Journal of Transport Geography 85 (2020) 102671

Contents lists available at ScienceDirect

Journal of Transport Geography

journal homepage: www.elsevier.com/locate/jtrangeo

Estimating the influence of crowding and travel time variability on accessibility to jobs in a large public transport network using smart card big data

Renato Arbex*, Claudio B. Cunha

Escola Politécnica da USP, University of São Paulo, São Paulo, Brazil

ARTICLE INFO

Keywords:
Public transport
Accessibility
Smart card data
In-vehicle crowding
Travel time reliability

ABSTRACT

Accessibility metrics are gaining momentum in public transportation planning and policy-making. However, critical user experience issues such as crowding discomfort and travel time unreliability are still not considered in those accessibility indicators. This paper aims to apply a methodology to build spatiotemporal crowding data and estimate travel time variability in a congested public transport network to improve accessibility calculations. It relies on using multiple big data sources available in most transit systems such as smart card and automatic vehicle location (AVL) data. São Paulo, Brazil, is used as a case study to show the impact of crowding and travel time variability on accessibility to jobs. Our results evidence a population-weighted average reduction of 56.8% in accessibility to jobs in a regular workday morning peak due to crowding discomfort, as well as reductions of 6.2% due to travel time unreliability and 59.2% when both are combined. The findings of this study can be of invaluable help to public transport planners and policymakers, as they show the importance of including both aspects in accessibility indicators for better decision making. Despite some limitations due to data quality and consistency throughout the study period, the proposed approach offers a new way to leverage big data in public transport to enhance policy decisions.

1. Introduction

As the more intense use of private cars for commuting in major cities generates problems and externalities such as congestion, pollution, and increased stress levels, public transport systems are becoming more important as a viable sustainable alternative to urban growth. To plan and design such systems, accessibility metrics are slowly gaining momentum towards being used as indicators to guide future decision making (Boisjoly and El-Geneidy, 2017). By calculating the number of opportunities reachable from each location throughout a city within a travel time threshold, planners can evaluate the benefits of new proposed transportation infrastructure and decide upon multiple improvement plans.

In addition, public transport systems in many cities around the world are nowadays very crowded, especially during peak hours. Moreover, deprived urban areas may suffer from the unreliability of travel times, forcing users to budget more travel time to safeguard their on-time arrivals (Arbex et al., 2016). Despite a relevant stream of research focusing on user time valuation of crowding discomfort and on travel time reliability, there remains a gap to evaluate how crowded operating conditions and travel time variability of public transport systems impact metrics related to accessibility to opportunities in cities.

In such a context, this paper aims to investigate how and to what extent travel time unreliability, the intensity of crowding discomfort and its spatial distribution affect accessibility to opportunities as perceived by users. Public transportation big data from smart cards and automatic vehicle location (AVL) for buses over a large period are used to achieve such objective. Months of data reveal crowding patterns regularity and travel time variations over time, while fine-grained crowding levels are used to measure perceived accessibility to jobs with crowding time valuation. Accessibility inequalities emerge from the results when including both crowding and travel time variability data. Finally, we discuss policy implications from such analysis.

Thus, this paper aims to address the following questions:

- How to estimate passenger load density in a very high spatio-temporal resolution?
- How to infer buffer travel time using public transport datasets, and what are the limitations?
- What are the impacts of including crowding discomfort valuation

* Corresponding author.
E-mail addresses: renatoarbex@usp.br (R. Arbex), cbcunha@usp.br (C.B. Cunha).

https://doi.org/10.1016/j.jtrangeo.2020.102671
Received 9 January 2019; Received in revised form 3 February 2020; Accepted 16 February 2020
Available online 23 April 2020
0966-6923/

and travel time variability in accessibility calculations for different city areas?

The remainder of the paper is organized as follows: in the next section the background and literature overview regarding smart card data use in transport research, accessibility, crowding valuation and travel time reliability are presented; Section 3 describes the proposed methodology. The results and discussions are presented in Section 4, followed by the concluding remarks in Section 5.

2. Background

As the aim of this paper is to delve into the impact of crowding and travel time reliability on accessibility to opportunities, four main streams of research are analyzed in the following subsections: (1) accessibility indicators; (2) smart card data use in transportation planning; (3) crowding discomfort in public transport; and (4) travel time reliability.

2.1. Accessibility

Accessibility in urban and transportation planning is a concept formulated long ago by Hansen (1959), defined as the potential of opportunities for interaction, representing the desire and ability to overcome spatial distances to access spatially distributed activities such as employment and social interactions. In-depth reviews of accessibility measures are found in Geurs and van Wee (2004) and Páez et al. (2012), including theoretical basis, interpretability, and communicability, as well as data requirements.

Of the main perspectives on accessibility measurement, the *location-based* one is related to the level of accessibility to spatially distributed activities, such as the number of jobs accessible within a specified time threshold (Geurs and van Wee, 2004). This metric is used by researchers to evaluate current and proposed transportation infrastructure. Hernandez (2018) uses a cumulative opportunity indicator to measure accessibility to jobs and education to explore the unequal access to urban opportunities among different social classes in Montevideo. Pereira (2019) evaluates accessibility for an on-going BRT project and discusses how different cut-off times for cumulative opportunity measures based on a single travel time threshold provides a distinct interpretation of results.

Although accessibility metrics have been extensively researched in the last decades, it is still marginalized in transportation planning practice (Boisjoly and El-Geneidy, 2017). From a comprehensive review of accessibility indicators of 32 recent metropolitan transport plans, Boisjoly and El-Geneidy (2017) suggest that there is “a trend toward a greater integration of accessibility objectives in transport plans, yet few plans have accessibility-based indicators that can guide their decision-making processes.”. Even so, accessibility analyzes in public transport networks are important as they help transit agencies and planners to identify areas that need improvements in transit service provision and transit investment priorities (Fayyaz et al., 2017). In this sense, recent initiatives aim to bring accessibility closer to transportation planning: for instance, Stewart (2017) proposes collaborative accessibility mapping for enhanced stakeholder engagement with a visualization tool, while Walker (2018) discusses how accessibility (“freedom”) should be a central evaluation criterion for transportation projects and equity analysis.

Some authors use automatic vehicle location data to address the problem of measuring actual accessibility as experienced by transit riders. For instance, based on AVL data to represent actual service, Stewart (2017) applies spatial analysis techniques to adjust accessibility metrics to account for the unreliability revealed in actual transit operations. Wessel and Farber (2019) compare accessibility measurements derived from schedule-based General Transit Feed System (GTFS) data with those obtained from GPS-based AVL data. They show that scheduled-based accessibility measures may overestimate net accessibility by 5–15% on average, and the relative difference between schedule and retrospective datasets decreases as one moves towards hourly measures of accessibility.

Other authors have focused on travel time variability in accessibility measures. For instance, Boisjoly and El-Geneidy (2016) derive time-sensitive analysis for different periods; their results show that the most commonly used accessibility measure (at 8 am) is representative of the relative accessibility (static or dynamic) throughout the day. Farber and Fu (2017) propose a new data object called travel time cube to explore the spatiotemporal patterns of public transit accessibility; demonstrating its use in three case studies that evaluate how different types of changes to the public transit network impact a variety of travel time characteristics in a region. Conway et al. (2018) describe a method to extend the concept of reliable accessibility to public transport; a Monte Carlo approach is applied to generate possible timetables that allow the estimation of percentiles of travel time to guarantee a certain desired probability of on-time arrival.

2.2. Smart card data

The adoption of smart cards in public transport systems in recent years has provided an opportunity for automatic generation of large data streams with great potential for improving planning and operation in urban public transport systems (Pelletier et al., 2011). Proper data analyses generate information such as travel patterns, and number of transfers (Bagchi and White, 2005), full trip inference (Trépanier et al., 2007) and route choice analysis (Wilson and Hemily, 2016). A comprehensive review of distinct aspects of smart card data use in public transit planning is provided by Pelletier et al. (2011).

One of the main uses of smart card information is the inference of origin-destination (OD) matrices, for which information on boarding and alighting location and respective times must be known, either recorded directly in data collection or through estimation methods. For transport systems that do not require tap outs of smart cards on the exit, the alighting location for each transaction should be inferred (Barry et al., 2002; Trépanier et al., 2007). One alternative commonly found in the literature is to apply trip chaining for destination inference, where the destination of each transaction is estimated to be near the next boarding location (see, for instance, Alsger et al., 2016; Gordon et al., 2013; He et al., 2015; Munizaga and Palma, 2012; Seaborn et al., 2009; Wang et al., 2011). For an in-depth review of trip chaining and transfer inference considerations, the reader is referred to Hickman (2016), while a comprehensive overview of destination inference models beyond trip chaining can be found in Li et al. (2018).

Although smart cards allow a richer assessment of public transport systems, specific challenges arise, due to factors such as erroneous software, erroneous data, and faulty hardware, which seem to have been overlooked in the literature (Robinson et al., 2014). Challenges may include data correction, data quality improvement, ability to process large volumes of data, distortion in longitudinal series, lack of data, and expansion of data samples (Wilson and Hemily, 2016).

2.3. Crowding

Crowding in public transportation systems affect urban commuters in many large cities in the world. As the number of passengers increases relative to the available capacity in public transport services, customers tend to become less willing to use it, which can be considered a user cost, just the same as the value of time for travel that users allocate (Hörcher et al., 2017). Crowding also generates multiple impacts on users' health and personal wellbeing, ranging from increased anxiety, stress, feelings of exhaustion, perceptions of risk and safety, feelings of invasion of privacy and propensity to arrive late at work, impacting organizational health and general quality of living (Tirachini et al., 2013). Contributing drivers of discomfort are dissatisfaction with standing and not being seated, fewer opportunities to make use of time during the journey and physical closeness of other travelers (Haywood et al., 2017). Also, benefits of a frequency increase on a congested transit line and a transformation from bus to tram can be underestimated when comfort is not incorporated in the demand modeling framework (Van Oort et al., 2015). Therefore, it is crucial to include crowding effects on travel time valuation for accessibility to opportunities calculation in transport planning.

Load factor and level of service (LOS) are measures used to describe crowding levels in public transport systems (Kittelson and Associates et al., 2013; Li and Hensher, 2013). Load factor is defined as the number of passengers per seat available in the vehicle, while LOS refers to thresholds based on load factor or passengers per area unit, ranging from A to F, representing the best quality of service to worst, respectively.

Regarding crowding perception, studies have focused on finding the valuation of time for standing passengers in crowding conditions based on the density of standees (Hörcher et al., 2017; Tirachini et al., 2016; Whelan and Crockett, 2009). Whelan and Crockett (2009) found crowding value of time multipliers based on passenger density and a load factor of transit systems. Hörsher et al. (2017) estimated the user cost of crowding in terms of equivalent travel loss with smart card data from Hong Kong: an additional passenger per square meter on average adds 11.9% to the travel time multiplier.

Nowadays, with smart card and AVL data available, it is possible to infer crowding patterns of transit lines as a direct output of the destination inference from trip chaining. It is feasible to estimate vehicle load profiles and distribution of crowding over time and space of public transport services for multiple periods, such as in Luo et al. (2018) and Sánchez-Martínez et al. (2018).

2.4. Travel time reliability

Travel time reliability indicates the consistency or dependability in travel times, as measured from day-to-day or across different times of the day (FHWA, 2010). It measures the extent of the unexpected delays travelers experience on their trips. High variation of travel times over multiple days for the same trip leads to unreliable travel times. Because of travel time unreliability, many drivers either adjust their schedules or budget extra time to allow for traffic delays (FHWA, 2010). Reliability metrics in the literature are measured in various ways, including percent of variation, buffer index, planning time index, misery index, among others (Pu, 2011). For public transport, some studies have focused on evaluating reliability (Chen et al., 2009; Kieu et al., 2015; Ma et al., 2015).

