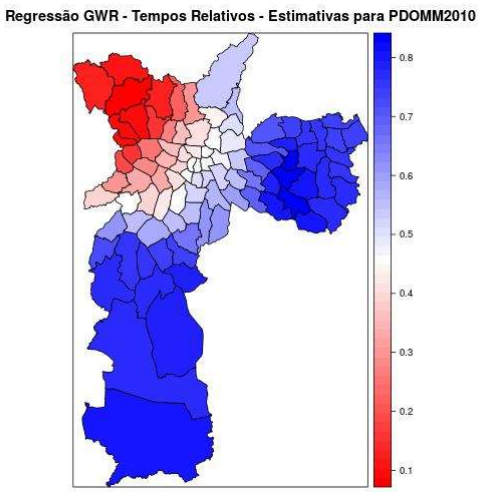


Figura 43: Coeficientes da taxa de motorização de motos do modelo GWR para R_t

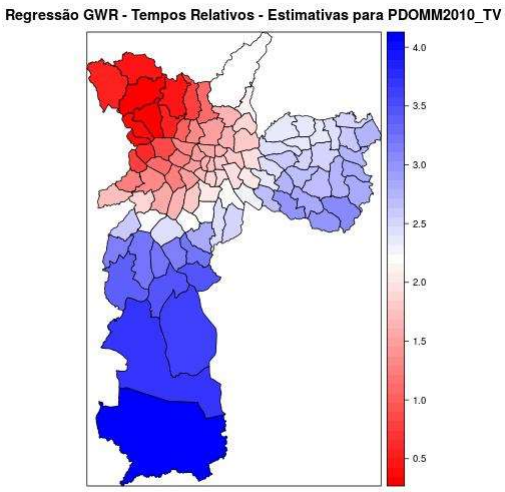


Figura 44: Valores T da taxa de motorização de motos do modelo GWR para R_t

A renda per capita dos distritos apresenta significância somente na zona leste, e está negativamente correlacionada com os R_t – quanto maior a renda do distrito menor o R_t. Outras regiões de SP com altos R_t e baixas rendas per capita - a zona sul e a zona norte – não

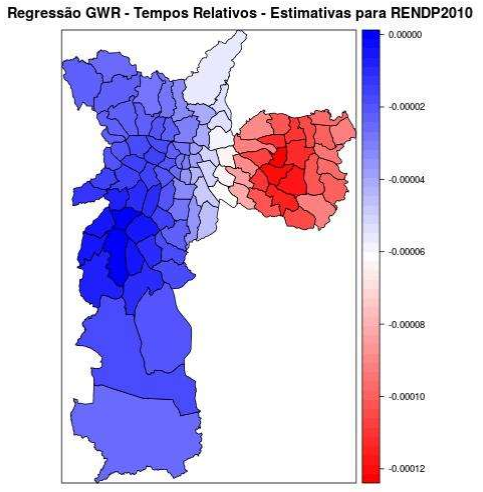


Figura 45: Coeficientes da renda per capita do modelo GWR para R_t

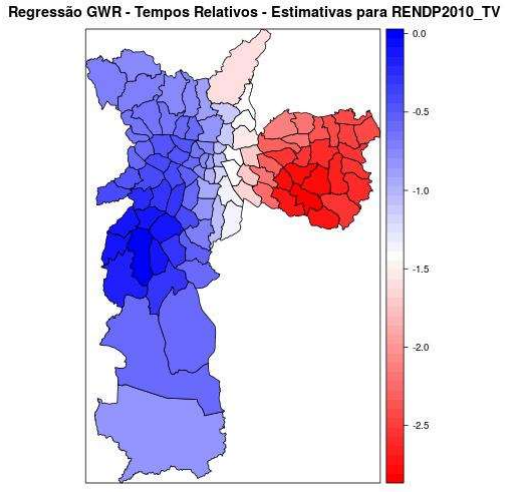


Figura 46: Valores T da renda per capita do modelo GWR para R_t

apresentam a mesma significância para essa variável. Uma possibilidade é que os distritos dessas outras zonas são mais afetados por outros fatores, é que a renda não distingue o R_t desses distritos, enquanto na zona leste distritos mais ricos estão associados a menores R_t's.

