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** (250 palavras)

**Palavras-chaves** (5 palavras)

# Introdução (750 palavras)

# Teoria (2.000 palavras)

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 deste. 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 desse 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 distâncias consideravelmente grande ao centro (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 sim nos anos 2000 os indicadores das periferias apresentavam números consideravelmente piores que 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 em muito dos municípios vizinhos terem desenvolvido importantes economias, 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 4o município de São Paulo (Aranha, 2005) - 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), o que gerou 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 processo 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. 5Ainda, 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 atual, 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-2). A dificuldade desses países de 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 outras (Letouzé e Jutting, 2014). Além disso, nesse cenário de fragilidade e coloca a crescente produção de dados e as novas formas de análises estatísticas baseadas nesse “movimento”: Big Data.

Uma das primeiras definições de Big Data está relacionada às caracterśiticas 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, conta com inovações computacionais e três grande conceitos definidores em relação ao dados envolvidos: Volume, Velocidade e Variedade (McAfee et al., 2012; Gandomi e Haider, 2015). De acordo com essa definição, o que caracteriza Big Data é não só o volume dos dados envolvidos, mas também a velocidade de produção de dados, com aplicações os análise de dados produzidos em tempo real, e a variedade de formatos, com o uso de dados estruturados e não estruturados . Ainda nessa direção 1existem definições que incluem 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). Existem outras definições de Big Data, que partem de outros pressupostos. Letouzé e Jütting (2014) o definem a partir de caracteŕisticas “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 automaticamente 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, e além disso de 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, é interessante que haja movimentos de integração entre as estatísticas oficiais e essas novas técnicas de análise.

De particular interesse para o presente trabalho, a produção de dados georreferenciados relativos a 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, com 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. 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 - 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 viéses deccorentes do uso de algoritmos de Big Data, que mesmo não sendo particularidades dessas estratégias, tem o potencial de interferência nos dados multiplicado pela dependencia dessas estratégias em uma grande quantidade de dados processados 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 descernir os possíveis viéses 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 8periféricos 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 oportunidade 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 ele 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 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 até o momento 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 em função do número de oportunidades W no local j dado dentro de uma função de atratividade g, multiplicada 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 a população de p.

Para os indicadores gravitacionais, a função g é uma função de atratividade do local j que é dada em função da 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 um 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 mesmas casa, 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. Além disso, 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 (1.000 palavras)

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

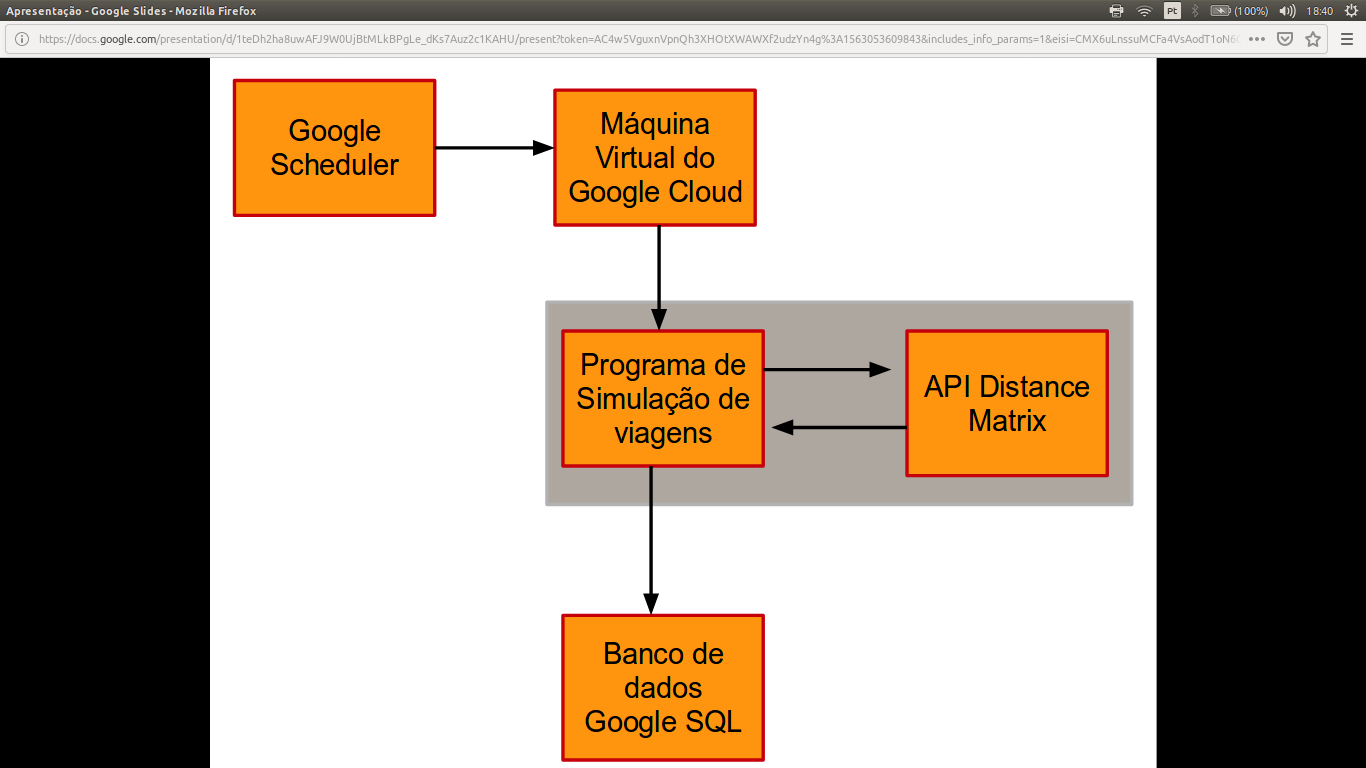
Simulação de viagens

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

* A relativa alta complexidade de simulações locais que consigam captar o comportamento em tempo real – com informações de trânsito – do trafêgo implicava o uso de alguma ferramenta de previsão de tempos de viagem já estabelecida e acessivel por meio remoto.
* A escolha feita (pela API DistanceMatrix da Google) implicou em um número limitado de requisições de viagens para manter o caráter gratuito das viagens
* Essa limitação implicou em escolhas para reduzir o numero de viagens “perdidas” na simulação, incorridas quando as coordenadas usadas na API não correspondiam ou não podiam ser aproximadas a endereços válidos
* 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. Essa consideração é importante pois o expediente mais recomendado seria uma distribuição aleatória, ou o mais aleatória possível. 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:

