FUNDAÇÃO GETÚLIO VARGAS  
ESCOLA DE ADMINISTRAÇÃO DE EMPRESAS DE SÃO PAULO

Programa Institucional de Bolsas de Iniciação Científica (PIBIC)

Relatório de conclusão de atividades

Aluno: Rodolfo Oliveira Lorenzo

Orientador: Eduardo de Rezende Francisco

Campo de estudo: Administração Pública e Estatística

São Paulo – SP

2019**Considerações iniciais**

O presente relatório conta com duas partes. A primeira parte consiste no relatório da pesquisa conduzida a partir do projeto apresentado ao início do processo de iniciação científica. A segunda parte consiste em um artigo aceito para publicação na revista RAE, resultado do trabalho conjunto com o mestrando da ESPM, André Insardi. A produção do artigo ocupou principalmente a primeira metade do período do projeto e foi essencial para adquirir os conceitos e habilidades que foram necessários para a realização do projeto proposta de início. Assim sendo, o trabalho resultante do artigo também é parte integrante do processo desenvolvido durante a pesquisa, o que justifica a sua inclusão nesse relatório.

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

# 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.

# 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]](#footnote-0). 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).

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 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).

# 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

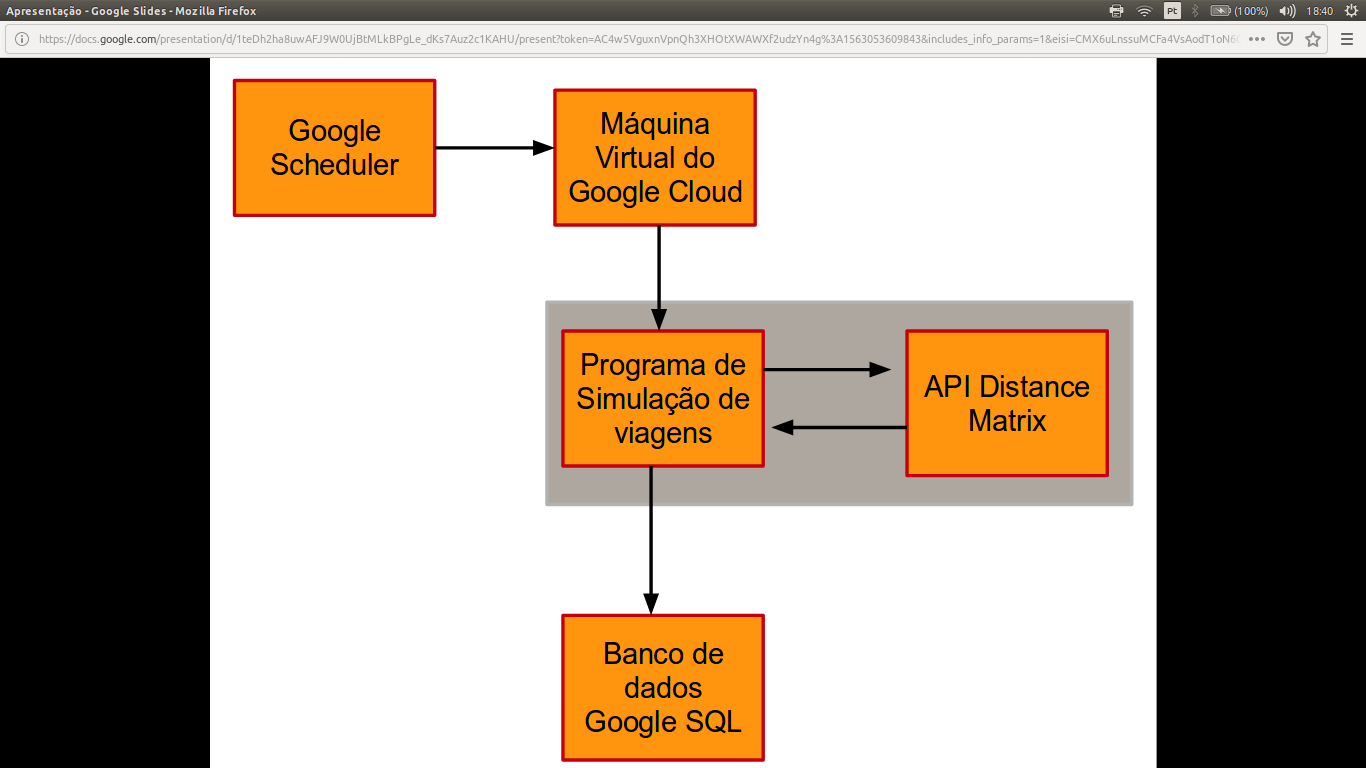
A simulação foi feita em duas etapas: primeiro a geração banco de endereços e depois a simulação das viagens propriamente ditas. De antemão algumas considerações precisam ser feitas:

* 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:

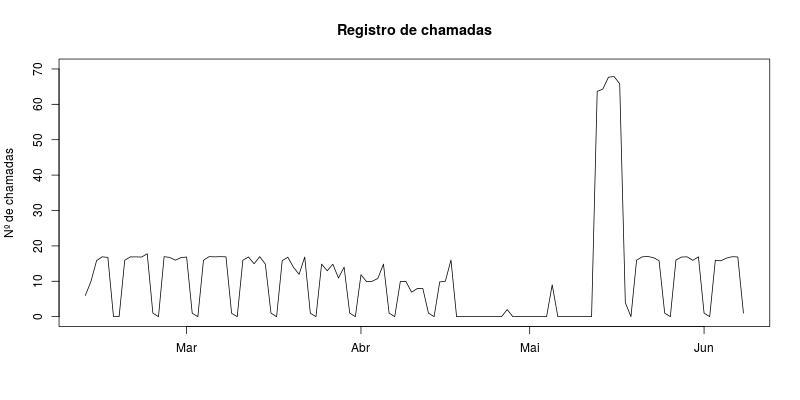
1. Usando o software Qgis, foi gerada, a partir do shapefile do município de São Paulo censo de 2010, uma grade com quadrículas de 500m
2. A partir da informação da população por setor censitário, foram calculadas as populações de cada quadriculas e foram retiradas as quadriculas com população zero.
3. 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.
4. 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:

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:

1. Abrir uma conexão com o banco de dados SQL da nuvem da Google
2. definir uma função para inserção dos dados das viagens no banco de dados
3. Carregar o banco de coordenadas (previamente carregado à máquina virtual)
4. Sortear dez coordenadas de origem e outras dez coordenadas de destino
5. Chamar a API Distance Matrix com as dez origens e os dez destinos, para viagens de transporte público
6. Processar os resultados devolvidos pela API e armazenava em um vetor auxiliar
7. Chamar novamente a API Distance Matrix com as mesmas dez origens e os dez destinos, para viagens de transporte privado
8. Processar os resultados devolvidos pela API e anexar ao vetor auxiliar
9. 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). 



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:

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:

Posteriormente no trabalho as medidas serão referidas como Dt e Rt. Além de explorar as distribuições de Dt e Rt 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 Dt e Rt 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 Dt e Rt 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:

em que Wn 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.

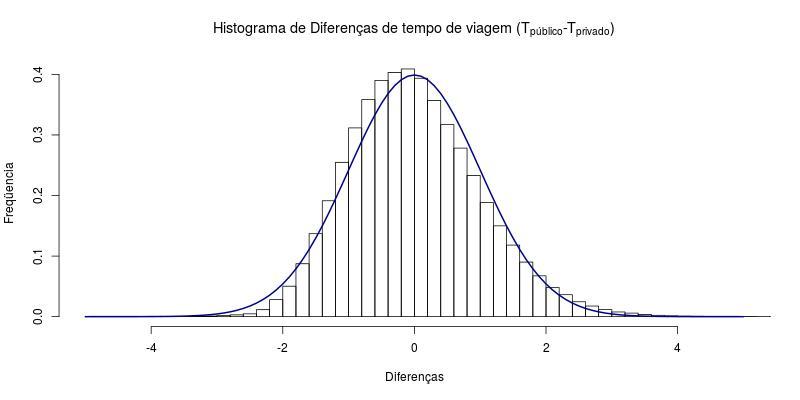
Para o cálculo do modelo foi utilizado o pacote *GWmodel* do software R. Foram feitos modelos para as duas medidas utilizando todas as variáveis selecionadas. A princípio, antes da calibragem, foi usado o critério da janela de seleção com os 20 vizinhos mais próximos; e a função distância selecionada 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)

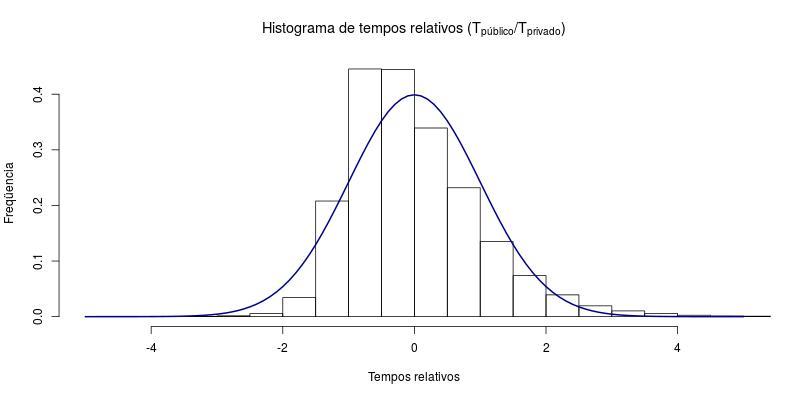
Os modelos foram comparados e a discussão das variáveis capturadas foi realizada a partir dessas comparações.

**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 |

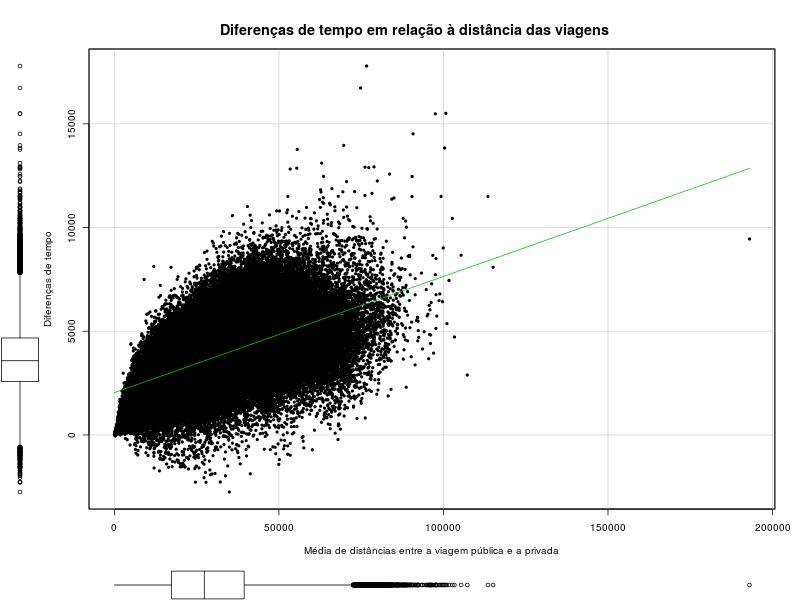
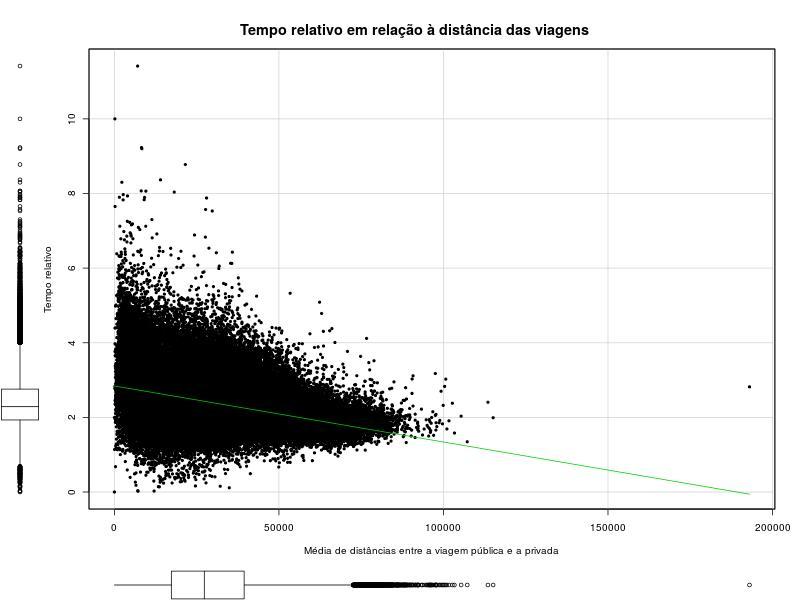
# Resultados

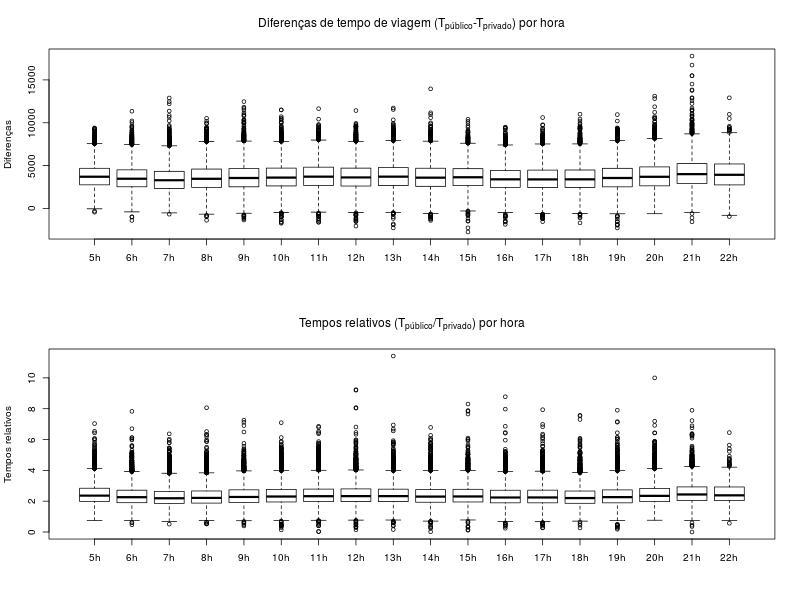
s.

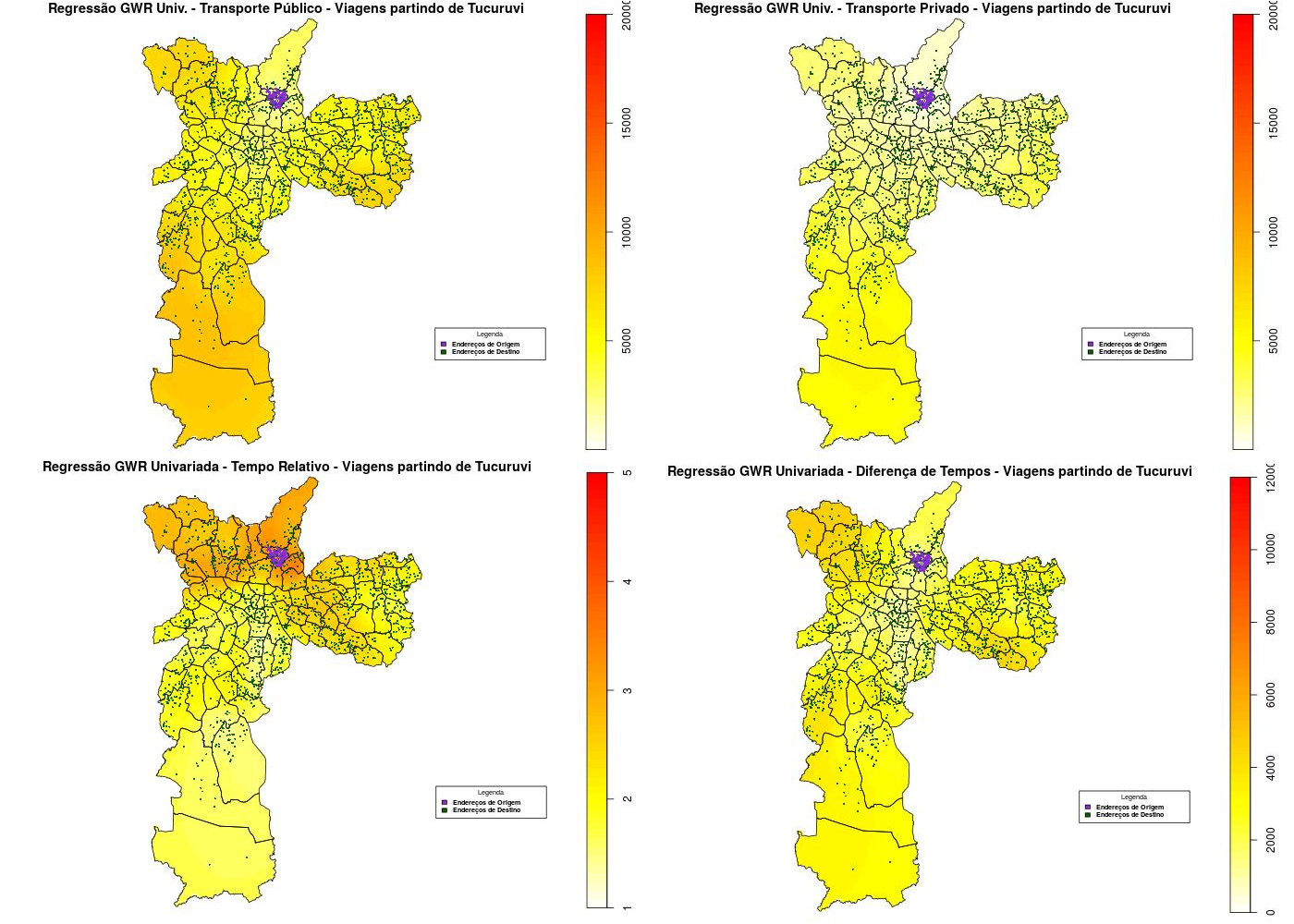


Nos dois casos importante notar que as distribuições indicam uma diferença efetiva entre o tempo de transporte público e de transporte privado; para a Dt, utilizando um teste T de diferença de médias obtemos com 95% de confiança que a média da diferença está entre 3680s e 3697s; um teste T para o Rt 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.

