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

Índice

Relatório Parcial

Pergunta de Pesquisa:

Usando ferramentas e algoritmos de Estatística Espacial, é possível utilizar a simulação de tempos de viagem para gerar informações relevantes sobre as viagens no município de São Paulo?

Objetivo Geral

Os dados de mobilidade são um insumo importante para o planejamento da expansão da infraestrutura urbana, tanto é assim que a empresa do Metropolitano de São Paulo é responsável por realizar a cada dez anos a Pesquisa de Origem e Destino (Pesquisa OD) sobre mobilidade na grande São Paulo. A pesquisa é referência para trabalhos acadêmicos e projetos de infraestrutura, mas devido a sua dimensão e custo, acontece a cada dez, apresentando um problema de periodicidade, mesmo sendo intercalada com uma pesquisa menor para atualizá-la, mas menos desagregada, no intervalo entre cada duas edições. Assim o presente trabalho buscará validar métodos de gerar duas das variáveis da Pesquisa OD, o tempo de viagem e a distância das viagens para diferentes modais, de forma que essa grandeza possa ser usada em futuras pesquisas, sendo geradas quando necessário.

Atividades realizadas

Atividades Previstas		**Estado**
Pesquisar referências		Em realização
Programa de cálculo de tempos de viagem	Sorteio de pontos	Feito
	Cálculo de viagens	Em realização
Coleta de dados OD 2017		Em espera

Levantamento de variáveis	A realizar
Modelagem de dados	A realizar
Análise de resultados	A realizar

Em relação ao cronograma inicial, o projeto está atrasado. Isso se deu principalmente em razão da maior complexidade encontrada para a formulação do programa de cálculo de viagens do que o esperado. Para que o programa consiga gerar dados úteis para a análise, ele precisa funcionar com uma periodicidade definida e prevista em suas chamadas; sua execução de ser ininterrupta e deve atingir períodos relevantes para a análise, como os dias e as horas em que ocorre a mobilidade urbana. A implementação dessas funcionalidades demonstrou ser mais complexa que o esperado. A idéia é utilizar ferramentas de computação em nuvem disponibilizadas em caráter gratuito por empresas de tecnologia para que o programa apresente essas características.

Em relação a levantamento teórico, conceitos relevantes à pesquisa foram encontrados em ainda precisam de maior referenciamento. Por essa razão, essa fase de trabalho também precisará de mais tempo de execução para que seja completada a contento.

As variáveis da pesquisa OD ainda não puderam ser analisadas em razão de não terem sido publicadas de forma desagregada até o momento. Caso isso não seja feito até uma data que permita a análise por esse projeto, serão usadas outras edições da pesquisa.

Escopo do projeto e referencial teórico

O referencial teórico está dividido nas temáticas relevante ao trabalho.

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 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

o 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.

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 de 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]. 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 aplicaçẽos análise de dados produzidos em tempo real, e a variedade de formatos, com o uso de dados estruturados e não estruturados . Ainda nessa direção

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

existem 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).

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 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):

$$A_{ik}^{p} = \sum_{j}^{\square} g(W_{jk}) f(c_{ij}^{p})$$

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, o 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 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).

Resultados parciais

Participação em artigo

Devido a proximidade do tema e do referencial teórico utilizado, foi realizada uma parceria para a escrita de um artigo. O tema é relativo ao uso de dados de espera de carros fornecido pela API da empresa Uber para calcular uma medida de acessibilidade. Os métodos utilizados para a obtenção de dados no artigo, a partir do uso de API em um ambiente de computação em nuvem, são em parte reproduzidos no presente trabalho. O artigo foi enviado a RAE e segue em análise; a versão enviada está em anexo.

Sorteio de pontos no município de São Paulo

O sorteio de pontos de coordenadas para os endereços de origem e destino das simulações foi feito a utilizando o software aberto QGIS e o shapefile dos distritos de São Paulo, disponível no portal Geosampa. A partir do shapefile, foi traçado uma grade dividindo o território do município em quadrículas de 1km². Usando a camada espacial da grade, foi sorteado um ponto dentro de cada uma das quadrículas, e para selecionar somente os pontos que se encontram dentro do município, foi realizada uma intersecção espacial entre o conjunto de pontos e o shapefile do município. Isso resultou em um conjunto de 1538 pontos, em coordenadas de latitude e longitude.

