FORUM

Submitted 10.19.2018. Approved 07.19.2019
Evaluated through a double-blind review process. Guest Scientific Editors: Eduardo de Rezende Francisco, José Luiz Kugler, Soong Moon Kang, Ricardo Silva, and Peter Alexander Whigham
Original version
DOI: http://dx.doi.org/10.1590/S0034-759020190606

MEASURING ACCESSIBILITY: A BIG DATA PERSPECTIVE ON UBER SERVICE WAITING TIMES

Medindo a acessibilidade: Uma perspectiva de Big Data sobre os tempos de espera do serviço da Uber

Medición de accesibilidad: Una perspectiva de Big Data sobre los tiempos de espera del servicio de la Uber

ABSTRACT

This study aims to relate information about the waiting times of ride-sourcing services, with specific reference to Uber, using socioeconomic variables from São Paulo, Brazil. The intention is to explore the possibility of using this measure as an accessibility proxy. A database was created with the mean waiting time data per district, which was aggregated to a set of socioeconomic and transport infrastructure variables. From this database, a multiple linear regression model was built. In addition, the stepwise method selected the most significant variables. Moran's I test confirmed the spatial distribution pattern of the measures, motivating the use of a spatial autoregressive model. The results indicate that physical variables, such as area and population density, are important to explain this relation. However, the mileage of district bus lines and the non-white resident rate were also significant. Besides, the spatial component indicates a possible relation to accessibility.

KEYWORDS | Accessibility, Big Data, Uber, space statistic, urban disparity.

RESUMO

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

PALAVRAS-CHAVE | *Acessibilidade, Big Data, Uber, estatística espacial, disparidade urbana.*

RESUMEN

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

PALABRAS CLAVE | *Accesibilidad, Big Data, Uber, estadística espacial, disparidad urbana.*

ANDRÉ INSARDI[1]
andre.insardi@espm.br
ORCID: 0000-0003-3782-3505

RODOLFO OLIVEIRA LORENZO[2]
rodolfolo@uol.com.br
ORCID: 0000-0003-4847-9201

[1]Escola Superior de Propaganda e Marketing, São Paulo, SP, Brazil

[2]Fundação Getulio Vargas, Escola de Administração de Empresas de São Paulo, São Paulo, SP, Brazil

INTRODUCTION

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

The increase in competition in a traditionally highly regulated market led to conflicts in many cities between the new services, the former suppliers of this market, taxis, and the local authorities. The difficulties faced by local authorities in framing these new services within existing legal frameworks is clearly illustrated in the case of San Francisco (Flores & Rayle, 2017).

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

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

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

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

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

In this context, the growth of global tech companies in mobility, such as Uber, Cabify and Lyft, has made them considerable players in this field. For instance, Uber has a daily average of 15 million rides across the world. In Brazil, the company provides its service in 100 cities with a network of 500 000 drivers and more than 20 million users.

Uber has developed, commercialized, and operates the application for smartphones that allows consumers to request rides from partnered drivers. In the process of requesting the rides, the Uber tool provides estimates of waiting time and travel cost on the user's app and in the web environment through public application programming interfaces (APIs).

These APIs generate estimates through the analysis of ride history in the user's region and the supply-and-demand curve of Uber's cars (Cohen et al., 2016) as can be verified in the available documentation (https://developer.uber.com).

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

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

As Uber's pricing algorithm follows the balance between supply and demand in order to influence driver's behavior (Hall, Horton, & Knoepfle, 2019), we assume that waiting times for Uber rides can reflect regional imbalances in the supply of cars. Following this line of thought, we have explored the relationship

  © RAE | São Paulo | **59(6)** | **November-December 2019** | 402-414 ISSN 0034-7590; eISSN 2178-938X

of waiting times with other variables associated with accessibility in similar literature (Hughes & MacKenzie, 2016; Lessa et al., 2019; Wang & Mu, 2018). Particularly, related to mobility and socioeconomic factors.