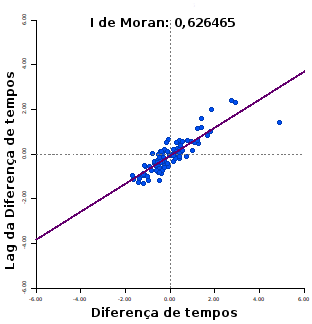
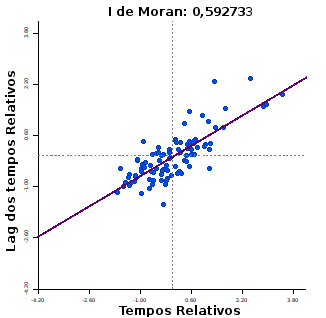
Essa suspeita pôde ser averiguada ao analisar a relação dessas medidas com a distância das viagens. Para cada uma das medidas foi feita uma comparação com a distribuição das médias entre as distâncias das viagens de transporte público e de transporte privado. Dt parece, de forma geral, crescer junto 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 Rt 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.

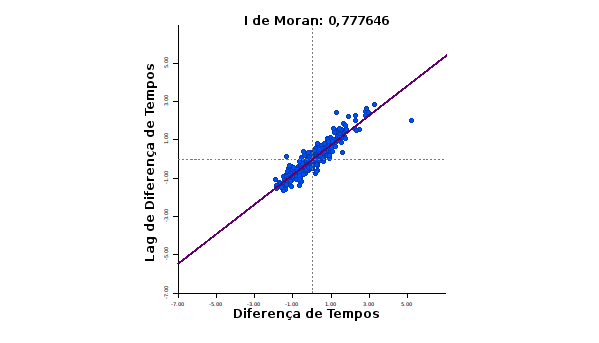
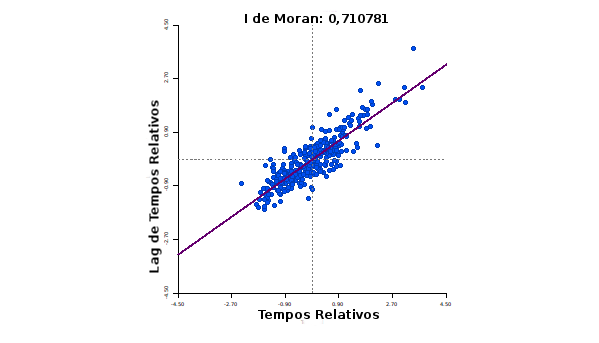


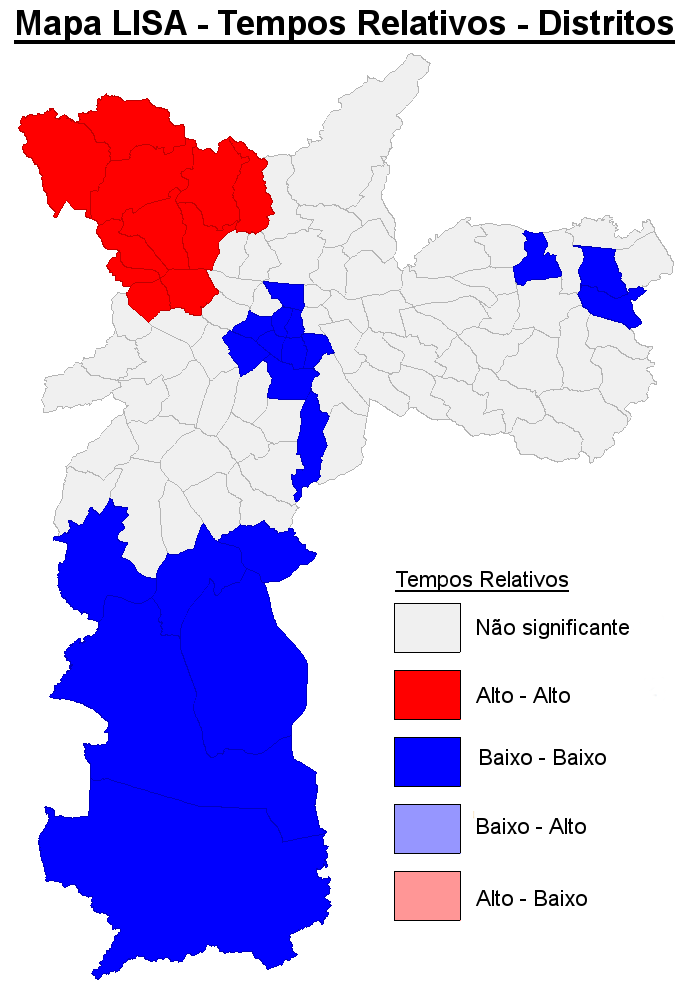
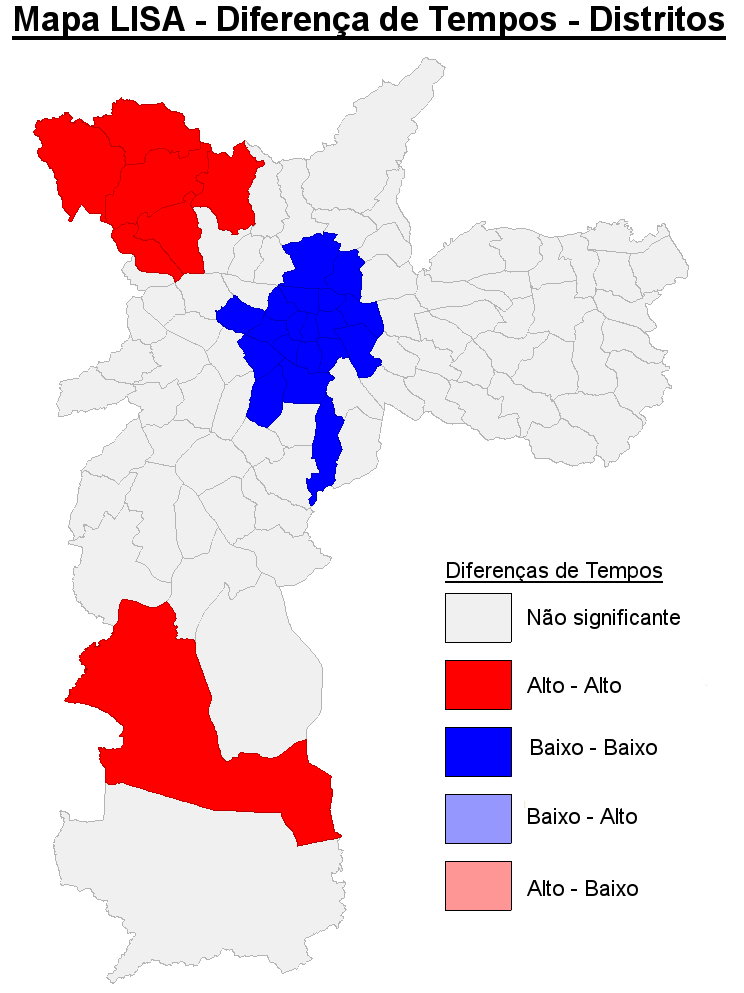
Explorando outras possíveis dependências dos dados em relação aos dias da semana e aos horários, foram elaborados gráficos para explicitar a distribuição das viagens de acordo com essas variáveis. Enquanto a comparação dos boxplots dos dias não parece indicar nenhuma diferença significativa, tanto para 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.

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 Dt e Rt. 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 Rt que não se repetem claramente na superfície de Dt. 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. 

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. 



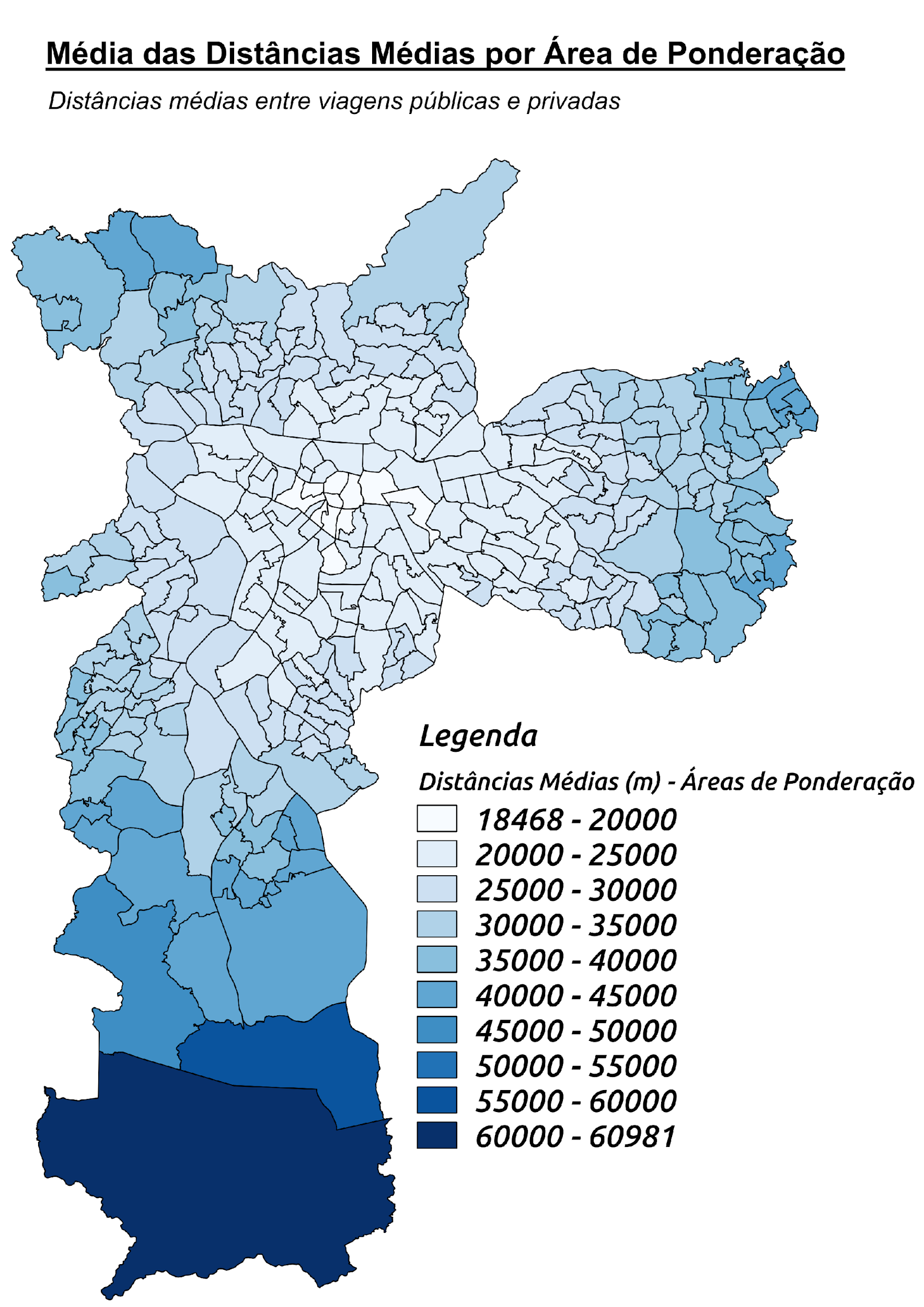


Os valores dos I’s de Moran Globais já indicam a presença de associação espacial significativa entre as medidas dos distritos e das áreas de ponderação, com valores consistentemente acima de 0,5. É notável que a associação é mais significativa nas áreas de ponderação (o que indica maior clusterização dos dados) que nos distritos, e mais fortes para a Dt que Rt.

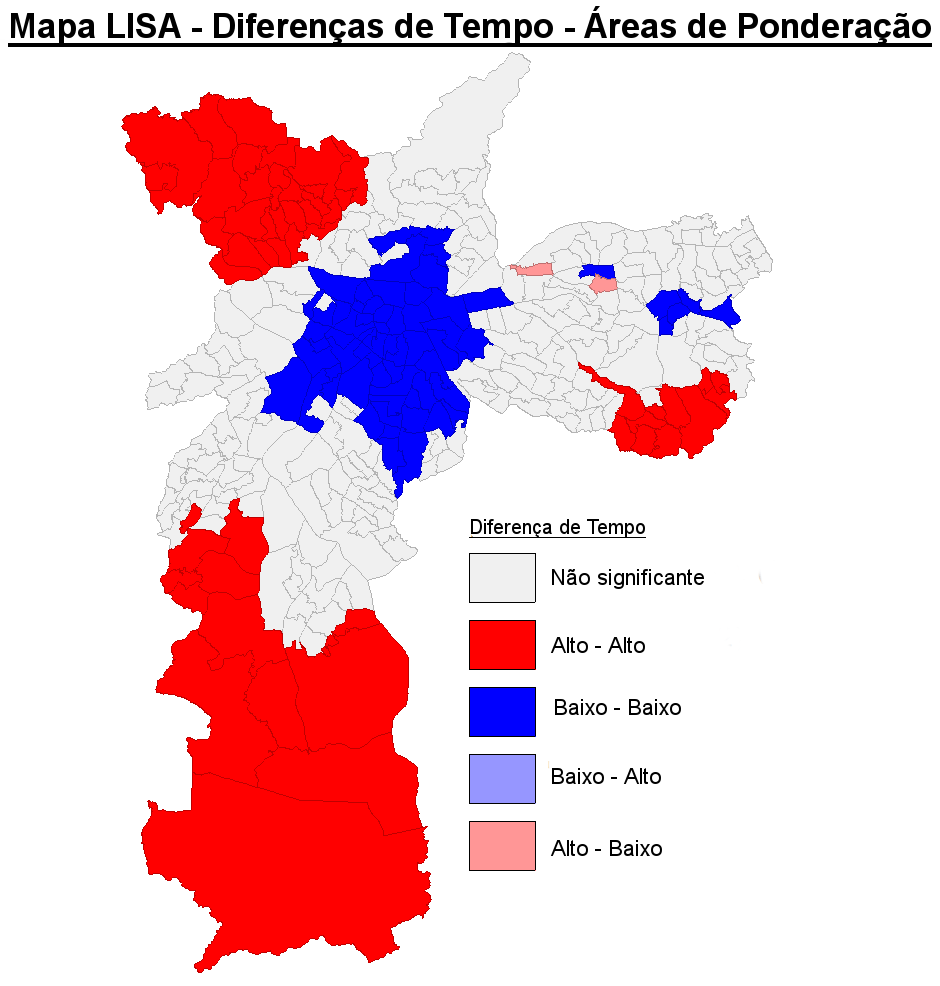
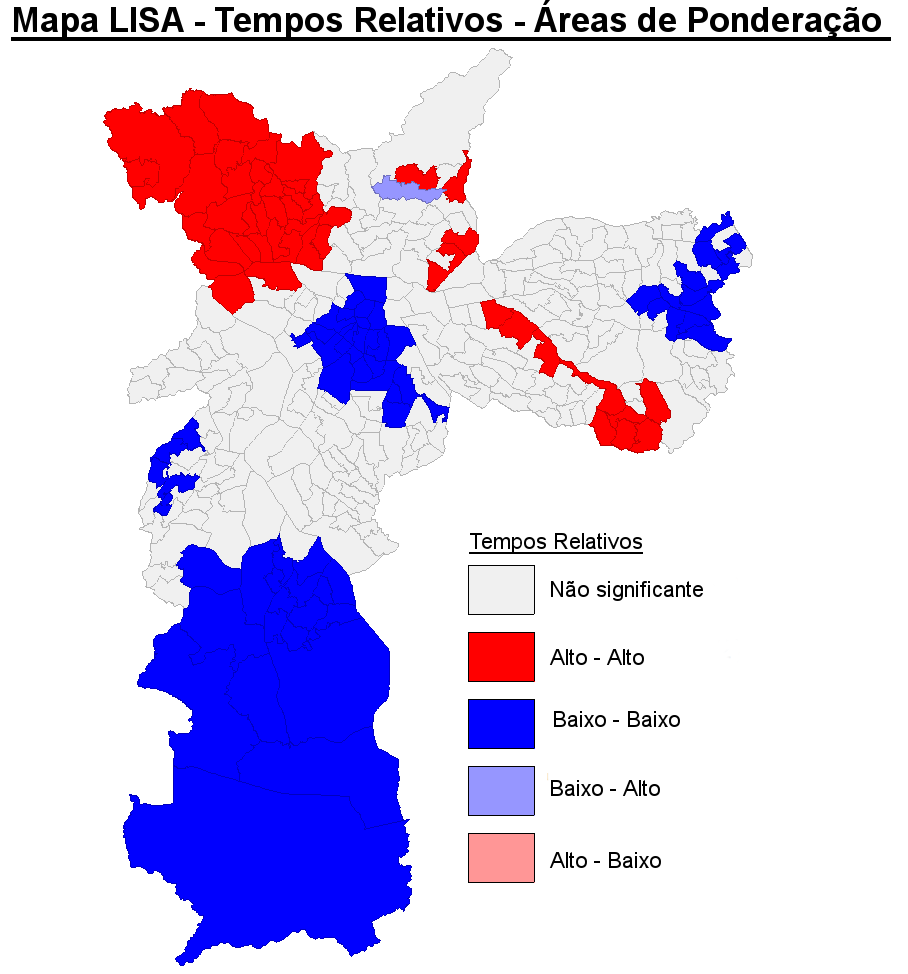
Os gráficos dos I’s de Moran Locais ressaltam a indicação de que há a associação espacial para as duas medidas nos dois níveis de análise. A diferença no I de Moran global entre as medidas pode ser interpretada pela diferente distribuição das observações nos quadrantes do gráfico dos I’s de Moran locais. 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: Dt tende a aumentar com o aumento das distancias das viagens enquanto Rt 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.