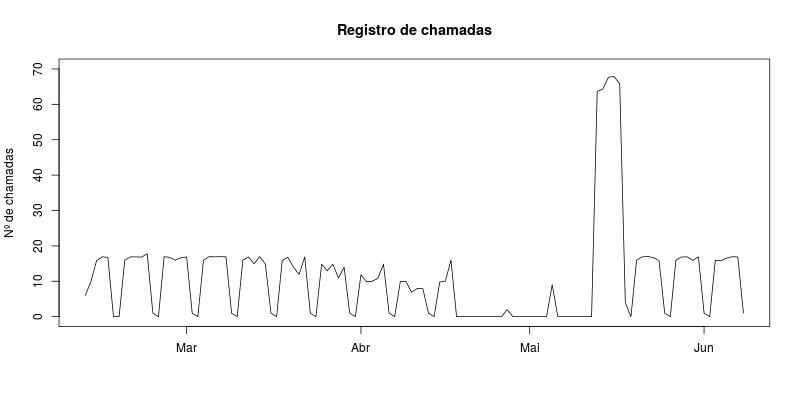
1. Usando o software Qgis, foi gerada, a partir do shapefile do município de São Paulo censo de 2010, uma grade com quadrículas de 500m
2. A partir da informação da população por setor censitário, foram calculadas as populações de cada quadriculas e foram retiradas as quadriculas com população zero.
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 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 da maquina virtual no ambiente em nuvem da Google; por sua vez a maquina 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 manha até às nove da noite. A intenção do espaçamento era obter uma amostragem de diferentes horário para comparar peridiocidades diárias e horárias nas viagens. Houve alguns problemas nesse agendamento, que serão abordados mais a frente. O programa iniciado pela máquina virtual segue o seguinte procedimento:

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 resultado 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 de 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 carregava 200 viagens pareadas (100 de transporte público e 100 de transporte privado). O periodo de simulação abrangeu o período entre 11 de Fevereiro de 2019 até 5 de Junho de 2019. Da metade do mês de março até a metade do mês de abril, como indicado no Gráfico 1, 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 (dez e cinco minutos antes das horas cheias, nas horas cheias e cinco minutos após as horas cheias). Nas duas ultimas 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 analise, sendo a primeira etapa uma análise do agregado de todas as viagens e a segunda etapa uma análise exploratório 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 presença de infraestrutura de transporte nos distritos do município de São Paulo. Foram montado modelos de regressão simples (OLS), modelos de regressão de autocorrelação 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 processamentode apresentam a seguinte estrutura:

| ID | Data | Hora | Dia | Latitude da origem | Longitude da origem | Endereço da origem |
| --- | --- | --- | --- | --- | --- | --- |
| Latitude do destino | Longitude do destino | Endereço do destino | Duração | Distância | 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 (publica e privada) 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 distancia 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 centrada no ponto a ser calculado e pela distância euclidiana até os pontos de destino de viagens próximos. 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 pra das áreas de ponderação. A partir dessas análises foi possivel verificar a clusterização dessas medidas no município. Os resultados para as duas medidas foram comparados assim como as diferenças entre os niveis de análise de distritos e de áreas de ponderação.

(explicar melhor os I’s de Moran e os mapas LISA)

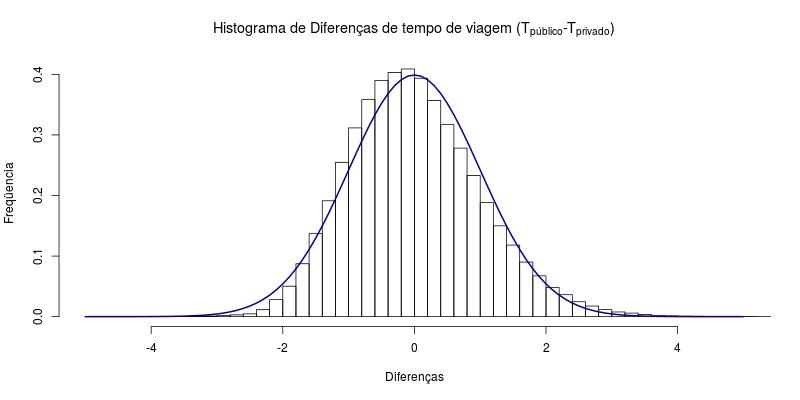
Por ultimo, foram realizadas as modelagem de regressões lineares. Para isso as medidas Dt e Rt foram agrupadas em torno dos distritos de origem e para cada distrito foi considerada a média das medidas que partiam do distrito. Isso forneceu para cada medida um conjuntos de dados georrefenciados no distrito de origem das viagens. Para a modelagem a partir desses conjuntos de dados foi usada uma base de dados semelhante à usada no artigo anexo, mas modificada; foram compilados dados socioeconomicos e de infraestrutura de transportes agregados por distrito – alguns dados relativos ao tempo de viagens foram descartados por apresentarem muita semelhança com os formato da variável dependente. As variáveis usadas estão descritas na tabela 1. A partir desses dados foi calculado um modelo de regressão linear, cujas variáveis foram reduzidas através de um proceso stepwise, seguida da retirada de variáveis ainda insignificantes e de variáveis colineares.

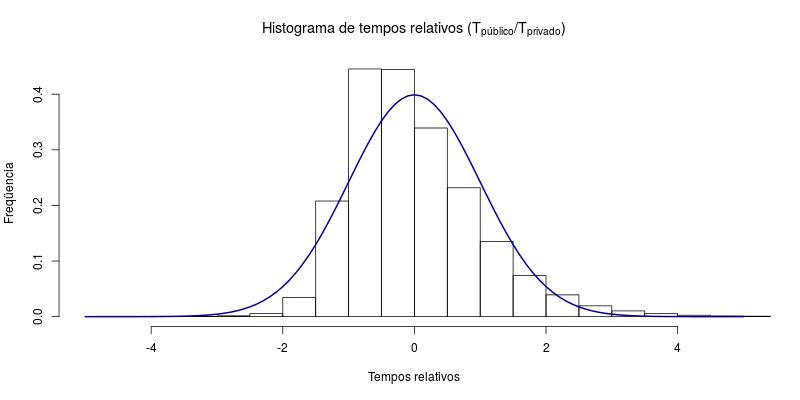
Também foram calculados dois modelos de regressão espacial: um de autocorrelação espacial (SAR), no qual uma matriz de vizinhança foi definida para os distritos e um peso associado aos valores dos viznhos foi adicionado à regressão, de forma que foram corrigidas as dependencias espaciais da medida. O segundo modelo usou a técnica de ponderação espacial (GWR), na qual foram calculados modelos lineares para cada distrito considerados as medidas de distritos viznhos ponderados pela distancia ao modelo do distrito.