Programa de calculos de viagens

O programa será feito usando conceitos de cloud computing. Estão sendo estudada a possibilidade de utilizar as ferramentas disponibilizadas pelo Google Cloud de forma gratuita para montar o programa.

Referências Bibliográficas

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 social1. *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.

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

Anexo 1

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

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.

RESUMO

The present paper aims to relate information of waiting times derived from car rental services, specifically Uber, with socioeconomic variables of the city of São Paulo, Brazil, with the intention of exploring the possibility of using this measure as a proxy of accessibility. A database with the averages of waiting times per district was created, which was aggregated with socioeconomic and transport infrastructure variables. From this base, a simple regression model (OLS) was constructed and with stepwise technique, the model was simplified to the most significant variables. The spatial distribution pattern of the measures was confirmed by Moran's I test, motivating the use of a special auto-regression model (SAR). The results indicate that physical variables as area and population density are important to explain this relation, but bus line length, minority rate and the spatial component suggests a possible relationship with accessibility.

PALAVRAS-CHAVE: Accessibility, Big Data, Uber, Space Statistic.

ABSTRACT

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 OLS, 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.

KEYWORDS: *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 OLS, 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, Estadística Espacial.*

INTRODUÇÃO

The aim of the present paper is to analyze mobility access in urban fabric by an exploratory study of a potential accessibility measure based on crowdsourcing tools, derived from Big Data approaches. The idea is, from data about waiting times for car rental services vehicles like Uber, explore the relationship of waiting times with variables usually associated in literature with accessibility and socioeconomic variables.

The results found for São Paulo indicate that three classes of variables are important to predict Uber access: The first class refers to the presence of transport infrastructure and transport behavior; the second class refers to the physical and demographic dimensions of São Paulo districts; the third class is the racial composition of districts. This last class correlates strongly with socioeconomic factors in São Paulo, like income and travel times that more or less follows a classic center periphery distribution (Francisco 2010).

From the observation that the waiting times data follows a clear spatial pattern, a spatial regression model was built considering the variables selected by relevance in a previous simple regression. The found model indicates that there is reason to consider Uber waiting times spatial distribution following the distribution of the indicated variables, and that there is an important relation between travel times and Uber access.

The findings indicate that with better characterization/design the data yield by crowdsourcing transport applications, particularly Uber, can become an important source of information about accessibility in urban environments.

The transport network distribution as an essential part of accessibility (Páez, Scott e Morency 2012) is a recurrent research subject among geographers, urban planners and social scientists. This field of studies is recently being transformed in relation to its methods and its approach 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 an useful concept to describe urban experience (Páez, Scott e Morency 2012).

Before the great evolution of the urban mobility market promoted by tech companies wielding crowdsourcing based tools, like Uber, is reasonable to imagine that the approximation of these two worlds has a great potential of mutual contribution and must be investigated.

Among this ambience of new approaches Letouzé e Jütting (2014) affirm that the official bodies

responsible for official statistics production must be aware to the evolutions of “Big Data”. Both to profit from new tools and approaches together with the scientific rigor of validation and analysis as to face the world of Big Data as an inestimable source for the advance of scientific research.

In this context, the creation of continental and global tech companies in mobility, like Uber, Cabify, 99, has produced expressive numbers in late years. For an 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 users1. Although Brazil’s recent crisis, both the service provided as the job opportunity in shared transportation appears to be well evaluated, collaborating to its maintenance even after the crisis (citation).

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 users app and in the web environment through public APIs (“Application Programming Interface”), a Big Data classifiable tool.

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

Thinking about the capacity to access the transport system, Wang& Mu (2018) propose that the estimates may be used as a measure of accessibility. As said, the debate about these new approaches in the area is recent. Even with the interferences of these new services in the transportation environment and the easy access to the tools Uber provides, there are few empirical studies using data from Uber (Hall & Krueger 2015; Hughes & Mackenzie 2016; Zhou, Wang & Li 2017; Wang & Mu 2018), opening an opportunity of empirical exploration of Uber’s fleet in light of accessibility theory.

LITERATURE REVISION AND RESEARCH QUESTION

The literature about Big Data and accessibility was revisited for this study. This section begins with the conceptualization of Big Data and its position in the current tech market. Then in light of Big Data theory Uber’s time estimation tool is characterized followed by a systematic review

of accessibility and its normative and positive dispositions.