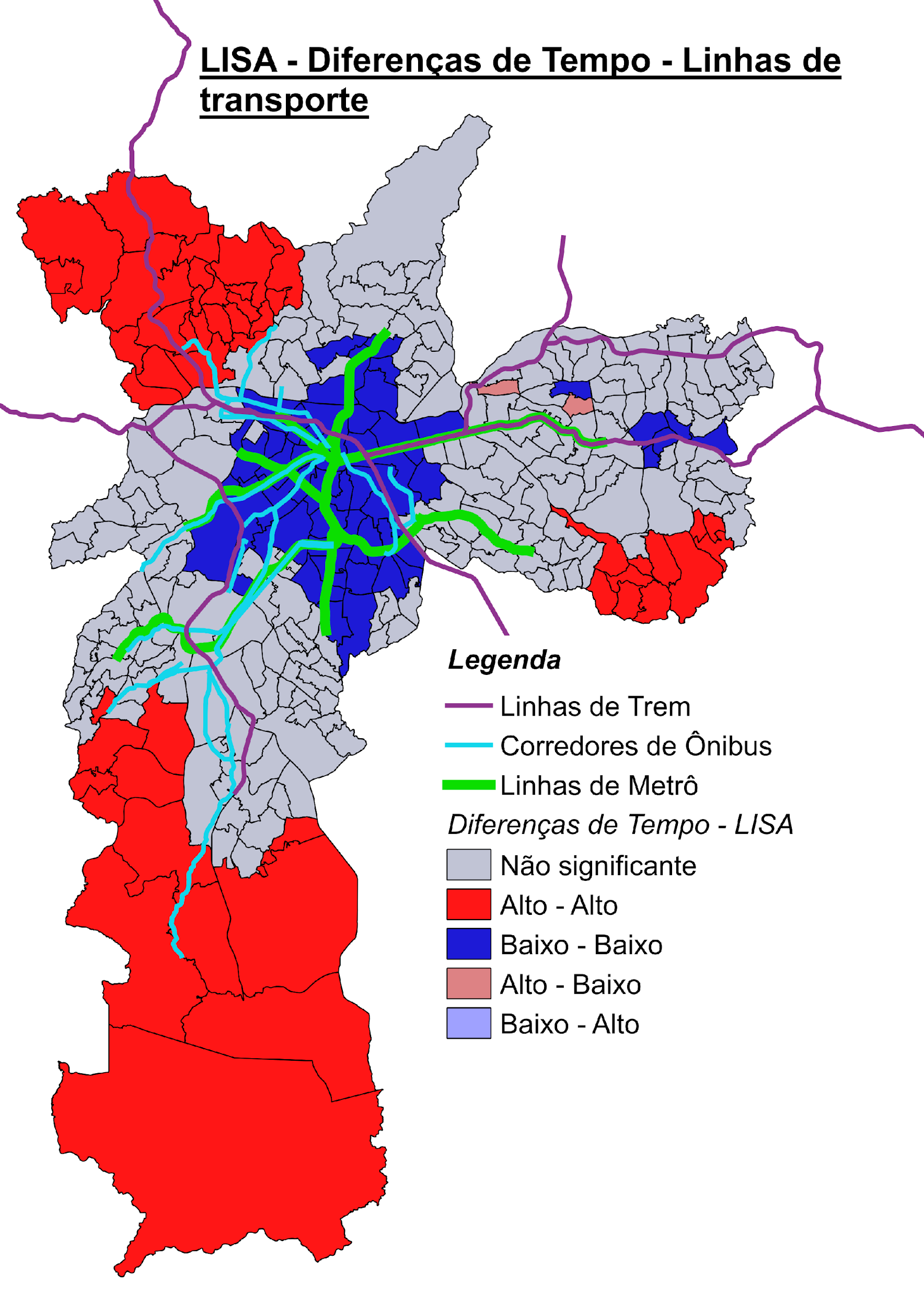


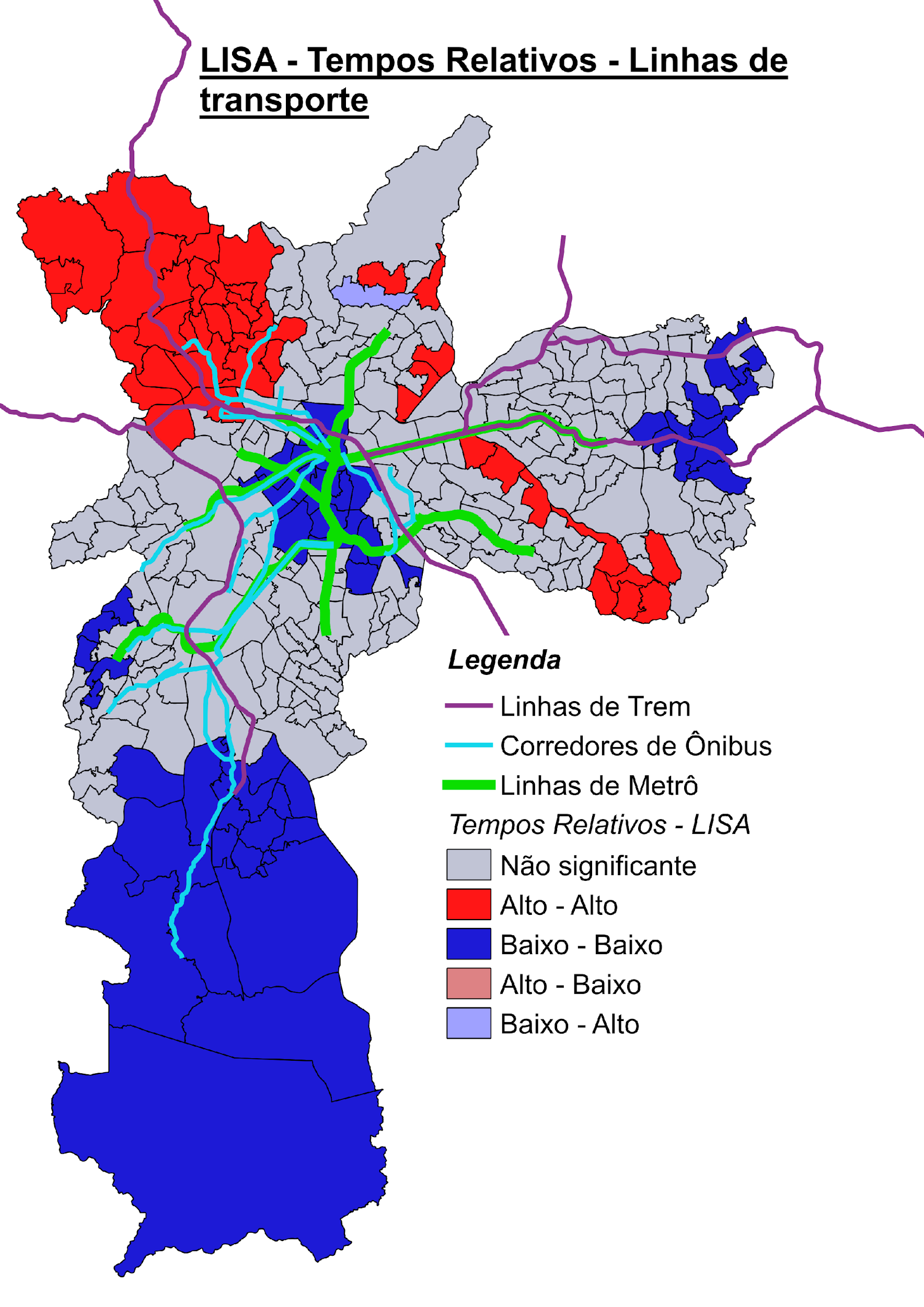
Os mapas LISA para as áreas de ponderação (figuras 16 e 17) apresentam um grau maior de granularidade e, como indicam os I’s de Moran globais e os gráficos de I’s de Moran locais, apresentam mais núcleos de clusters que os distritos.



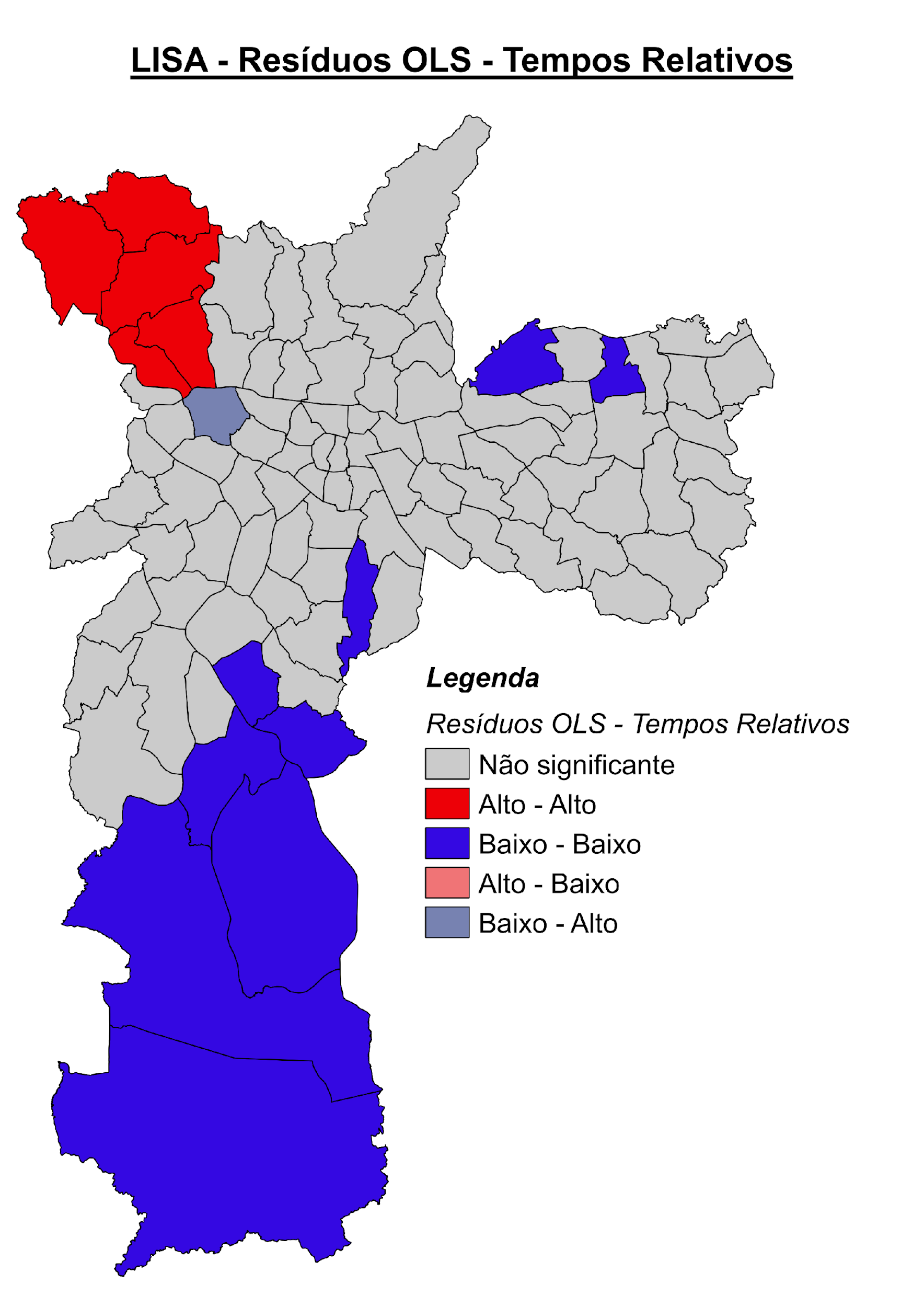
Os mesmos núcleos são visíveis nos dois mapas; no mapa de Rt os clusters de baixo valor no centro e ao sul e o cluster de altos valores a noroeste estão presentes, enquanto no mapa de Dt os núcleos de altos valores a noroeste e ao sul e de baixos valores no centro também se repetem. Porém são novos no mapa de Rt 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 Dt, 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 Rt dos distritos).

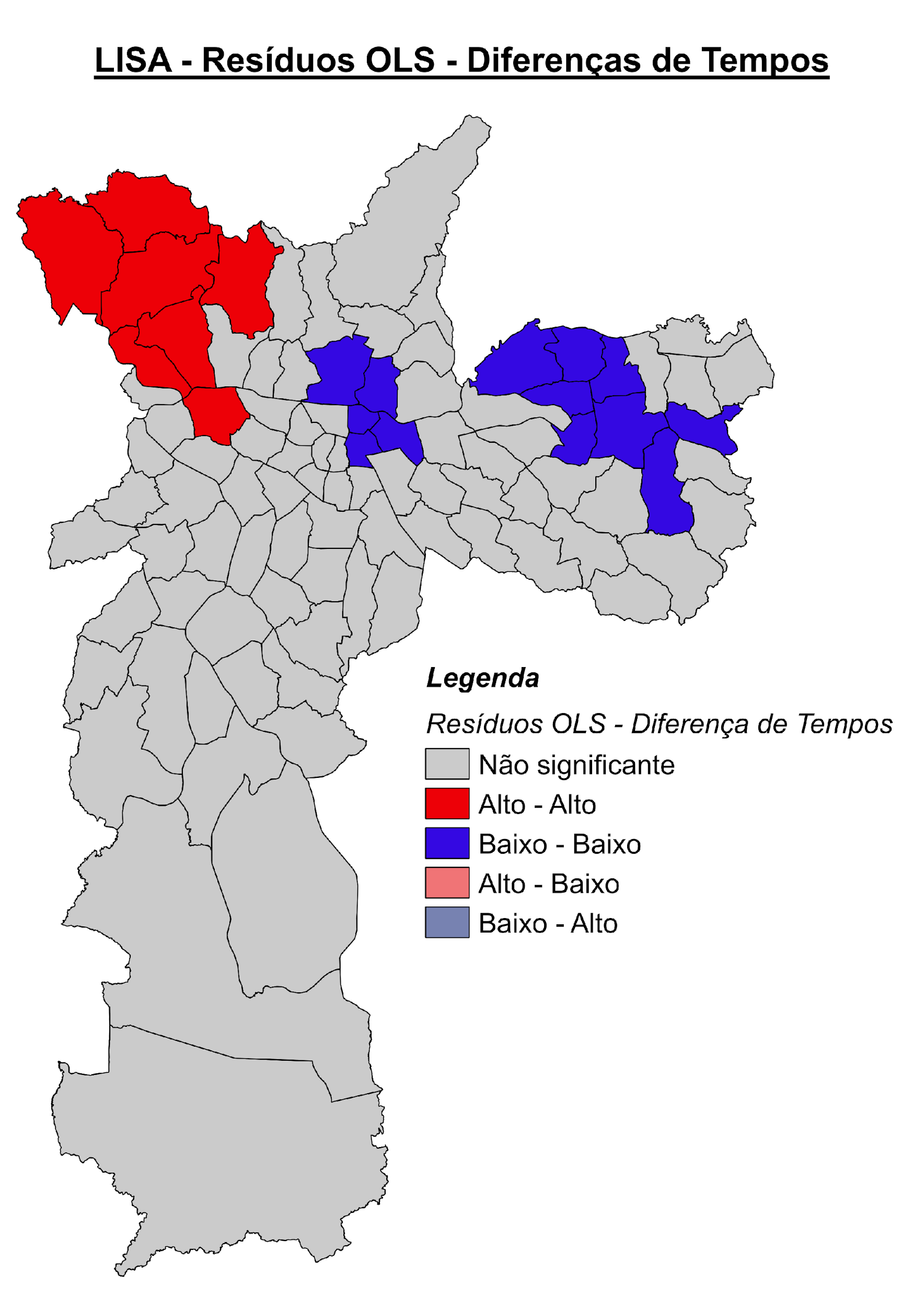
Uma explicação para a diferença do número de clusters entre os níveis de análise é que como os distritos apresentam uma área maior, há uma compensação no interior deles entre tendências diferentes das medidas. Dois exemplos podem ser dados: na comparação entre os mapas de Dt, o pequeno cluster de valores 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 Dt. 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.



Modelagem dos dados

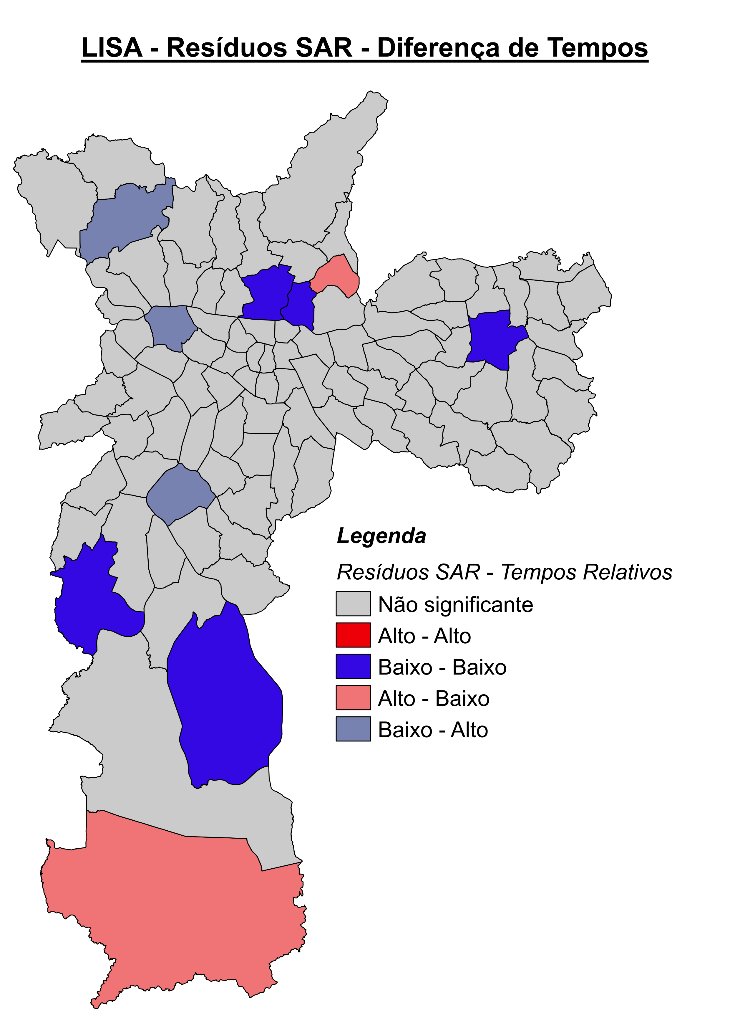
A primeira modelagem dos dados feita foi uma OLS simples, a partir das variáveis da Tabela 2 para as duas medidas. Os resultados das duas regressões estão reproduzidos no anexo (Anexos II e III). Os modelos finais após a retirada de variáveis colineares e variáveis não significantes apresentam R2 relevantes, 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.

