**Título**

**RESUMO**

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

# 

# **1. INTRODUÇÃO**

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

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

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

**1.1. Problema da Pesquisa**

**1.2. Justificativas**

**2. REVISÃO DE LITERATURA**

**2.1. Mobilidade Urbana e Acessibilidade**

**-passar por algumas abordagens de mobilidade**

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

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

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

**2.2. Indicadores de acessibilidade**

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

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

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

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

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

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

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

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

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

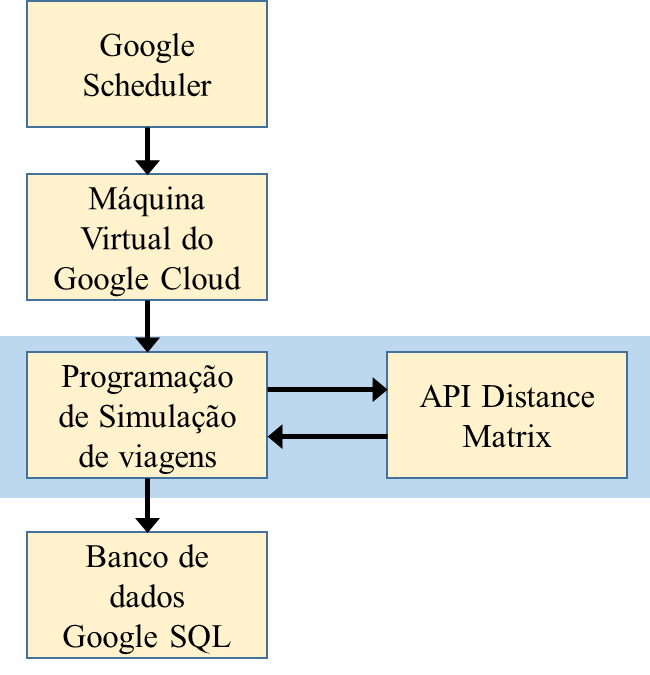
Como o movimento de *Big Data* se apresenta como um vetor de mudança na sociedade moderna, é interessante que haja movimentos de integração entre as estatísticas oficiais e essas novas técnicas de análise. Kim (xxxx) compara os usos e desafios do *Big Data* no setor privado e no setor público; além do uso em estatísticas oficiais, o *Big Data* pode ser usado para melhorar a capacidade dos Estados de fornecer serviços públicos, aumentar sua tranparência e combater fraudes. De particular interesse para o presente trabalho, a produção de dados georreferenciados relativos à mobilidade é essencial para captar a distribuição da mobilidade no tecido urbano. Dentro dos meios de Big Data, os dados gerados pela utilização dos celulares – ainda mais no contexto em que volume da rede móvel supera o volume de rede fixa (Lee & Kang, 2015) - já fornece um enorme volume de dados georreferenciados e, dependendo do uso de aplicativos, informações sobre os meios de transporte. Essa produção massiva de dados permite inclusive o uso desses dados para análises em tempo real, como os erviços de mapas para calcular rotas de transporte. Também pelo lado das estatísticas oficiais a produção de dados georreferenciados para entender os problemas urbanos, inclusive de mobilidade, é corrente e importante para embasar a adoção de políticas públicas específicas para cada localidade. A compreensão da dimensão geográfica dos problemas e da distribuição da infraestrutura presente e dos serviços ajudam a diagnosticar ineficiências e priorizar esforços, além de fornecer uma visão sistêmica dos indicadores sociais. Essa visão pode ajudar a escolher combinações de formas diferentes de intervenção pública (Torres et al., 2003, Torres e Oliveira, 2001). Como a produção desses dados através de pesquisas empíricas de validade estatística, como a Pesquisa OD (METRO, 2008), tende a ser bem custosa, o acesso a dados derivados dos novos aplicativos sociais que usam a localização podem permitir o acesso a informações de mobilidade de maneira muito mais barata, ainda que contendo algum grau de viés (Kwan, 2016). E enquanto 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) a disponibilidade de dados e técnicas utilizando Big Data deve ser vista com cautela. Kwan (2016) alerta para vieses decorrentes do uso intensivo de algoritmos em qualquer tipo de análise. Mesmo que esses vieses não sejam particularidades de estratégias de *Big Data*, o uso intensivo de algoritmos no processamento e análise de dados tem o potencial de gerar interferência nos dados sem que seja possível ao pesquisador acompanhar os dados que serão usados, dado o seu volume. Por essa razão a importância da validação de estratégias de Big Data junto a estratégias tradicionais é importante para discernir os possíveis vieses introduzidos pelo processamento de dados.