One classical definition of the Big Data movement takes in consideration the characteristics of the data produced in the virtual environments of massive users presence: Volume, Variety and velocity, the three Vs (McAfee et al, 2012).

According to this definition the data generated by the new forms of technology uses and applications constitute relatively big databases, up to Petabytes (1000Gb) or even Exabytes (1000Pb). At the same time, the many applications that generate data do so in various formats (photos, commentaries, reactions, videos…), generating heterogeneous databases in contrast with structures data. Also the sheer productions speed of this data can be very fast, to the limit that real time processing is necessary.

This is not the only conceptual approach to Bog Data. A more sociological view tries to describe the movement with three “Cs”: Crumbs, Capacities and Communities (Letouzé e Jütting, 2014).

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 programing 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 member share ideas with specific language and common validation methods. These communities can be established in open and collaborative environments, like OpenSource community, or in more restrictive ones, like tech groups in big corporations with access to big databases (Letouzé e Jütting, 2014).

Uber estimation tool uses in the present article is by both definitions a Big Data tool. In relation to the three Vs an enormous volume of rides’ results data constantly feeds the tool’s algorithm with high speed interpretations of spatial non-structured data.

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).

To discuss the access of individuals to the city urban mobility 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 et al, 2012) and approaches close to the present one, using social media data to trace mobility behavior (Noulas et al, 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) what 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 term of space and social contexts (Morandi et al, 2015). But 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,2005 Stelder 2016, Paez, 2012). 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 quantity 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 (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 it is as a temporal measure (time to access) (Páez, Scott e Morency 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 analyzed critically 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, Scott and Morency (2012) define two epistemic approaches to studying 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 the travels expectations while a positive one bases itself on the actual travel experience. The present study will take consider the positive approach as the data refers to actual Uber waiting times in São Paulo

The following hypothesis were elaborated based on the revision:

H1 – The estimated waiting time when requesting a ride in Uber's platform can be used as a proxy for accessibility in São Paulo.

H2 – Uber waiting time distribution relates to socioeconomic indicators' polarization in São Paulo.

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 OLS 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 calculated, as suggested by Wang & Mu (2018).

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 form São Paulo.

Uber Data

The data was collected from the APIs present in Uber Developers portal, accessible from the company's site (https://developer.uber.com), and covers the estimated waiting times of all Uber products during the of 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, 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 2,500,00 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 alocated in the Azure cloud computing service provided by Microsoft. Its architecture follows diagram 1.

Socioeconomic Data

The information was collected from national, State (State of São Paulo) and municipal (São Paulo Municipality) levels. The socioeconomic data at the district level were 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 Georreferenced information (http://geosampa.prefeitura.sp.gov.br).

The decennial Demographic census is one of Brazil's most important statistical products. The 2010 census present 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. The geographic distribution is made as to make inference possible inside “weighting areas” (310 for São Paulo) defined by IBGE, but following where possible municipal administrative divisions. (IBGE, 2010).

The RAIS is a Ministry’s of Work an Employment governmental tool for management of Brazilian work relations. The data is compiled from statements filled by the business about their and their employees work situation. The declarations are mandatory for a series of business nominated in the law (BRASIL, 2016).

SEADE foundation is a nationally recognized statistical institution known for its technical capacity and it is 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 georrefereced data about a row of subjects, including the distribution of transport infrastructure.

All data was searched or aggregated to the district level (Wang & Mu, 2018; Francisco, 2010). 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 (Torres et al, 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 & Mu (2018).

Descriptive Data Analysis

In the exploration of the 2,500,000 calls made to the API two patterns were noted: First, the lack of estimates for some of Uber’s products, and second, the 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 extreme south comprises a mangrove area that does not have access roads. The northern strip and the other spots over the city are regions that Uber classifies as risk zones and does not provide services.

The cover of each service and its availability can seen in table 2 It can be noted that the UberX, the most popular service Uber provides, has the larger coverage with almos 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 reasonig 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 OLS models were create, the first having the average waiting time as a dependent variable and the second having the standard deviation as the dependent variable.

Simple Linear Regression OLS

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, 2005).

For the construction of the OLS 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 (Km^2)
- 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.