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

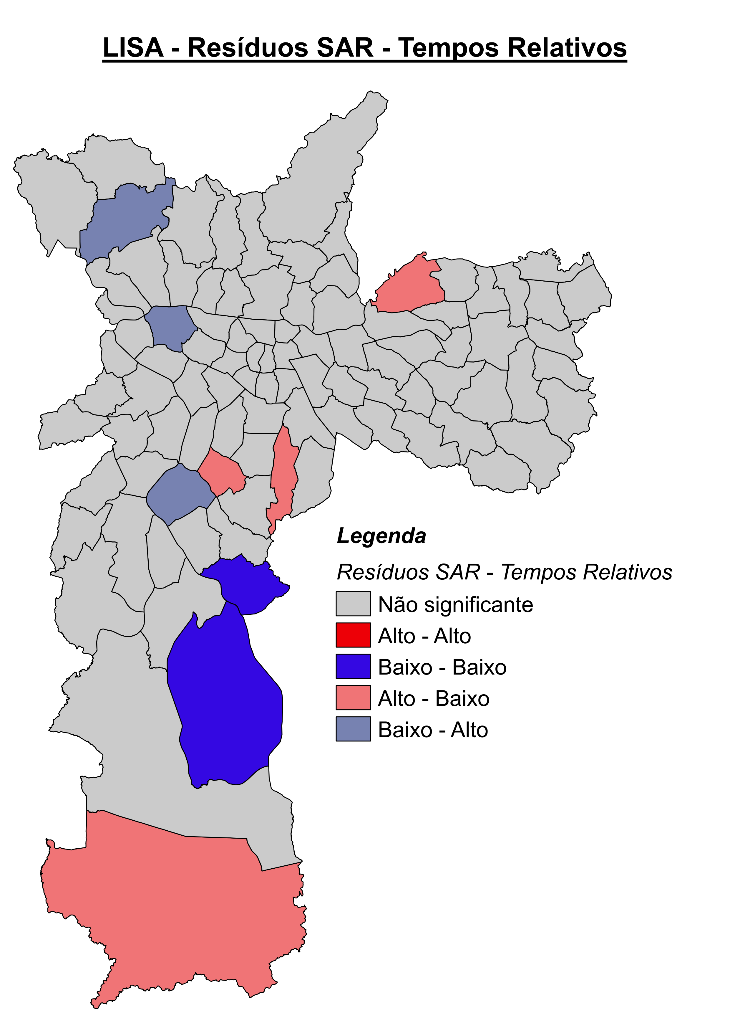
**Tabela 3: Variáveis selecionadas nos modelos SAR e GWR**

|  |  |  |  |
| --- | --- | --- | --- |
| Dt - SAR | Rt - SAR | Dt - GWR | Rt - 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 R2 aumentou seu valor – de 0.48947 para 0.760621 em Dt e de 0.304930 para 0.669616 em Rt – mas para Dt 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 Dt**



**Figura 23: Mapa LISA para resíduos do modelo SAR de Rt**

Os coeficientes selecionados no modelo SAR de Dt 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 Dt 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 Rt 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 Rt 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.

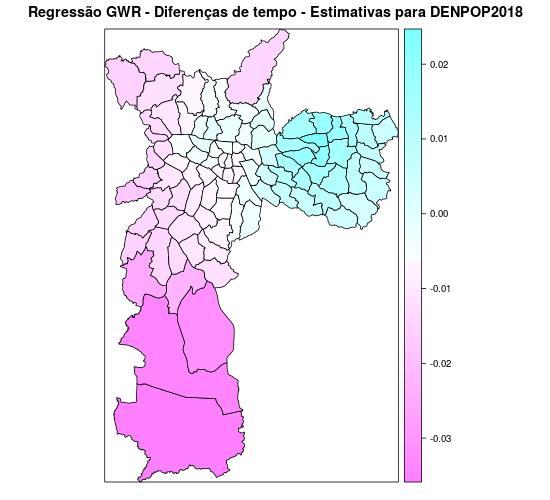
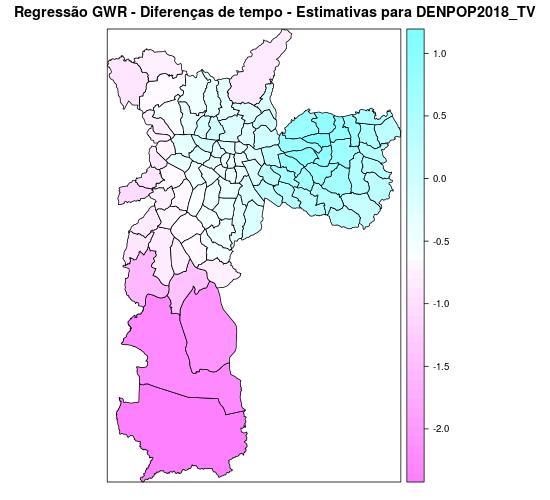
A diferença das medidas pode também trazer informações sobre os modelos. As duas medidas apresentam particularidades em sua variação. Dt, como mostrado, apresenta valores pequenos para viagens curtas e um crescimento linear em função da distância, enquanto Rt apresenta valores altos em viagens curtas, 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 Dt  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 Dt 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).

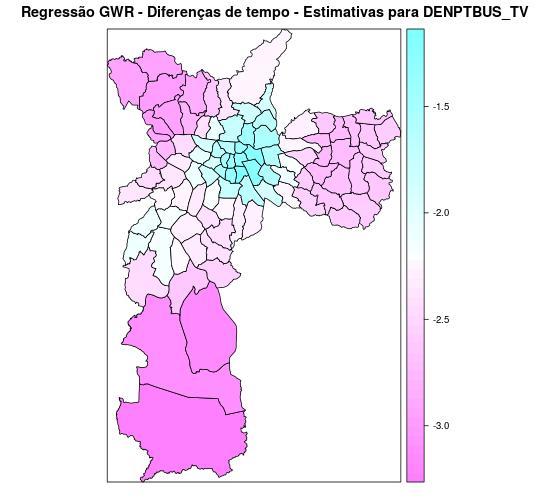
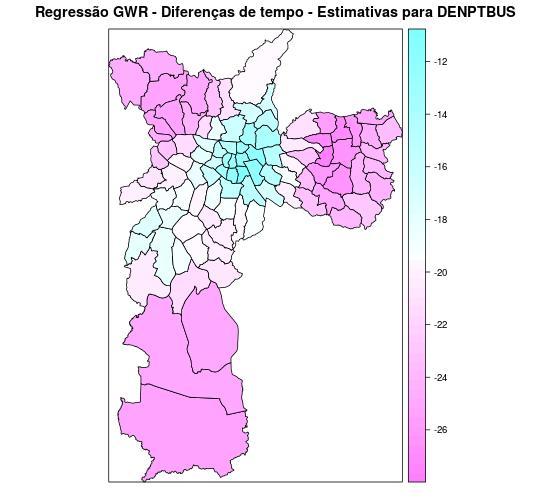
Como Rt não é sensível à distância, a distribuição das médias distritais não segue ao padrão centro-periferia. Ao mesmo tempo, apesar de ser muito sensível a pequenas distâncias (nas quais o valor de Rt é alto), ao agregar as viagens em torno das médias dos distritos de origem a estrutura da simulação 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.

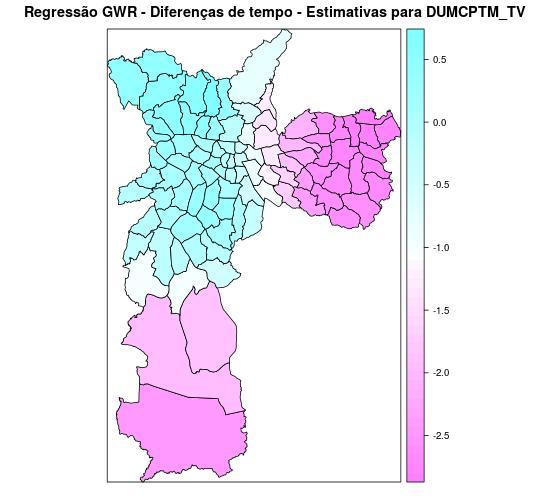
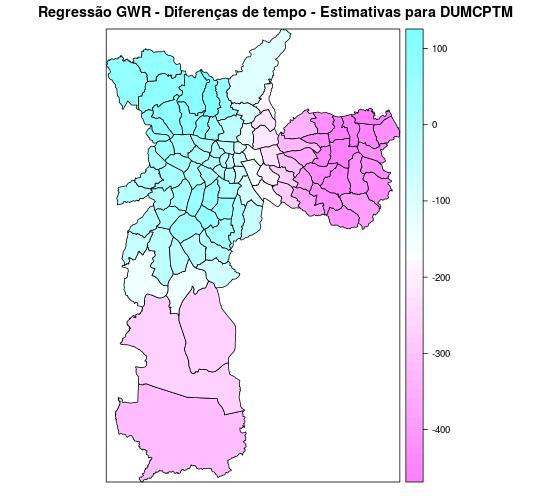
A modelagem GWR apresenta um conjunto de figuras para sua descrição. Cada modelo GWR apresenta um conjunto de regressões lineares para cada um dos distritos. As saídas do programa de cálculo da função estão no anexo (Anexo VI eVII). A análise das variáveis desse modelo é mais complexa que a do modelo 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 R2 dos dois modelos são maiores que os dos modelos OLS simples, mais menores que dos modelos SAR: 0,704 para Dt e 0,626 para Rt..

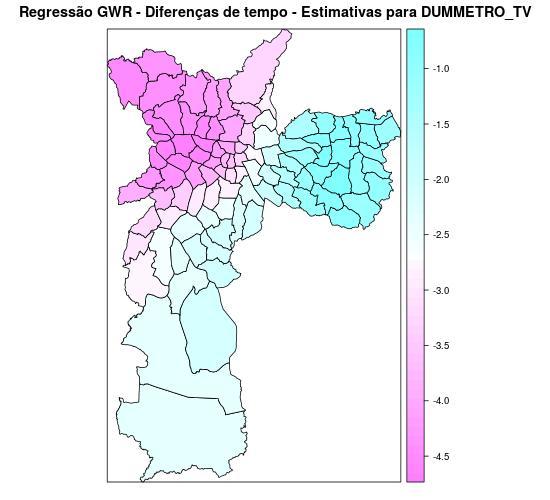
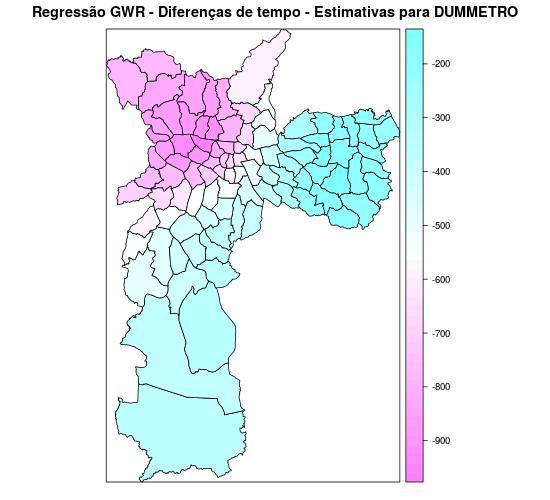
Para Dt as Figuras 24 a 34 mostram os valores dos coeficientes e a os valores T deles em cada distrito, além do R2 em cada distrito. A significância é dada pelo valor de um teste T para cada distrito e está representada no mapa de TV (T-values). Uma aproximação de valores de T que tornam os coeficientes 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 (Dt) e 35 (Rt) vizinhos).

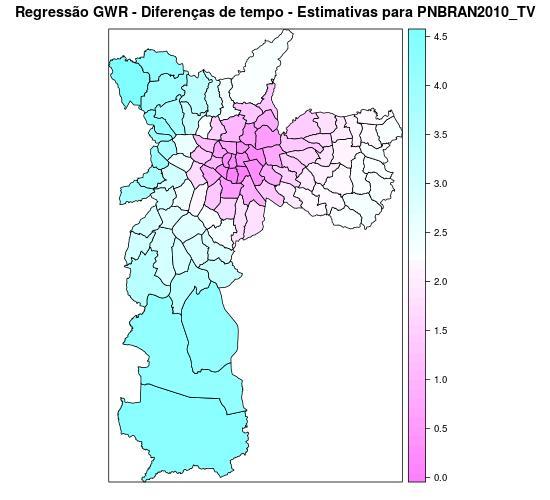
Uma das características do modelo é a possibilidade de variação dos coeficientes da regressão no espaço. Isso também implica que certas variáveis explicativas podem ser significantes em algumas regiões e não em outras. Na figura 25, vemos que a densidade populacional apresenta valores T maiores que o 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.

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 Dt 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.

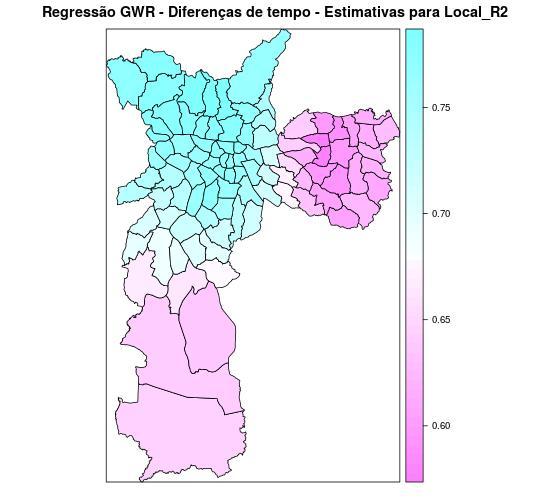


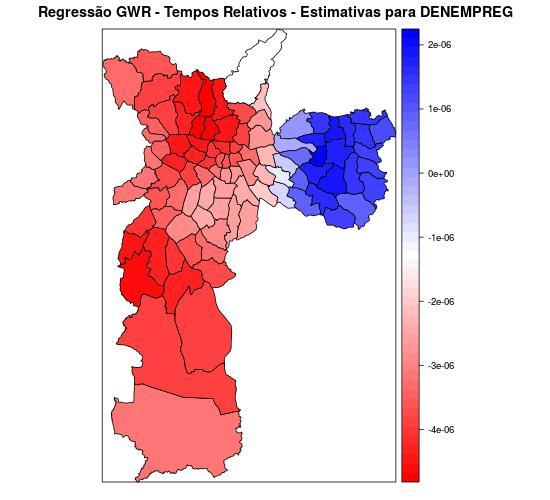
O acesso à CPTM (Figuras 28 e 29) é significante principalmente na zona leste, com a zona sul apresentando Marsilac dentro da margem de significância. Onde é significante, a presença de estações da CPTM está associada a redução de Dt. Vale ressaltar que como a variável DUMCPTM é uma dummy de acesso, 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 Dt na região. 

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 Dt. 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 Dt.

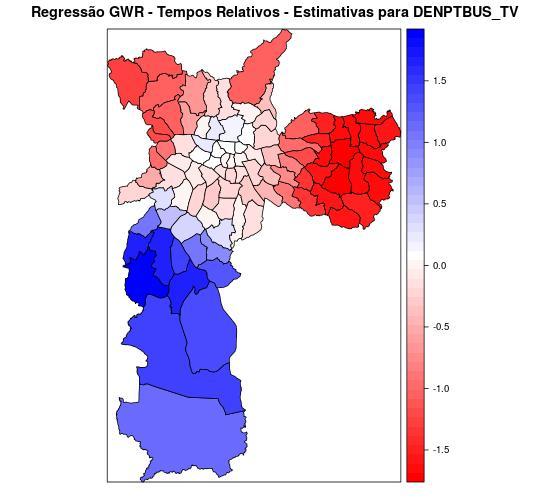
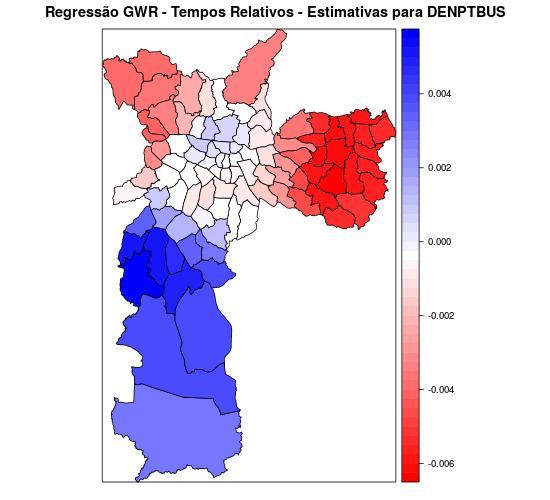
A variável de proporção da população não branca (Figuras 32 e 33) não apresenta significância no centro (onde também seus valores são baixos) e na parte mais próxima da zona leste, mas é significante para o resto do município. Seu coeficiente, onde é significante, é positivo – há uma associação positiva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

A Figura 34 indica o desempenho em termos de R2 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.

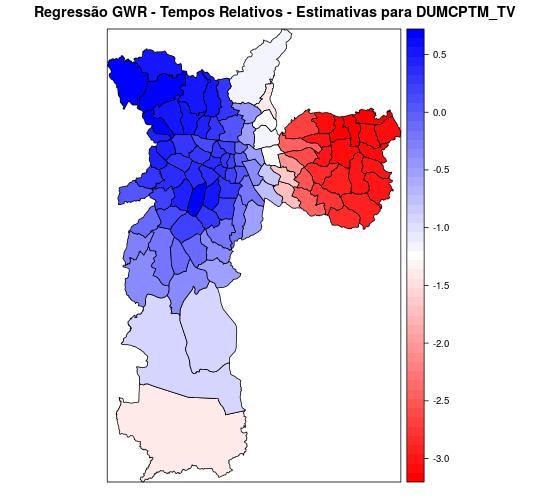
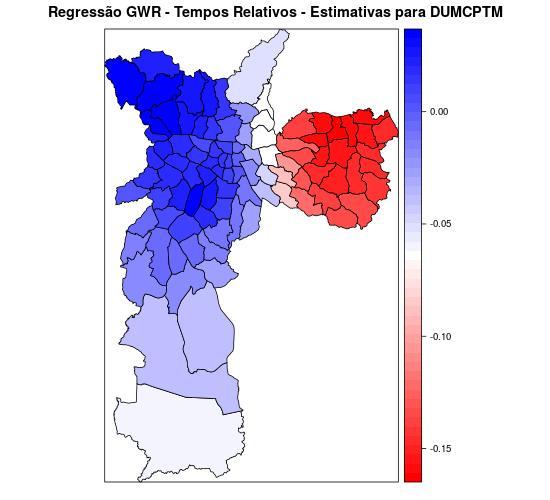


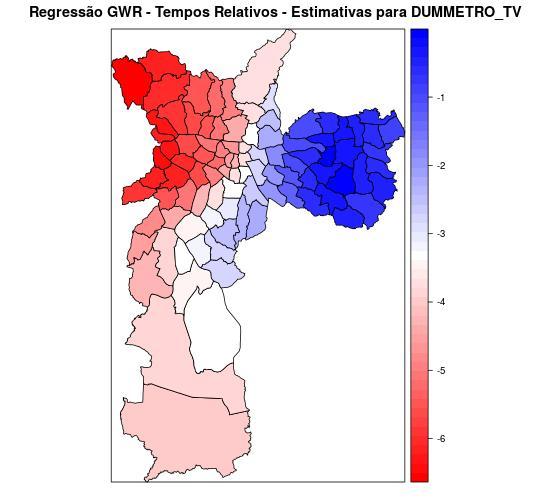
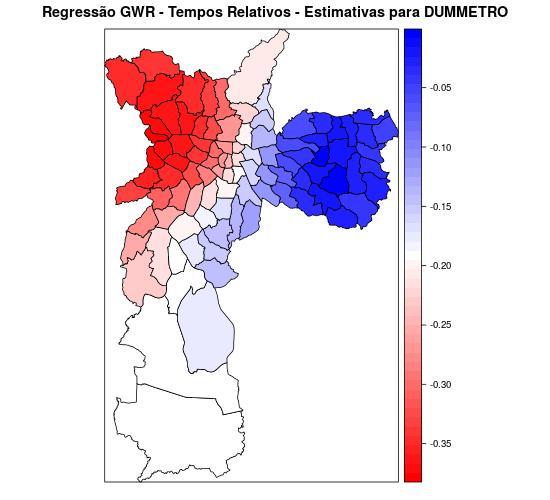
Para Rt as figuras 35 a 47 mostram os valores dos coeficientes e a significância deles em cada distrito, além do R2 em cada distrito. 