# **3. MÉTODO**

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

**3.1. Simulação de viagens**

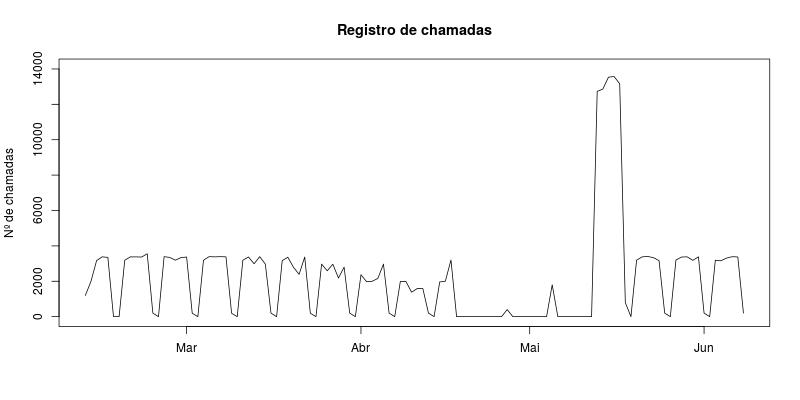
A simulação foi feita em duas etapas. Primeiro foi gerado um banco de endereços parcialmente aleatorizado no município de São Paulo utilizando o software QGis; a partir dessa base de endereços foi feita a simulação das viagens propriamente ditas utilizando a API de roteamento do *Google*. A escolha dessa ferramenta levou em consideração a possibilidade de simular viagens considerando o efeito em tempo real do tráfego nas viagens, fator considerado um diferencial da estratégia adotada. Porém isso implicou em um número limitado de requisições de viagens, por questões de custo; essa limitação levou a uma série de escolhas para reduzir o número de viagens “perdidas” na simulação (quando as coordenadas usadas na API não correspondiam ou não podiam ser aproximadas a endereços válidos, como no caso de coordenadas dentro das represas de São Paulo). Para tentar obter diferentes padões de mobilidade ao longo do tempo, foi decidido distribuir as viagens simuladas ao longo dos dias úteis (segunda-feira à sexta-feira, das 5h às 21h). Isso levou à opção por usar computação em nuvem para a simulação, para garantir a frequência das chamadas à *API*. Foram adotadas, por isso, medidas para de reduzir o esforço computacional a fim de evitar problemas relacionados ao desempenho. Essas limitações definiram o processo de escolhados endereços das viagen. Procurou-se sortear endereços em regiões mais densamente povoadas para evitar possíveis perdas, seguindo um modelo de densidade de probabilidade da população a partir dos setores cenitários do censo. Ao mesmo tempo, para reduzir o esforço computacional, foi montada uma base de coordenadas *offline*, que foi usada para sortear os endereços das viagens. O processo de geração dessa base de endereços foi feito nas seguintes etapas: 1) O mapa do município de São Paulo (em formato *shapefile*) foi dividido por uma grade com quadrículas de 500 metros de lado; 2) Foram calculadas as populações de cada quadrícula com dados do Censo Demográfico de 2010 e retiradas as quadrículas com população igual a zero; 3) As quadrículas foram divididas em quintis de densidade populacional, e sorteados aleatoriamente pontos geográficos dentro de cada quadrícula, de acordo com o quintil: 5 pontos para o quintil mais populoso, 4 para o 2º quintil, 3 para o 3º, 2 para o 2º e 1 ponto para o quintil menos populoso. 4) O conjunto de pontos resultante foi usado como base para o sorteio dos endereços de origem e destino. Essa primeira etapa foi realizada utilizando bases cartográficas abertas do município de São Paulo e o software aberto QGIS. A simulação das viagens foi feita a partir de um programa desenvolvido em Python, executado no serviço de computação em nuvem da Google. A estrutura do programa seguiu a arquitetura apresentada na Figura 1. Foi usada uma ferramenta de agendamento (*Google Scheduler*) de ativação ligada a uma máquina virtual no ambiente em nuvem da Google, para que a chamada ao serviço de viagem do Google Maps fosse realizada nos dias úteis da semana, a cada hora cheia, das cinco da manhã até às nove da noite. O programa seguiu as etapas descritas na Figura 1.