Variability of travel times for public transport passengers are experienced during multiple trip stages: access times, waiting times, transfer times, and in-vehicle travel times. To account for those delays, passengers must also budget additional time to ensure on-time arrival most of the time. The buffer time concept is easily calculated and may relate well to the way travelers make decisions, as it represents the number of minutes of extra travel time that is needed to arrive on time (Lomax et al., 2003). To be late for work no more than one day per month (considering about 20 workdays) the buffer time would be the difference between the average travel time and the 95th percentile travel time (FHWA, 2010). Some studies also use buffer time as the difference between the median travel time and the 90th (Lee et al., 2016) or the 95th percentile travel time (Kwon et al., 2011).

With rich, continuous data, reliability metrics can be more easily calculated. In public transport, the use of GPS technology in buses' AVL systems now enables planners to be better informed of commercial bus speeds throughout the network (Cortés et al., 2011). However, for public transportation origin-destination (OD) pairs, the travel time variability is harder to obtain as both transfer times and users' real routing decisions should also be taken into account.

2.5. Summary of the literature review

The review of the related literature above evidences that some gaps, which have led to distinct research streams have been conducted somewhat separately, without a broader interaction of concepts for enhancement of the transportation planning process as a whole. Firstly, although notorious research has been conducted in crowding valuation for public transport systems, the spatial impact of the increased travel time valuation due to crowding on accessibility metrics is yet to be known. Accessibility is typically measured based on travel duration and distance, and few studies have used generalized travel cost for accessibility calculations (El-Geneidy et al., 2016). Cui and Levinson (2018) extend accessibility analysis to incorporate the full cost of travel that comprises the following components: travel time, safety, emissions, and monetary costs. Nonetheless, to the best of our knowledge, no studies have analyzed the negative effects of the spatial distribution of crowding and travel time variability on accessibility indicators of public transportation systems. As Hernandez (2018) proposes, discussions on the level of service of public transport should be included. The negative effects of the spatial distribution of crowding and travel time variability on accessibility indicators of public transportation systems is a literature gap this paper aims to fill.

Secondly, while the variability of travel times does cause users to budget extra travel time for their trips, the spatial distribution of this effect, and how it impacts accessibility to opportunities in cities is an important aspect to be explored. As users protect themselves from the effect of transport network unreliability, they must raise their total travel time budget, with the outlying results hidden from typical public transportation surveys.

Addressing these two gaps may also contribute to accessibility research in a way to better evaluate the influence of capacity constraints of transit vehicles on accessibility indicators. If only travel times are used to build accessibility maps, passenger capacity is not considered, thus affecting policy decisions. Comparing the accessibility of two different proposed transit systems that operate with the same speed but with distinct vehicle capacities will not show any difference for place-based access using only travel time. However, when the capacity constraint is added, the one with higher capacity yields a lower passenger density, which, in turn, translates into a better quality of service and lower generalized cost. Accessibility metrics, therefore, should include crowding and travel time variability effects to enhance their importance as an indicator for properly designing public transport networks.

3. Methodology

The proposed approach aims to evaluate the spatial distribution of the impact of both crowding discomfort and travel times variabilities on perceived accessibility to jobs and discuss transportation planning policy implications. We use the term "*perceived accessibility*" to represent the potential influence of the higher travel time valuation by users deriving from crowding and travel time variability on their decisions to justify difficult access to certain locations.

Spatiotemporal variability of travel demand was analyzed by processing smart card and AVL data from buses to build a dataset on buffer times and fine-grained crowding occurrence in São Paulo's public transport network. Firstly, data filtering was applied to select viable workdays. Then, general travel demand characteristics were calculated based on trip chaining methodology that was applied to multiple days. To calculate buffer times, inferred origin-destinations mean travel time distributions for a couple of months were used. For crowding estimation, the day that corresponded to the most stable travel demand was used as a typical workday. From there, networks in the form of GTFS data were produced, which incorporated discomfort travel time disutility by using crowding travel time multipliers with added buffer times to allow mapping the inequalities of accessibility.

We first describe the datasets used in this study and initial data

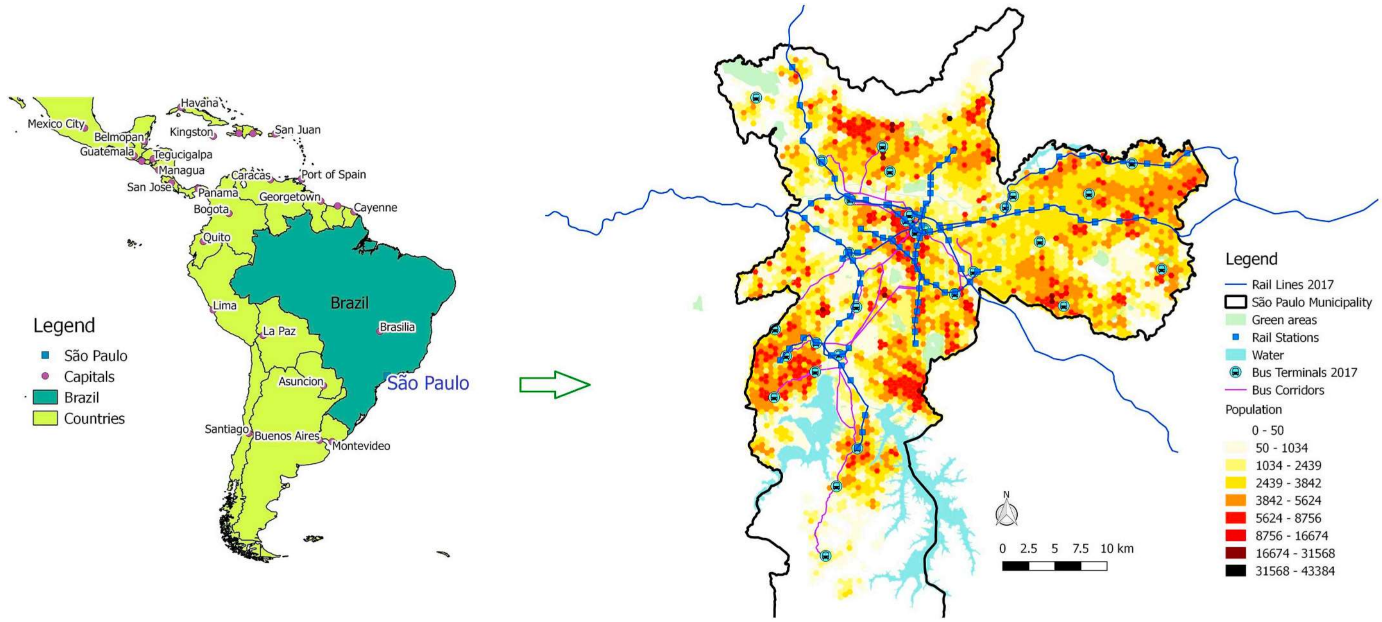

Fig. 1. Left: São Paulo location in a regional context.

filtering. We then describe the proposed procedures for trip chaining and transfer inference, crowding, and accessibility estimation.

3.1. Datasets

The city of São Paulo is the largest metropolis in Brazil and is one of the most populated areas in the world, with about 12 million inhabitants unevenly distributed, as shown in Fig. 1. The public transport bus network in the city of São Paulo is quite extensive, being one of the largest in the world, comprising 1335 bus lines (or routes) that are operated by a fleet of approximately 14,300 vehicles and about 20,000 bus stops (SPTrans, 2018). Metro and rail systems consist of 13 lines with 174 stations. About 13 million smart card transactions are recorded daily from *Bilhete Único*, the city public transport smart card, which is roughly divided into approximately 10 million transactions on buses and other 3 million on the metro and urban rail systems. The smart card system in São Paulo has a high penetration rate: about 96% of all transactions in a typical working day for the bus system; only 4% is still paid cash onboard (Arbex and da Cunha, 2018).

AVL monitoring data from the entire city bus fleet was used for this research, which is also managed by SPTrans (an acronym for São Paulo Transportes, the public company responsible for managing the bus system in São Paulo). Each bus generates a geocoded location record about every 45 s on average, producing a total of 27 million AVL records in a typical workday. There are about 180,000 vehicle trip departures per day throughout the city.

A summary of the two datasets, their variables, and data volume is presented in Table 1. Public transport network data, AVL records, bus trips, and the smart card dataset, without any personal information, were provided by SPTrans. Speed profiles were provided by Scipopulis (Scipopulis, 2018) after processing of raw AVL records. Existing jobs and population information were obtained from open governmental databases (IBGE, 2018; MTE, 2018), spatially represented in a hexagonal city grid of 4898 units after processing. Apart from jobs and population data, which were fixed for the period of analysis, all other datasets comprised a one-year interval, from July 1, 2017, to June 30, 2018.

Right: São Paulo municipality's main transit lines and population distribution.

In buses, the smart card system in São Paulo records a transaction when the user taps the card at the ticketing device, a small equipment located near the turnstile. The turnstile is located inside buses near the front door of each vehicle. Between the front door and the turnstile, there is a small area with a few seats, mainly dedicated to the elderly and those with priority needs. The existence of this area provides a faster boarding process as it enables the bus to depart while users are still waiting for their turn to pass their smart cards and go through the turnstile.

Two important issues to highlight are that, firstly, the so-called *Bilhete Único* smart card dataset for bus rides in São Paulo records neither the boarding nor the alighting location, since users are not required to tap-out; therefore, the proposed methodology comprises the estimation of boarding location, as well as alighting time and location estimations using trip chaining method (Trépanier et al., 2007).

Table 1
Datasets used in this study.

Category	Attributes	Data volume
Public Transport Network Data	Stops, Stations, Locations, Lines, Stop sequences, Fleet Sizes and Standing Area Data (both buses and rail)	20,306 bus stops and stations ~15,000 buses
Smart Card Transactions	Card ID, Route ID, Direction, Time, Vehicle/Station ID	~ 3,989,000,000/year ~260,000,000 processed
AVL Location records (GPS)	Vehicle ID, Line, Time, Latitude, Longitude	~ 8,742,000,000/year ~ 541,000,000 processed
Bus Trips	Vehicle ID, Line, Trip Start, Duration	~ 3,600,000 processed records (20 days)
Speed Profiles	Stop ID pair, 24 h Speed Profiles (processed from GPS)	510,140 records (20 days)
Jobs	Hex grid job distribution from government data (RAIS)	4,949,966 jobs in 4898 cells
Population	Hex grid population distribution from government data (IBGE)	11,809,666 people in 4898 cells

Secondly, the transaction validation point location may be different between bus types, as it is not positioned immediately on vehicle entrance. Therefore, the transaction record location may not reflect the real boarding location of the user. As a result, during the data cleaning phase, part of the transactions needed to be discarded, as they might wrongly indicate that the user did not board that actual stop. In the next subsection, we describe the initial data filtering and the selection of eligible workdays.

3.2. Data filtering

Travel time for trips between all OD pairs for several months was used to estimate travel time variability. To identify valid workdays, initially, the three months with the highest demand in terms of total transactions were selected. We discarded national holidays and weekends. Outliers were identified and removed for each month using the interquartile range method applied to the distribution of total daily transactions for workdays of each month. The interquartile range has a variable upper and lower limit for the dataset as shown in Eq. (1). Days for which the total transaction count was lower than L_{down} and higher than L_{up} were considered as outliers for that month.

$$L_{up} = Q_3 + 2 \times |Q_3 - Q_1|$$

$$L_{down} = Q_1 - 2 \times |Q_3 - Q_1|$$

(1)
where:
L_{up} is the upper limit for data filtering;
L_{down} is the lower limit for data filtering;
Q_i is the i-th quartile of the total smart card transactions for that month for all workdays.