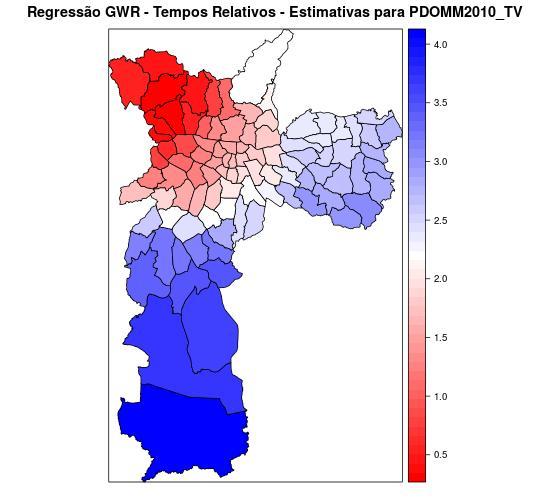
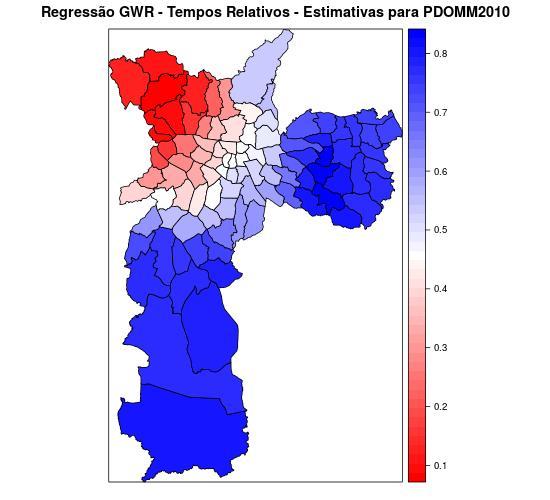
A variável de densidade de empregos (Figuras 35 e 36) é pouco significante para quase todo o município de São Paulo. As regiões onde seus valores t estão próximos do limiar da confiança são parte da região sul e a região ao norte do centro, onde os valores de seu coeficiente indicam que a densidade de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

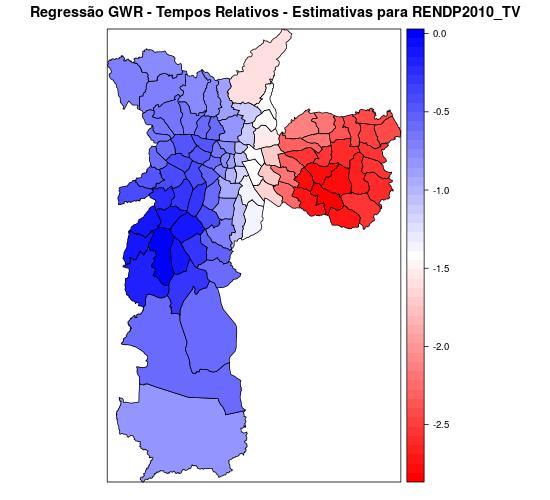
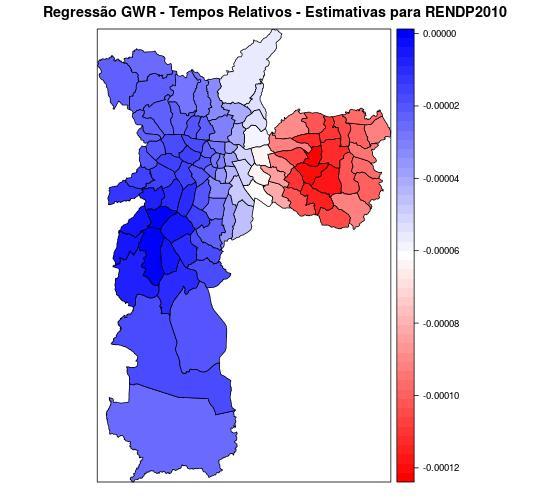
A densidade de pontos de ônibus (Figuras 37 e 38), assim como a densidade de empregos, apresenta pouca significância local nos distritos. A área onde ela mais se aproxima da significância é na zona Sudoeste, onde a maior densidade de pontos de ônibus está associada a maiores Rt.

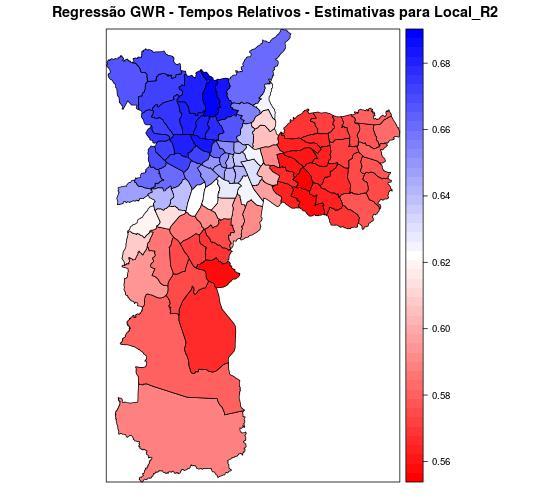
A variável de acesso à CPTM (Figuras 39 e 40) apresentou significância para a zona leste, onde ela está associada à redução dos Rt. Essa relação está de acordo com o modelo de Dt, com exceção da zona sul, que naquele modelo é significante. Essa diferença pode estar relacionada à já discutida dependência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

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 Rt, 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 Dt: como elas são dummies de acesso, sua significância também vale como a associação da ausência de acesso nos Rt.



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 Rt, com coeficientes bastante expressivos. Uma interpretação para o comportamento dessa variável é que regiões onde o Rt é alto apresentam um diferencial importante de tempo entre os meios de transporte, o que constituiria um incentivo para a posse de motos. 

A renda per capita dos distritos apresenta significância somente na zona leste, e está negativamente correlacionada com os Rt – quanto maior a renda do distrito menor o Rt. Outras regiões de SP com altos Rt e baixas rendas per capita - a zona sul e a zona norte – não 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 Rt desses distritos, enquanto na zona leste distritos mais ricos estão associados a menores Rt’s. 

A qualidade dos modelos em termos de R2 para Rt parece ser pior que para os modelos de Dt. Ainda assim, os padrões espaciais de variação do R2 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. 

# 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 Dt | OLS Rt | SAR Dt | SAR Rt | GWR Dt | GWR Rt |
| R2 | 0,489470 | 0,304930 | 0,760621 | 0,669616 | 0,7040348 | 0,6257745 |
| R2 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 R2 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 R2 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 Dt que em Rt – 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 Dt), 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 R2 e significâncias consideráveis, apresentam poucas outras garantias de qualidade – por exemplo, só o modelo SAR de Rt 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.

# 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.

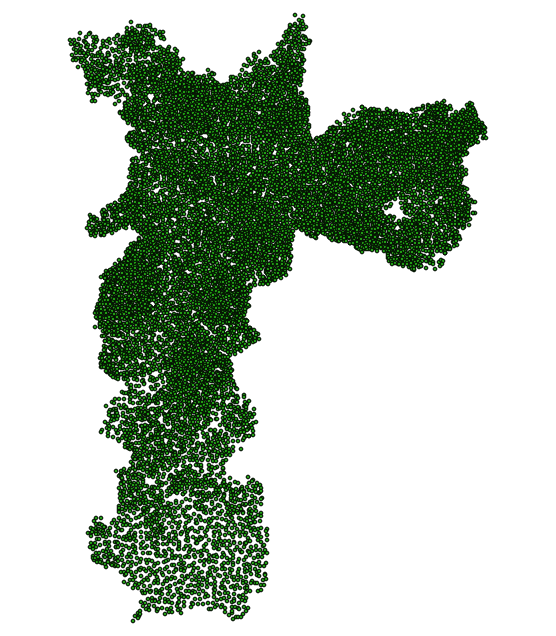
TRIBBY, C. P., ZANDBERGEN, P. A. High-resolution spatio-temporal modeling of public transit accessibility. Applied Geography, 34, 345-355, 2012.

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.

# Anexos (350 palavras)

Anexo I – Mapa da base dos endereços



Anexo II - Saída da regressão OLS simples para Dt.

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 Rt

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 Dt.

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 Rt

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

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**UBER SERVICE WAITING TIME AS A MEASURE OF ACCESSIBILITY: AN EXPLORATORY STUDY IN SÃO PAULO CITY VIEWED THROUGH THE LENS OF BIG DATA**

***Tempo de espera da uber como medida de acessibilidade: um estudo exploratório na cidade de São Paulo sob a ótica do big data***

***Tiempo de espera para tener en cuenta como medida de acceso: se trata de una exploración en la ciudad de São Paulo con una gran cantidad de datos.***

**ABSTRACT**

The present paper aims to relate information about the waiting times ride-sourcing services. Specifically, Uber, with socioeconomic variables from the city of São Paulo, Brazil. The intention is to explore the possibility of using this measure as an accessibility proxy.

A database was created with the mean waiting times for service data per district, which was aggregated to a set of socioeconomic and transport infrastructure variables. From this database, a Multiple linear Regression Model (MLR) was built. Also, the stepwise method selected the most significant variables of the model. The spatial distribution pattern of the measures was confirmed by Moran’s I test, motivating the use of a Spatial Autoregressive model (SAR). The results indicate that physical variables, such as area and population density, are important to explain this relation. However, the mileage of bus lines in the district and the rate of non-white residents was also found to be significant. Besides, the spatial component indicates a possible relation with accessibility.

**KEYWORDS:** Accessibility, Big Data, Uber, Space Statistic.

***RESUMO***

*O presente artigo busca relacionar informações sobre o tempo de espera de serviços de aluguel de carro, especificamente Uber, com variáveis socioeconômicas da cidade de São Paulo com a intenção de explorar a possibilidade uso dessas medidas como um proxy de acessibilidade. Foi montada uma base com a média dos dados de tempo de espera do serviço por distrito, que foi agregada a um conjunto de variáveis socioeconômicas e de infraestrutura de transporte. A partir dessa base foram elaborados modelos de regressão linear multipla (RLM), e utilizando o método stepwise foram selecionadas as variáveis mais significativas do modelo. Foi verificado padrão espacial das variáveis através do teste I de Moran, que motivou a elaboração de um modelo espacial autoregressivo (SAR). Os resultados indicam que variáveis físicas são importantes para essa relação, como área e densidade populacional, mas a quilometragem de linhas de ônibus no distrito a taxa de residentes não brancos, além do componente espacial, indica uma possível relação com acessibilidade.*

***PALAVRAS - CHAVE:*** *Acessibilidade, Big Data, Uber, Estatística Espacial.*

***RESUMEN***

*El presente artículo busca relacionar informaciones sobre el tiempo de espera de servicios de alquiler de coches, específicamente Uber, con variables socioeconómicas de la ciudad de São Paulo con la intención de explorar la posibilidad de utilizar esas medidas como un proxy de accesibilidad. Se ha montado una base con la media de los datos de tiempo de espera del servicio por distrito, que se ha agregado a un conjunto de variables socioeconómicas y de infraestructura de transporte. A partir de esta base se elaboraron modelos de regresión lineal MLR, y utilizando el método stepwise se seleccionaron las variables más significativas del modelo. Se verificó el patrón espacial de las variables a través de la prueba I de Moran, que motivó la elaboración de un modelo espacial autoregresivo (SAR). Los resultados indican que las variables físicas son importantes para esa relación, como el área y la densidad de población, pero el kilometraje de líneas de autobús en el distrito, la tasa de residentes no blancos, además del componente espacial, indica una posible relación con accesibilidad.*

***PALABRAS CLAVE:*** *Accesibilidad, Big Data, Uber, Estadísti*

1. **INTRODUCTION**

The introduction of ride-sourcing companies in the private urban mobility market has considerably changed the habits of many urban inhabitants - its traditional users, private car owners and commuters alike.

The increase in competition in a traditionally very regulated market led to conflicts in many cities between the new services, the former suppliers of this market, taxis, and the local authorities, that had difficulties in framing these new services within existing legal frameworks - as the case in San Francisco illustrates well (Flores & Rayle, 2017).

In this context, ride-sourcing companies tried to build an environmentally friendly image – selling themselves as ride-sharing services (Flores & Rayle, 2017). Beyond that, they claimed its capacity to reduce the number of vehicles on the street (and the emission that came from it) and to offer better and cheaper services to former neglected areas by the taxis.

In this debate, a number of studies tried to shed light in the effects of the resulting transformations. Jin, Kong, Wu, and Sui (2018) did a review on recent literature regarding the effects of the ride-sourcing services. Among the issues discussed there are economic efficiency and social equity of these services. Broadly, there seems to be evidence to support the increase in private transportation service in peripheral areas and low-income areas formerly neglected by taxis. Despite that, the customer profile of the service is identified to be younger, richer and better educated than the mean individual of the population; the article points to a possible exclusion from the service along the lines of a “digital divide” - both in terms of generation and income (Jin et al., 2018).

However, another essential aspect of the ride-sourcing services is its capacity to generate high quality Big Data that can be used to evaluate several questions about the service. The access to this data allows for rich analysis with a great detail and can further reveal the interplay between other modes of transportation and ride-sourcing (Jin et al., 2018).

In particular, the possibility of analyzing urban accessibility with this data source is interesting. The transport network distribution as an essential part of accessibility (Páez, Scott, & Morency, 2012) is a recurrent research subject among geographers, urban planners and social scientists. Recently, this field of studies is being transformed regarding its methods and its approach. It happened through the influence of among other things Big Data, opening dialogs with other disciplines (Schwanen, 2016). At the same time, the field continues to make itself relevant: whilst distance friction is a reality, accessibility will continue to be a useful concept to describe urban experience (Páez et al., 2012).

Particularly considering Big Data approaches, Letouzé and Jütting (2015) affirm that the official bodies responsible for official statistics and indicators production – including academic bodies - must be aware to the evolutions in “Big Data”. Both to profit from new tools and approaches, together with the scientific rigor of validation and analysis as well as to face the world of Big Data as an inestimable source of data for the advance of scientific research. The potential of Big Data tools is also of considerable relevance to public managers and public policy makers (Kim, Trimi, & Chung, 2014). In a multimodal transport network, the usage of different big data tools can help to better regulate the private transport supply and to better deliver public transportation according to user needs (Kim et al., 2014; Lessa, Lobo, & Cardoso, 2019).

In this context, the growth of global tech companies in mobility, such as Uber, Cabify and Lyft, has made them considerable players in this field. For instance, Uber has a daily average of 15 million rides over the world. In Brazil, the company provides its service for 100 cities with a network of 500.000 drivers and more than 20 million users.

Uber develops, commercialize and operates the homonym application for smartphones that allow consumers to demand rides to partnered drivers. In the process of demanding the rides, the Uber’s tool provides estimates of waiting times and travels cost on the user’s app and in the web environment through public APIs (“Application Programming Interface”).

Those APIs generate estimates through the analysis of rides history in the region of the solicitor and the supply-and-demand curve of Uber’s cars (Cohen et al., 2016) as can be verified in the available documentation ([https://developer.uber.com](https://developer.uber.com/)).

Wang and Mu (2018) and Hughes and Mackenzie (2016) propose the usage of these estimates as a possible measure for accessibility. The interferences of these new services in the transportation environment and the easy access to the tools Uber provides already motivated some studies (Hall & Krueger, 2016; Hughes & MacKenzie, 2016; Wang & Mu, 2018; Zhou, Wang, & Li, 2017). It opens an opportunity of empirical exploration of Uber’s fleet in light of accessibility theory.

The aim of the present paper is to use Big Data tools developed by Uber, one of the largest ride-sourcing providers in São Paulo, to generate data to conduct an exploratory study of a potential accessibility measurement.

As Uber’s pricing algorithm follows the balance between supply and demand in order to influence driver’s behavior (Hall, Horton, & Knoepfle, 2019), we assume that waiting times for Uber rides can reflect regional imbalances in the supply of cars. Following this line of t thought, we have explored the relationship of waiting times with other variables associated in similar literature with accessibility (Hughes & MacKenzie, 2016; Lessa et al., 2019; Wang & Mu, 2018), variables related to mobility and socioeconomic variables.

The results found are in contrast with previous studies (Hughes & MacKenzie, 2016; Wang & Mu, 2018), regarding the variables found to be relevant in the context of São Paulo. Having in mind the warning in (Schwanen, 2016) about generalizing conclusions with Big Data to different local contexts, further studies can elaborate comparisons between different cities to elucidate how the local context can be better taken into account in similar approaches. Also, further validation of this data source can build a new tool capable of helping in decision making of transport planners and public policy managers in cities (Kim et al., 2014; Letouzé & Jütting, 2015).

1. **LITERATURE REVISION AND RESEARCH QUESTION**

This section begins with the conceptualization of “Big Data” and its positioning in the current technology market, conceptualizing Uber's estimation tool in light of the “Big Data” concept. It is followed by a revision of the concept of accessibility, its normative and positivist dispositions and its managerial context.

* 1. Big Data

A classical definition of the Big Data movement takes in consideration the features of the data produced in the virtual environments of massive users’ presence: Volume, Variety and Velocity, the three Vs (Brynjolfsson & McAfee, 2012).

According to this definition, the data generated by the new ways of using technology and applications generates relatively large databases, arriving at the scale of Petabytes or Exabytes; the various forms of data generation (such as photo posts, comments on social networks, reactions to comments from others, videos and audios, etc.) are responsible for generating heterogeneous databases in contrast to structured databases; and the timing of the generation of these databases is very fast, demanding in some cases real time processing. These aspects, make the techniques capable of handling the data in this valuable timeframe.