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

LITERATURE REVISION AND RESEARCH QUESTION

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

Big Data

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

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

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

"Crumbs" is 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, constitute the databases to be analyzed in Big Data. "Capacities" are the techniques, both statistical and programming, used to manipulate these data and extract information. The third concept, "Communities," refers to the behavior patterns of the producers of Big Data environments inside specific communities whose members share ideas with specific language and common validation methods. These communities can be established in open and collaborative environments, such as OpenSource communities, or in more restrictive ones, like tech groups in big corporations with access to big databases (Letouzé & Jütting, 2015).

The Uber API in the Big Data context

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

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

Furthermore, Pääkkönen and Pakkala's (2015) description of Big Data software architecture suits Uber's API architecture and technology as described by its documentation (https://developer.uber.com) very well.

Accessibility

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

  © RAE | São Paulo | **59(6)** | **November-December 2019** | 402-414 ISSN 0034-7590; eISSN 2178-938X

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

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

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

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

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

In discussing accessibility, Páez et al. (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 positivism and is defined in terms of observed accessibility parameters, or how much people do travel. The normative approach analyzes travel expectations while positivism bases itself on the actual travel experience. This study will take into consideration the positivist approach as the data refers to actual Uber waiting times in São Paulo.

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

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

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

> H_1: The estimated waiting time when requesting a ride on Uber's platform can be used as a proxy for accessibility;
>
> H_2: Uber waiting time distribution relates to socioeconomic indicators' polarization.

METHODOLOGY

This section begins with a description of the data collected 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 multiple linear regression (MLR) models with stepwise method is explained. As the significant variables selected by this method were found by Moran's I test to have high spatial dependency, a final spatial autoregressive (SAR) model was calculated, as suggested by Wang & Mu (2018).

   ISSN 0034-7590; eISSN 2178-938X

Data and variables

To verify the hypotheses elaborated above, two sets of data were collected. The first is the data from Uber waiting times, and the second, the spatialized socioeconomic data from São Paulo.

Uber Data

The data were collected from the APIs present in Uber's Developers portal, accessible from the company's site (https://developer.uber.com), which cover the estimated waiting times of all Uber products during August 2018.

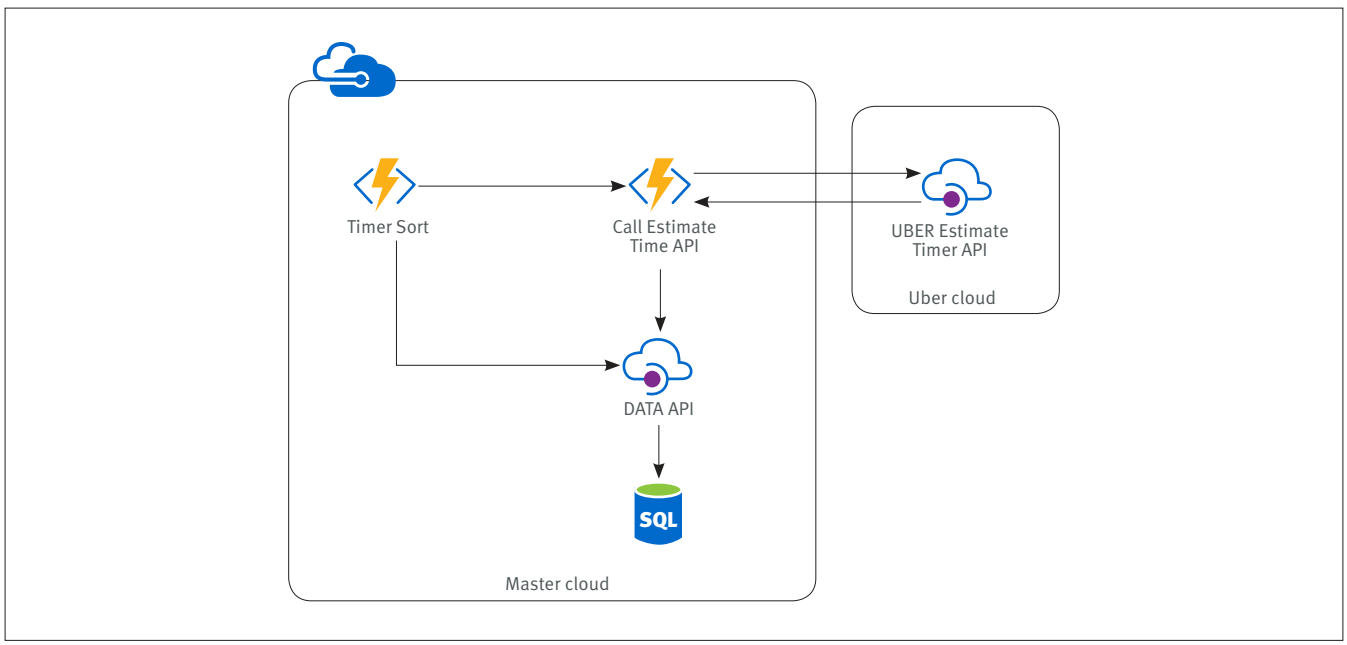
The city of São Paulo is divided into 96 districts, a territorial and administrative division that provides the local administration a certain degree of autonomy (Francisco, 2010). Following the methodology proposed by Wang and Mu (2018), the districts were used as the spatial unit of analysis. To guarantee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