Fig. 2 depicts the monthly variation of smart card transactions, revealing the months with the lowest daily demands as January and July, due school vacation, and the months with highest: September, October, and November 2017. Eleven days were removed from these three months, due to an instability of bus AVL data collection in some regions of the city, thus resulting in 42 workdays. Other 22 days had to be removed due to the unavailability of smart card transactions for very few rail stations in which the integration with the bus network is possible, such that the city-wide results would not be undermined. Finally, the trip chaining methodology was applied to the resulting 20 valid workdays as described in the following subsection.

3.3. Trip chaining and transfer inference

We used trip chaining to infer destination locations as riders are not required to tap the smart card when alighting in São Paulo. A transfer inference process distinguishes transfers from short activities and is needed to group transactions (trip legs or stages) to build trips, enabling calculation of demand characteristics and OD matrices.

The overall approach for this stage is depicted in Fig. 3. Firstly, it was necessary to estimate the boarding location for each transaction, which was accomplished using AVL data from buses or the location of rail and metro stations and BRT transfer terminals. The location of the corresponding alighting stop was then inferred using the next boarding location; in other words, it was assumed that the passenger ends a transit trip leg in the vicinity of the location identified as the boarding location of the next trip leg. To find the destination of the trip leg that corresponds to the last transaction of a day, the origin of the first transaction of that day was considered, in a similar way to previous works in the literature that apply trip chaining (e.g., Munizaga and Palma, 2012; Trépanier et al., 2007). This assumption reflects that most users departs from and returns to home when they use public transport during the day. If the number of stops traveled by the user was negative (i.e., when the direction of the bus line provided by AVL might be wrong, thus resulting in a spatial inconsistency regarding the inferred boarding and the alighting location), this procedure was reapplied considering the reverse direction of the line, as backward movement is inconsistent. Once trip chaining was concluded, the procedure splits into two parts: one to calculate loads, crowding and statistics at the line and stop levels and the other to distinguish activities to build full trips and the OD matrices.

This process was applied to all 20 selected valid workdays, using the Python language data analysis libraries: Pandas and Numpy (NumPy, 2018; Pandas, 2018). A single regular PC desktop computer equipped

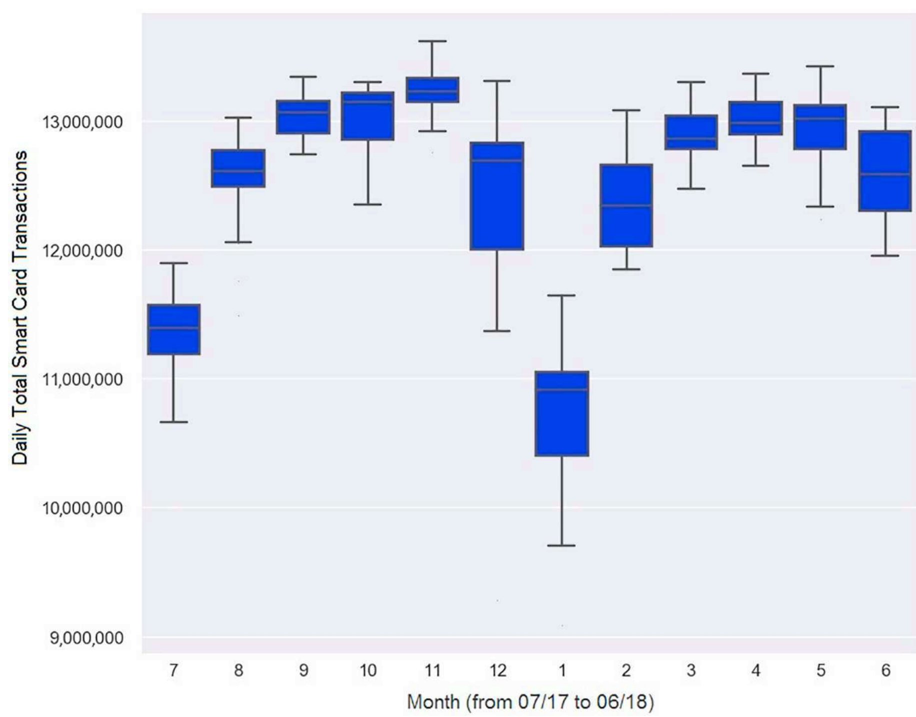

Fig. 2. Volume of smart card transactions distribution for every month in the study period.

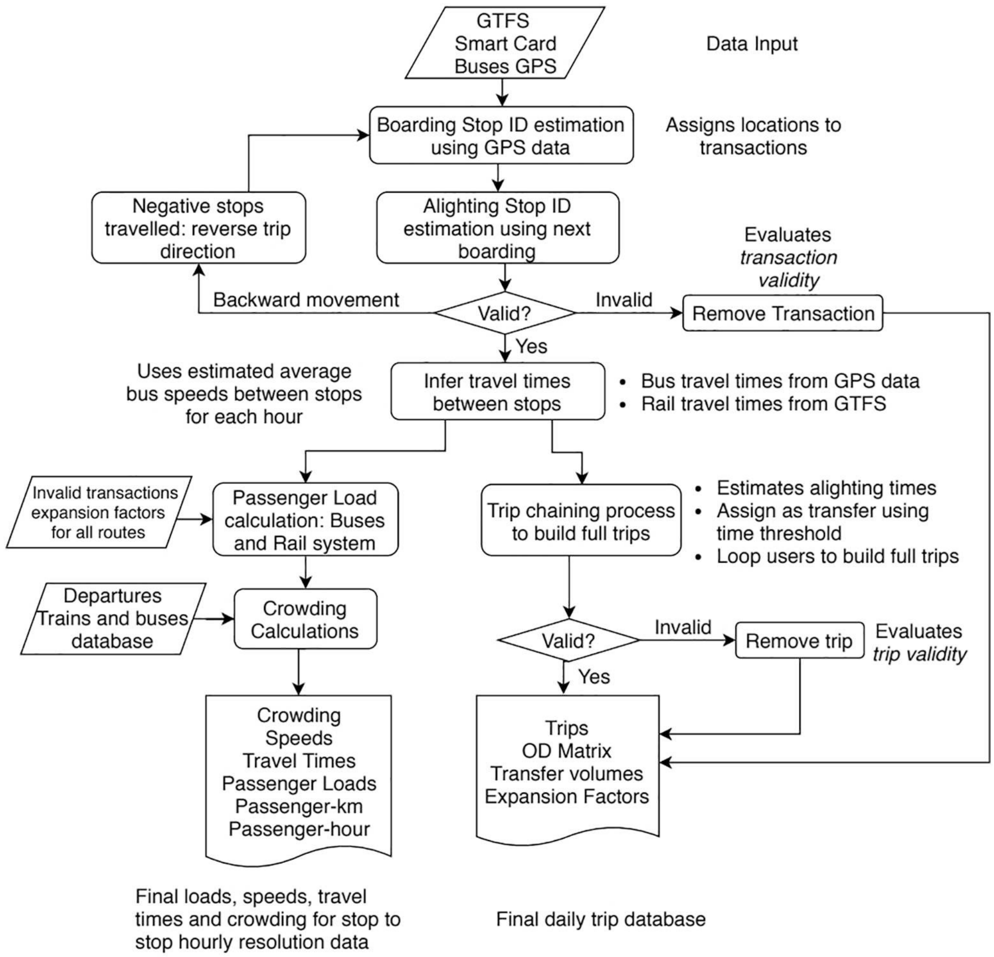

Fig. 3. Trip chaining and crowding calculations flowchart methodology.

with an Intel processor (i7 4770 CPU @ 3.4GHz with 32 GB RAM) was used to process the data for three days, which was deemed as a feasible timeframe for this analysis on a typical regular desktop computer..

3.3.1. *Boarding stop estimation*

The boarding location was estimated by comparing each transaction time with all AVL records of the corresponding vehicle ID for that day. The vehicle location record that yielded the minimum absolute time difference with respect to the boarding transaction time was assigned to that transaction. As both datasets are composed of millions of records, this imposes a challenging processing problem. Therefore, in our approach, we looped vehicle by vehicle and compare the time of all transactions that occurred in that vehicle with the time of every location record for that vehicle. After the location of each transaction had been assigned, the nearest bus stop ID of that route was then determined.

It should be noted, however, that in São Paulo, the exact boarding location may not be exactly where the boarding transaction occurred, as there is a small standing area between the bus door and the turnstile, where the ticketing machine is positioned inside the bus. Therefore, the location corresponding to the instant that the passenger passes through the bus turnstile is assigned as the boarding location. The previous assumption indeed might bring some errors to the boarding location estimation and is adequately treated later, during *transaction validity* check. On the other hand, turnstiles discourage fare evasion, which is very low in São Paulo and thus has not influenced the results.

For metro, urban rail, and BRT stations, the process of assigning the boarding location is straightforward, as, for every smart card transaction that takes place corresponds to a fixed location turnstile whose code is properly identified and registered.

3.3.2. *Alighting stop estimation*

For alighting location and alighting stop estimation, a trip chaining model was applied. The alighting stop ID was assumed as the nearest stop of the subsequent boarded vehicle. To avoid inconsistencies that might arise, mainly due to data quality, one of the methodological challenges in processing smart cards as highlighted by ***Wilson (2016), in this paper we used two types of validity checks to search for erroneous and inconsistent data: *transaction validity check* and *trip validity check*, which are, respectively, the first and second validity checks depicted in Fig. 3. The transaction validity check was applied after the destination inference stage and filtered for the problems listed below:

- Untraceable cards. For example, smart cards used by operators when users pay cash onboard (about 4%);
- There is only one transaction for that card ID throughout the day;
- No location data inferred for the transaction. This might occur because a vehicle had a malfunctioning AVL;
- Distance from the transaction location to the nearest bus stop of the corresponding line is greater than 500 m. This threshold was chosen as the acceptable error for boarding location estimation for this study;
- Time difference between transaction timestamp and AVL location record is greater than 2 min. For example, for an average speed of 15 km/h, a reasonable value for buses in urban traffic as our data shows, this would represent a 500 m error to the estimated boarding location;
- Coded line in AVL equipment different from the coded line in the

vehicle fare equipment; distance from estimated alighting stop to the next boarding stop is greater than 2000 m. This is higher than most of the previous studies in the literature (e.g., 1000 m as in Munizaga and Palma, 2012) once the user might have made the validation transaction after the bus departed due to turnstile location in buses;
- The number of stops traveled per transaction equals zero. This condition occurs whenever users taps the card right before alighting, which happens for users that held back their validation transactions, mainly near terminals or on crowded vehicles in which they are incapable to pass the turnstile earlier.

The resulting filtered valid transactions account for about 75% of all transactions. Most removals (about 16% loss) are due to the turnstile location issues, adjusted with the last two transaction validity check filters. From this step on, a transaction expansion factor for each bus route and direction was determined. This transaction expansion factor was used in crowding estimation (lower left portion of Fig. 3) as input data to expand transactions with known start and end stop IDs to all transactions recorded throughout the day for each route and direction.

After the trip chaining process, all remaining valid transactions proceed to travel time calculation. In this process, each transaction is assigned an estimated travel time by using hourly average bus speeds between adjacent stops that are calculated from AVL data processing. The road network distance between consecutive positions along the trip path is considered for computing the average speed between AVL consecutive positions, which may be interpolated to determine the average speed between consecutive bus stops, making it easier to calculate travel times (Monteiro et al., 2015). For rail systems (metro and train lines) travel times from GTFS is used. For all possible station-to-station OD pairs, the shortest paths are calculated considering the generalized costs on the rail network (Arbex and Da Cunha, 2017), which accounts for waiting, in-vehicle, and transfer times. It should be noted that the size of the rail network is limited; consequently, for many OD pairs, there is only one possible feasible path.