This is not the only conceptual approach to Big Data. A more sociological view tries to describe the movement with three “Cs”: Crumbs, Capacities and Communities (Letouzé & Jütting, 2015).

The Crumbs are a reference to the nature of the data collected in relation to the behavior of users in these new applications. These users leave behind traces of their activities while interacting and these traces, or crumbs, which constitute the databases to be analyzed in Big Data. Capacities are the techniques both statistical and programming used to manipulate these data and extract information. The third concept, Communities, refers to the behavior patterns of the producers of Big Data environments inside specific communities whose members share ideas with specific language and common validation methods. These communities can be established in open and collaborative environments, like OpenSource community, or at more restrictive ones, like tech groups in big corporations with access to big databases (Letouzé & Jütting, 2015).

* 1. Uber API in the Big Data context

The Uber’s estimation tool used in this article is by both definitions a Big Data tool. In relation to the three Vs, a huge volume of rides’ results data constantly feeding the tool’s algorithm with high speed interpretations of spatial non-structured data (Cohen et al., 2016).

At the same time, the data that feeds the algorithm are traces of drivers’ and users’ activity, or crumbs. The tools that process the data received online at a high rate are also tools included in the term “Capacities” developed by Big Data environment. In addition, the community of operators in the system of transport startups, particularly Uber, fits the “Communities concept”.

Also, the description of Big Data software architectures made by Pääkkönen and Pakkala (2015) suits very well with Uber’s APIs architecture and technology as described by its documentation ([https://developer.uber.com](https://developer.uber.com/)).

* 1. Accessibility

Discussing individuals' access to urban mobility in the city is an important subject. There are a number of different approaches from the use of traditional statistical surveys (METRO, 2008) to computational simulations of urban mobility behavior in a given urban territory (Krajzewicz, Erdmann, Behrisch, & Bieker, 2012). Also, approaches close to the present one, using social media data to trace mobility behavior (Noulas, Scellato, Lambiotte, Pontil, & Mascolo, 2012).

In big urban centers of size and regional importance comparable to São Paulo, knowing people mobility is essential to understand urban dynamics as the sheer size of the city feeds its capacity to attract more people. (Aranha, 2005). This fact has implications for mobility solutions for the locomotion demand. One of the main factors for measuring individuals’ capacity of moving inside the city is their routine travel times.

In São Paulo, this variable reflects socioeconomic contexts, indicating that access to mobility is differentially distributed in terms of space and social contexts (Morandi et al., 2016) . However, mobility as capacity of moving in the city is limited in the sense of not considering the conditions of locomotion.

A more comprehensive concept in this sense is accessibility (Litman, 2003; Páez et al., 2012; Stelder, 2016). It is possible to define accessibility as the potential to access spatial distributed opportunities (Wang & Mu, 2018). Hansen (1959) defines accessibility by the manner in which people interact with places. To illustrate, the amount of roads measured in kilometers in a determined region is a usual measure. Similarly in geography, accessibility is the measure which a person participates in determined activity (M. Kwan, 1998; Weibull, 1980).

Accessibility reflects spatial development that consists in transport network and distribution of opportunities, materialized in soil uses and occupations. A possible interpretation for this, is as a temporal measure (time to access) (Lessa et al., 2019; Páez et al., 2012). As a practical example, it is possible to consider the travel time to work as a comparative measure to understand the balance in job occupation and the racial, economic and gender disparities contained in urban area distributions (Preston & Mclafferty, 1999; Tribby & Zandbergen, 2012).

Geographers and social scientists have critically analyzed the economy and the inequality in transport and its correlations with socioeconomic inequalities. Schwanen (2016) arguments that transport distribution has sociospatial polarization intensified under capitalism dynamics as the transport infrastructure is an asset to attract capital and investments, bringing job opportunities, more efficiency and competition.

In discussing accessibility, Páez et al. (2012) define two epistemic approaches to study the concept: the first one is a normative approach defined in terms of which accessibility parameters are to be considered as reasonable, or in other words, how much is reasonable for a person to travel. The second approach is a positive one and is defined in terms of observed accessibility parameters, or how much people do travel. The normative approach analyses travel expectations while a positive one bases itself on the actual travel experience. The present study will take into consideration the positive approach as the data refers to actual Uber waiting times in São Paulo.

Studies such as that of Hughes and MacKenzie (2016), Lessa et al. (2019) and Wang and Mu (2018) address this positivist approach. In a more traditional fashion, Lessa (2019) investigates the relations between travel times collected in origin-destination surveys and transit data and the distribution of public transport network infrastructure. Hughes and MacKenzie (2016) approach to accessibility uses waiting times for Uber services and socioeconomic data in order to explore the spatial correlations between the dispersion of Uber's service and Seattle's socio-economic disparities. It can be measured by variables such as population density, average income per capita and percentage of non-whites. Wang and Mu (2018) do a cross-sectional study creating a spatial lag regression model that tests the relation of Uber waiting time with socioeconomic and transport infrastructure variables of the city of Atlanta.

As Hughes and MacKenzie (2016), Wang and Mu (2018) alerts, these preliminary studies are subject to economic and cultural influences from the researched location. In addition Letouzé and Jütting (2015), Schwanen (2016) and Kwan (2016) warn us about the possible existence of bias in databases from Big Data tools. This is extremely relevant to the discussion of the accessibility and use of databases of Big Data the investigation of Uber waiting time as a proxy for accessibility in São Paulo, since the particular economic and cultural context is relevant for the outcomes of the analysis.

Inspired by these studies, this article discusses the following hypotheses in the context of São Paulo:

H1: The estimated waiting time when requesting a ride in Uber’s platform can be used as a proxy for accessibility;

H2: Uber waiting time distribution relates to socioeconomic indicators’ polarization.

1. **METHODOLOGY**

This section begins with a description of the data collection of estimated Uber waiting times and socioeconomic data followed by a descriptive analysis to define the data clippings for final analysis (Wang & Mu, 2018). Then, the construction of the MLR models with stepwise method is explained. As the significant variables selected by this method were found by a Moran’s I test to have high spatial dependency, a final SAR model was calculated, as suggested by Wang & Mu (2018).

* 1. **Data and variables**

To verify the hypothesis elaborated two sets of data were collected. First, the data from Uber waiting times. Second, the spatialized socioeconomic data from São Paulo.

* + 1. **Uber Data**

The data was collected from the APIs present in Uber's Developers portal, accessible from the company’s site ([https://developer.uber.com](https://developer.uber.com/)). Moreover, covers the estimated waiting times of all Uber products during August 2018.

The city of São Paulo is sectioned in 96 districts, a territorial and administrative division that provides the local administration a certain degree of autonomy (Francisco, 2010). Following the methodology proposed by Wang and Mu (2018), the districts were used as the spatial unit of analysis. To guarantee that each district had at least one random sample point the software developed for the data collection followed the logic below:

* Section the city in quadricules of 1km², amounting to 1720 quadricules;
* Randomly sample a point in each quadricule;
* The sampled coordinates are used to consult Uber waiting times API;
* If the return is successful, the result is stored;
* If an error flag is raised, three more attempts with the same coordinate are made before the attempt is stored as a null.

This process was repeated each 30 minutes during August 2018. More than 2,528,400 calls to the API were stored being a little more than 2,240,000 valid calls.

To process the data collected, a C# software was developed following the concepts of cloud computing to access the Uber’s APIs. This program was allocated in the Azure cloud computing service provided by Microsoft. Its architecture follows diagram 1.

* + 1. **Socioeconomic Data**

The choice of socioeconomic indicators was based on similar studies (Hughes & MacKenzie, 2016; Wang & Mu, 2018). The analysis units chosen were the districts of São Paulo. Comparable variables were found for population density, employment density, minority rate, mean income per capita, motorization rate (for cars and motorcycles), public transport infrastructure and mean travel time to work. The socioeconomic data at the district level was collected from 2010 IBGE’s (Brazilian Institute of Geography and Statistics) demographic census, the National Ministry of Work an Employment (RAIS – Anual Report of Social information), the SEADE’s São Paulo State municipalities Indicators web portal and the São Paulo Municipal portal of Georeferenced information (http://geosampa.prefeitura.sp.gov.br).

The decennial Demographic census is one of Brazil's most important statistical products. The 2010 census presents two sets of data: the universe, that ideally comprehends every Brazilian, and the Sample, in which the respondents (a statistical fraction of the populations) are asked a more detailed survey. The sample fraction varies between municipalities: in São Paulo around 5% of the households were includes in the Sample, In whole Brazil 10,7% of the households were selected, or 6,192,332 households (I. BRASIL, 2010).

The RAIS is a Ministry’s of Work and Employment tool for management of Brazilian work relations. The data is compiled from statements filled by businesses about the working situation between them and their employees. The declarations are mandatory for a series of businesses nominated in law (M. do T. BRASIL, 2016).

SEADE foundation is a nationally recognized statistical institution. The foundation is known for its technical capacity. It is also responsible as part of National Statistic System for producing data, and aggregating existing data in the interests of São Paulo State and its municipalities (Francisco, 2010).

The São Paulo’s municipal online platform GEOSAMPA provides a series of geo-referenced data on a range of issues, including the distribution of transportation infrastructure.

All data was searched or aggregated to the district level (Francisco, 2010; Hughes & MacKenzie, 2016; Lessa et al., 2019; Wang & Mu, 2018). The 2010 census Sample was used to account for the percentage of population for each district that is in different classes of travel time to work/school, and also for the percentage of households in each district that possess a car and a motorcycle.

The socioeconomic data distribution follows a clear center periphery pattern with minor variations and is associated with the distribution of infrastructure and public policies (Francisco, 2010; Torres, Marques, Ferreira, & Bitar, 2003). The relations of this pattern with the distribution of Uber waiting times can shed some light in the relation of accessibility and sociospatial composition as attempted by Wang and Mu (2018).

* 1. **Descriptive Data Analysis**

The exploration of the 2,528,400 API calls showed two patterns: First, the general lack of estimates for some of Uber’s products, and second, the complete absence of calls in some regions of the city, as is shown in figure 2. It is possible to note the absence in the extreme south of the city and in a strip to the north, besides some spots spread over the city.

The identified causes were the presence of a mangrove area in the south that does not have road access and for the northern strip and the other spots over the city. Those are regions that Uber classifies as risk zones and does not provide service.

The cover of each service and its availability can be seen in table 2. It can be noted that the UberX, the most popular service Uber provides, has the largest coverage with almost 100% of time estimate responses registered. Consequently, it is the service that better reflects Uber’s fleet spatial distribution. It was considered coherent with this reasoning to adopt UberX waiting times for the purpose of measuring accessibility.

Filtering only data from the UberX product, it can be noted a great amplitude in the averages and the standard deviations between districts as can be seen in Figure 3. It is reasonable to consider that a district with a bigger fleet and more accessible shows less variation in waiting times. We could say more accessible districts show lower standard deviations (Wang & Mu, 2018). By this reason, two MLR models were created. The first having the average waiting time as a dependent variable, and the second having the standard deviation as the dependent variable.

* 1. **Multiple Linear Regression MLR**

From the principle of inference statistics, it is possible to make statements about characteristics of a population with a sample of it. The regression analysis is the term that describes a family of methods that permits exploring and infer the relation between two or more variables (Francisco, 2010; Hair, 2006).

For the construction of the MLR models, only the data from UberX aggregated by districts was used. The dependent variable for model 1 was the average waiting times and the dependent variable for model 2 was the standard deviation. The independent variables used were:

* Area (Km²)
* Population
* Population density
* Income per capita – Demographic Census (in reais)
* Jobs (Commerce, Services, Transformation Industry, Civil Construction)
* Employers (Commerce, Services, Transformation Industry, Civil Construction)
* Proportion of non-white residents (Black, *Pardos* and Indians)
* Travel time
* Rate of household car and motorcycle motorization
* Number of bus points
* Bus lines length
* Quantity of bus lines
* Number of Metro stations

The software R and its extensions “stats” and “car” were used for the confection of the regressive models.

* 1. **Spatial Auto Regressive Model – SAR**

Francisco (2010) suggests that before creating a spatial auto regressive model it is convenient to verify the spatial auto-correlation of the dependent variable. The literature uses a measure established by Moran. The Moran index is an indicator of the correlation between the value of the observed variable in a spatial unit of analysis and the values of that variable on the unit’s region (its neighbors).

After the verification of the geographic auto-correlation of Uber waiting times through the Moran’s I, the highly significant variables of the MLR model were selected and the spatial auto-regressive model (SAR) was calculated. Francisco (2010) defines SAR as a regression model capable of incorporate the spatial neighbors’ matrix (or spatial proximity) as a part of the explanatory variables.

The GeoDa software version 1.12 was used to make the SAR model.

**RESULTS**

Table 3 resumes the results of the MLR models with dependent variable as the average of UberX waiting times and standard deviation of waiting times. It is possible to note that average time model has a better degree of explanation as it has a R² of 0.893 against a R² of 0.717 for the standard deviation model.

A stepwise MLR model was created analyzing the collinearity of the independent variables of the final model, it was considered convenient to work with VIF's of less than 5 as recommended by Batterham et al. (1997). Besides that, the model selected was the mean of waiting time, due to its greater power of explanation. Table 4 shows the final result of the model. Meanwhile, Table 5 shows the collinearity analysis of the selected independent variables for this final model. We can observe a high degree of significance for bus lines length, district area, population density, jobs, percentage of non-whites and travel times over 120 minutes, with a R-squared 0.879. These elements corroborate with the positivist view of studies Lessa et al. (2019) and Páez et al. (2012).

Figure 4 shows the geographical dependence of the average Uber X waiting time, with a Moran I of 0.59, which shows the geographic slope of the dependent variable. In Figure 5, we note the distribution of the variable and its geographical dependence through the neighborhood as we can notice uniform areas with below time and low attendance.

From the variables in Table 4 we generated a spatial regression model for the average Uber X waiting time. The results presented in Table 6 show a high degree of significance for the variables bus lines length, district area, population density and percentage of non-whites. With the incorporation of the geographic factor in the model R-squared jumps to 0.89.

Observing the coefficients of the final SAR model, we can say that the number of bus lines (-0.073) and the population density (-0.0023) negatively influence Uber waiting time, as it is expected according to Wang and Mu (2018); the logic is that the highest concentration of population and its transport flow would attract the Uber drivers to these regions, boosting supply and thus providing a shorter waiting time for service.

In contrast, it is worth noting the coefficient of the nonwhite percentage (211.03) variable, which shows an increase in waiting time when the concentration of minorities is identified in the region. The possibility of this relationship - worsening travel and waiting times versus minority distribution - has been approached on a recurring basis in the literature on accessibility (Flores & Rayle, 2017; Hughes & MacKenzie, 2016; Páez et al., 2012; Wang & Mu, 2018). As suggested by Wang and Mu (2018), the economic cultural differences between the city of São Paulo and the city of Atlanta can be the motivator of the discrepancy between the results obtained. This occurrence motivates us to continue the debate on the regional idiosyncrasies of accessibility.

When analyzing the results, we cannot confirm H1, that is, we cannot affirm that the estimation of service waiting time can be used as an accessibility proxy. We assume it, since for some variables the relation with Uber X waiting time was not significant.

We can confirm H2 for the distribution of the quantity and number of bus lines, which corroborates (Lessa et al., 2019) as a measure of accessibility, district area, population density and percentage of nonwhites. In studies of Hughes and MacKenzie (2016), Wang and Mu (2018), the relationship between Uber's waiting time and the distribution of minorities was not significant.

1. **DISCUSSION AND CONCLUSION**

The final model presents different variables from similar studies. Minority rate, for instance, presents a high significance in the model with coefficient 0.00, in contrast with the cases of Seattle (Hughes & MacKenzie, 2016) and Atlanta (Wang & Mu, 2018). The relevance of local context is important to explain the findings in the article. As waiting times can have a relation to consistent supply and demand of cars (Hall et al., 2019), one possible explanation is that, as the socioeconomic spatial pattern (Torres et al., 2003) also goes along with minority rates and peripheral status, the service tends to be less accessible to these regions – in which its demand could be lower. In the case of São Paulo, the correlation between minority rates and income encourage us to at least think on this possibility, as the affordability of Uber can be relatively lower than in the contexts of Seattle or Atlanta. Future analysis using data from Uber cost estimate can help exploring this pattern.

On the other hand, some variables as population density (Hughes & MacKenzie, 2016; Lessa et al., 2019; Wang & Mu, 2018) and road density (Wang & Mu, 2018), find some resonance in the discussion. Even their being far from enough to explain accessibility and being aware of the local nature of the problem (Schwanen, 2017), it is possible to propose them as variables that can help to describe local variables. This possibility exists since their relation to accessibility appear to be consistent.

The need to promote intuitive and highly communicable accessibility measures is latent between researchers (Páez et al., 2012). Likewise, the construction of an transport accessibility measure from Big Data tools as Uber’s API can contribute to the communication and understanding of this indicator for the general public as the theory gets close to a service of mass consumption. The expansion of the same analysis to other service providers and a more precise identification of this market share in the transportation field in São Paulo can also help to identify possible biases in using Uber’s tool.

Public management can make use of similar tools to develop a “basket” of indicators for different modes of transportation and measurements regarding their interrelations. These can be used to better regulate existing activity – by maintaining a real time measurement of the effect of public regulation in aspects of the transport system, it could be possible to better adjust intervention in the interests of the users – and to tailor public service delivery in mobility - as with a more intimate integration between public and private modes (Jin et al., 2018), given that well balanced indicators can reflect user behavior and needs of users (Lessa et al., 2019; Páez et al., 2012)

However, some limitations of the study deserve attention, as they provide indications for future research. It is important to underline the importance of more comprehensive analysis. This study limited itself to Uber’s fleet, in cities like São Paulo where there are at least one more player with a significant fleet. It will be interesting to reproduce the methodology to other Companies that provide similar service. Second, the replication of the study in other Brazilian capitals makes itself necessary as the economic and cultural factors that affect the disparities can be analyzed. Likewise, as their contexts are much more similar than Atlanta and Seattle, for an instance. Third, a deepening in the Uber’s time estimate tool understanding to identify possible biases that could have suppressed any relation. Comparing it to other services can be useful in this sense. A better comprehension of this data can be a major advance in tools for municipalities to better serve their populations transportation needs. Also, it is worth revisiting this study after the publication of 2020 IBGE Census. It is valid since many of the variables used in the study are from 2010 Census and projection from it.