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

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

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

Figure 1. **Diagram of the Uber data collection application**



Socioeconomic data

The choice of socioeconomic indicators was based on similar studies (Hughes & MacKenzie, 2016; Wang & Mu, 2018). The analysis units chosen were the districts of São Paulo. Comparable variables were found for population density, employment density, minority rate, mean income per capita, motorization rate (for cars and motorcycles), public transport infrastructure, and mean travel time to work. The socioeconomic data at the district level was collected from the 2010 IBGE (Brazilian Institute of Geography and Statistics) demographic census, the National Ministry of Work and Employment (RAIS – Anual Report of Social information), the SEADE's São Paulo State municipalities Indicators web portal, and the São Paulo Municipal portal of Geo-referenced information (http://geosampa.prefeitura.sp.gov.br) and is summarized in Exhibit 1.

© RAE | São Paulo | **59(6)** | **November-December 2019** | 402-414 ISSN 0034-7590; eISSN 2178-938X

Exhibit 1. **São Paulo's socioeconomic data**

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

The decennial Demographic census is one of Brazil's most important statistical products. The 2010 census presents two sets of data: the universal, which 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 included in the Sample. Across the whole of Brazil 10,7% were selected, or 6,192,332 households (IBGE, 2010).

The RAIS is the Ministry of Work and Employment's tool for management of Brazilian work relations. The data is compiled from statements made by businesses about the working situation between them and their employees. The declarations are mandatory for a series of businesses nominated by law (Ministério do Trabalho, 2016).

The SEADE Foundation is a nationally recognized statistical institution. The foundation is known for its technical capacity. It

   ISSN 0034-7590; eISSN 2178-938X

is also responsible, as part of the National Statistic System for producing data, for aggregating existing data in the interests of São Paulo State and its municipalities (Francisco, 2010).

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

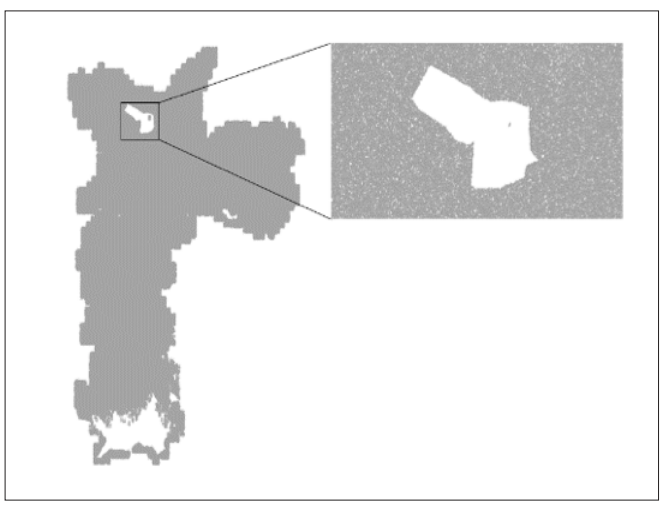
All data was searched or aggregated to the district level (Francisco, 2010; Hughes & MacKenzie, 2016; Lessa et al., 2019; Wang & Mu, 2018). The 2010 census Sample was used to account for the percentage of the population of each district that is in different classes of travel time to work/school, and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