After all travel times had been estimated, the main trip chaining process started (lower right portion of Fig. 3). It loops through every single card ID, merging multiple boardings into a single full trip whenever the time difference between each alighting and the next transaction is below a specified threshold. We used 30 min in this study, as the OD matrices with 30 min allowable transfer time are slightly more accurate compared to those with 60 and 90 min (Alsger et al., 2016). This process is illustrated graphically in Fig. 4, denoting a user who made five transactions and three full trips.

The *trip validity check* filters only for valid trips, as below:

- Total distance traveled must be higher than zero;
- The number of transfers should be between 0 and 6. This threshold was chosen as most trips can be made with up to 6 transfers, and trips with 7 or more might actually have been 2 trips with a short activity in between;
- First boarding Stop ID should be different from the last alighting Stop ID;
- Total trip travel time should be between 1 min and 6 h (time between first boarding and last alighting);
- Average trip speed should be between 0.01 and 80 km/h;
- The minimum haversine distance between first boarding and last alighting stop should be 300 m;
- The trip circuitry factor is below 3. Trip circuitry factor is the network distance traveled by the user distance divided by the haversine distance between the first boarding and last alighting stop. The maximum threshold of 3 was chosen as higher values might imply that the user made short activities instead of a transfer.

At last, a full-trip level expansion factor was built to make up for the invalid trips. It is based on smart card users as the penetration rate is very high, about 96% in São Paulo (Arbex and da Cunha, 2018). This factor was determined by simply dividing the number of unique card IDs (including untraceable cards) observed for that day by the number of unique card IDs that only generated valid trips. Though this assumption could be revised and refined using a large metropolitan travel survey, this trip expansion factor procedure showed satisfactory for the aims of our study.

3.4. Crowding estimation

After all previous steps had been performed, the left portion of the approach depicted in Fig. 3 was conducted with the final aim to reach a highly detailed resolution of crowding data, that is, the load factor and passenger density (passenger per square meter in standing areas), hourly, for each bus and rail line, between every pair of subsequent stops. This data enabled the calculation of public transportation level of services for all segments, including the bus network as well as metro and train rail services. The load factor is defined as the passenger load (i.e., the number of passengers going through a pair of stops within a specific time interval) divided by the number of seats supplied in that same period (Kittelson and Associates et al., 2013). This section details how this crowding was calculated for both buses and rail systems.

Regarding the bus system, the information required corresponds to the fleet database, with vehicle capacities and standing areas, and bus lines departures for each day, with their start time. For buses, both load factor and passenger/m^2 are calculated for all routes and directions, while for the rail system, only passenger density is calculated for all services. Table 2 illustrates the average capacities of vehicle types used in São Paulo transit system. Table 3 represents Level of Services (LOS) based on load factors (bus system) and passenger density (rail system) adopted in this study, which is used to map and communicate crowding spatial distribution. Load factor thresholds for buses consider local crowding distribution and rail LOS are adopted from Sarkar and Jain (2017). Passenger/m^2 is used to apply crowding discomfort valuation for rail, as this metric is comparable between all trains' internal layouts.

For the rail crowding estimation, the first step was to construct passenger load for all segments and time intervals throughout the day. Initially, the shortest path algorithm was applied (as the rail network is not large and does not allow many multiple paths), considering generalized costs within rail network to build all OD pairs shortest paths and eligible transfers within the network. Then, a destination station distribution matrix was built for each hour of the day and each origin station, considering station destinations inferred from the trip chaining. In a second step, the total number of boardings for each station were organized for each hour of the day, based on total smart card transactions and a station-based expansion factor that accounts for other forms of payments for each station (i.e., cash and the metropolitan smart card in São Paulo metro area). From there, passenger OD pairs were assigned to the shortest generalized cost path to achieve hourly passenger loads. The final results are passengers per square meters in standing areas for all rail segments between subsequent stations along the day after incorporating fleet composition data, standing area per rail vehicle type, and scheduled hourly frequency.

3.5. Accessibility

We adopt the cumulative opportunities accessibility metric, which has been widely used because of its easy communication, interpretability, and calculation (El-Geneidy and Levinson, 2006; Geurs and van Wee, 2004). It is calculated as in Eq. 2, by counting the number of jobs reachable A_i from location i to all other locations j, reducing the number of jobs available in j, O_j, by an impedance function $f(t_{ij})$ of time cost from i to j. Jobs are added if they can be reached within a specified travel time threshold as in Eq. 3 and 4, considering or not travel time variability, respectively.

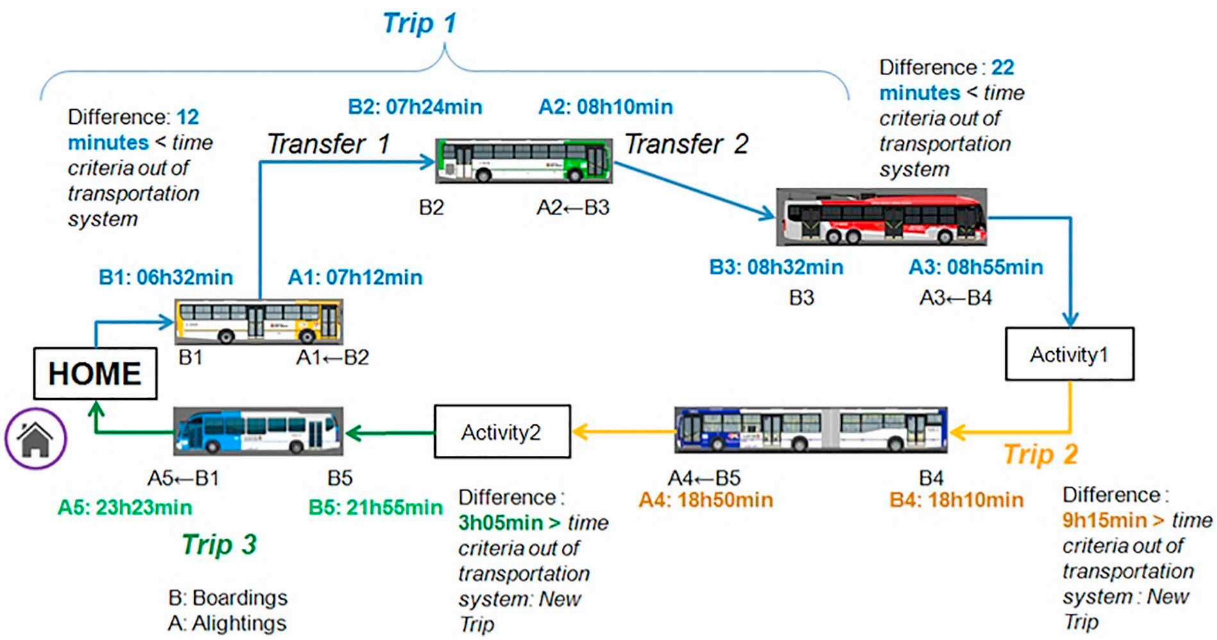

Fig. 4. A graphical example of user transactions and estimated full trips for a day.

Table 2
Average seat capacity and available standing area in main vehicle types.

Vehicle type name (average capacity of main vehicle types)	Seat capacity	Standing area available [m^2]
Microbus	21	2.9
Minibus	20	3.4
Midibus	25	4.9
Basic	35	6.5
Padron 13 m	32	8.9
Padron 15 m	38	10
Articulated bus 18.5 m	37	15.2
Articulated bus 23.0 m	57	18.8
Biarticulated buses	47	25
Metro trains (average)	271	213
Urban rail trains (average)	406	267

Table 3
Level of Service (LOS) used in this study (adapted from Sarkar and Jain (2017)).

LOS	Load Factor (Buses)	Passenger Density (Rail)
	Passengers/seat	Passenger/m^2
A	< 1.0 (Seated Passengers)	= 0
B	≥ 1.0–1.5	> 0–2.0
C	≥ 1.5–2.0	≥ 2.0–4.0
D	≥ 2.0–2.5	≥ 4.0–6.0
E	≥ 2.5–3.0	≥ 6.0–8.0
F	≥ 3.0	≥ 8.0

$$A_i = \sum_{j=1}^{n} O_j \bullet f(t_{ij}) \tag{2}$$

$$f_1(t_{ij}) = \begin{cases} 1, \; t_{ij} \leq T_i^{95th} \\ 0, \; t_{ij} > T_i^{95th} \end{cases} \tag{3}$$

$$f_2(t_{ij}) = \begin{cases} 1, \; (t_{ij} + bt_{ij}) \leq T_i^{95th} \\ 0, \; (t_{ij} + bt_{ij}) > T_i^{95th} \end{cases} \tag{4}$$

This study proposes a variable travel time threshold defined as T_i^{95th}, which is the 95th percentile of all observed full trips travel times departing from location *i* for a chosen time period, as this better reflects current travel behaviors of worker access to employment opportunities. The average travel time of trips observed from smart card data for zones located further away from city central areas is two to three times higher than those of central zones. Therefore, using a single value such as $T_i =$ 45 min for travel time threshold (usual for cumulative opportunity accessibility calculations) would hinder insights about crowding influence on accessibility for distant areas, as some zones have an average travel time of more than 100 min. Nonetheless, to support the adoption of the T_i^{95th} approach, we also perform a comparison with both the cumulative opportunity using 30, 60 and 90 min as well as gravity-based negative exponential impedance function. We used all observed full trips for the 95th travel time calculation, as there is no information regarding trip purpose on smart card data and using other methods to infer trip purpose would add more uncertainty to the data. Although the observed travel time does not include access or first waiting times due to the nature of the smart card dataset, it still represents inter-region propensity to accept higher travel times as jobs are located further away than users' residence.

To include travel time variability in accessibility calculations, the buffer time needed to guarantee on-time arrival when traveling from location *i* to *j* 95% of times, bt_{ij} is added to travel time t_{ij} in Eq. 4. In this research, as our travel time data is estimated from the dataset and not directly measured, we used a median-based buffer time, defined as the difference between the T_{ij}^{95th} and the T_{ij}^{50th} travel time between locations *i* and *j* considering multiple days of data.

Eq. 5 denotes total travel time: the sum of access time from walking t_{ij}^{walk}, waiting time t_{ij}^{wait}, in-vehicle travel times $t_{ij}^{in-vehicle}$ and transfer times $t_{ij}^{transfer}$ if the trip requires any transfers. To assess the difference in accessibility due to crowding, a α coefficient is applied to the in-vehicle travel time component in Eq. 4. The α coefficient is defined in Eq. 6, in which passenger density between two pairs of subsequent stations or stops $density_{segment}$ is considered to raise in-vehicle travel time perception due to discomfort. This coefficient was calculated for each stop-to-stop segment, hourly, for all bus/metro/rail routes. Eq. 6 is defined based on values proposed in Whelan and Crockett (2009), due to the absence of local studies. While we acknowledge that the effect of crowding on perceived travel time might also vary depending on if a passenger has a seat or not, for the sake of simplicity, crowding discomfort is applied to all users traveling an overcrowded route/

direction/h, regardless whether they were standing. Although trip generalized costs would require weighting other travel components such as walking and waiting times, we decided to weight only in-vehicle travel time and add buffer times to focus on those effects as evaluating the spatial impact of crowding conditions and travel time variability on perceived accessibility to jobs is the main objective of this paper.

$$t_{ij} = t_{ij}^{walk} + t_{ij}^{wait} + \alpha \bullet t_{ij}^{in-vehicle} + t_{ij}^{transfer} \tag{5}$$

$$\alpha = \begin{cases} 0.085 \bullet density_{segment} + 1.5321, \textit{ with crowding discomfort} \\ 1, \textit{ without crowding discomfort} \end{cases} \tag{6}$$

4. Results and discussion

4.1. Variability of travel demand characteristics

In order to evaluate the spatial influence of crowding discomfort and travel time variability on accessibility to jobs, crowding patterns have been calculated using a representative workday, along with buffer time estimated from multiple workdays of public transportation big data. To evaluate the stability of travel demand and support using a representative workday for crowding patterns, this section reports results based on a 20-workday analysis. Firstly, Table 4 reports the results of the evaluation of the variability of demand travel patterns using general trip statistics after the application of the methodology for the 20 qualified workdays, considering a 24-h period. Most importantly, it can be seen that the coefficients of variation of all reported attributes are low, which means that public transport demand is stable throughout the 20 regular workdays from a city-wide point-of-view. These results support the use of crowding patterns evaluated for a representative day to measure its spatial impact on accessibility.