|  |  |  |
| --- | --- | --- |
| Dado do distrito | Fonte | Data |
| Área (km2) | Dados do IBGE. Retirado do portal de indicadores dos municípios paulistas (IMP) Fundação SEADE | 2009 |
| População | 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 | 2010/2018 |
| Densidade | Cálculo a partir dos dados de Área e População | 2010/2018 |
| 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 | 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 |
| Empregos (Comércio, Serviços, Indústria de Transformação, Construção Civil) | Portal Infocidade do município de São Paulo. Fonte original dos dados: Ministério do Trabalho e Emprego. Relação Anual de Informações Sociais – Rais. | 2010/2016 |
| Estabelecimentos (Comércio, Serviços, Indústria de Transformação, Construção Civil) | Portal Infocidade do município de São Paulo. Fonte original dos dados: Ministério do Trabalho e Emprego. Relação Anual de Informações Sociais – Rais. | 2010/2016 |
| % de não brancos (pretos, pardos e Indígenas) | Dados do IBGE. Censo 2010 | 2010 |
| Tempo de viagem | Amostra Censo IBGE. A proporção das pessoas, ponderadas pelo peso, em cada classe de tempo de viagem do questionário, por distrito. | 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 |
| Quantidade de pontos de ônibus | Portal Geosampa | 2018 |
| Extensão de linhas de ônibus (em quilômetros) | Portal Geosampa | 2018 |
| Número de Estações de Metrô | Portal Geosampa | 2018 |

# Resultados (2.000 palavras)

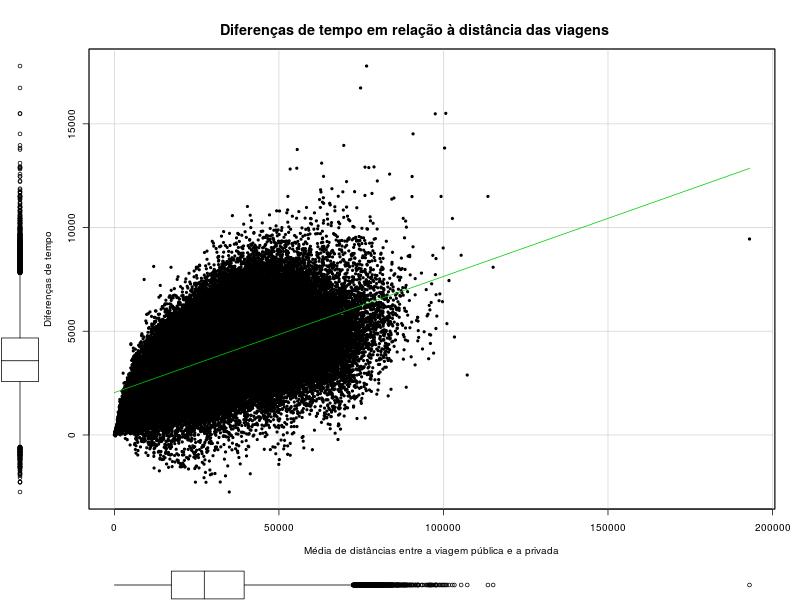
Os gráficos 1 e 2 representam a distribuição de freqüencias dos valores normalizados das diferenças de tempos e dos “Tempos Relativos”. Os dois histogramas indicam uma distribuição razoavelmente próxima da distribuição T. Particularmente no caso dos “Tempos relativos” a distribuição parece ser ligeiramente deslocada para a esquerda, mas foi assumido para fins de análise a normalidade dos dados.

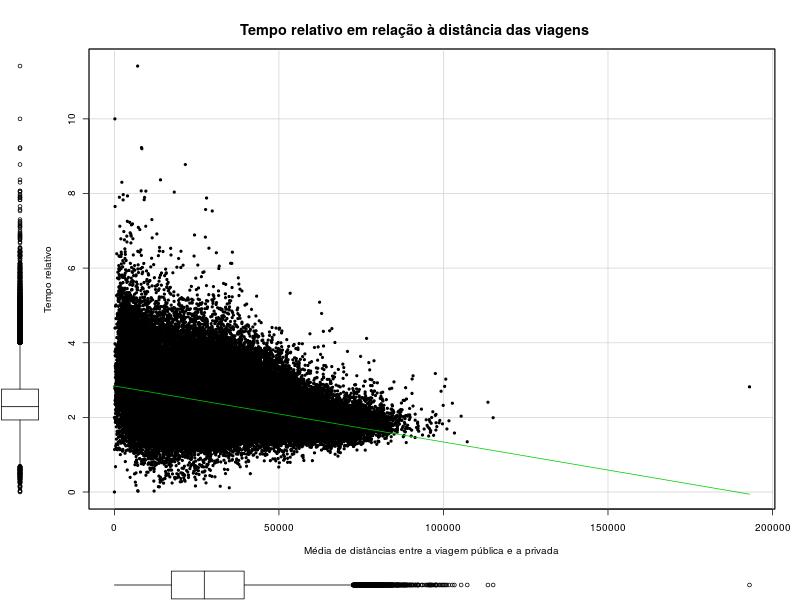




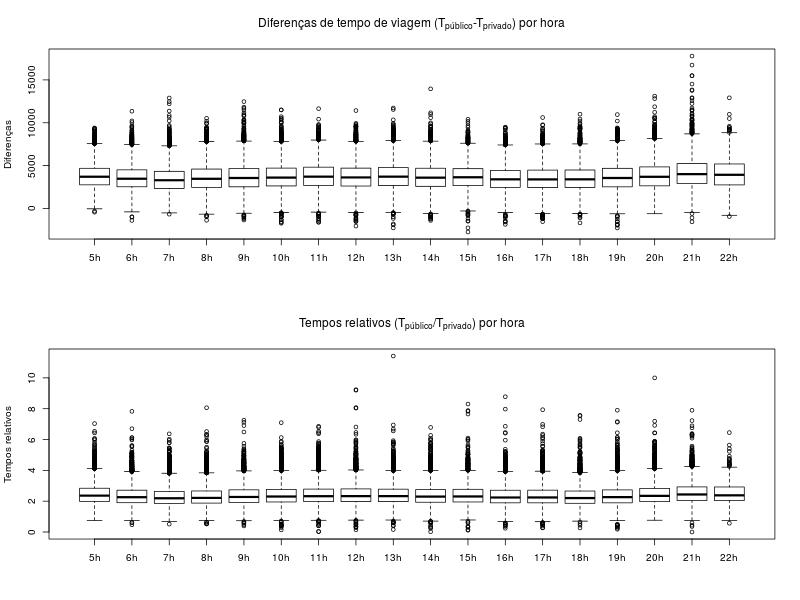
Nos dois casos, apesar de claro, é importante notar que as distribuições indicam uma diferença efetiva entre o tempo de transporte público e de transporte privado; de fato, para a diferenças entre tempos, utilizando um teste T de diferença de médias obtemos com 95% de confiança que a média da diferença está entre 3680s e 3697s; um teste T para o tempo relativo 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 que nas previsões de transporte público são incluidos 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 disttribuição das médias entre as distâncias das viagens de transporte público e de transporte privado. A diferença dos tempos 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. Esse dado indica que tamanho das viagens (refletida no tempo absoluto delas) tem alguma proporção com a diferença de tempos entre modais, ou seja, a velocidade do transporte público é menor. A distribição do “Tempo Relativo” apresenta algumas informações novas. A correlação é de aproximadamente -0,359, não muito significativa. E apesar da correlação e da linha de regressão linear simples indicar uma relação negativa entre as distribuições, visualmente os valores de distancias maiores parecem tender a um valor pŕoximo à média da distribuição. A concepção dessa medida – a razão entre os tempos de viagem dos diferentes modais – seria, por principio, menos variante em função da distância do que a diferença entre os tempo 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, mas ainda sim há muita variação na distribuição. 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 trechos pedestres aumentam consideravelmente o tempo total.