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

Descriptive data analysis

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

Figure 2. **Coordinate points for Uber waiting time estimates**



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

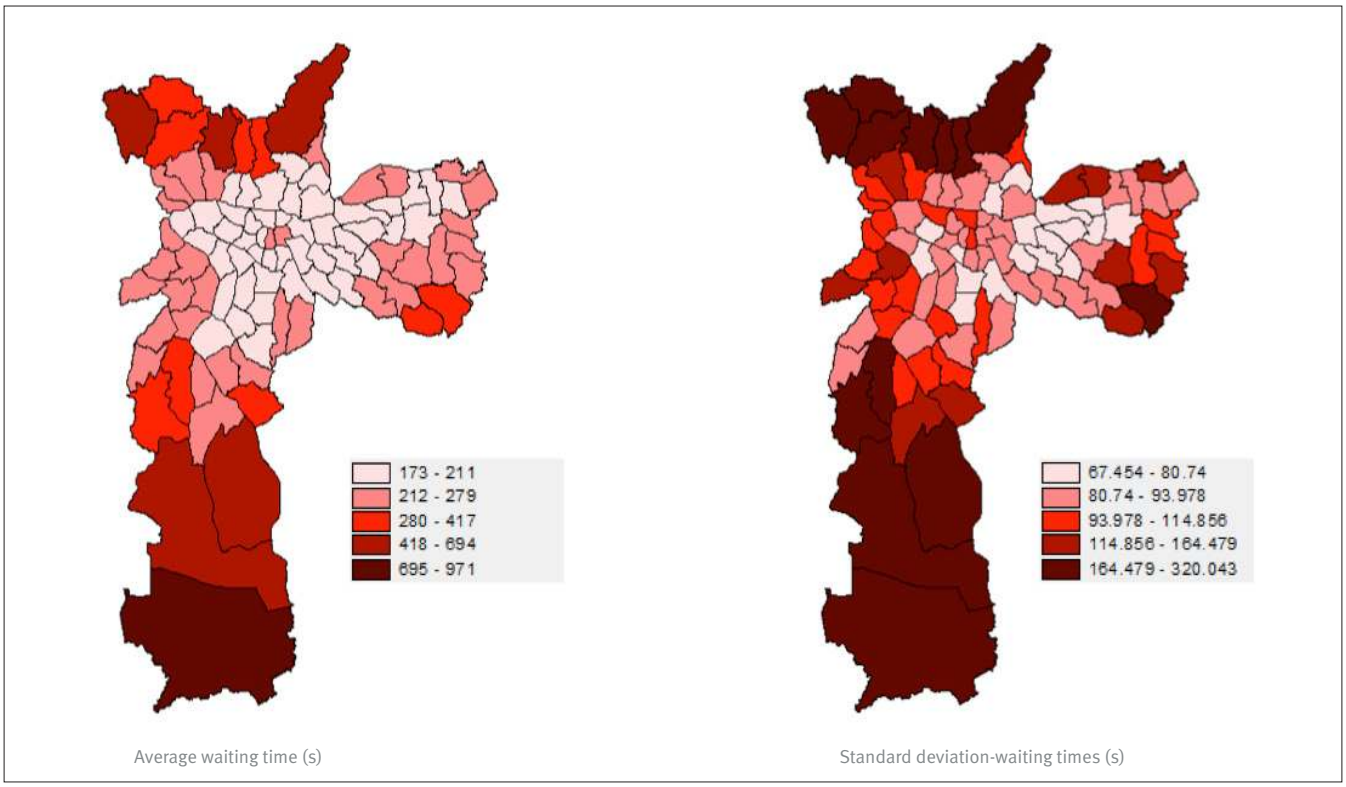
The cover of each service and its availability can be seen in Table 1. UberX, the most popular service Uber provides, has the largest coverage with almost 100% of time estimate responses registered. Consequently, it is the service that better reflects Uber's fleet spatial distribution. Therefore, UberX waiting times were adopted for the purpose of measuring accessibility.