Table 5 shows estimated buffer time distribution (needed to guarantee on-time arrival when traveling from location i to j for 95% of the times, given by bt_{ij}) considering the partition of the city of São Paulo into 32 administrative districts, also referred to as Subprefectures. The median is inferior to 3 min for all OD pairs with at least 30 trips for all 20 days, indicating that the analyzed OD pairs tend to have relatively stable travel times. Nonetheless, some OD pairs revealed significantly higher buffer times, with the maximum reaching 13.6 min.

Fig. 5 illustrates how variability in travel time between OD pairs allows the estimation of buffer times using the proposed methodology. In this figure, the blue dashed lines represent the minimum and maximum mean travel times for all days comparing two chosen OD pairs. Buffer times are calculated as T_{ij}^{95th}-T_{ij}^{50th} considering the median values for all days. OD pair 20–14 (Parelheiros to Santo Amaro) yields a high buffer time of 8.7 min, while for OD pair 04–12 (Casa Verde to Vila Mariana) corresponds to only 2.3 min. As this method relies on

Table 4
Travel demand characteristics and variability.

		Interday Stats [20 work days]		
Daily Trip Stats	Time Period	Mean	SD	CV
Volume of Trips	24 h	8,208,096	159,519	1.9%
Trip distance (m) [median]	24 h	9462	138	1.5%
Number of Transfers [mean]	24 h	0.88	0.01	1.5%
Travel Time (min) [median]	24 h	37.3	0.6	1.5%
Trip speed (km/h) [median]	24 h	15.7	0.2	1.4%
Load Factor [weighted average]	24 h	1.42	0.03	2.4%
% of Passenger-hour on LOS A or B	24 h	59.7%	1.5%	2.6%

Table 5
Subprefecture OD pairs and buffer times.

Area	Zoning Type	Subprefectures
	Time Period	7 h–10 h
	City Partitions	32
	Possible OD Pairs	1024
OD Pairs	With Trips All 20 work days	897
	At Least 100 Trips (Every 20 days)	523
Estimated Buffer Time for OD Pairs (minutes) (30+ Trips/20 days)	10th Percentile	0.7
	50th Percentile	2.3
	90th Percentile	4.8
	Maximum	13.6

observed trips, there is indeed a certain trade-off between zone size and trip sample. We adopted the size for a sample equal to 30 trips, thus resulting in 563 OD pairs for which we could estimate buffer times. OD pairs without sufficient observable trips did not include a buffer time addition.

By combining the results obtained so far, we could use crowding data based on the most representative day without loss of generality. The most regular day was chosen in a similar way to the methodology proposed by Liu et al. (2019) that considers the mean error of passenger travel demand of each OD pair. For this study, the day with the least mean error for OD pairs trip volumes among the 20 workdays, considering a travel matrix from a regional partition resulting in 81 OD pairs, is September 4, 2017, which was used to obtain the results that follow.

4.2. Crowding spatiotemporal distribution

Using data from the day with the least mean error for OD pairs trip volumes (i.e., September 4, 2017), the spatiotemporal distribution of crowding levels could be estimated. Considering the temporal dimension, Fig. 6 shows passenger-hour by LOS for each hour of that day on the entire bus system. It can be observed that the passenger-hour by the level of service estimation is a good measure of crowding distribution, as it weights passenger loads on segments by their travel times for each specific hour. Fig. 6 also evidences an intense peak of passenger-hour during peak hours between 6:00 and 8:59 in the morning.

A GTFS network with travel time multiplied by crowding discomfort perception (considering conditions between 7:00 and 7:59 am), buffer time from 20 days morning peak trip (6:00–8:59) with a departure time of 7:00 was used in the following accessibility comparisons.

Bus passenger loads and LOS are depicted in Fig. 7 for the 7:00–7:59 am period, in which thickness denotes passenger loads, and the colors represent LOS (as presented in Table 3). Fig. 7 shows that the more central areas have a higher level of service for the bus system, while the transportation corridors that connect the suburban areas to the more central areas suffer from worse crowding conditions.

Passenger loads and LOS for the rail system (metro and metropolitan train) are depicted in Fig. 8. As the volumes of passenger load are of a different magnitude, it was necessary to present them in a different figure. The maximum load is the central black LOS F segment with about 70,000 passengers/h. Bus loads are made transparent for system-wide comparisons. During this period, most rail segments operate near capacity, indicating high usage of the system.

4.3. Accessibility to jobs

Cumulative opportunity accessibility metric is a well-known way to measure how a public transport network serves its purpose: promoting urban dwellers access to multiple types of opportunities. In this subsection, we use cumulative accessibility to jobs to evaluate how commuters' access to employment is spatially affected by crowding

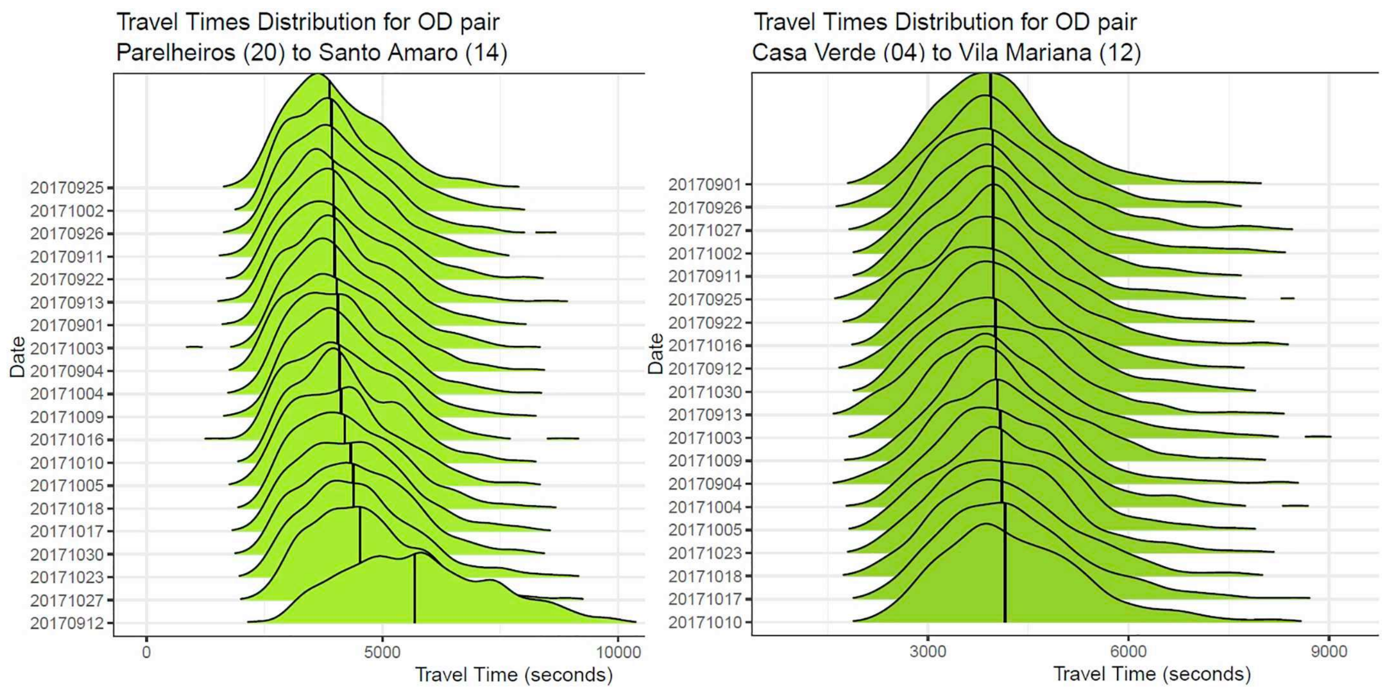

Fig. 5. Travel Time Distribution for two OD pairs with a high (left) and low (right) variability between days.

discomfort and variability of travel time. To apply Eq. 1 (presented in subsection 3.5) to measure accessibility, a matrix of all-pairs travel time must be calculated.

We use four scenarios for comparisons, representing travel time perception of users, based on an exact 7 a.m. departure. It should be noted that this is a simplification that was adopted to reduce processing times. Averaging accessibility over the departure period is ideal (Owen and Levinson, 2015); however, it would require days of processing with OTP due to the size of the resulting travel time matrix with the large spatial resolution required for this study. The scenarios are as follows:

(a) Network with bus speeds from 7 a.m.;
(b) Network with bus speeds from 7 a.m. + travel time variability;
(c) Network with bus speeds from 7 a.m. + crowding discomfort;
(d) Network with bus speeds from 7 a.m. + travel time variability + crowding discomfort.

To increase the spatial resolution of results, 4898 centroids of a 300-

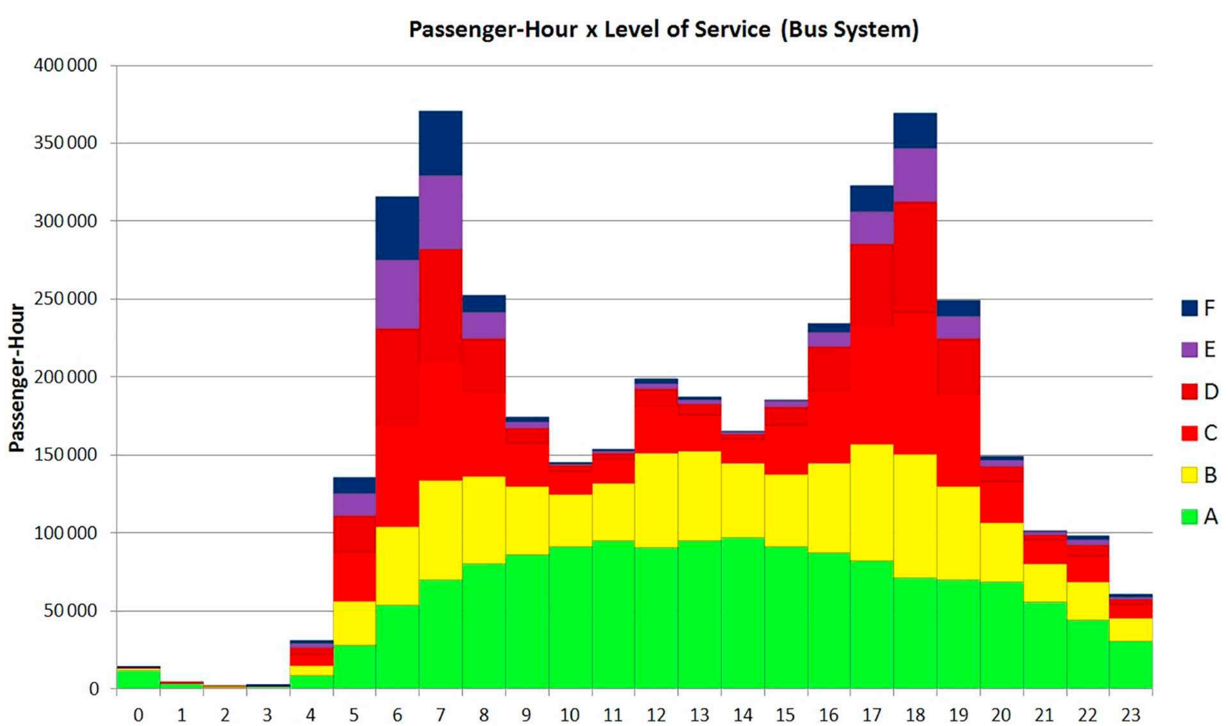

Fig. 6. Hourly Passenger-hour distribution by Level of Service.