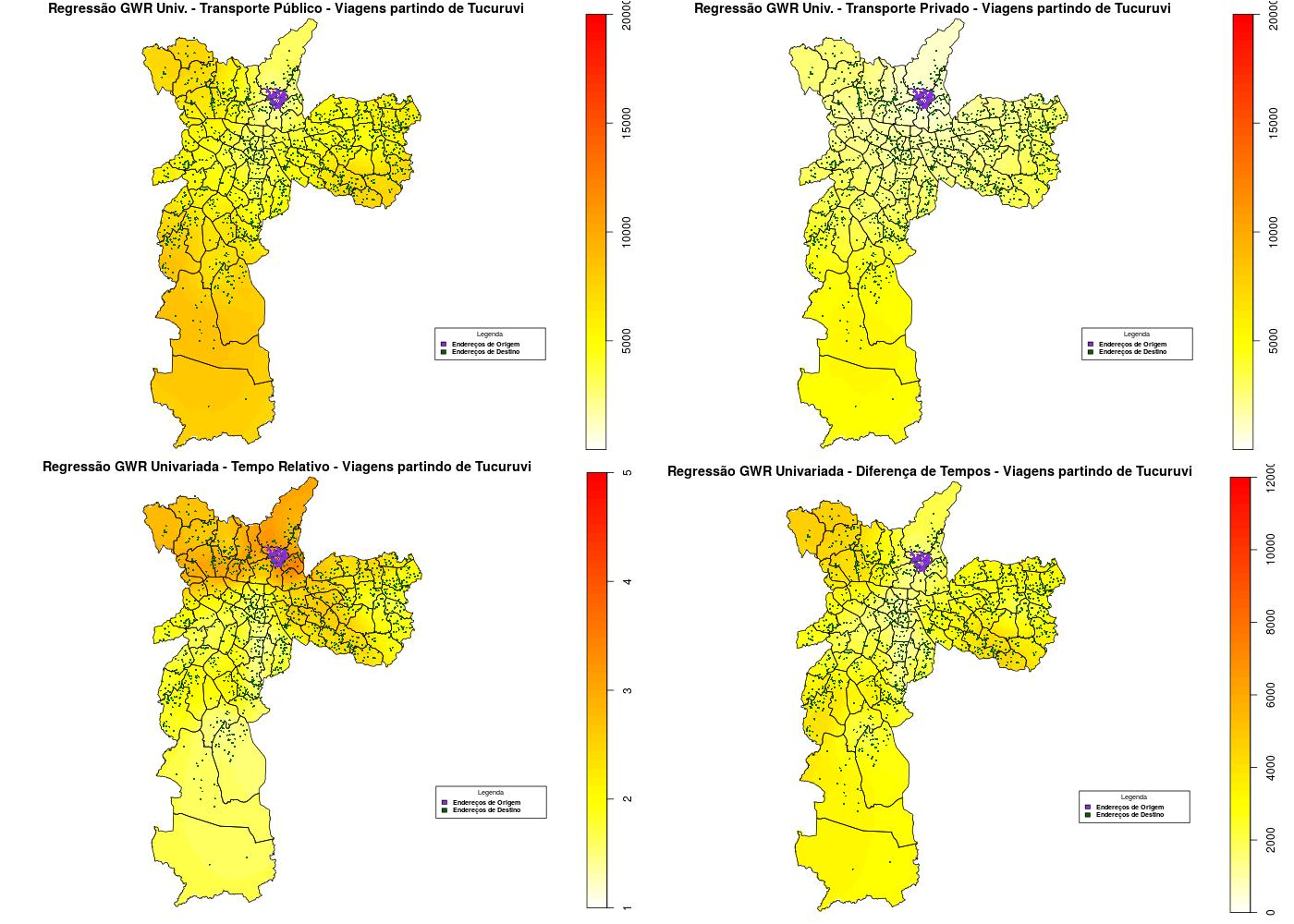




Explorando outras possíveis dependências dos dados em relação aos dias da semana e aos horários, foram elaborados gráficos para explicitar a distribuição das viagens de acordo com essas variáveis. Enquanto a comparação dos boxplots dos dias não parece indicar nenhuma diferença significativa, tanto para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

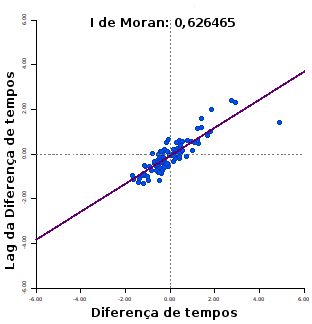


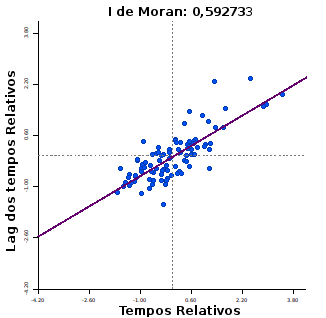
A análise exploratória espacial a partir das superficies suavizadas foi feita tanto para os distrito como para as áreas de ponderação. O resultado foram superficies suavizadas das medidas analisada para todo o município; a qualidade dessas superfícies é diretamente relacionada a densidade de pontos nas próximidades das médias estimadas, o que implica que as superficies para os distritos apresentam maior estabilidade das estimativas por serem calculadas a partir de um conjunto maior de pontos. A partir desses mapas foi possível identificar alguns padrões interessantes e marginais à analise do conjuto 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 da diferença de tempos de viagem e nos “Tempos Relativos”. Ao mesmo tempo a comparação das duas medidas permite entender melhor o comportamento das viagens. Em relação a zona leste é visivel uma zona mais clara no extremo leste (acompanhando os corredores de transporte público), mas há na zona que conecta ao centro um espaço de maiores “Tempos Relativos” que não se repete claramente nas diferenças de tempo. Isso pode ser uma indicação de que apesar da diferença de tempo dos modais se manter para viagens para essa região, o valor absoluto do tempo dos dois modais caiu em uma proporção semelhante.