**REFERENCES**

Aranha, V. (2005). *Mobilidade pendular na metrópole paulista*. 96–109. https://doi.org/http://dx.doi.org/10.1590/S0102-88392005000400006

Batterham, A. M., Tolfrey, K., George, K. P., Alan, M., Tolfrey, K., & Keith, P. (1997). Nevill’s explanation of Kleiber’s 0.75 mass exponent: an artifact of collinearity problems in least squares models? *Journal of Applied Physiology*, (3), 693–697.

BRASIL, I. (2010). Censo Demogŕafico. *Notas Metodológicas–Microdados Da Amostra. Rio de Janeiro*.

BRASIL, M. do T. (2016). Relação Anual de Informações Sociais - RAIS ano-base 2016. *Portaria n. 1464 de 30 de Dez. 2016. Aprova Instruções Para a Declaração Da Relação Anual de Informações Sociais - RAIS Ano-Base 2016*.

Brynjolfsson, E., & McAfee, A. (2012). Big Data : The Management Review. *Harvard Business Review*, (October), 1–12. Retrieved from http://tarjomefa.com/wp-content/uploads/2017/04/6539-English-TarjomeFa-1.pdf

Cohen, P., Metcalfe, R., Angrist, J., Chen, K., Doyle, J., Farber, H., … Hahn, R. (2016). Using big data to estimate consumer surplus: The case of uber. *National Bureau of Economic Research.*

Flores, O., & Rayle, L. (2017). How cities use regulation for innovation: The case of Uber, Lyft and Sidecar in San Francisco. *Transportation Research Procedia*, *25*, 3756–3768. https://doi.org/10.1016/j.trpro.2017.05.232

Francisco, E. de R. (2010). Indicadores de renda baseados em consumo de energia elétrica: abordagens domiciliar e regional na perspectiva da estatística espacial. *Doctoral Dissertation*.

Hair, J. F. (2006). *Multivariate data analysis.* (P. E. India., Ed.).

Hall, J. V, Horton, J. J., & Knoepfle, D. T. (2019). *Pricing Efficiently in Designed Markets : The Case of Ride-Sharing*. 1–76.

Hall, J. V, & Krueger, A. B. (2016). An analysis of the labor market for Uber’s driver-partners in the United States. In *National Bureau of Economic Research*. https://doi.org/https://doi.org/10.1177/0019793917717222

Hansen, W. G. (1959). How accessibility shapes land use. Journal of the American Institute of planners. *Journal of the American Institute of Planners*.

Hughes, R., & MacKenzie, D. (2016). Transportation network company wait times in Greater Seattle, and relationship to socioeconomic indicators. *Journal of Transport Geography*, *56*, 36–44. https://doi.org/10.1016/j.jtrangeo.2016.08.014

Jin, S. T., Kong, H., Wu, R., & Sui, D. Z. (2018). Ridesourcing , the sharing economy , and the future of cities. *Cities*, *76*(January), 96–104. https://doi.org/10.1016/j.cities.2018.01.012

Kim, B. G., Trimi, S., & Chung, J. (2014). Big-Data Applications in the Government Sector. *Communications of the ACM*. https://doi.org/http://dx.doi.org/10.1145/2500873

Krajzewicz, D., Erdmann, J., Behrisch, M., & Bieker, L. (2012). Recent development and applications of SUMO-Simulation of Urban MObility. *International Journal On Advances in Systems and Measurements*.

Kwan, M. (1998). Space-time and integral measures of individual accessibility: A com- parative analysis using a point-based framework. *Geographical Analysi*.

Kwan, M. P. (2016). Algorithmic geographies: Big data, algorithmic uncertainty, and the production of geographic knowledge. *Annals of the American Association of Geographers*.

Lessa, D. A., Lobo, C., & Cardoso, L. (2019). Accessibility and urban mobility by bus in Belo Horizonte / Minas Gerais – Brazil. *Journal of Transport Geography*, *77*(May 2018), 1–10. https://doi.org/10.1016/j.jtrangeo.2019.04.004

Letouzé, E., & Jütting, J. (2015). Official Statistics, Big Data and Human Development: Towards a New Conceptual and Operational Approach. *Data Pop Alliance White Paper Series*.

Litman, T. (2003). Measuring Transportation: Traffic Mobility and Accessibility. *Victoria Transport Policy Institute*.

METRO. (2008). Pesquisa Origem-Destino 2007. *Secretaria de Transportes Metropolitanos*.

Morandi, E., Ribeiro, R., Hernandes, E., Camara, B., Spinola, L., & Francisco, E. D. R. (2016). *Análise Geoespacial da Relação entre Transporte Público sobre Trilhos, Renda e Tempo Médio de Deslocamento*.

Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., & Mascolo, C. (2012). A Tale of Many Cities : Universal Patterns in Human Urban Mobility. *PloS One*, *7*(5). https://doi.org/10.1371/journal.pone.0037027

Páez, A., Scott, D. M., & Morency, C. (2012). Measuring accessibility: Positive and normative implementations of various accessibility indicators. *Journal of Transport Geography*, *25*, 141–153. https://doi.org/10.1016/j.jtrangeo.2012.03.016

Pekka, P., & Daniel, P. (2015). Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems. *Big Data Research*, Vol. 2, pp. 166–186. https://doi.org/10.1016/j.bdr.2015.01.001

Preston, V., & Mclafferty, S. (1999). Spatial mismatch research in the 1990s : progress and potential. *Papers in Regional Science*, *402*, 387–402.

Schwanen, T. (2016). Geographies of transport I: Reinventing a field? *Progress in Human Geography*, *40*(1), 126–137. https://doi.org/10.1177/0309132514565725

Schwanen, T. (2017). Geographies of transport II: Reconciling the general and the particular. *Progress in Human Geography*, *41*(3), 355–364. https://doi.org/10.1177/0309132516628259

Stelder, D. (2016). Regional Accessibility Trends in Europe : Road Infrastructure , 1957 – 2012. *Regional Studies*, *50*(6), 983–995.

Torres, H. da G., Marques, E., Ferreira, M. P., & Bitar, S. (2003). Pobreza e espaço : padrões de segregação em São Paulo. *Estudos Avançados*, *17*(47), 97–128. https://doi.org/http://dx.doi.org/10.1590/S0103-40142003000100006

Tribby, C. P., & Zandbergen, P. A. (2012). Author ’ s personal copy High-resolution spatio-temporal modeling of public transit accessibility. *Applied Geography Journal*, *34*, 345–355. https://doi.org/10.1016/j.apgeog.2011.12.008

Wang, M., & Mu, L. (2018). Spatial disparities of Uber accessibility: An exploratory analysis in Atlanta, USA. *Computers, Environment and Urban Systems*, *67*, 169–175. https://doi.org/10.1016/j.compenvurbsys.2017.09.003

Weibull, J. W. (1980). On the numerical measurement of accessibility. *Environment and Planning A*.

Zhou, X., Wang, M., & Li, D. (2017). From stay to play–A travel planning tool based on crowdsourcing user-generated contents. *Applied Geography*.

Table 1: São Paulo’s Socioeconomic data

|  |  |  |
| --- | --- | --- |
| District data | Source | Date |
| Area (km²) | Data from IBGE. Collected from the portal “Indicadores dos Municípios Paulistas” (IMP) – SEADE Foundation | 2009 |
| Population | Original data from IBGE’s demographic census annualy readjusted by SEADE Foundation. Collected from the portal “Indicadores dos Municípios Paulistas” (IMP) – SEADE Foundation | 2010/2018 |
| Population Density | Computed from area and population data | 2010/2018 |
| Permanent particular households | Original data from IBGE’s demographic census annualy readjusted by SEADE Foundation. Collected from the portal “Indicadores dos Municípios Paulistas” (IMP) – SEADE Foundation | 2010/2018 |
| Income per Capita - Demographic Census (In current reais) | Original data from IBGE’s demographic census. Collected from the portal “Indicadores dos Municípios Paulistas” (IMP) – SEADE Foundation | 2010 |
| Jobs (Commerce, Services, Transformation Industry, Civil Construction) | São Paulo’s municipal portal “Infocidade”. Original data source: Ministry of Work and Employment - Annual Report of Social Information (RAIS) | 2010/2016 |
| Employers (Commerce, Services, Transformation Industry, Civil Construction) | São Paulo’s municipal portal “Infocidade”. Original data source: Ministry of Work and Employment - Annual Report of Social Information (RAIS) | 2010/2016 |
| % of non-whites (Black, Pardos, Indians) | IBGE’s 2010 Demographic Census | 2010 |
| Travel times | IBGE’s Demographic Census Sample. Proportion of people, weighed, in each of the survey’s class of travel time, by district. | 2010 |
| Household motorization rate | IBGE’s Demographic Census Sample. Motorization rate by district (cars and motorcycles) from the weighed households of the Sample | 2010 |
| Bus Stops | Geosampa Portal | 2018 |
| Extension of Bus Lines (Km) | Geosampa Portal | 2018 |
| Metro Stations | Geosampa Portal | 2018 |

Tabela 2: Returns of calls of Uber eaiting time estimates, by product

|  |  |  |  |
| --- | --- | --- | --- |
| Service | Number of successful calls | Average of estimated waiting time (seconds) | Standard deviation of estimated waiting time (seconds) |
| Bag | 1,327,710 | 400 | 187.2 |
| Bike Rack | 58,758 | 598 | 280.6 |
| Black | 987,605 | 491 | 215.5 |
| Black Bag | 549,919 | 506 | 220.6 |
| Pool | 361,166 | 200 | 87.9 |
| Select | 1,713,733 | 340 | 170.6 |
| Uber X | 2,233,720 | 377 | 277.9 |

Table 3: Average and standard deviation regressions result for UberX

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| DV | | Average Uber X | | |  | Std. Deviation Uber X | | |
| Coeficient | Std. Error | p-value |  | Coeficient | Std. Error | p-value |
| (Intercept) |  | 943.00 | 750.40 | 0.213 |  | 39.35 | 515.30 | 0.939 |
| QTLINBUS2018 | Bus lines' quantity | 0.13 | 0.19 | 0.514 |  | 0.12 | 0.13 | 0.353 |
| KMLINBUS2018 | Bus lines' legth (Km) | -0.12 | 0.07 | 0.080 |  | -0.09 | 0.05 | 0.057 |
| QTPONTBUS2018 | Number of bus stops | -0.12 | 0.13 | 0.361 |  | 0.01 | 0.09 | 0.902 |
| QTESTMETRO2018 | Quantity of Metro stations | 0.91 | 5.22 | 0.862 |  | -1.14 | 3.58 | 0.751 |
| RENDP2010 | Income per Capita | -0.00 | 0.01 | 0.922 |  | 0.00 | 0.01 | 0.515 |
| ARE1 | Area (km2) | 3.55 | 0.31 | 0.00 |  | 0.74 | 0.21 | 0.001 |
| POP2018 | Population | 0.00 | 0.00 | 0.492 |  | -0.00 | 0.00 | 0.894 |
| DENPOP2018 | Population density | -0.00 | 0.00 | 0.007 |  | -0.01 | 0.00 | 0.000 |
| DOMP2018 | Number of particular permanent households | -0.00 | 0.00 | 0.646 |  | 0.00 | 0.00 | 0.546 |
| ESTAB2016 | Employers | -0.00 | 0.01 | 0.970 |  | -0.00 | 0.00 | 0.791 |
| EMP2016 | Jobs | 0.00 | 0.00 | 0.384 |  | 0.00 | 0.00 | 0.547 |
| PNBRAN2010 | Proportion of non-whites | 300.20 | 77.69 | 0.000 |  | 202.20 | 53.35 | 0.000 |
| TEMP2010\_5MIN | Travel time - up to 5 minutes | -587.50 | 1,166.00 | 0.616 |  | 111.00 | 800.50 | 0.890 |
| TEMP2010\_30MIN | Travel time - from 6 to 30 minutes | -978.90 | 785.70 | 0.217 |  | 37.68 | 539.50 | 0.945 |
| TEMP2010\_60MIN | Travel time - from 31 to 60 minutes | -278.40 | 831.70 | 0.739 |  | 180.80 | 571.10 | 0.752 |
| TEMP2010\_120MIN | Travel time - from 61 to 120 minutes | -1,046. | 805.20 | 0.198 |  | -141.10 | 552.90 | 0.799 |
| TEMP2010\_121MIN | Travel time - more than 121 minutes | -1,443. | 992.80 | 0.150 |  | -215.90 | 681.80 | 0.752 |
| TEMP2010\_0MIN | No travel | -821.30 | 765.00 | 0.286 |  | 10.86 | 525.30 | 0.984 |
| PDOMC2010 | Car - rate of household motorization | -139.10 | 367.70 | 0.706 |  | -164.10 | 252.50 | 0.518 |
| PDOMM2010 | Motorcycle - rate of household motorization | 37.73 | 82.42 | 0.648 |  | 25.17 | 56.60 | 0.658 |

Table 4: Stepwise regression results for UberX average waiting times

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| DV | | Average Uber X | | |
| Coeficient | Std. Error | p-value |
| (Intercept) |  | 171.00 | 18.17 | 0.000 |
| KMLINBUS2018 | Bus lines' legth (Km) | - 0.0786 | 0.02 | 0.002 |
| ARE1 | Area (km2) | 3.634 | 0.23 | 0.000 |
| DENPOP2018 | Population density | - 0.0035 | 0.0009 | 0.000 |
| EMP2016 | Jobs | 0.00021 | 0.0001 | 0.097 |
| PNBRAN2010 | Proportion of non-whites | 295.70 | 44.57 | 0.000 |
| TEMP2010\_121MIN | Travel time - more than 121 minutes | - 1,239,00 | 422.50 | 0.004 |

Table 5: Variance inflation factors analysis

|  |  |  |
| --- | --- | --- |
| VIF | | Average Uber X |
| Coeficient |
| KMLINBUS2018 | Bus lines' legth (Km) | 1.181254 |
| ARE1 | Area (km2) | 2.139971 |
| DENPOP2018 | Population density | 1.317234 |
| EMP2016 | Jobs | 1.894758 |
| PNBRAN2010 | Proportion of non-whites | 2.462318 |
| TEMP2010\_121MIN | Travel time - more than 121 minutes | 2.615456 |

Table 6: Spatial lag regression results of Uber X

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| DV | | Average Uber X | | |
| Coeficient | Std. Error | p-value |
| W\_MEDIA |  | 0.406 | 0.092 | 0.000 |
| CONSTANT |  | 114.573 | 23.246 | 0.000 |
| KMLINBUS2018 | Bus lines' legth (Km) | - 0.073 | 0.022 | 0.001 |
| ARE1 | Area (km2) | 2.455 | 0.263 | 0.001 |
| DENPOP2018 | Population density | - 0.002 | 0.001 | 0.005 |
| PNBRAN2010 | Proportion of non-whites | 211.038 | 44.297 | 0.000 |

Figure 1: Diagram of the Uber data collection application.

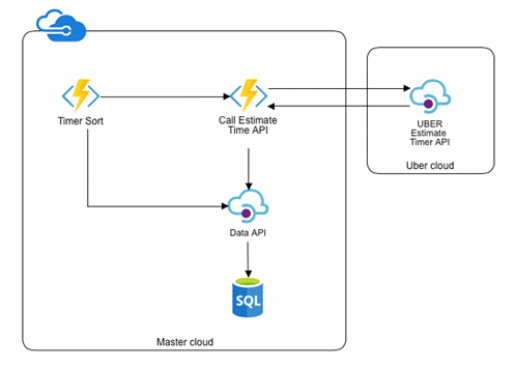


Figure 2: Coordinate points for Uber waiting time estimates

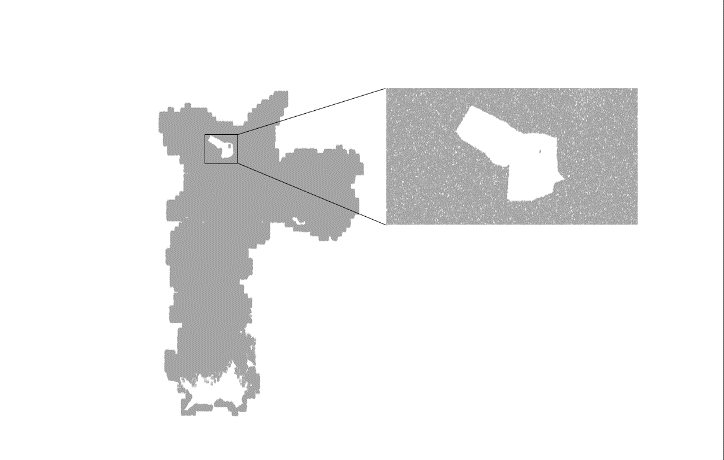
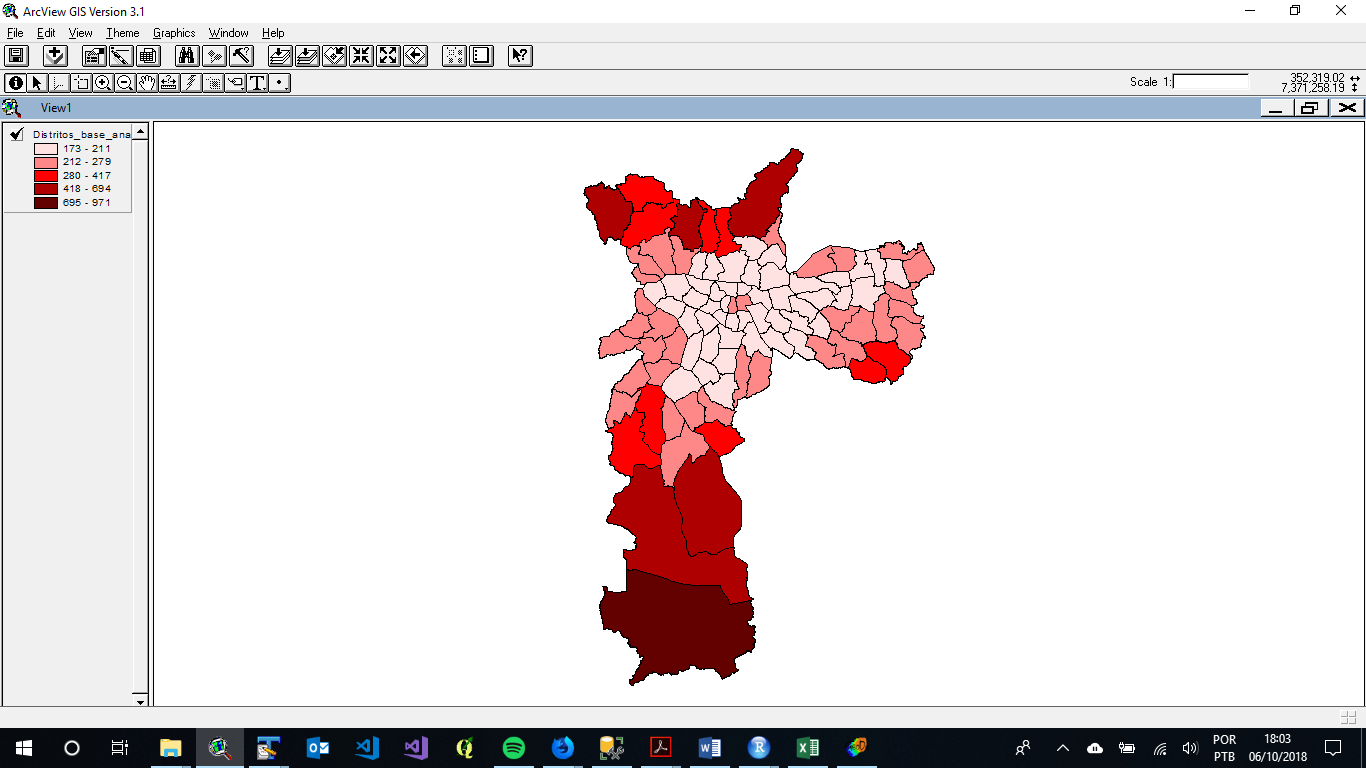