* Abrir uma conexão com o banco de dados SQL da nuvem da Google
* Carregar o banco de coordenadas
* Sortear dez coordenadas de origem e dez coordenadas de destino
* Chamar a API Distance Matrix com as dez origens e os dez destinos, para viagens de transporte público
* Processar os resultados devolvidos pela API e armazenar em um vetor auxiliar
* Chamar novamente a API Distance Matrix com as mesmas dez origens e os dez destinos, para viagens de transporte privado
* Processar os resultados devolvidos pela API e anexar ao vetor auxiliar
* Submeter o vetor auxiliar à função que insere os dados no Banco de Dados hospedado na nuvem.

**Figura 1.** Etapas e Estrutura funcional do programa elaborado para simular os dados  
Fonte: Elaborado pelo autor

Cada chamada da API Distance Matrix retornou uma lista com duzentas viagens com dados de horário e dia da semana, coordenadas da origem e do destino, endereços da origem e do destino da viagem, duração, distância e tarifa da viagem. Para cada par origem-destino houve registro de viagem de carro e de transporte público. O período de simulação foi entre os dias 11 de fevereiro de 2019 a 5 de junho de 2019. O total de viagens armazenadas no banco de dados nesse período foi de 257.400 viagens, sendo 253.450 viagens válidas – 128.700 (100% de aproveitamento) das viagens de carro e 126.725 das viagens transporte público (98,47% de aproveitamento).

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



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

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

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

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

Dada uma viagem OD, foram selecionadas todas as viagens simuladas com a origem dentro de um *buffer* de 1500m (escolhido por incluir 95% dos trajetos a pé para acessar os transportes); dessas viagens foi selecionado ainda o subconjunto de viagens simuladas em que os pontos de destino também se encontrassem dentro de um *buffer* de 1500m do ponto de destino da viagem OD. A partir desse conjunto de viagens simuladas foi calculada a média da duração simulada (considerando se a viagem comparada era de transporte público ou privado), que por sua vez foi subtraída da soma da duaração da viagem OD, do tempo andando da origem e do tempo até o destino da viagem OD, convertidos em segundos. Esse algoritmo foi repetido para todas as viagens OD e a diferença calculada foi registrada para cada viagem. Uma variante desse algoritmo foi realizada também, adicionando mais uma restrição para a seleção das viagens: além dos *buffers* espaciais, foi utilizado um *buffer* temporal, filtrando as viagens simulada que começassem duas horas antes ou depois da viagem OD.

**3.3. Análise dos dados**

A analsie dos dados verificou se é possivel identificar uma diferença consistente entre as medidas simuladas de tempo e as medidas registradas na OD e em caso de haver essa diferença, procurar identificar quais variáveis associadas as viagens poderiam explicar a diferença entre as medidas e se existe alguma dependencia espacial envolvida.

Para verificar a existencia de diferença foi feito um teste T de diferença de médias entre as medidas de tempo da OD e calculadas a partir da simulação. Os teste foram realizados separadamente para o conjunto de viagens motorizadas privadas e para as viagens de transporte público e para os cenários com e sem o buffer de restrição de tempo. Foram verificadas diferenças nesses cenários e o passo seguine foi elaborar modelos lineares (OLS) utilizando as variaveis descritas na Tabela X como variáveis explicativas. Para selecionar as variáveis significantes foi utilizado um processo stepwise atraves do pacote *lembrarQual* do software R. Para a análise de dependencias espaciais, as medidas foram espacializadas dentro das zonas OD, calculando-se a média em torno das zonas de origem e de destino. As variáveis selecionadas pelo OLS foram espacializadas (explicar como) e foram feitos modelos de autocorrelação espacial para os oito cenários possíveis de agragação e de condições (publico e privado X origem e destino X com e sem restrição de tempo).

# **4. RESULTADOS E ANÁLISE**

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

# **REFERÊNCIAS BIBLIOGRÁFICAS**