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. FO this 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 OLS model were selected and the spatial auto-regressive model (SAR) was calculated. Francisco defines SAR as a regression model capable of incorporate the spatial neighbors matrix (or spatial proximity) as a part of the explanatory variables.

For the confection of the SAR model the software GeoDa version 1.12 was used.

RESULTS

Table 3 resumes the results of the OLS 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 R^2 of 0.893 against a R^2 of 0.717 for the standard deviation model.

Because of its higher degree of explanation, the average waiting time was used to build a stepwise regression model to find which of the independent variables shown sufficient

significant relations with UberX waiting times. Table 4 shows high significance for length of bus lines, district's area, population density, jobs, non-white rate and proportion of travel times superior to 121 minutes, with an R^2 of 0.879.

The Figure 4 shows the spatial dependency of UberX waiting times averages, with a Moran's I of 0.59, showing high spatial dependency of the dependent variable. In Figure 5 the distribution of the variable and its spatial dependency; through some of neighborhoods there are uniform areas of low-low waiting times.

With the variables presented in Table 4 a spatial regression model was generated for UberX average waiting times and the results in Table 5 show a high degree of significance for the variables length of bus lines, district's area, population density and non-white rate. With the spatial factor in the model, the R^2 goes to 0.89.

DISCUSSION AND CONCLUSION

The crowdsourcing transport market is indeed innovative and can be considered as a new transport system at the same time it does not suffer the traditional regulation in its distribution (Wang & Mu, 2018). The analysis of Uber's fleet distribution under the lens of accessibility has allowed the identification of a strong relation between UberX distribution and some socioeconomic variables. Even if H1 does not proceed that instigate the search for an understanding of the dynamics behind these new services.

One of the main findings of this article is the relation between Uber waiting times and the proportion of non-white residents. The fact that this relation was shown different in Wang & Mu (2018) study about Atlanta encourages the continuity of the investigation, as the distribution of minorities is significantly associated to accessibility measures (Páez, Scott e Morency 2012), opening space for a comparative approach.

The need to promote intuitive and highly communicable accessibility measures is latent between researchers (Páez, Scott e Morency 2012) and the construction of an transport accessibility measure from crowdsourcing tools like Uber 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. Letouzé e Jütting (2014) corroborates this affirmation when they say that there must an approximation of Big Data tools to the Academy to promote new indicators and

official statistics.

However, deserves some attention some limitations of the study, as it provide indications for future researches. First, 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 the other Companies in the future. Second, the replication of the study in other Brazilian capitals makes itself necessary because the economic and cultural factors that affect the disparities can be analyzed. Third, a deepening in the Uber's time estimate tool understanding to identify possible biases that could have suppressed any relation. It is worth revisiting this study after the publication of 2020 IBGE Census.

REFERÊNCIAS

Aranha, V. (2005). Mobilidade pendular na metrópole paulista. São Paulo em perspectiva, 19(4), 96-109.

Batty, M. (2016). How disruptive is the smart cities movement?.

BRASIL, IBGE. Censo Demográfico. (2010). Notas Metodológicas–Microdados da Amostra. Rio de Janeiro, 2010.

BRASIL, Ministério do Trabalho. Portaria n. 1464 de 30 de dez. 2016. Aprova instruções para a declaração da Relação Anual de Informações Sociais - RAIS ano-base 2016. Brasília, DF, dez. 2016.

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

Cohen, P., Hahn, R., Hall, J., Levitt, S., & Metcalfe, R. (2016). Using big data to estimate consumer surplus: The case of uber (No. w22627). National Bureau of Economic Research.

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

Hall, Jonathan V.; Krueger, Alan B. An analysis of the labor market for Uber's driver-partners in the United States. ILR Review, p. 0019793917717222, 2015.

Hansen, W. G. (1959). How accessibility shapes land use. Journal of the American Institute of planners, 25(2), 73-76.

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.

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, 5(3&4).

Letouzé, E., & Jütting, J. (2014). Official statistics, big data and human development: towards a new conceptual and operational approach. Data Pop Alliance and PARIS21.

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

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data: the management revolution. Harvard business review, 90(10), 60-68.

Morandi, E. et al. Análise Geoespacial da Relação entre Transporte Público sobre Trilhos, Renda e Tempo Médio de Deslocamento, 2015.

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), e37027.