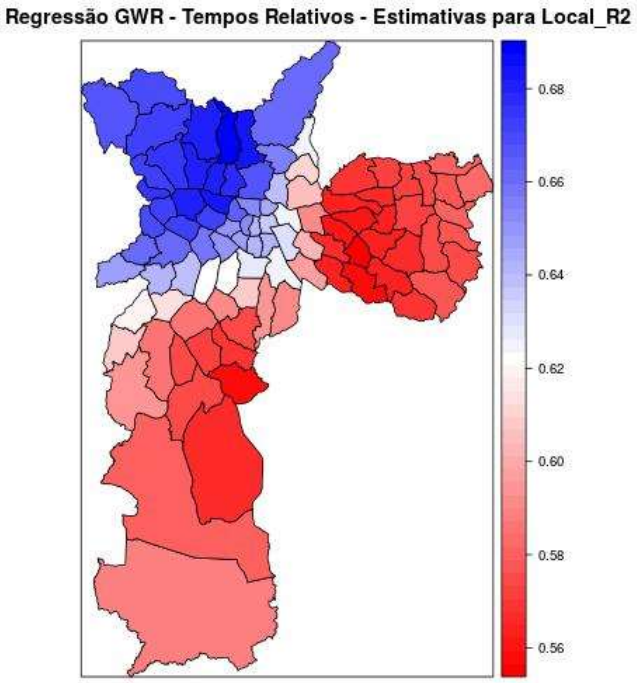


Figura 34: Valores de R^2 do modelo GWR para R_t

A qualidade dos modelos em termos de R^2 para R_t parece ser pior que para os modelos de D_t. Ainda assim, os padrões espaciais de variação do R^2 apresentam semelhanças, com modelos mais capazes de lidar com a variância dos dados no centro e na região norte do que na região Sul e Leste. Isso reforça a possibilidade de que a ausência de variáveis importantes para essas regiões prejudica o modelo.

5. Discussão e Conclusão

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 a 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 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 que 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.

As estatísticas descritivas indicaram pouca variação dos dados em função da hora e dos dias da simulação, contrariamente ao que era esperado, ao menos em relação aos horários de pico. Parte dessa “invisibilização” dos picos de tempo pode ser explicada pela natureza pouco intensiva da simulação: como a cada hora só eram simuladas 100 viagens, sendo possível que as regiões de picos de tempos de viagens tenham sido ignoradas. Ainda levando em consideração que o algoritmo de caminhos do Google Maps evita o congestionamento, a simulação pode refletir melhor a experiência de viagens dos usuários de serviços de rotas, não conseguindo captar a experiência de motoristas que não fazem uso dessa ferramenta.

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 na pesquisa Origem e Destino pode ajudar a "calibrar" esse viés com o "viés" 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.

Tabela 4: Comparação dos modelos regressivos

	OLS D_t	OLS R_t	SAR D_t	SAR R_t	GWR D_t	GWR R_t
R^2	0,489470	0,304930	0,760621	0,669616	0,7040348	0,6257745
R^2 ajustado	0.467029	0,274377	-	-	0,6386598	0,5475248
Nº de variáveis	4	4	4	3	5	6
Variáveis	DUMCPTM DENESTAB PNBRAN2010 DENPOP2018	PDOMC2010 DUMMETRO PNBRAN2010 RENDP2010	DUMCPTM DENLINBUS PNBRAN2010 DENPOP2018	DUMCPTM DUMMETRO PDOMM2010	DUMCPTM DUMMETRO PNBRAN2010 DENPOP2018 DENPTBUS	DUMCPTM DUMMETRO PDOMM2010 DENPTBUS RENDP2010 DENEMPREG
Fator espacial	Não há		Coeficiente espacial dependente da vizinhança – contribui para retirar a perturbação da autocorrelação espacial das outras variáveis		Cada unidade espacial de análise tem os seus coeficientes, permitindo que eles variem no espaço	
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		Apesar de não ser um modelo muito intuitivo, o GWR permite boas visualizações de seus resultados. A sua capacidade de variar os coeficientes e a significância das variáveis permite que ele dê conta de incluir variáveis não estacionárias sem dobrar o modelo.	
Desvantagens	Não consegue lidar com as dependências espaciais dos dados. Mesmo selecionando variáveis semelhantes aos modelos espaciais, seus 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.		Os modelos GWR são mais complexos, com mais variáveis, e a sua interpretação global é mais difícil que os outros dois modelos.	