Table 1. **Returns of calls of Uber waiting 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

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

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



Average waiting time (s)

Standard deviation-waiting times (s)

Multiple linear regression (MLR)

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

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

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

The software R and its extensions "stats" and "car" were used for the computation of the regressive models.

Spatial auto regressive model (SAR)

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

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

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

ISSN 0034-7590; eISSN 2178-938X

RESULTS

Table 2 summarizes the results of the MLR models with the dependent variable as the average of UberX waiting times and standard deviation of waiting times. It is possible to note that the average time model has a better degree of explanation as it has an R^2 of 0.893 against an R^2 of 0.717 for the standard deviation model.

Table 2. **Average and standard deviation regressions result for UberX**

DV		Average Uber X			Std. Deviation Uber X		
		Coefficient	Std. Error	p-value	Coefficient	Std. Error	p-value
(Intercept)		943.00	750.40	0.213	39.35	515.30	0.939
QTLINBUS2018	Bus lines' quantity	0.13	0.19	0.514	0.12	0.13	0.353
KMLINBUS2018	Bus lines' length (Km)	-0.12	0.07	0.080	-0.09	0.05	0.057
QTPONTBUS2018	Number of bus stops	-0.12	0.13	0.361	0.01	0.09	0.902
QTESTMETRO2018	Quantity of Metro stations	0.91	5.22	0.862	-1.14	3.58	0.751
RENDP2010	Income per Capita	-0.00	0.01	0.922	0.00	0.01	0.515
ARE1	Area (km^2)	3.55	0.31	0.00	0.74	0.21	0.001
POP2018	Population	0.00	0.00	0.492	-0.00	0.00	0.894
DENPOP2018	Population density	-0.00	0.00	0.007	-0.01	0.00	0.000
DOMP2018	Number of particular permanent households	-0.00	0.00	0.646	0.00	0.00	0.546
ESTAB2016	Employers	-0.00	0.01	0.970	-0.00	0.00	0.791
EMP2016	Jobs	0.00	0.00	0.384	0.00	0.00	0.547
PNBRAN2010	Proportion of non-whites	300.20	77.69	0.000	202.20	53.35	0.000
TEMP2010_5MIN	Travel time - up to 5 minutes	-587.50	1,166.00	0.616	111.00	800.50	0.890
TEMP2010_30MIN	Travel time - from 6 to 30 minutes	-978.90	785.70	0.217	37.68	539.50	0.945
TEMP2010_60MIN	Travel time - from 31 to 60 minutes	-278.40	831.70	0.739	180.80	571.10	0.752
TEMP2010_120MIN	Travel time - from 61 to 120 minutes	-1,046.	805.20	0.198	-141.10	552.90	0.799
TEMP2010_121MIN	Travel time - more than 121 minutes	-1,443.	992.80	0.150	-215.90	681.80	0.752
TEMP2010_0MIN	No travel	-821.30	765.00	0.286	10.86	525.30	0.984
PDOMC2010	Car - rate of household motorization	-139.10	367.70	0.706	-164.10	252.50	0.518
PDOMM2010	Motorcycle - rate of household motorization	37.73	82.42	0.648	25.17	56.60	0.658

A stepwise MLR model was created analyzing the collinearity of the independent variables of the final model, it was considered convenient to work with variance inflation factors (VIF's) of less than 5 as recommended by Batterham, Tolfrey, and George (1997).