Pääkkönen, P., & Pakkala, D. (2015). Reference architecture and classification of technologies, products and services for big data systems. Big Data Research, 2(4), 166-186.

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.

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

Schwanen, T. (2016). Geographies of transport I: Reinventing a field?. Progress in Human Geography, 40(1), 126-137.

Stelder, D. (2016). Regional accessibility trends in Europe: Road infrastructure, 1957–2012. Regional Studies, 50(6), 983-995.

Torres, H. D. G., Marques, E., Ferreira, M. P., & Bitar, S. (2003). Pobreza e espaço: padrões de segregação em São Paulo. Estudos avançados, 17(47), 97-128.

Tribby, C. P., & Zandbergen, P. A. (2012). High-resolution spatio-temporal modeling of public transit accessibility. Applied Geography, 34, 345-355.

Wang, M., & Mu, L. (2018). Spatial disparities of Uber accessibility: An exploratory analysis in Atlanta, USA. Computers, Environment and Urban Systems, 67, 169-175.

Weber, J., & Kwan, M. P. (2003). Evaluating the Effects of Geographic Contexts on Individual Accessibility: A Multilevel Approach1. Urban Geography, 24(8), 647-671.

Zhou, X., Wang, M., & Li, D. (2017). From stay to play–A travel planning tool based on crowdsourcing user-generated contents. Applied Geography, 78, 1-11.

Table 1: São Paulo's Socioeconomic data

District data	Source	Date
Area (km^2)	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,	São Paulo's municipal portal "Infocidade". Original data source: Ministry of Work and Employment - Annual Report	2010/2016

Services, Transformation Industry, Civil Construction)	of Social Information (RAIS)	
% 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 distric (cars and motorcycles) from the weighed households of the Sample	2010
Bus Stops	Portal Geosampa	2018
Extension of Bus Lines (Km)	Portal Geosampa	2018
Metro Stations	Portal Geosampa	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)								

		943.000	750.400,00	0,213		39.350,00	515.300,00	0,939
QTLINBUS2018	Bus lines' quantity	0,1254	191,30	0,514		122,80	131,40	0,353
KMLINBUS2018	Bus lines' legth (Km)	-0,1195	67,38	0,080		-89,49	46,27	0,057
QTPONTBUS2018	Number of bus stops	-0,1217	132,40	0,361		11,21	90,92	0,902
QTESTMETRO2018	Quantity of Metro stations	0,9081	5.216,00	0,862		-1.143,00	3.582,00	0,751
RENDP2010	Income per Capita	-0,001017	10,30	0,922		4,62	7,07	0,515
ARE1	Area (km2)	3,551	309,20	-		744,30	212,30	0,001
POP2018	Population	0,0006892	1,00	0,492		-0,09	0,69	0,894
DENPOP2018	Population density	-0,004164	1,49	0,007		-5,01	1,02	0,000
DOMP2018	Number of particular permanent households	-0,001331	2,89	0,646		1,20	1,98	0,546
ESTAB2016	Employers	-0,000269	7,23	0,970		-1,32	4,97	0,791
EMP2016	Jobs	0,0002896	0,33	0,384		0,14	0,23	0,547
PNBRAN2010	Proportion of non-whites	300,2	77.690,00	0,000		202.200,00	53.350,00	0,000
TEMP2010_5MIN	Travel time - up to 5 minutes	-587,5	1.166.000,00	0,616		111.000,00	800.500,00	0,890
TEMP2010_30MIN	Travel time - from 6 to 30 minutes	-978,9	785.700,00	0,217		37.680,00	539.500,00	0,945
TEMP2010_60MIN	Travel time - from 31 to 60 minutes	-278,4	831.700,00	0,739		180.800,00	571.100,00	0,752
TEMP2010_120MIN	Travel time - from 61 to 120 minutes	-1046	805.200,00	0,198		-141.100,00	552.900,00	0,799
TEMP2010_121MIN	Travel time - more than 121 minutes	-1443	992.800,00	0,150		-215.900,00	681.800,00	0,752
TEMP2010_0MIN	No travel	-821,3	765.000,00	0,286		10.860,00	525.300,00	0,984
PDOMC2010	Car - rate of household motorization	-139,1	367.700,00	0,706		-164.100,00	252.500,00	0,518
PDOMM2010	Motorcycle - rate of household	37,73						