A tabela 4 faz uma breve comparação entre os modelos utilizados no trabalho. As modelagens espaciais para as duas medidas refletiram razoavelmente 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.

Assim, 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 afete o desempenho do transporte privado. Isso porque as medidas analisadas também são influenciadas pelas variações nos tempos de viagens privadas. Os modelos utilizados 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 análise. Em particular, os modelos utilizados, apesar de apresentarem R^2 e significâncias consideráveis, apresentam poucas outras garantias de qualidade – por exemplo, só o modelo SAR de R_t garantidamente não apresenta heterocedasticidade em seus resíduos.

Uma próxima etapa de validação desses dados pode ser realizada a partir dos microdados da pesquisa de Origem e Destino 2017 do Metrô. Esse conjunto de dados permitirá uma comparação mais direta em relação não só às medidas elaboradas aqui para análise, mas também em relação aos tempos absolutos de viagens. Da mesma forma, a pesquisa OD traz informações importantes sobre o comportamento de mobilidade que pode balizar decisões de simulação de viagens, como as densidades de sorteios de endereços.

A possibilidade de abordagens como essa substituírem as pesquisas empíricas como a OD ainda estão distantes. As medidas de demandas de viagens, com informações estatisticamente relevantes sobre escolhas do modal, objetivos da viagem, divisões socioeconômicas, entre outros detalhes, são difíceis de se simular. Mas isso não quer dizer que os resultados da OD não podem ser enriquecidos com esse tipo de abordagem, que se alimentada com informações da própria pesquisa pode fornecer estimativas uteis de tempos de viagens – para o caso em

particular. 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.

Quanto ao uso desses dados para alimentar processos de decisão dos entes da administração pública: As indicações do presente trabalho apontam que os dados simulados refletem informações importantes acerca da mobilidade em São Paulo. Mas como adverte Kwan (2016), é preciso conhecer as limitações e os vieses derivados do uso intensivo de algoritmos para que se possa conhecer a qualidade dos dados gerados. A praticidade e a tempestividade dessa estratégia, para que seja plenamente aproveitada, deve passar por esforços importantes 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.

6. Referências (10 a 30 referências, 500 palavras)

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

BADDELEY, A., TURNER, R. Spatstat: an R package for analyzing spatial point patterns. Journal of statistical software, 12(6), 1-42, 2005.

BADDELEY, A. Analysing spatial point patterns in R. Technical report, CSIRO, 2010. Version 4. Fevereiro de 2008. URL https://research. csiro. au/software/r-workshop-notes.

CÂMARA, G., MONTEIRO, A. M., FUCKS, S. D., CARVALHO, M. S. Spatial analysis and GIS: a primer. National Institute for Space Research. Brasil, 2004.

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

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.

HÄGERSTRAAND, T. What about people in regional science?. Papers in regional science, v. 24, n. 1, p. 7-24, 1970.

JÚNIOR, J. A. O. Direito à mobilidade urbana: a construção de um direito social. Revista dos Transportes Públicos-ANTP-Ano, 33, 1o, 2011.

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.

LITMAN, T. Measuring Transportation: Traffic Mobility and Accessibility. Victoria Transport Policy Institute, 2003.

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.

PÁEZ, A., SCOTT, D. M., MORENCY, C. Measuring accessibility: positive and normative implementations of various accessibility indicators. Journal of Transport Geography, 25, 141-153, 2012.