Essa relação das medidas se dá pela própria forma como elas foram construidas: 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.

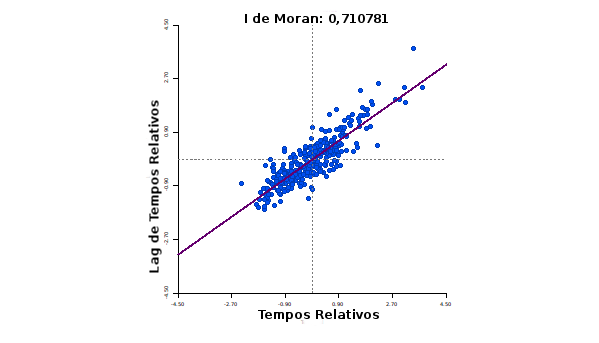
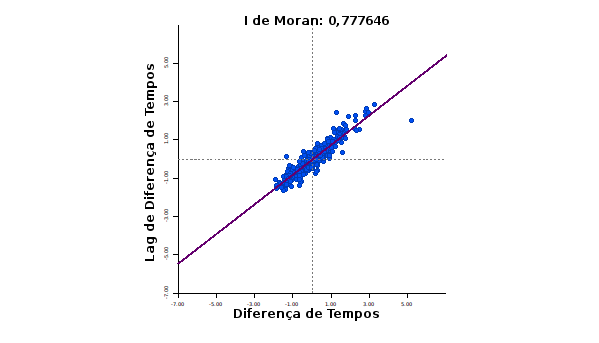
A próxima etapa da análise espacial foi a identificação de clusters das medidas nos distritos e nas áreas de ponderação de São Paulo. Para isso as medidas foram agregadas à divisão geográfica da origem das viagens – para cada zona de origem foi calculada a média das medidas relativas à zona. A partir dessa agregação foram calculados a partir do software Geoda os I’s de Moran (Figura K1 e K2) e os mapas de associação espacial.

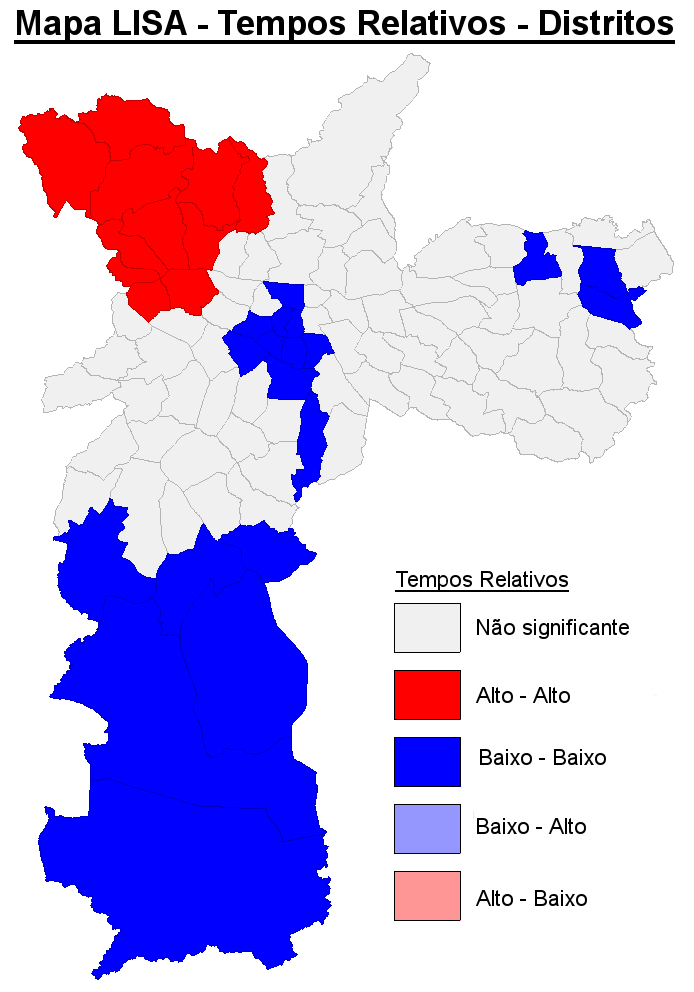
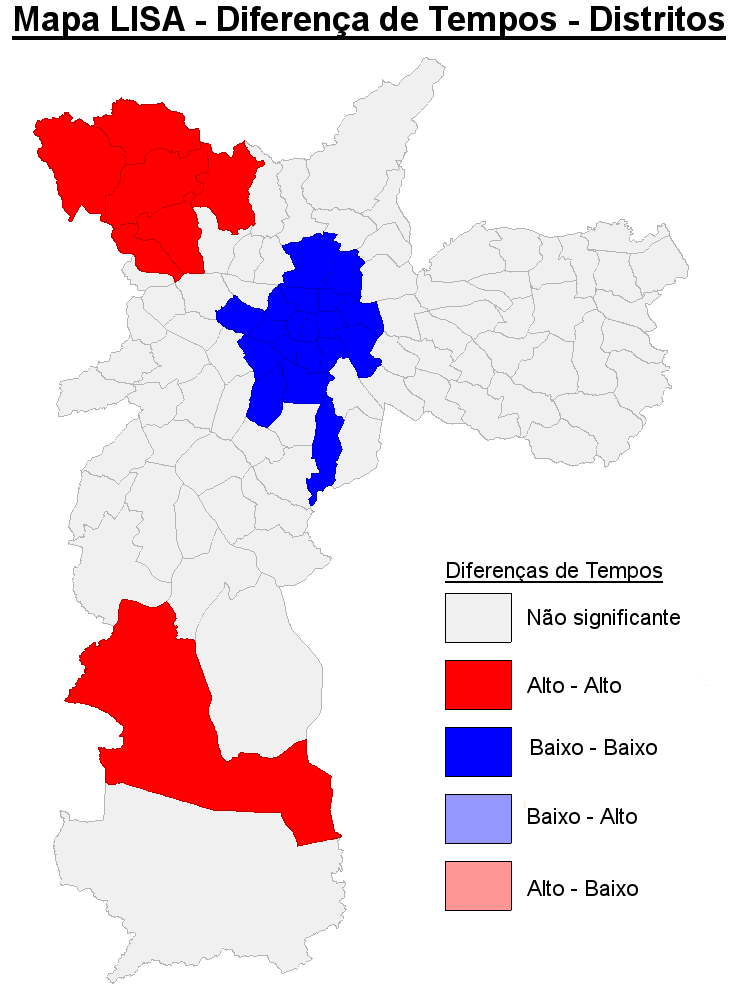