Besides that, the model selected was the mean of waiting time, because of its greater power of explanation. Table 3 shows the result of the model. Meanwhile, Table 4 shows the collinearity analysis of the selected independent variables for this final model. We can observe a high degree of significance for bus line length, district area, population density, jobs, percentage of non-whites, and travel times over 120 minutes, with an R-squared of 0.879. These elements corroborate the positivist view of Lessa et al.'s (2019) and Páez et al.'s (2012) studies.

Table 3. **Stepwise regression results for UberX average waiting times**

DV		Average Uber X		
		Coefficient	Std. Error	p-value
(Intercept)		171.00	18.17	0.000
KMLINBUS2018	Bus lines' length (Km)	- 0.0786	0.02	0.002
ARE1	Area (km²)	3.634	0.23	0.000
DENPOP2018	Population density	- 0.0035	0.0009	0.000
EMP2016	Jobs	0.00021	0.0001	0.097
PNBRAN2010	Proportion of non-whites	295.70	44.57	0.000
TEMP2010_121 MIN	Travel time - more than 121 minutes	- 1,239,00	422.50	0.004

Table 4. **Variance inflation factors analysis**

VIF		Average Uber X
		Coefficient
KMLINBUS2018	Bus line length (Km)	1.181254
ARE1	Area (km²)	2.139971
DENPOP2018	Population density	1.317234
EMP2016	Jobs	1.894758
PNBRAN2010	Proportion of non-whites	2.462318
TEMP2010_121MIN	Travel time - more than 121 minutes	2.615456

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

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

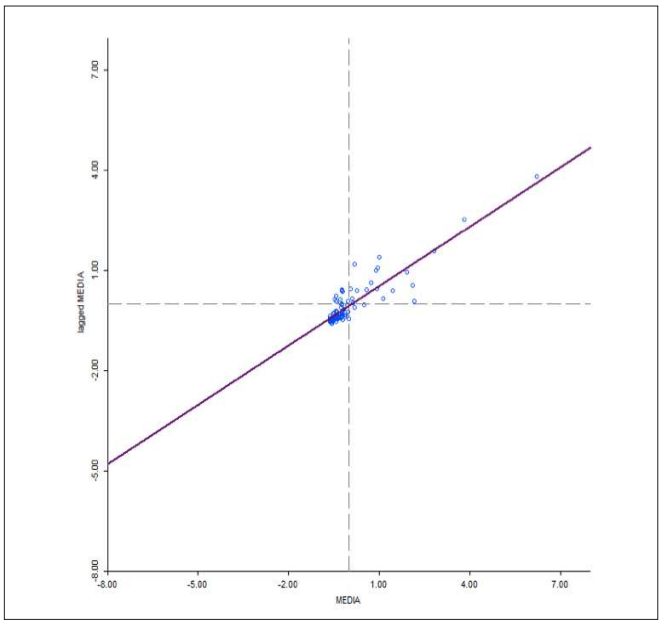
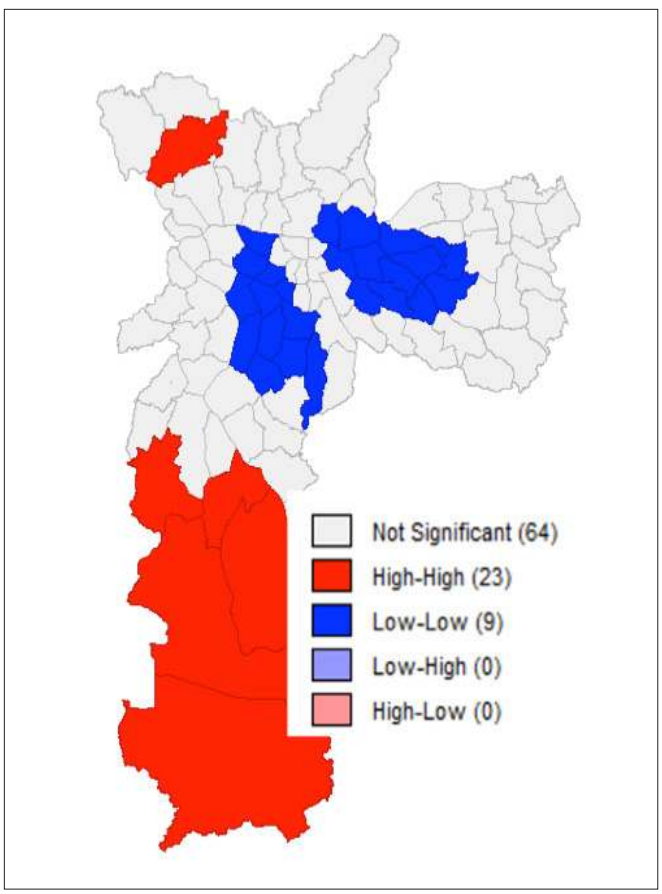