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.

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.

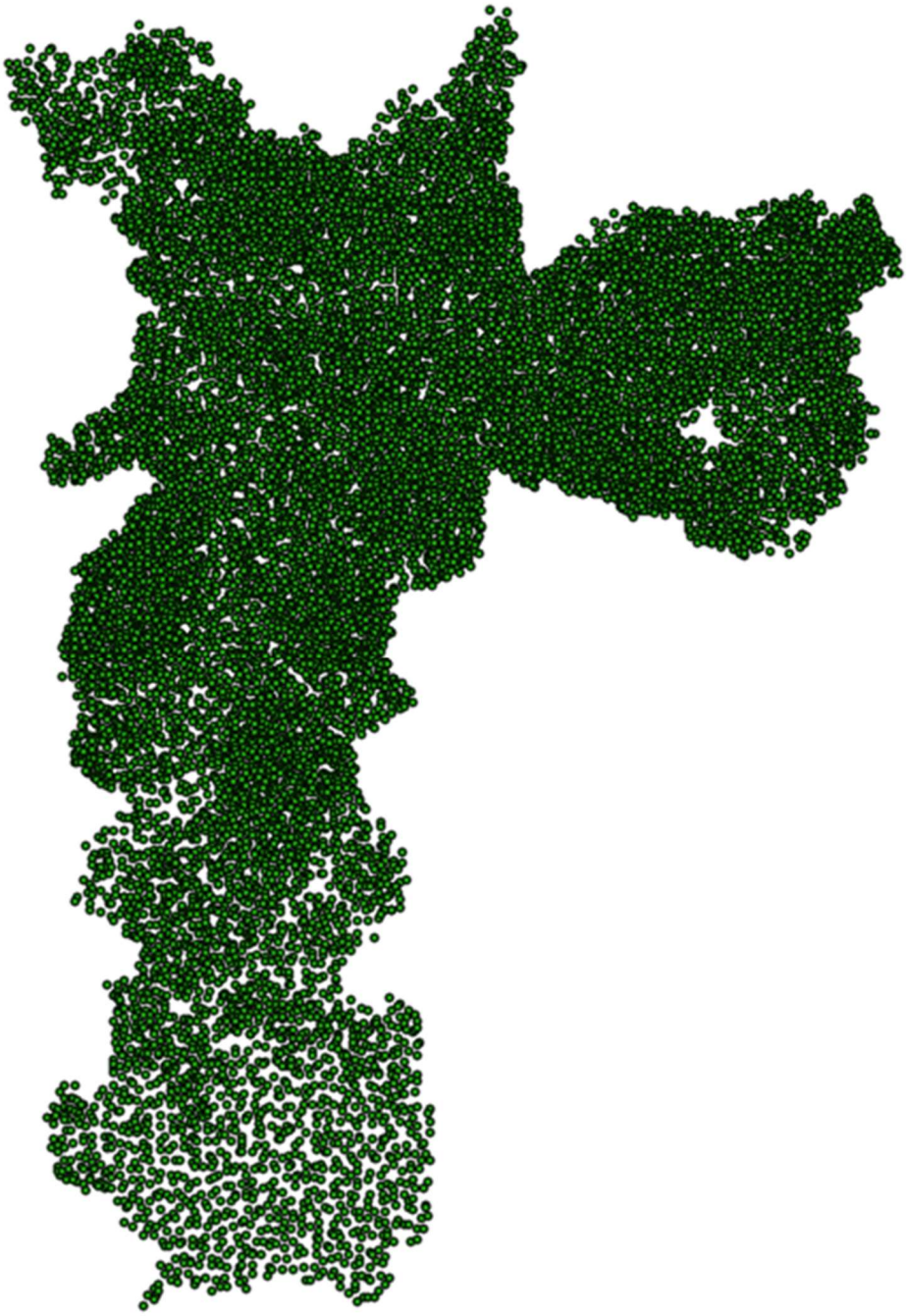
TRIBBY, C. P., ZANDBERGEN, P. A. High-resolution spatio-temporal modeling of public transit accessibility. Applied Geography, 34, 345-355, 2012.