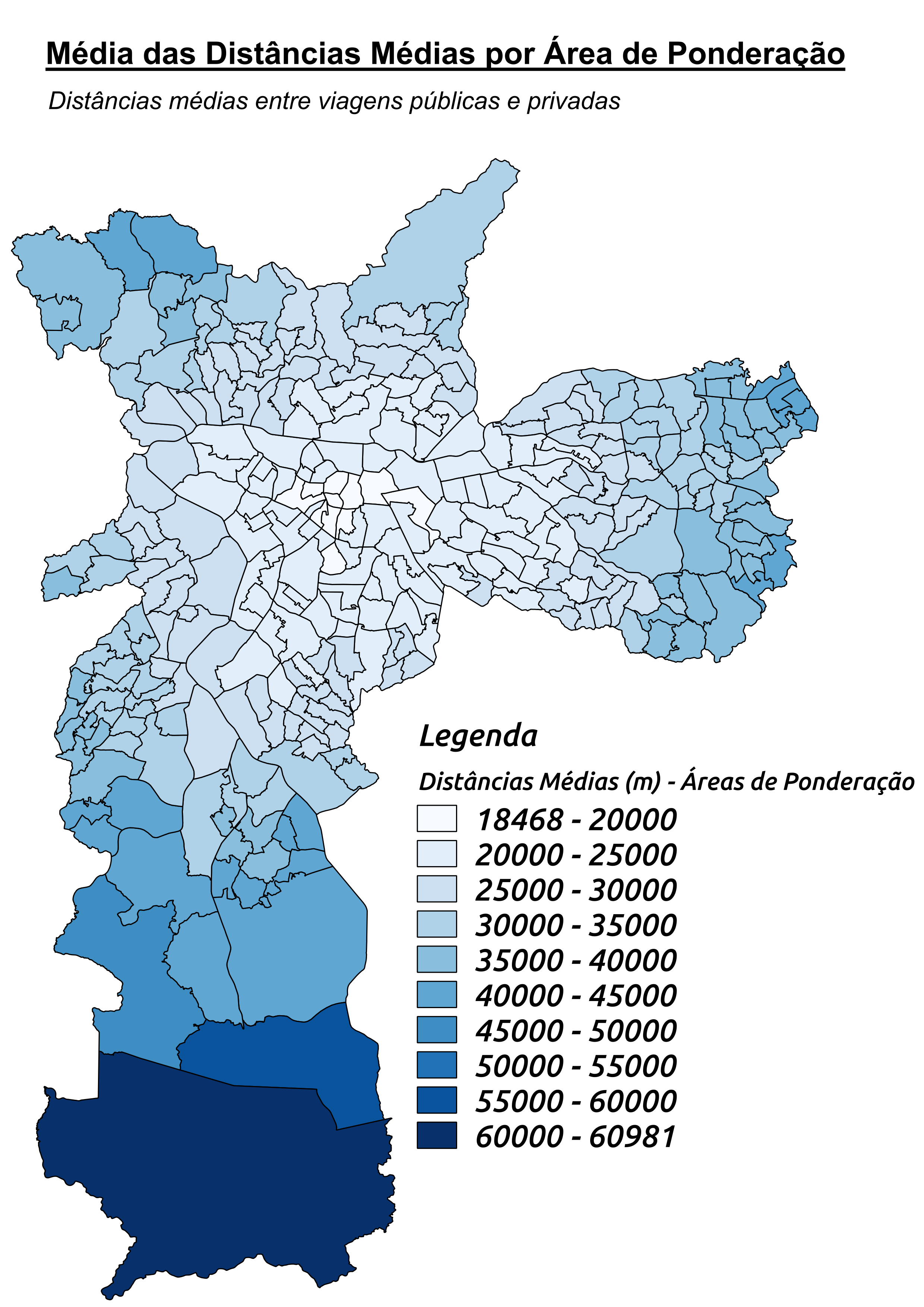
Os valores dos I’s de Moran Globais já indicam a presença de associação espacial significativa entre as medidas dos distritos e das áreas de ponderação, com valores consistentemente acima de 0,5. É notável que a associação é mais signifivativa nas áreas de ponderação (o que indica maior clusterização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

Os graficos dos I’s de Moran Locais ressaltam a indicação de que há a associaçaão 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.