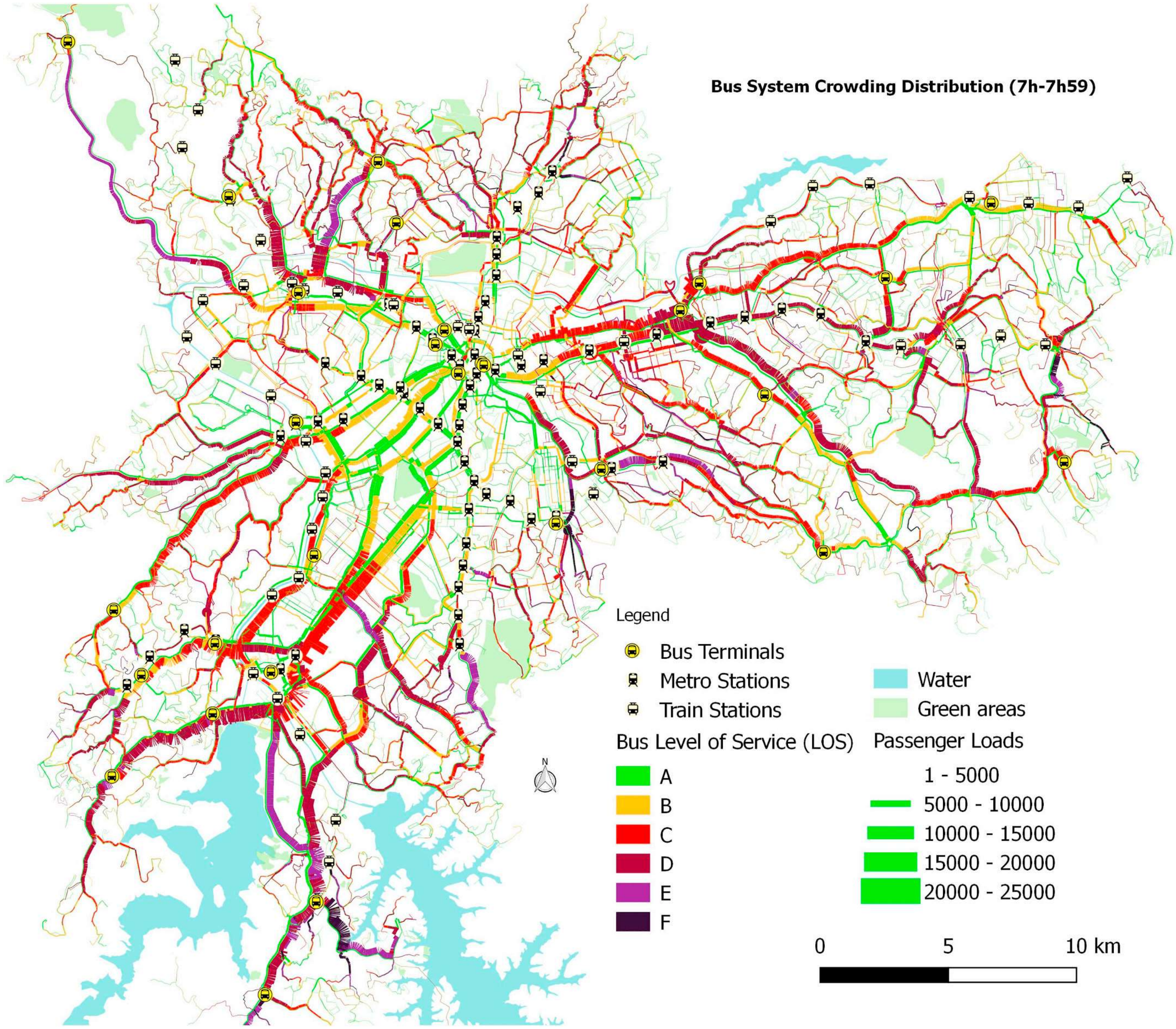

Fig. 7. Bus System Level of Service for 7 h-7 h59 period.

m size hex grid that cover the whole urbanized area of the city of São Paulo were used as departing locations. Hexagon division was chosen because of the advantages of being more similar in shape to circles compared to squares, which potentially lessens bias due to edge effects (Birch et al., 2007).

Travel time estimation for each OD pair for accessibility calculations is performed using software OpenTripPlanner OTP (OpenTripPlanner), 2019), a multi-modal trip planner from a group of open-source software projects that aims to provide transportation network analysis services (Conveyal, 2018). A similar use of OTP for accessibility calculations had also been made by Widener et al. (2017) and El-Geneidy et al. (2016). OTP requires as inputs a GTFS format transit network file, an OpenStreetMap (OSM) database file, and some simple routing parameters. The transit network files we used are the four modified General Transit Feed System (GTFS) (Google, 2019) files, which comprise scenarios (a) to (d). The output files of OTP analysis are travel time matrices with total time (including access, waiting, in-vehicle travel times and transfer times) between all OD pairs.

The only file in our study that needed to be modified is *stop_times.txt*, as it comprises route and stop level travel times. We modified travel times for all pairs of stops for each route and direction to represent each scenario. The schedule is included as part of the GTFS information with hourly headways for all services.

We use the number of jobs as a proxy for opportunities due to spatial data availability. The location of jobs is estimated by distributing city jobs to hex grids based on their street location, as available at the free-access governmental database RAIS obtained from Brazil's Ministry of Work (MTE, 2018). In the RAIS database, for each company, the street postal code is available. We distributed job count from each company to a buffer of 50 m around the street where it is located. Then, jobs were aggregated to the hex grid layer based on area share. There is a small location error associated with this procedure, as the government does not share the exact known location of each company within the street in this public dataset. However, the precise location is not required as the spatial resolution of our analysis is based on a 300-m hex grid.

To aggregate results into the larger city regions, the population is considered to weight accessibility. The population count for each hex grid is estimated using data from the latest national census conducted in 2010 (IBGE, 2018) after distributing the population by proportional area share of census tract shapes to the hex grid layer. As census tracts

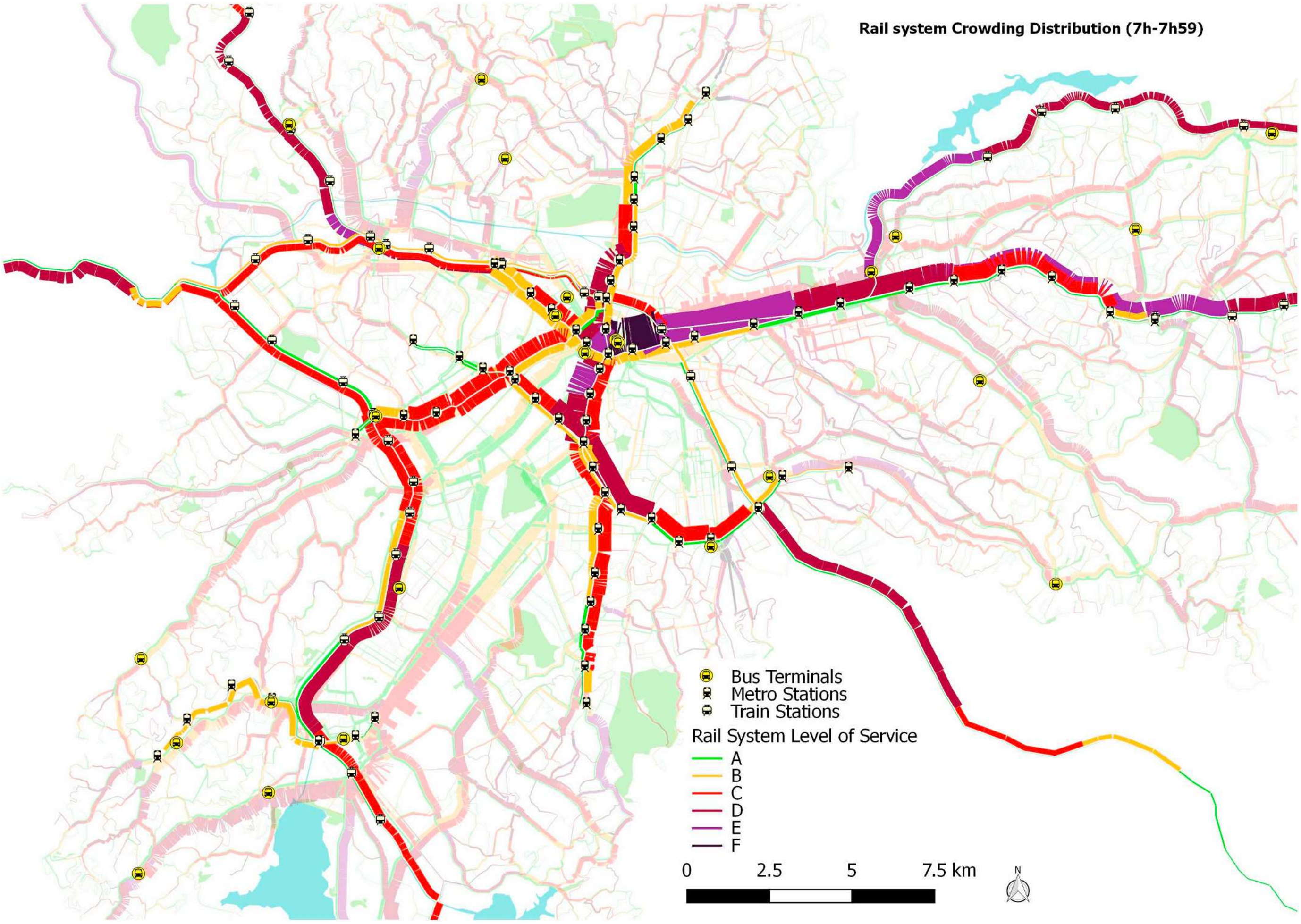

Fig. 8. Rail System Loads and Levels of Service for the 7 h-7 h59 period.

are small in area (about the size of a hex grid) so that the spatial distribution of the population is kept representative.

To allow research replicability, we have made the code used in this section freely available on Github repository platform (http://bit.ly/38Bb9ip) along with some sample data that is required to generate a population-weighted average accessibility by regions. The code includes calculation procedures for all accessibility metrics used in this paper.

Table 6 shows the results of population-weighted accessibility averages by city regions for all four scenarios (a-d). With the addition of perceived travel times due to crowding discomfort only, the reduction of accessibility considering travel time as perceived by users is acute. The results show that the crowding discomfort valuation influences access to jobs, particularly for zones located further away from the central area. Travel time variability has a hidden impact on some specific areas of the city. On average, the commuters' weighted average reductions in perceived accessibility to jobs in a workday morning peak are 56.8% due to crowding discomfort, 6.2% due to travel time variability and 59.2% when both are combined. When taken together, crowding and travel time variability have a stronger impact of reducing perceived accessibility to job opportunities in a congested public transportation network due to heightened travel time perception and how users perceive the quality of their travels. This outcome clearly evidences the impending need for including both crowding and travel time reliability aspects on accessibility calculations to properly reflect user perception.

In order to verify the validity of the new approach that uses a variable travel time threshold (defined as the 95th percentile of all observed full trips travel times T_i^{95th}), we show in Table 7 a comparison with accessibility metrics that are more commonly used in practice.