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.

7. **Anexos** (350 palavras)

Anexo I – Mapa da base dos endereços



Anexo II - Saída da regressão OLS simples para D_t.

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            : mapa_Origem_tempo_relativo_distritos
Dependent Variable  :  Tempo_dif  Number of Observations:  96
Mean dependent var  :    3579.12  Number of Variables   :  5
S.D. dependent var  :    876.211  Degrees of Freedom    :  91

R-squared           :   0.489470  F-statistic           :   21.8115
Adjusted R-squared  :   0.467029  Prob(F-statistic)     : 1.20723e-12
Sum squared residual: 3.76279e+07  Log likelihood       :   -754.406
Sigma-square        :     413494  Akaike info criterion :    1518.81
S.E. of regression  :    643.035  Schwarz criterion     :    1531.63
Sigma-square ML     :     391958
S.E of regression ML:    626.065

------------------------------------------------------------------------------
    Variable     Coefficient     Std.Error    t-Statistic   Probability
------------------------------------------------------------------------------
      CONSTANT      3690.87       223.681       16.5006    0.00000
      DENESTAB     -0.38491      0.137601      -2.7973    0.00629
    DENPOP2018   -0.0559716      0.013862      -4.03776    0.00011
    PNBRAN2010      2484.44       474.926       5.23121    0.00000
       DUMCPTM     -316.795       138.412      -2.28878    0.02441
------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   7.714877
TEST ON NORMALITY OF ERRORS
TEST              DF         VALUE           PROB
Jarque-Bera          2          99.9301         0.00000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST              DF         VALUE           PROB
Breusch-Pagan test    4          30.7424         0.00000
Koenker-Bassett test  4           9.9755         0.04084
================================ END OF REPORT ================================
```

Anexo III - Saída da regressão OLS simples para R_t

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            : mapa_Origem_tempo_relativo_distritos
Dependent Variable  :   Temp_rel  Number of Observations:   96
Mean dependent var  :    2.42342  Number of Variables   :   5
S.D. dependent var  :   0.250656  Degrees of Freedom    :   91

R-squared           :   0.304930  F-statistic           :   9.98051
Adjusted R-squared  :   0.274377  Prob(F-statistic)     : 9.65476e-07
Sum squared residual:    4.19232  Log likelihood        :    14.0744
Sigma-square        :  0.0460695  Akaike info criterion :   -18.1487
S.E. of regression  :   0.214638  Schwarz criterion     :   -5.32699
Sigma-square ML     :    0.04367
S.E of regression ML:   0.208974

------------------------------------------------------------------------------
    Variable     Coefficient     Std.Error    t-Statistic   Probability
------------------------------------------------------------------------------
    CONSTANT      2.60741      0.160771       16.2182     0.00000
    PDOMC2010     4.11534       1.2916        3.18622     0.00198
   PNBRAN2010    -1.00598      0.258965       -3.88462     0.00019
    DUMMETRO    -0.202657     0.0542508       -3.73557     0.00033
   RENDP2010  -8.63621e-05   3.58425e-05      -2.40949     0.01799
------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   17.152181
TEST ON NORMALITY OF ERRORS
TEST              DF         VALUE          PROB
Jarque-Bera        2         56.3643        0.00000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST              DF         VALUE          PROB
Breusch-Pagan test 4         17.7595        0.00138
Koenker-Bassett test 4        7.0129        0.13521
================================ END OF REPORT ================================
```