	motorization		82.420,00	0,648		25.170,00	56.600,00	0,658

Table 4: Stepwise regression results for UberX average waiting times

DV		Average Uber X		
		Coeficient	Std. Error	p-value
(Intercept)		171.000,00	18.170,00	0,000
KMLINBUS2018	Bus lines' legth (Km)	-78,60	23,99	0,002
ARE1	Area (km2)	3.634,00	232,60	0,000
DENPOP2018	Population density	-3,35	0,92	0,000
EMP2016	Jobs	0,21	0,13	0,097
PNBRAN2010	Proportion of non-whites	295.700,00	44.570,00	0,000
TEMP2010_121MIN	Travel time - more than 121 minutes	- 1.239.000,00	422.500,00	0,004

Table 5: 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

Figura 1: Diagram of the Uber data collection application.

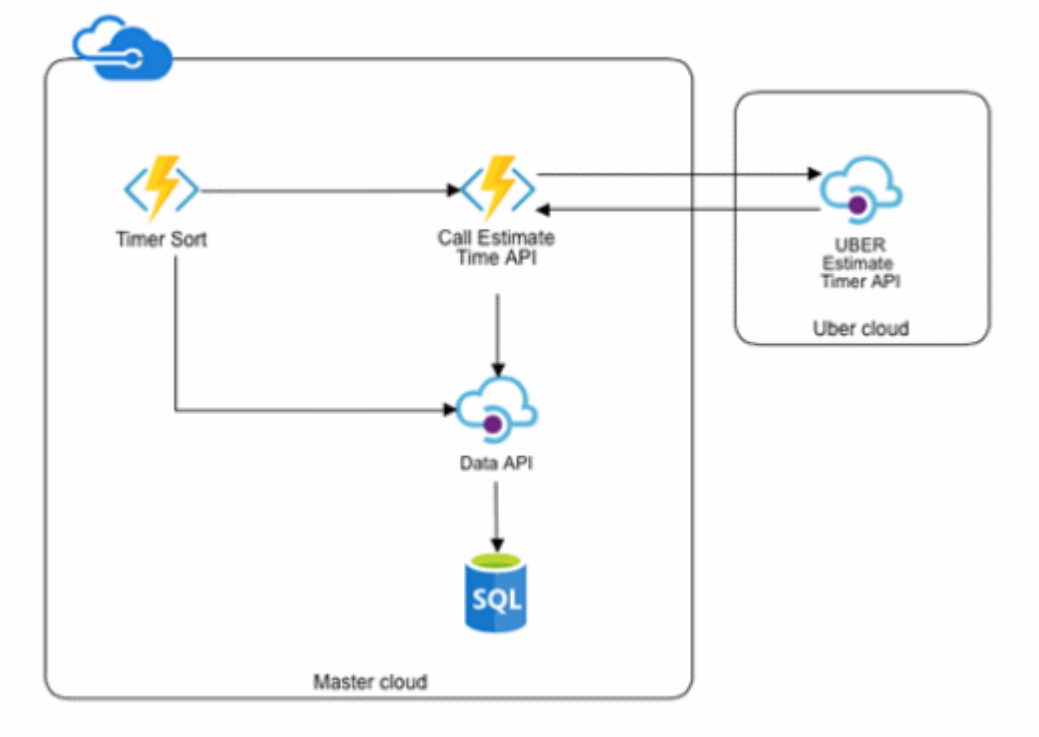

Figure 2: Coordinate points for Uber waiting time estimates

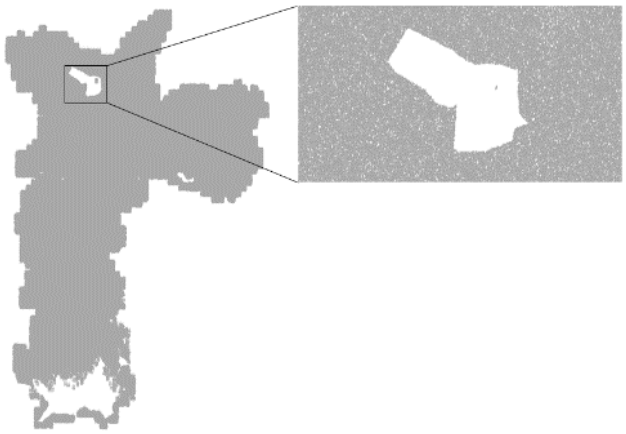

Figure 3. Comparison between average and standard deviation of UberX estimate waiting times