Figure 5. **Uber X average waiting time for neighborhood clusters**



   ISSN 0034-7590; eISSN 2178-938X

From the variables in Table 3 we generated a spatial regression model for the average Uber X waiting time. The results presented in Table 5 show a high degree of significance for the variables bus line length, district area, population density, and percentage of non-whites. With the incorporation of the geographic factor into the model, the R-squared jumps to 0.89.

Table 5. **Spatial lag regression results of Uber X**

DV		Average Uber X		
		Coefficient	Std. Error	p-value
W_MEDIA		0.406	0.092	0.000
CONSTANT		114.573	23.246	0.000
KMLINBUS2018	Bus lines' length (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

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

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

When analyzing the results, we cannot confirm H_1, that is, we cannot affirm that the estimation of service waiting time can be used as an accessibility proxy. We assume this, because for some variables the relation with Uber X waiting time was not significant.

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

DISCUSSION AND CONCLUSION

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

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

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

Public management can make use of similar tools to develop a "basket" of indicators for different modes of transportation and measurements regarding their interrelations. These can be used to better regulate existing activity and to

   ISSN 0034-7590; eISSN 2178-938X

tailor public service delivery in mobility, as with a more intimate integration between public and private modes. By maintaining a real time measurement of the effect of public regulation in aspects of the transport system, it could be possible to better adjust intervention in the interests of the users (Jin et al., 2018), given that well balanced indicators can reflect user behavior and needs (Lessa et al., 2019; Páez et al., 2012).

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

REFERENCES

Aranha, V. (2005). *Mobilidade pendular na metrópole paulista*. *São Paulo em Perspectiva*, *19*(4), 96-109. doi:10.1590/S0102-88392005000400006

Batterham, A. M., Tolfrey, K., George, K. P. (1997). Nevill's explanation of Kleiber's 0.75 mass exponent: An artifact of collinearity problems in least squares models? *Journal of Applied Physiology*, *82*(2), 693-697. doi:10.1152/jappl.1997.82.2.693

Cohen, P., Metcalfe, R., Angrist, J., Chen, K., Doyle, J., Farber, H., Hahn, R. (2016). Using big data to estimate consumer surplus: The case of Uber. *National Bureau of Economic Research* (Working paper nº 22627). Retrieved from https://www.nber.org/papers/w22627 doi: 10.3386/w22627

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

Francisco, E. de R. (2010). *Indicadores de renda baseados em consumo de energia elétrica: abordagens domiciliar e regional na perspectiva da estatística espacial* (Doctoral Thesis, Fundação Getulio Vargas). Retrieved from https://bibliotecadigital.fgv.br/dspace/handle/10438/8158

Hair, J., Babin, B., Money, A., & Samouel, P. (2005). *Fundamentos de métodos de pesquisa em administração*. Porto Alegre, RS: Bookman Companhia Ed.

Hall, J. V, Horton, J. J., & Knoepfle, D. T. (2019). *Pricing efficiently in designed markets: The Case of Ride-Sharing*. 1-76. Retrieved from http://john-joseph-horton.com/papers/uber_price.pdf

Hall, J. V, & Krueger, A. B. (2016). An analysis of the labor market for Uber's driver-partners in the United States. In *National Bureau of Economic Research* (Working paper nº 22843). Retrieved from https://www.nber.org/papers/w22843. doi:10.3386/w22843.