Anexo IV - Saída da regressão SAR para D_t.

```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : mapa_Origem_tempo_relativo_distritos
Spatial Weight      : mapa_Origem_tempo_relativo_distritos
Dependent Variable  :  Tempo_dif  Number of Observations:  96
Mean dependent var  :   3579.12  Number of Variables  :  6
S.D. dependent var  :   876.211  Degrees of Freedom   :  90
Lag coeff.  (Rho)   :  0.756164

R-squared           :  0.760621  Log likelihood       :  -726.107
Sq. Correlation     : -          Akaike info criterion :   1464.21
Sigma-square        :   183783  Schwarz criterion     :   1479.6
S.E of regression   :   428.699

------------------------------------------------------------------------------
    Variable     Coefficient     Std.Error      z-value    Probability
------------------------------------------------------------------------------
    W_Tempo_dif     0.756164     0.0638933       11.8348     0.00000
       CONSTANT      1013.64       265.464       3.81836     0.00013
      DENLINBUS     -5.87178       2.79589      -2.10015     0.03572
     PNBRAN2010       1183.8       348.151       3.40026     0.00067
        DUMCPTM     -260.749       92.4829      -2.81943     0.00481
     DENPOP2018    -0.0246901    0.00937455      -2.63374     0.00845
------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                           DF     VALUE       PROB
Breusch-Pagan test              4       34.5974     0.00000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX : mapa_Origem_tempo_relativo_distritos
TEST                           DF     VALUE       PROB
Likelihood Ratio Test           1       55.3101     0.00000
================================ END OF REPORT ================================
```

Anexo V - Saída da regressão SAR para R_t

```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : mapa_Origem_tempo_relativo_distritos
Spatial Weight      : mapa_Origem_tempo_relativo_distritos
Dependent Variable  :   Temp_rel  Number of Observations:   96
Mean dependent var  :    2.42342  Number of Variables   :   5
S.D. dependent var  :   0.250656  Degrees of Freedom    :   91
Lag coeff.  (Rho)   :   0.745058

R-squared           :   0.669616  Log likelihood        :    42.0364
Sq. Correlation     : -           Akaike info criterion :   -74.0728
Sigma-square        :  0.0207574  Schwarz criterion     :   -61.2511
S.E of regression   :   0.144074

------------------------------------------------------------------------------
    Variable     Coefficient     Std.Error      z-value    Probability
------------------------------------------------------------------------------
     W_Temp_rel     0.745058     0.0671267       11.0993     0.00000
       CONSTANT     0.529982      0.172999       3.0635     0.00219
       PDOMM2010    0.341544      0.115849       2.94817    0.00320
       DUMMETRO    -0.146594     0.0333705      -4.39294    0.00001
       DUMCPTM     -0.0638153    0.0304845      -2.09337    0.03632
------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                           DF     VALUE       PROB
Breusch-Pagan test              3       2.2426    0.52361

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX : mapa_Origem_tempo_relativo_distritos
TEST                           DF     VALUE       PROB
Likelihood Ratio Test           1      62.2511    0.00000
================================ END OF REPORT =================================
```