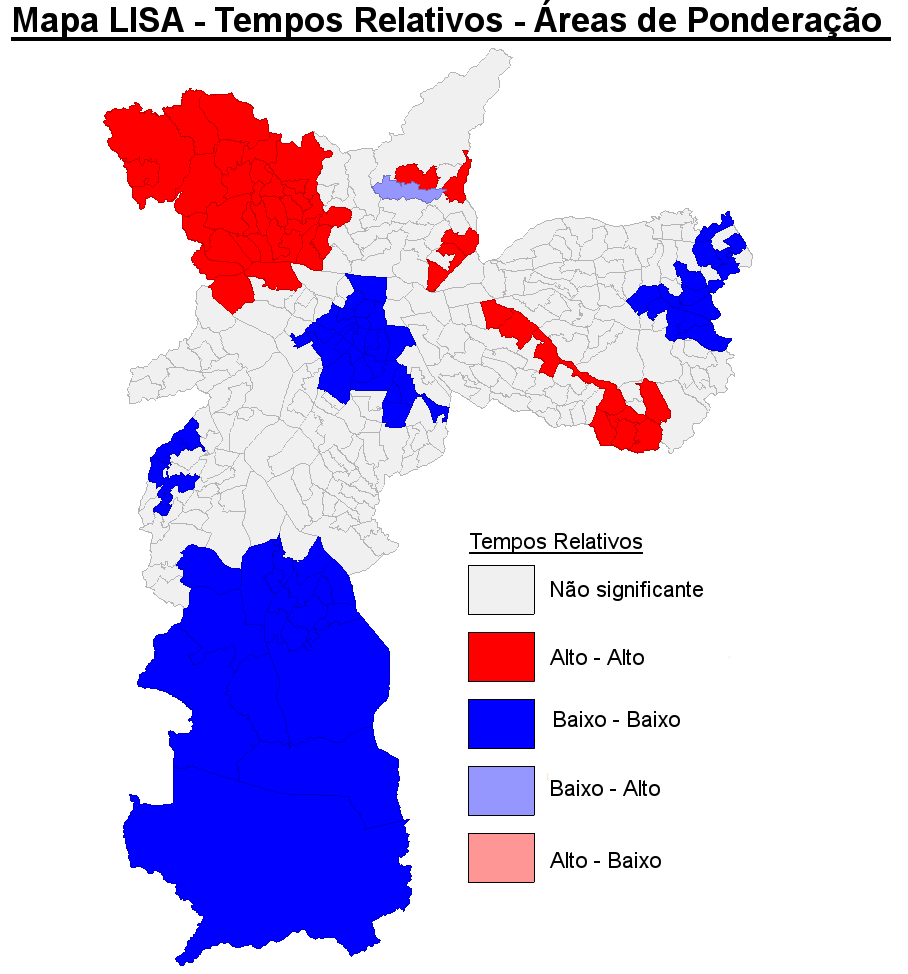
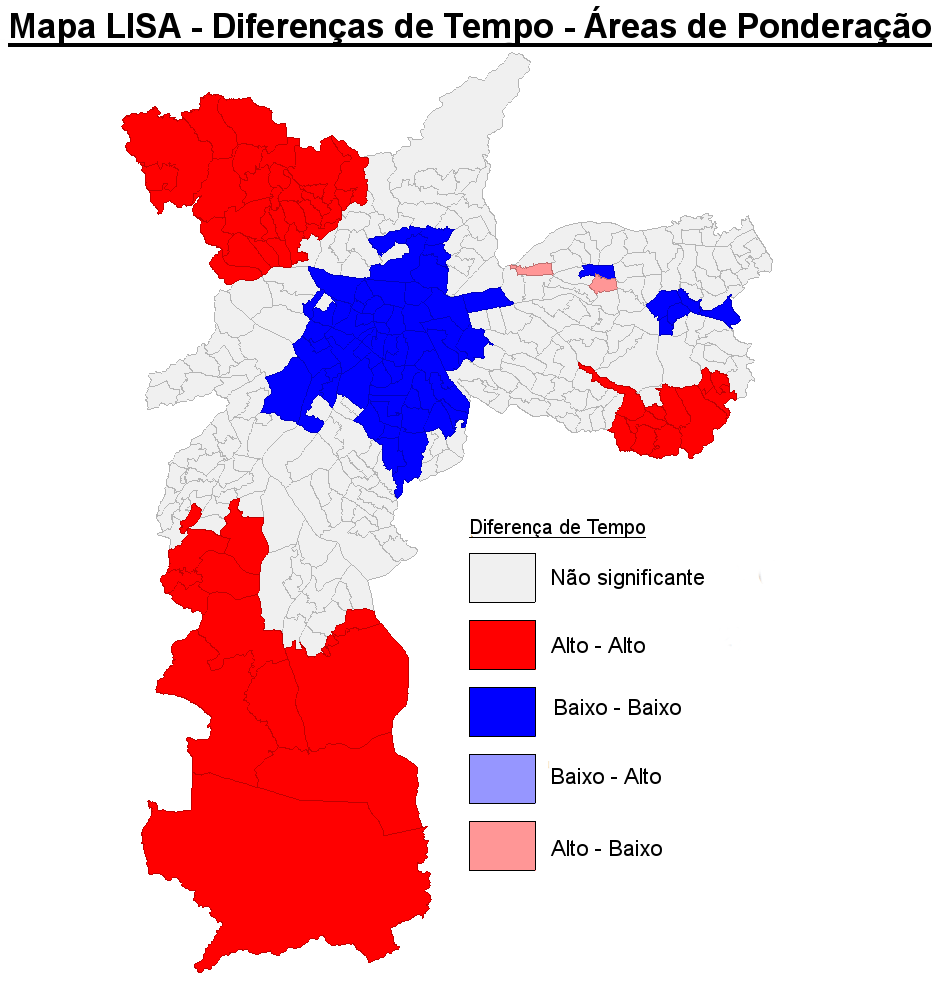


A análise dos mapas de associação espacial (LISA) feitos a partir da mesma análise derivada dos I’s de Moran permite visualizar melhor as relações de clusterização espacial das zonas analisadas. No caso dos distritos as duas medidas apresentam dois núcleos de clusters em comum: uma região de altos tempos relativos e alta diferenças de tempo na zona noroeste do municipio 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: As diferenças tendem a aumentar com o aumento das distancias das viagens enquanto os “Tempos Relativos” diminuem. Como ambas as regiões, por estarem relativamente longe do agregado das zonas dos municípios, apresentam uma proporção maior de viagens longas. A figura K mostra o a distribuição da média de distancia das viagens pareadas (média da distancia da viagen por trasnporte público e da viagen por trasnporte 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.