Hansen, W. G. (1959). How accessibility shapes land use. *Journal of the American Institute of planners, 25*(2), 73-76. doi:10.1080/01944365908978307

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. doi:10.1016/j.jtrangeo.2016.08.014

Instituto Brasileiro de Geografia e Estatística. (2010). Censo demográfico. *Notas Metodológicas–Microdados da Amostra*. *Rio de Janeiro*. Retrieved from ftp://ftp.ibge.gov.br/Censos/Censo_Demografico_2010/Resultados_Gerais_da_Amostra/Microdados/Documentacao.zip

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

Kim, B. G., Trimi, S., & Chung, J. (2014). Big-Data applications in the government sector. *Communications of the ACM*, *57*(3), 78-85. doi:10.1145/2500873

Krajzewicz, D., Erdmann, J., Behrisch, M., & Bieker, L. (2012). Recent development and applications of SUMO-Simulation of Urban MObility. *International Journal On Advances in Systems and Measurements, 5*(3-4), 128-138.

Kwan, M. (1998). Space-time and integral measures of individual accessibility: A comparative analysis using a point-based framework. *Geographical Analysis, 30*(3), 191-216. doi:10.1111/j.1538-4632.1998.tb00396.x

Kwan, M. P. (2016). Algorithmic geographies: Big data, algorithmic uncertainty, and the production of geographic knowledge. *Annals of the American Association of Geographers, 106*(2), 274-282. doi:10.1080/00045608.2015.1117937

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

Letouzé, E., & Jütting, J. (2015). Official statistics, Big Data and human development: towards a new conceptual and operational approach. *Data Pop Alliance White Paper Series*. Retrieved from https://paris21.org/sites/default/files/WPS_OfficialStatistics_June2015.pdf

Litman, T. (2003). Measuring transportation: Traffic mobility and accessibility. *Victoria Transport Policy Institute. ITE Journal*, *73*(10), 28-32.

McAfee, A., & Brynjolfsson, E. (2012). Big Data: The Management Revolution. *Harvard Business Review*. Retrieved from https://hbr.org/

Metrô. (2008). Pesquisa Origem-Destino 2007. *Secretaria de Transportes Metropolitanos*. Retrieved from http://www.metro.sp.gov.br/pesquisa-od/arquivos/OD_2007_Sumario_de_Dados.pdf

Ministério do Trabalho (2016). Relação Anual de Informações Sociais - RAIS ano-base 2016. *Portaria n. 1464 de 30 de Dez. 2016. Aprova Instruções Para a Declaração Da Relação Anual de Informações Sociais - RAIS Ano-Base 2016*. Retrieved from http://www.rais.gov.br/sitio/sobre.jsf

Morandi, E., Ribeiro, R., Hernandes, E., Camara, B., Spinola, L., & Francisco, E. D. R. (2016). *Análise Geoespacial da Relação entre Transporte Público sobre Trilhos, Renda e Tempo Médio de Deslocamento*. Paper presented on XL EnANPAD, Salvador, BA. Retrieved from http://www.anpad.org.br/~anpad/eventos.php?cod_evento=1&cod_edicao_subsecao=1302&cod_evento_edicao=83&cod_edicao_trabalho=21023

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*(9). doi:10.1371/annotation/ca85bf7a-7922-47d5-8bfb-bcdf25af8c72

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

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. doi:10.1016/j.bdr.2015.01.001

Preston, V., & McLafferty, S. (1999). Spatial mismatch research in the 1990s: Progress and potential. *Papers in Regional Science*, *78*(4), 387-402. doi:10.1111/j.1435-5597.1999.tb00752.x

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

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

Stelder, D. (2016). Regional accessibility trends in Europe: Road Infrastructure, 1957-2012. *Regional Studies*, *50*(6), 983-995. doi:10.1080/00343404.2014.952721

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

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

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

Weibull, J. W. (1980). On the numerical measurement of accessibility. *Environment and Planning A: Economy and Space, 12*(1), 53-67. doi:10.1068/a120053

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. doi:10.1016/j.apgeog.2016.10.002