Anexo VI - Saída da regressão GWR para Diferenças de Tempos

```
***********************************************************************
*                   Package   GWmodel                                 *
***********************************************************************
Program starts at: 2019-07-27 19:00:32
Call:
gwr.basic(formula = Tempo_dif ~ PNBRAN2010 + DENPTBUS + DUMMETRO +
 DUMCPTM + DENPOP2018, data = dados, bw = bw_def, kernel = kernel_type,
 adaptive = TRUE, F123.test = TRUE)

Dependent (y) variable:  Tempo_dif
Independent variables:  PNBRAN2010 DENPTBUS DUMMETRO DUMCPTM DENPOP2018
Number of data points: 96
***********************************************************************
*                Results of Global Regression                         *
***********************************************************************

Call:
 lm(formula = formula, data = data)

Residuals:
   Min      1Q  Median      3Q     Max
-1540.63  -392.74   -27.65   257.91  2781.23

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4082.01944  256.05224  15.942  < 2e-16 ***
PNBRAN2010  2058.15828  521.77321   3.945 0.000158 ***
DENPTBUS     -23.94905    8.95680  -2.674 0.008905 **
DUMMETRO    -365.13992  161.78312  -2.257 0.026430 *
DUMCPTM     -302.19110  135.80239  -2.225 0.028565 *
DENPOP2018    -0.03556    0.01628  -2.184 0.031570 *

---Significance stars
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 629 on 90 degrees of freedom
Multiple R-squared: 0.5169
Adjusted R-squared:  0.49
F-statistic: 19.26 on 5 and 90 DF,  p-value: 5.489e-13
***Extra Diagnostic information
Residual sum of squares: 35607171
Sigma(hat): 615.467
AIC:  1517.512
AICc:  1518.785
***********************************************************************
*        Results of Geographically Weighted Regression                *
***********************************************************************

*********************Model calibration information*********************
Kernel function: gaussian
Adaptive bandwidth: 31 (number of nearest neighbours)
Regression points: the same locations as observations are used.
Distance metric: Euclidean distance metric is used.

***************Summary of GWR coefficient estimates:******************
                 Min.     1st Qu.      Median     3rd Qu.      Max.
Intercept  3418.509072 3522.797259 3728.280999 3881.040055 4096.1557
PNBRAN2010 -635.433851  843.463965 1497.333789 1724.214862 2814.9595
DENPTBUS    -36.167378  -25.330630  -20.007130  -16.440848   -7.4621
DUMMETRO   -988.591302 -770.268771 -494.586461 -274.670870  -64.3393
DUMCPTM    -511.807549 -305.507731  -32.707978   24.501113  158.6499
DENPOP2018   -0.035883   -0.011074   -0.004379    0.010311    0.0364
************************Diagnostic information*************************
Number of data points: 96
Effective number of parameters (2trace(S) - trace(S'S)): 17.18775
```

```
Effective degrees of freedom (n-2trace(S) + trace(S'S)): 78.81225
AICc (GWR book, Fotheringham, et al. 2002, p. 61, eq 2.33): 1493.534
AIC (GWR book, Fotheringham, et al. 2002,GWR p. 96, eq. 4.22): 1470.842
Residual sum of squares: 21813708
R-square value:  0.7040348
Adjusted R-square value:  0.6386598
******************F test results of GWR calibration*******************
---F1 test (Leung et al. 2000)
 F1 statistic Numerator DF Denominator DF    Pr(>)
    0.69959         Inf           90 0.004711 **
---F2 test (Leung et al. 2000)
 F2 statistic Numerator DF Denominator DF Pr(>)
     3.1163     -1.8509           90    NA
---F3 test (Leung et al. 2000)
         F3 statistic Numerator DF Denominator DF     Pr(>)
Intercept       0.90516     38.37514          Inf  0.637778
PNBRAN2010      2.35656     42.36888          Inf 1.534e-06 ***
DENPTBUS        1.78162     33.49680          Inf  0.003552 **
DUMMETRO        4.56934     42.67692          Inf < 2.2e-16 ***
DUMCPTM         3.45815     63.08890          Inf < 2.2e-16 ***
DENPOP2018      1.85101     25.18193          Inf  0.005820 **
---F4 test (GWR book p92)
 F4 statistic Numerator DF Denominator DF   Pr(>)
    0.61262     78.81225           90 0.01349 *

---Significance stars
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
****************************************************************************
```