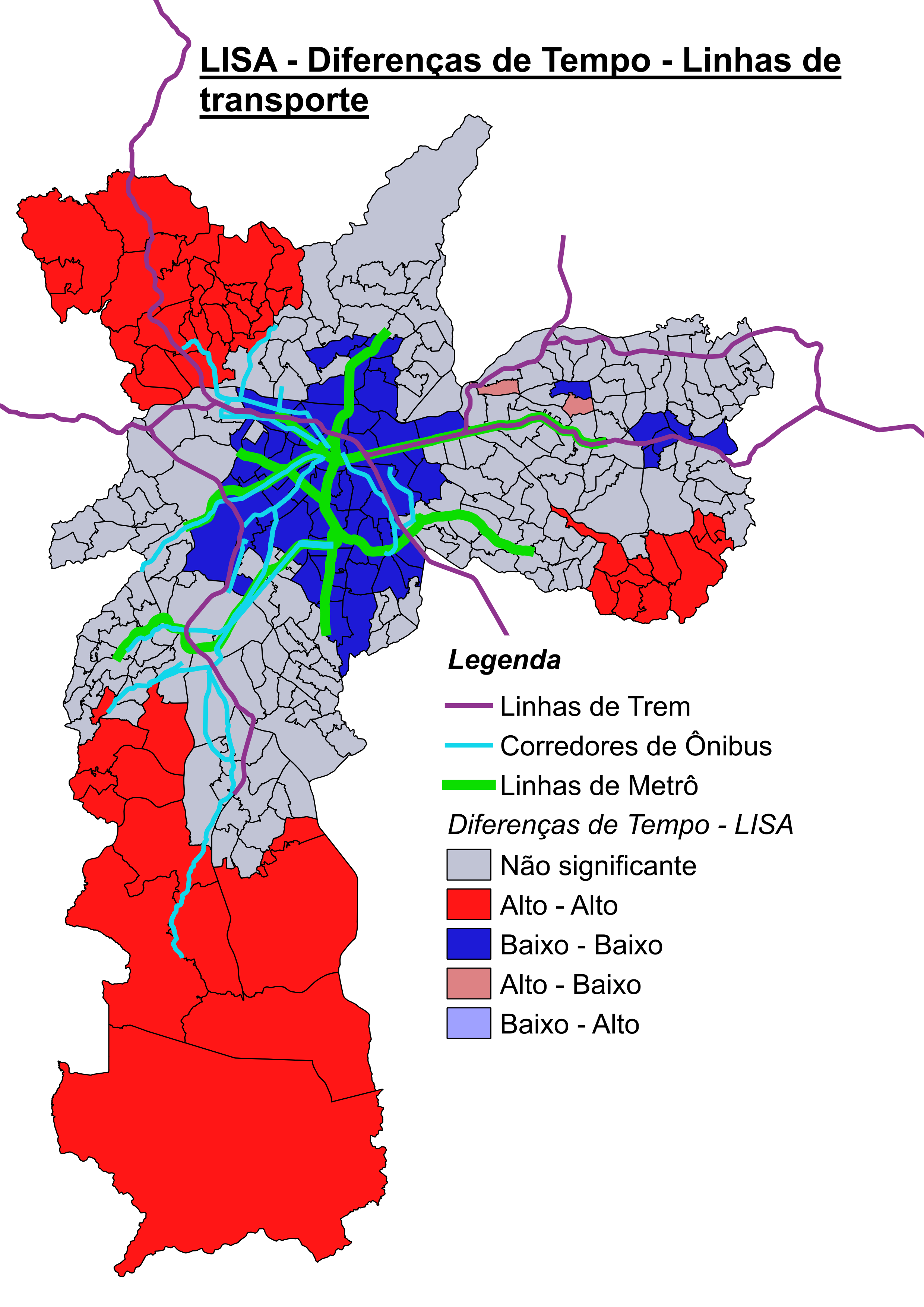
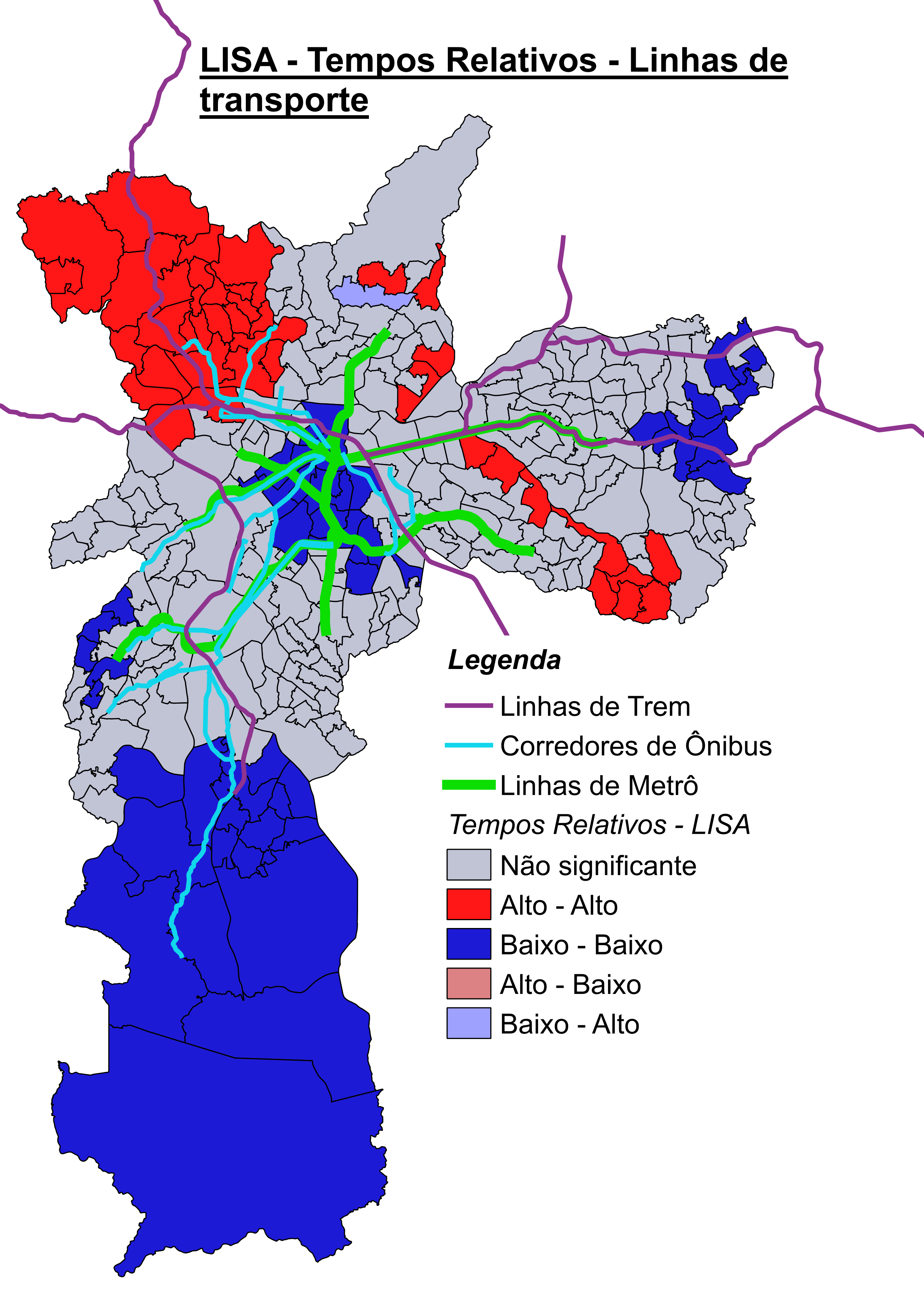


Os mapas LISA para as áreas de ponderação apresentam um grau maior de granulamento e, como indicam os I’s de Moran globais e os gráficos de I’s de Moran locais, apresentam mais núcleos de clusters que os distritos.

Os mesmos núcleos são vísiveis nos dois mapas; nos “tempos relativos” os clusters de baixo valor no centro e ao sul e o cluster de altos valores a noroeste estão presentes, enquanto no mapa de diferenças de tempo 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 “tempos relativo” os nucleos de baixo valores à sudoeste e um núcleo delgado de valores altos que se estende dos sul da zona leste até a zona norte entre os núcleos de valores baixo no centro e na zona leste. Enquanto o primeiro não aparece no mapa de diferenças de tempo, tanto 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 “tempos relativos” dos distritos).

Uma explicação para a diferença do número de clusters entre os níveis de análise é que pelo fato de os distritos apresentarem 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 dinamica 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 unidaades 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 k e z) encoraja a idéia de que há uma identificação clara entre os agrupamentos baixos e a proximidade do Metro e do trem; ao mesmo tempo, como os dados são simulações afetadas diretamente pela presença dessas infraestruturas, é de se esperear que o resultado reflita essa distribuição. A partir da do constraste visual, a presença do metrô parece fortemente relacionada à redução das medidas de análise, e consequentemente dos tempos de viagem do trasnporte público – nenhum dos clusters de valores altos de ambas as medias apresenta ma 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 baixo 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 fate de eles starem associados às outras estruturas de trasnporte. Mas onde há somente os corredores de ônibua, não parece haver impacto significativo.



# Conclusão (1000 palavras)

# 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)

1. http://www.reuters.com/article/2010/11/05/ozatp-ghana-economy-idAFJOE6A40BG20101105 6 [↑](#footnote-ref-2)