Table 6
Accessibility to Jobs aggregated by Subprefectures.

Area Data (City Regions)				Jobs Accessible (Scenarios of Perceived Travel Time)				Comparisons		
ID	Name	Avg. 95th Percentile Travel Time (Morning Peak) [min] [6 h-8 h59]	Morning Peak Public Transport Trips [6 h-8 h59]	(a)	(b)	(c)	(d)	(c) vs (a)	(b) vs (a)	(d) vs (a)
1	CENTRAL	68	90,661	3,878,228	3,792,160	3,302,188	3,200,416	−15%	−2%	**−17%**
2	NORTH	94	324,950	2,509,171	2,357,440	840,719	779,079	−66%	−6%	**−69%**
3	WEST	79	239,812	2,736,942	2,597,693	1,531,143	1,447,089	−44%	−5%	**−47%**
4	EAST	97	522,337	2,104,872	1,947,195	660,113	626,006	−69%	−7%	**−70%**
5	SOUTH	87	440,233	2,377,274	2,222,429	1,091,344	1,016,678	−54%	−7%	**−57%**

Table 7
Accessibility to Jobs aggregated by city regions computed for scenarios (a-d), considering the proposed 95th percentile-based and commonly used accessibility metrics.

Scenario	Region	95th Percentile	Gravity-based	Cumul. 30 min	Cumul. 60 min	Cumul. 90 min
(a) Network with bus speeds from 7 a.m.	CENTRAL	3,878,228	492,315	992,506	3,356,586	4,650,091
	NORTH	2,509,171	178,401	80,318	763,605	2,449,489
	WEST	2,736,942	256,145	266,145	1,443,315	3,100,288
	EAST	2,104,872	148,397	83,659	574,893	1,811,380
	SOUTH	2,377,274	206,306	161,995	1,072,427	2,584,308
(b) Network with bus speeds from 7 a.m. and travel time variability	CENTRAL	3,792,160	478,048	933,139	3,356,586	4,650,091
	NORTH	2,357,440	169,193	70,509	763,605	2,449,489
	WEST	2,597,693	245,051	247,162	1,443,315	3,100,288
	EAST	1,947,195	141,034	75,929	574,893	1,811,380
	SOUTH	2,222,429	195,985	144,946	1,072,427	2,584,308
(c) Network with bus speeds from 7 a.m. and crowding discomfort	CENTRAL	3,302,188	431,586	826,255	2,727,310	4,392,476
	NORTH	840,719	88,398	44,868	255,072	939,713
	WEST	1,531,143	175,702	187,879	877,177	1,950,818
	EAST	660,113	72,494	55,611	270,683	701,599
	SOUTH	1,091,344	125,686	101,729	541,598	1,415,690
(d) Network with bus speeds from 7 a.m. + travel time variability + crowding discomfort	CENTRAL	3,200,416	419,475	771,833	2,634,136	4,320,389
	NORTH	779,079	84,062	41,064	234,452	879,119
	WEST	1,447,089	168,815	176,572	832,289	1,873,206
	EAST	626,006	69,243	51,339	254,871	667,907
	SOUTH	1,016,678	119,954	93,511	508,831	1,347,223
(c) vs (a)	CENTRAL	−15%	−12%	−17%	−19%	−6%
	NORTH	−66%	−50%	−44%	−67%	−62%
	WEST	−44%	−31%	−29%	−39%	−37%
	EAST	−69%	−51%	−34%	−53%	−61%
	SOUTH	−54%	−39%	−37%	−49%	−45%
(b) vs (a)	CENTRAL	−2%	−3%	−6%	−3%	−1%
	NORTH	−6%	−5%	−12%	−10%	−6%
	WEST	−5%	−4%	−7%	−7%	−4%
	EAST	−7%	−5%	−9%	−8%	−6%
	SOUTH	−7%	−5%	−11%	−8%	−5%
(d) vs (a)	CENTRAL	−17%	−15%	−22%	−22%	−7%
	NORTH	−69%	−53%	−49%	−69%	−64%
	WEST	−47%	−34%	−34%	−42%	−40%
	EAST	−70%	−53%	−39%	−56%	−63%
	SOUTH	−57%	−42%	−42%	−53%	−48%

More specifically, we also calculate population-weighted accessibility averages by city regions for all scenarios (a-d) and comparisons between them considering a gravity-based impedance function as well as using fixed travel time thresholds of 30, 60 and 90 min for cumulative opportunity accessibility.

These results in Table 7 evidence that single cut-off time thresholds are not suited for the proposed application, as low cut-offs such as 30 min do not properly represent the average travel time required for the peripheral zones to reach most city jobs. For example, accessible jobs using 30 min cut-off for the east region correspond to less than 5% of the 90 min cut-off. However, larger values such as 90 min or more improperly inflate accessibility levels for central zones. From our analysis of smart card data, only 5% of travelers have travel times superior to 68 min from the central region (Table 6). Therefore, due to the nature of very distinct travel times distributions for different city regions, there is not a single appropriate cut-off time.

On the other hand, gravity-based accessibility metrics build upon spatial interaction modeling theory and are deemed as an alternative to avoid an arbitrary selection of cut-off times for accessibility calculations. For the gravity-based metrics, we calibrated a negative exponential impedance function using travel time following the methodology proposed by Iacono et al. (2010) that uses the percentage of trips by travel time interval. We used data from a large recent regional household survey published in 2018 from Metrô SP (2018) to filter transit and walk trips, and the resulting impedance function (as in Eq. 2) is $f(t_{ij}) = \alpha\, e^{-\beta t_{ij}} = 0.355\, e^{-0.029 t_{ij}}$. As presented in Table 7, while gravity-based indeed provides better results than 30-min cut-off fixed threshold, using the assumption of a fixed α and β parameters for the impedance function for all zones does not properly represent the acute differences in trip duration frequency distribution throughout city regions. While only 2% of trips from central zones take more than 90 min, about 17% of those from the eastern region take more than that value. Therefore, we proposed the 95th percentile of travel times, calculated for subprefecture level (32 city divisions) using smart card data, for our accessibility calculations.

Fig. 9 compares reachability for scenarios (a), (c) and (d) considering two points for a trip departing at 7 h. On the left, a point in Parelheiros subprefecture (id = 20) and, on the right, one in Pinheiros (id = 11). The average combined reduction due to crowding and travel time variability on perceived accessibility to jobs is 87% for Parelheiros subprefecture and only 20% for Pinheiros. While for the former accessibility to key job locations in more central areas are affected, for the latter, the influence of crowding does not appear to hinder how users perceive accessibility to central jobs.

Most importantly, Fig. 10, which represents the spatial influence of both crowding discomfort and travel time variability on accessibility to jobs, reveals that both aspects highly impact the outskirts and peripheries. On the other hand, the accessibility from more centrally located areas are not much influenced by crowding and travel times variabilities, since, as seen in Fig. 7, users have access to unused capacity on bus lines to reach nearby jobs. This uneven impact is of most relevance for city planners and policymakers aiming to find support to showcase how improving frequencies, capacities and implementing more dedicated lanes for public transport systems for reliability will bring residents closer to city opportunities.

5. Conclusion and future work

We have proposed a methodology to measure and evaluate the spatial impact of travel time variability and crowding discomfort on

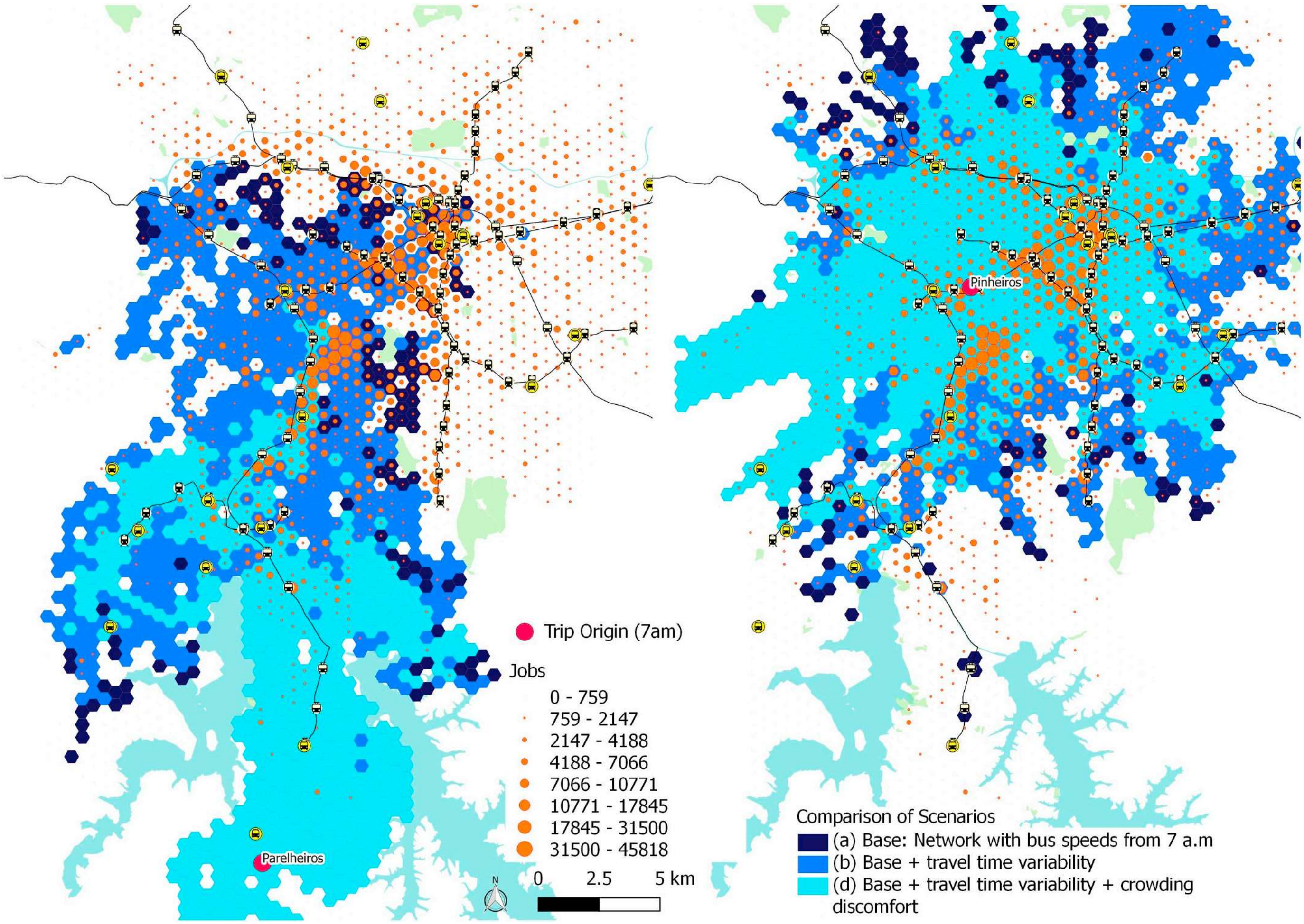

Fig. 9. Areas reached from selected locations.

accessibility to jobs considering travel time as perceived by users in a large congested public transport network using multiple months of smart card and AVL big data. Crowding distribution was revealed for a representative workday from stable travel patterns among 20 workdays, enabling the creation of a public transport networks with embedded crowding time valuation for all routes and added buffer time for OD pairs. Buffer time was estimated using 95th travel time from revealed trips for OD pairs with a minimum representative number of trips, with limitations such as zone sizes and lack of trip purpose information. We analyzed accessibility comparing scenarios that considered (a) speeds from AVL data, (b) speeds and travel time reliability, (c) speeds and crowding, and (d) all three together.