Anexo VII - Saída da regressão GWR para Tempos Relativos

```
*********************************************************************
*                   Package   GWmodel                               *
*********************************************************************
Program starts at: 2019-07-27 18:51:51
Call:
gwr.basic(formula = Temp_rel ~ DUMMETRO + PDOMM2010 + DUMCPTM +
 DENPTBUS + RENDP2010 + DENEMPREG, data = dados, bw = bw_def,
 kernel = kernel_type, adaptive = TRUE, F123.test = TRUE)

Dependent (y) variable:  Temp_rel
Independent variables:  DUMMETRO PDOMM2010 DUMCPTM DENPTBUS RENDP2010 DENEMPREG
Number of data points: 96
*********************************************************************
*                Results of Global Regression                       *
*********************************************************************

Call:
 lm(formula = formula, data = data)

Residuals:
   Min      1Q  Median      3Q     Max
-0.43016 -0.12556 -0.03633  0.11446  0.78266

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.184e+00  1.226e-01  17.813  < 2e-16 ***
DUMMETRO    -1.921e-01  5.764e-02  -3.332  0.00126 **
PDOMM2010    8.229e-01  2.466e-01   3.337  0.00124 **
DUMCPTM     -8.071e-02  4.707e-02  -1.715  0.08987 .
DENPTBUS    -1.474e-03  3.186e-03  -0.463  0.64476
RENDP2010   -4.572e-05  3.756e-05  -1.217  0.22676
DENEMPREG   -5.780e-07  3.558e-06  -0.162  0.87131

---Significance stars
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.2207 on 89 degrees of freedom
Multiple R-squared: 0.2814
Adjusted R-squared: 0.233
F-statistic:  5.81 on 6 and 89 DF,  p-value: 3.839e-05
***Extra Diagnostic information
Residual sum of squares: 4.33397
Sigma(hat): 0.2147232
AIC:  -8.958767
AICc:  -7.303594
*********************************************************************
*        Results of Geographically Weighted Regression              *
*********************************************************************

**********Model calibration information**********
Kernel function: gaussian
Adaptive bandwidth: 35 (number of nearest neighbours)
Regression points: the same locations as observations are used.
Distance metric: Euclidean distance metric is used.

**********Summary of GWR coefficient estimates:**********
               Min.     1st Qu.      Median     3rd Qu.    Max.
Intercept  1.9533e+00  2.2430e+00  2.2951e+00  2.4228e+00  2.7840
DUMMETRO  -3.7857e-01 -2.9860e-01 -1.9177e-01 -6.1354e-02 -0.0043
PDOMM2010  7.9279e-02  4.0409e-01  5.6079e-01  7.6351e-01  0.8339
DUMCPTM   -1.6293e-01 -9.1942e-02 -6.2775e-03  1.8454e-02  0.0349
DENPTBUS  -6.3765e-03 -3.8033e-03 -7.8687e-04  1.9618e-04  0.0056
RENDP2010 -1.2268e-04 -8.4675e-05 -2.7105e-05 -1.6482e-05  0.0000
DENEMPREG -4.7520e-06 -3.8696e-06 -3.1399e-06 -5.3954e-07  0.0000
**********Diagnostic information**********
Number of data points: 96
```

```
Effective number of parameters (2trace(S) - trace(S'S)): 16.42901
Effective degrees of freedom (n-2trace(S) + trace(S'S)): 79.57099
AICc (GWR book, Fotheringham, et al. 2002, p. 61, eq 2.33): -51.69775
AIC (GWR book, Fotheringham, et al. 2002,GWR p. 96, eq. 4.22): -73.62622
Residual sum of squares: 2.257146
R-square value:  0.6257745
Adjusted R-square value:  0.5475248
**********F test results of GWR calibration**********
---F1 test (Leung et al. 2000)
 F1 statistic Numerator DF Denominator DF     Pr(>)
    0.58252         Inf           89 3.054e-05 ***
---F2 test (Leung et al. 2000)
 F2 statistic Numerator DF Denominator DF Pr(>)
     4.5231     -1.2675           89    NA
---F3 test (Leung et al. 2000)
        F3 statistic Numerator DF Denominator DF     Pr(>)
Intercept       4.9200     39.3425          Inf < 2.2e-16 ***
DUMMETRO       11.3773     45.3285          Inf < 2.2e-16 ***
PDOMM2010       2.0503     32.2580          Inf 0.0004017 ***
DUMCPTM         4.9195     62.6547          Inf < 2.2e-16 ***
DENPTBUS        2.7801     35.3075          Inf 7.877e-08 ***
RENDP2010       4.2588     30.8051          Inf 2.371e-14 ***
DENEMPREG       1.8813     30.5957          Inf 0.0022385 **
---F4 test (GWR book p92)
 F4 statistic Numerator DF Denominator DF    Pr(>)
     0.5208     79.5710           89 0.001673 **

---Significance stars
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
**********************************************************************
```