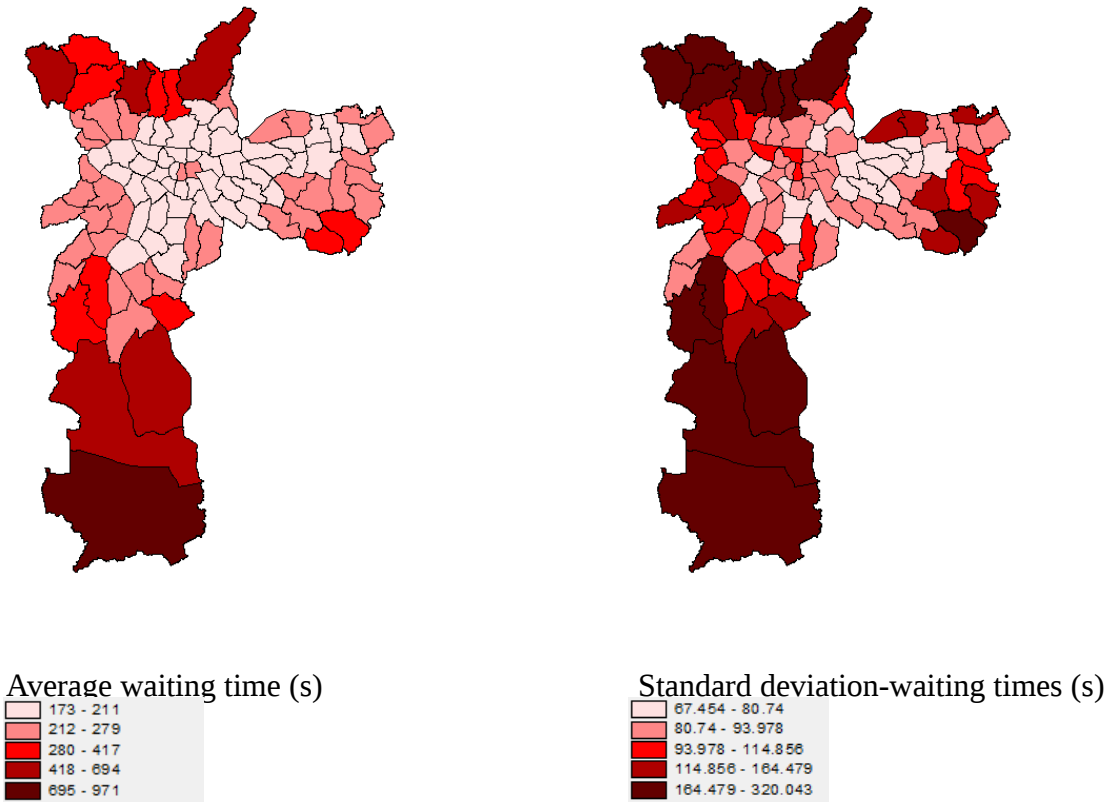

Figure 4. Uber X average waiting time Moran's I

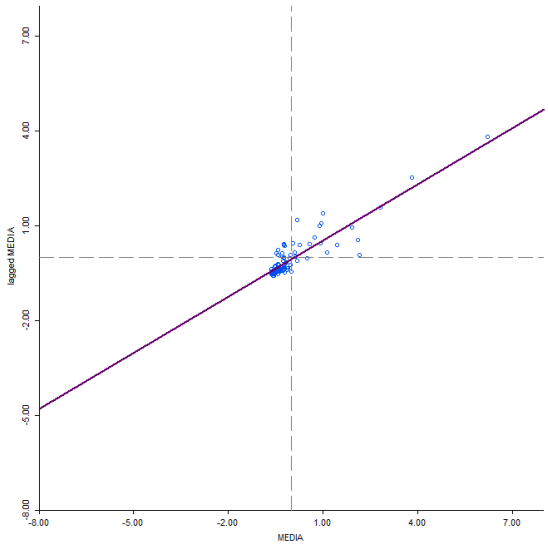

Figura 5. Uber X average waiting time neighbourhood clusters