Figure 3. Comparison between average and standard deviation of UberX estimate waiting times



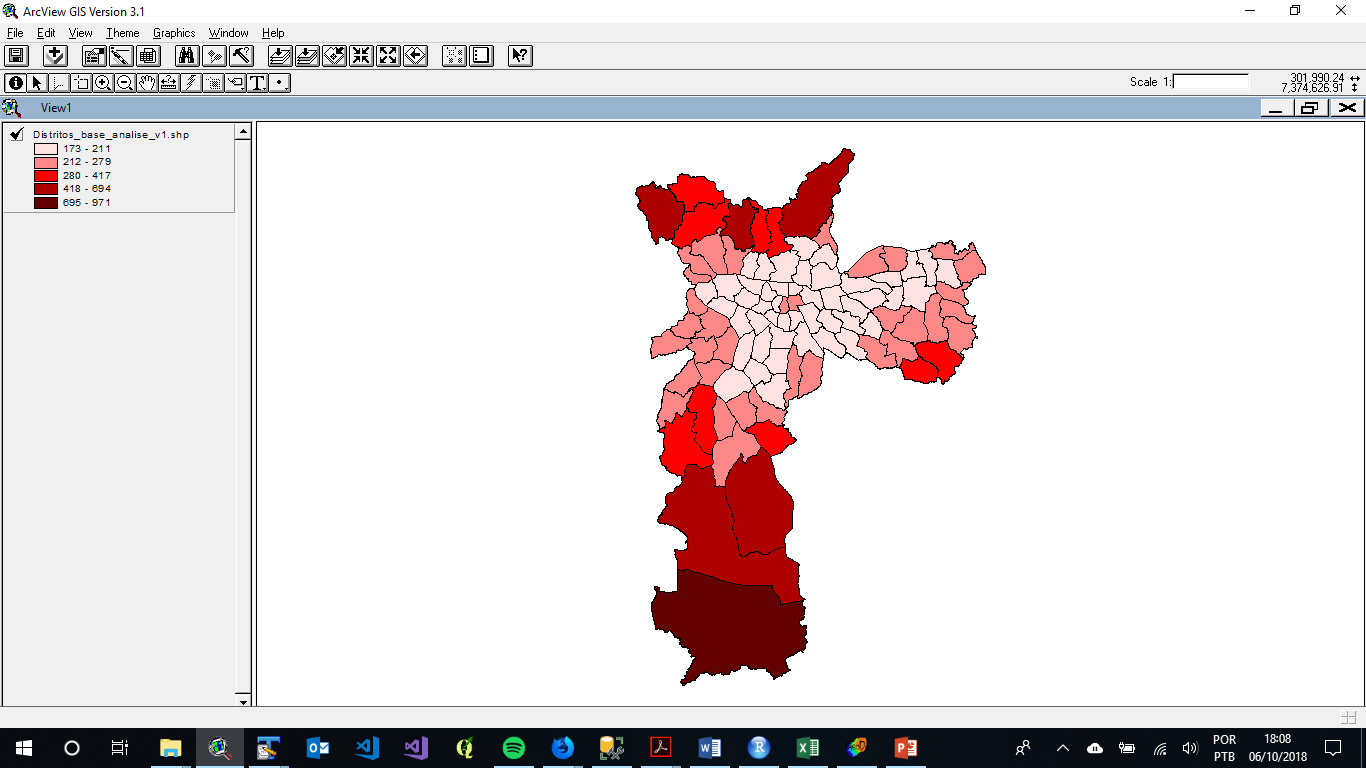
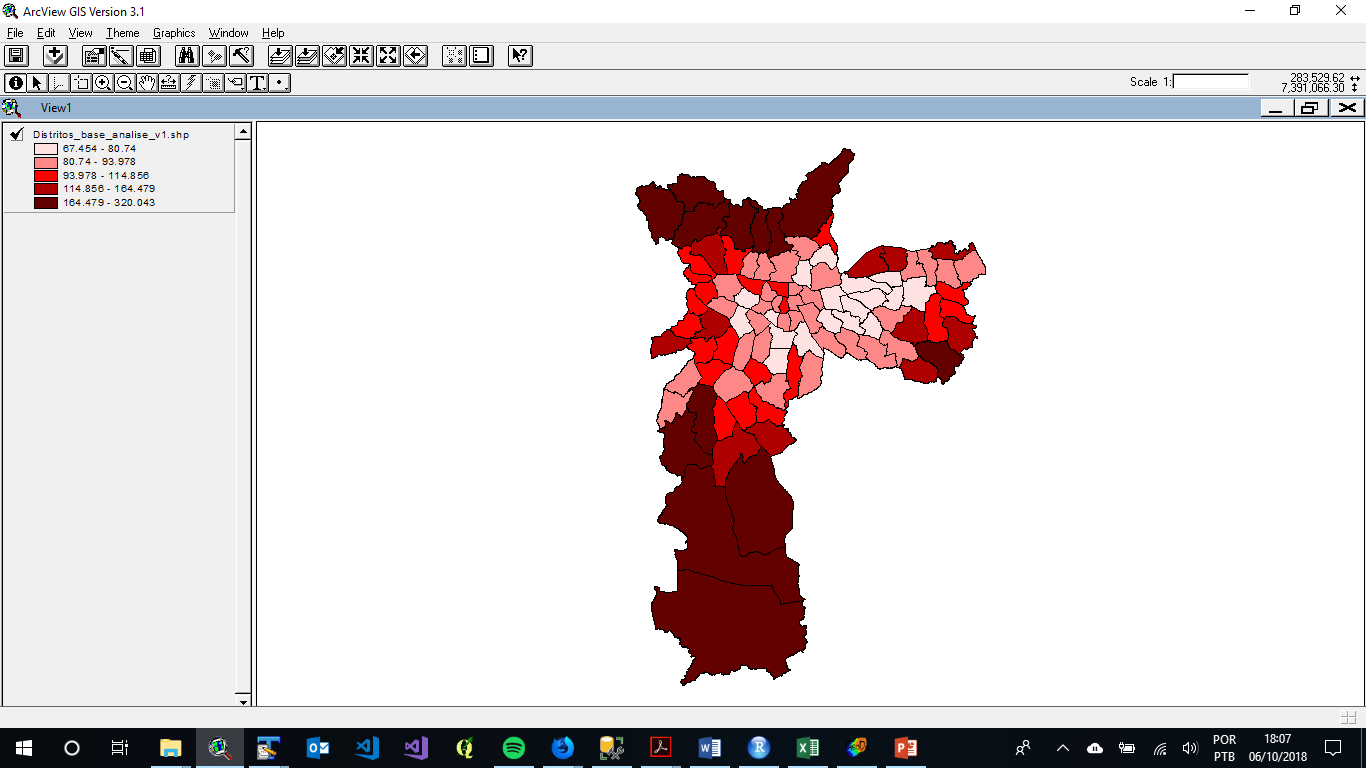
Average waiting time (s) Standard deviation-waiting times (s)

Figure 4. Uber X average waiting time Moran’s I

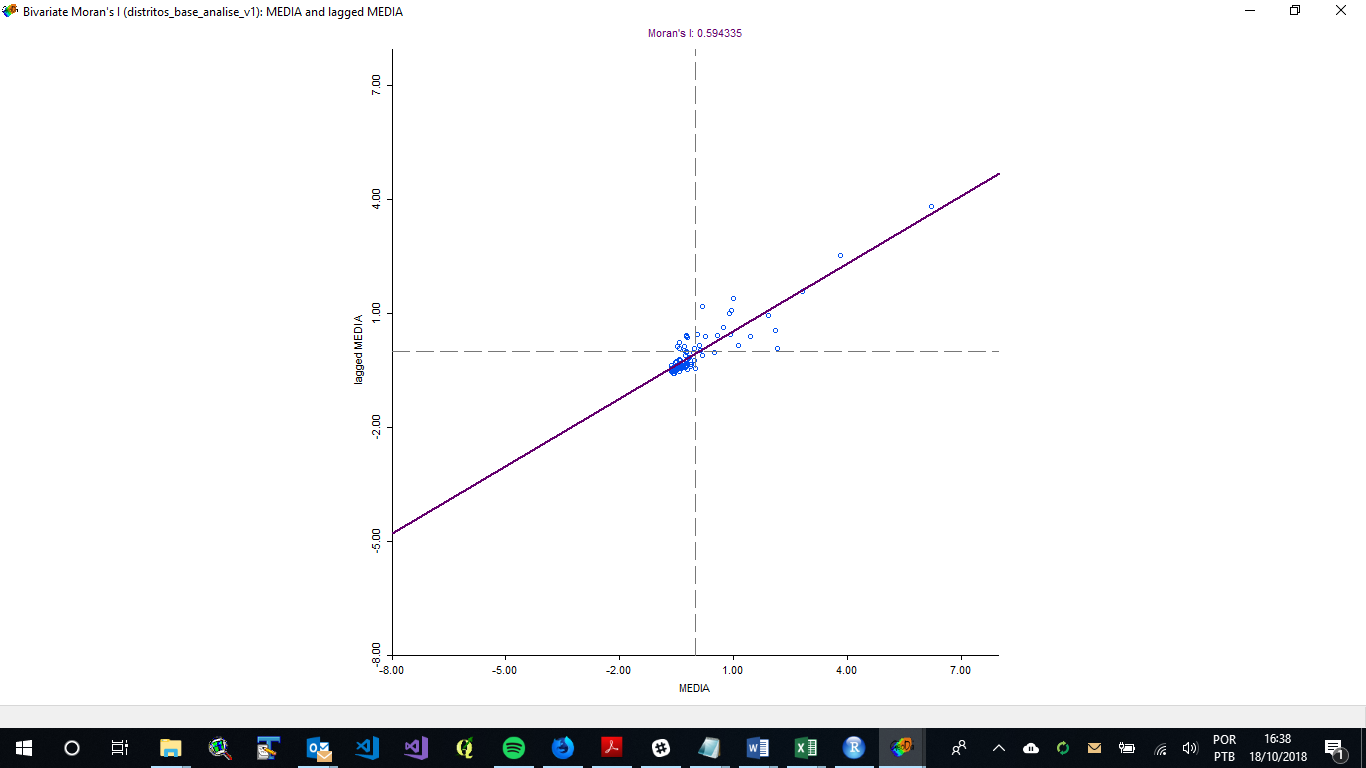
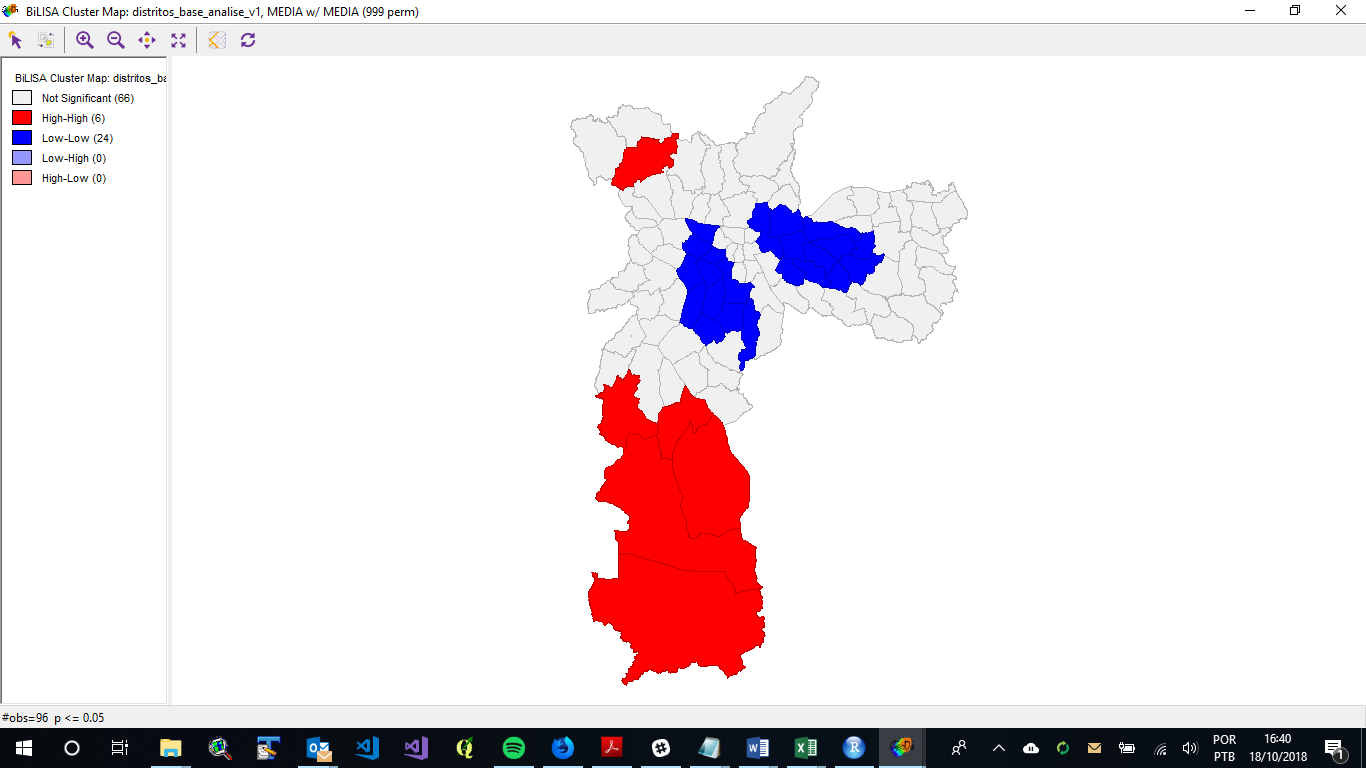
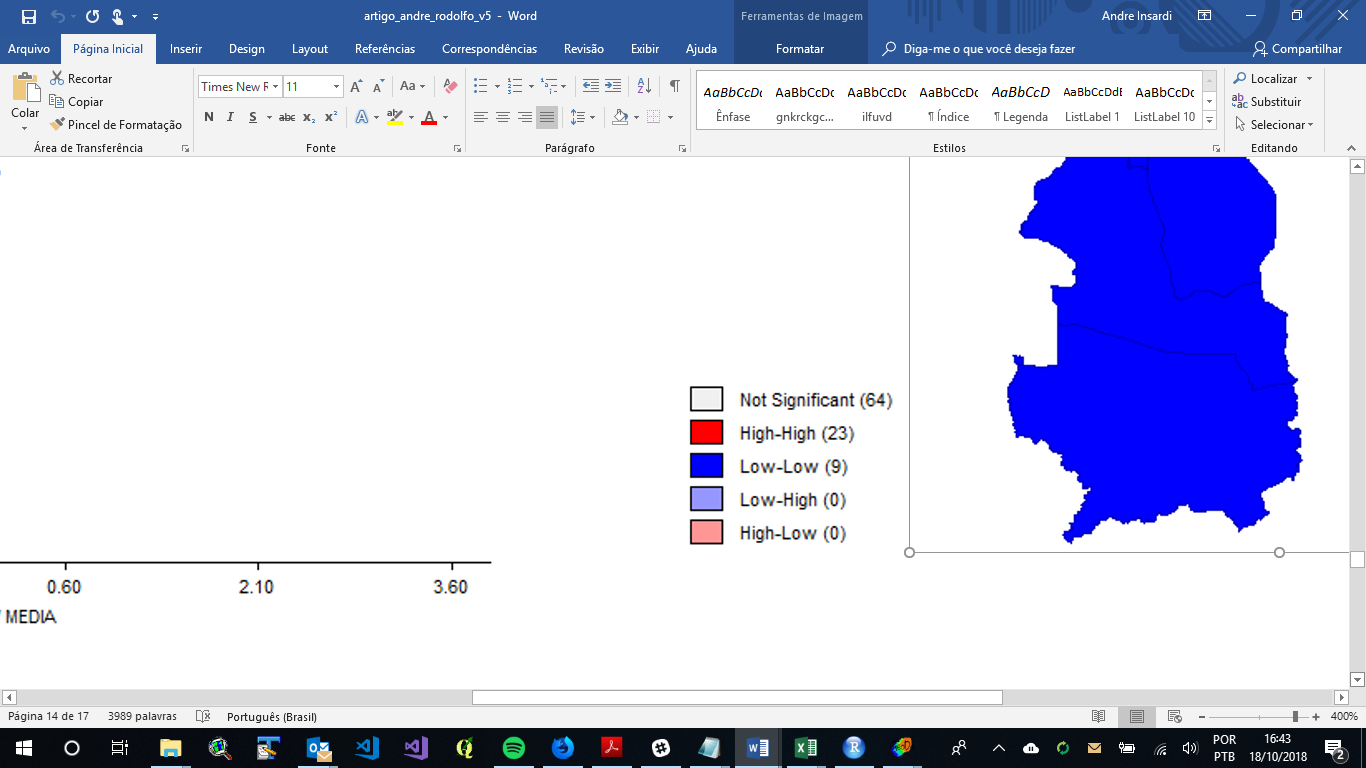


Figura 5. Uber X average waiting time neighbourhood clusters





1. http://www.reuters.com/article/2010/11/05/ozatp-ghana-economy-idAFJOE6A40BG20101105 6 [↑](#footnote-ref-0)