The results we have obtained show that there is a spatially uneven and significant influence of the higher perceived travel times and trip buffer times on reduction of accessibility considering travel time as perceived by users throughout the city of São Paulo, Brazil. While higher perceived travel times do not reduce accessibility per se, crowding and travel time variability may be used to justify harder access to certain areas when travelers make decisions where to look for jobs. Literature has shown the relevance of crowding discomfort on user perception and travel time unreliability in trip buffer times, and this research aimed to translate those into accessibility assessments. In this paper, we have shown the importance of how crowding and travel time variability influence accessibility and propose future research paths to include those aspects for policy evaluations using accessibility metrics.

For public transport providers, the policy implications of this research are two-fold. Firstly, the proposed methodology enables insights to unveil city areas that are more influenced by crowding and travel time reliability effects, thus enabling investment priority in regions which the population will sense higher gains in perceived travel time. Secondly, the proposed approach allows the evaluation of different intervention scenarios in public transport systems: in the case where higher capacity vehicles are proposed, traditional travel-time only accessibility metrics are not able to capture benefits as perceived by users. Therefore, using the proposed methodology while evaluating higher capacity vehicle plans, the reduction of crowding and users' perceived travel time is revealed. Transit agencies should consider that plans that ought to increase only speeds and improve general accessibility will not thrive if the system has capacity constraints that hamper attracting new users.

Data quality and availability issues have emerged from the analysis as there was a need to remove otherwise valid workdays from our analysis due to data loss. Therefore, the examination of data quality should be done carefully, evaluating potential data losses and errors to enable large panel analysis. Although these data issues may arise due to the multiple and diverse datasets that we have had to analyze, comprising billions of records, we acknowledge our efforts to maintain the most precise analysis and plan for further validations with a large travel survey recently made available for researchers. We also plan to investigate how to improve the precision of boarding stop identification obtained from GPS data and estimated alighting stop identification using the aforementioned travel survey and other local field surveys on buses. Other limitations of our work have also to be noticed: the lack of information on real-time travel speeds and times for the rail system, as well as not including access or first waiting time in trips inferred from smart card data. However, this is a known issue in methodologies for

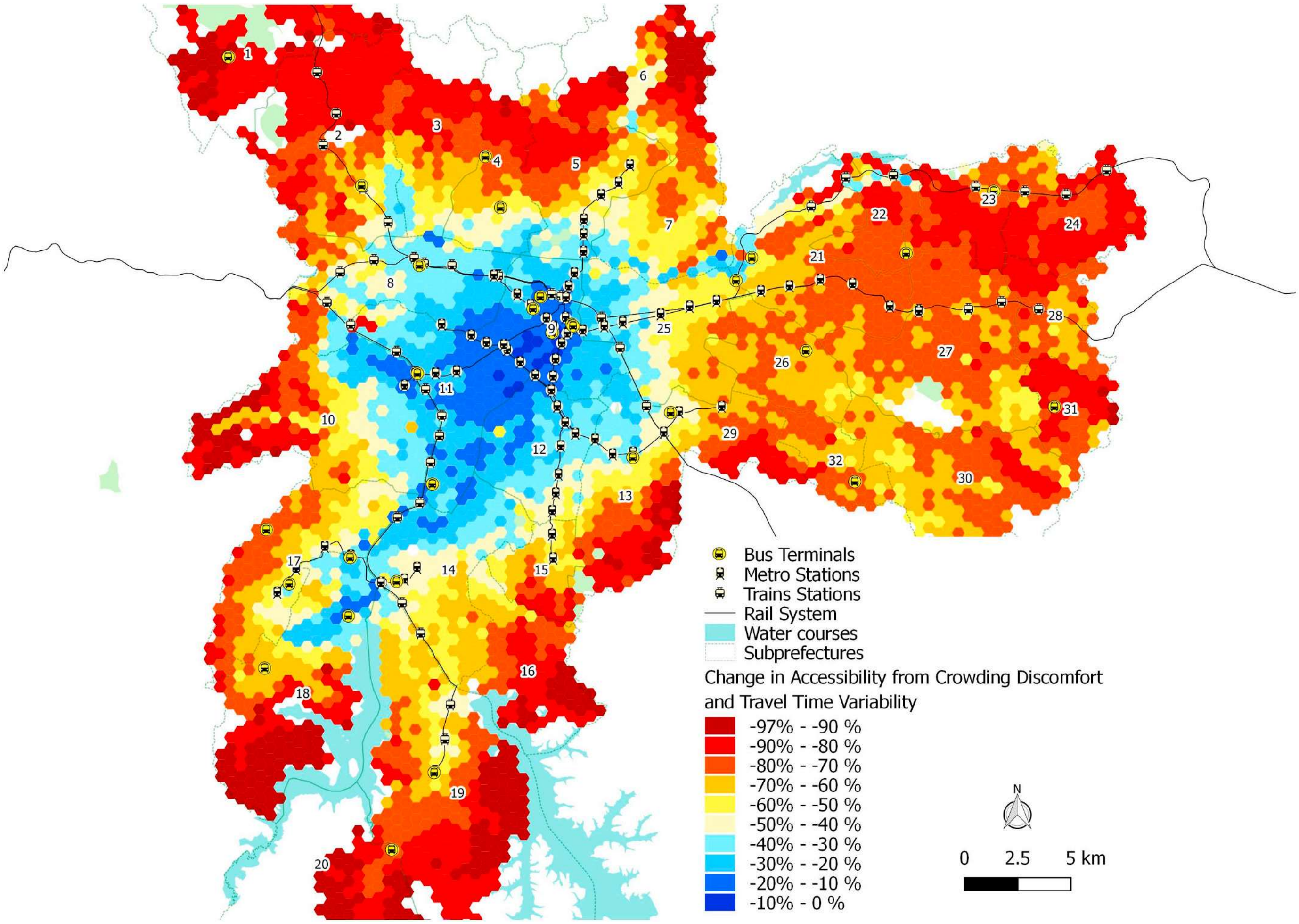

Fig. 10. Influence of crowding and travel time variability on accessibility to jobs considering travel time as perceived by users for a 7 a.m. departure.

smart card data processing.

Multiple future research opportunities arise from this research. One is related to developing an econometric model to evaluate the impact on accessibility incorporating socio-economic variables of city districts, a topic still unexplored in the literature. Fine-tuning the effect of crowding on perceived travel time differently for seating and standing passengers is another possible future improvement. Future research could also assess the distribution of crowding and travel time variability impacts on accessibility along the day for complete profiling of transportation proposals. Other future research direction may comprise analyzing day-to-day variability of crowding influence. Including increased travel time whenever users are unable to board vehicles due to crowding is another possibility of future research.

Calibrating crowding valuation for local characteristics is also an important future research path. Future research should apply a stated preference survey with socio-economic data to measure how users evaluate crowding by presenting pictures of crowding levels (Batarce et al., 2016; Tirachini et al., 2017) or by further analyzing smart card data itself (Yap et al., 2017). Ultimately, future work accessibility should consider a complete generalized cost function for impedance, also including weighted walking and waiting times based on the local perception of travel time. The approach we have proposed to analyze the effect of crowding and travel time variability on accessibility metrics are transferrable and thus can be applied in other cities; it is certainly useful to improve public transport systems, particularly in large metropolis in developing countries, whose network coverage and LOS usually do not match those found in more developed countries. Policymakers can use finer measures for evaluating distinct transportation projects and proposals for equity impacts on perceived travel time savings with proper crowding valuation. In this sense, future work should include the proposed accessibility metrics as one of the main objective functions in the problem of designing public transport network, as accessibility is a core concept that translates the target of promoting access to opportunities. Bringing public transport plans evaluation metrics closer to user perception of service quality will contribute to attractive public transportation systems. In the long run, this will increase public transport as a viable option, promoting a modal shift from cars and thus creating more livable environments for future generations.

Acknowledgments

The authors gratefully acknowledge the data provided by SPTrans (São Paulo Transportes) and Scipopulis for this research paper. The first author thankfully acknowledge the financial support received from Brazil's Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Finance Code 001; the second author acknowledges the support of Brazil's CNPq (National Council for Scientific and Technological Development) [grant number 309424/2018-6].

References

Alsger, A., Assemi, B., Mesbah, M., Ferreira, L., 2016. Validating and improving public transport origin-destination estimation algorithm using smart card fare data. Transp. Res. Part C Emerg. Technol. 68, 490–506. https://doi.org/10.1016/j.trc.2016.05.004.

Arbex, R.O., Da Cunha, C.B., 2017. Estimação da matriz origem-destino e da distribuição espacial da lotação em um sistema de transporte sobre trilhos a partir de dados de bilhetagem eletrônica. Transportes 25, 166–179. https://doi.org/10.14295/transportes.v25i3.1347.

Arbex, R.O., da Cunha, C.B., 2018. Uma avaliação do impacto do desabastecimento de combustível durante a greve dos caminhoneiros no sistema de transporte público de São Paulo: efeitos na lotação, na velocidade das viagens e na regularidade dos usuários, in: Anais Do XXXII Congresso de Pesquisa e Ensino Em Transporte Da ANPET. Gramado, RS. pp. 842–853.

Arbex, R.O., Alves, B.B., Giannotti, M.A., 2016. Comparing accessibility in urban slums using smart card and bus GPS data. In: TRB 95th Annual Meeting Compendium of Papers. Transportation Research Board, Washington, DC p. 17p.

Bagchi, M., White, P.R., 2005. The potential of public transport smart card data. Transp. Policy 12, 464–474. https://doi.org/10.1016/j.tranpol.2005.06.008.

Barry, J., Newhouser, R., Rahbee, A., Sayeda, S., 2002. Origin and destination estimation in New York City with automated fare system data. Transp. Res. Rec. J. Transp. Res. Board 1817, 183–187. https://doi.org/10.3141/1817-24.

Batarce, M., Muñoz, J., de Dios Ortúzar, Juan, Raveau, S., Mojica, C., Ríos, R.A., 2016. Valuing crowding in public transport: implications for cost-benefit analysis. Transp. Res. Part A Policy Pract. 91, 358–378. https://doi.org/10.1016/j.tra.2016.06.025.

Birch, C.P.D., Oom, S.P., Beecham, J.A., 2007. Rectangular and hexagonal grids used for observation, experiment and simulation in ecology. Ecol. Model. 206, 347–359. https://doi.org/10.1016/j.ecolmodel.2007.03.041.

Boisjoly, G., El-Geneidy, A., 2016. Daily fluctuations in transit and job availability: a comparative assessment of time-sensitive accessibility measures. J. Transp. Geogr. 52, 73–81. https://doi.org/10.1016/j.jtrangeo.2016.03.004.

Boisjoly, G., El-Geneidy, A.M., 2017. How to get there? A critical assessment of accessibility objectives and indicators in metropolitan transportation plans. Transp. Policy 55, 38–50. https://doi.org/10.1016/j.tranpol.2016.12.011.

Chen, X., Yu, L., Zhang, Y., Guo, J., 2009. Analyzing urban bus service reliability at the stop, route, and network levels. Transp. Res. Part A Policy Pract. 43, 722–734. https://doi.org/10.1016/j.tra.2009.07.006.

Conveyal, 2018. Conveyal - We Help to Plan Better Cities [WWW Document]. URL. https://www.conveyal.com/.

Conway, M.W., Byrd, A., Van Eggermond, M., 2018. Accounting for uncertainty and variation in accessibility metrics for public transport sketch planning. J. Transp. Land Use 11, 541–558. https://doi.org/10.5198/jtlu.2018.1074.

Cortés, C.E., Gibson, J., Gschwender, A., Munizaga, M., Zúñiga, M., 2011. Commercial bus speed diagnosis based on GPS-monitored data. Transp. Res. Part C Emerg. Technol. 19, 695–707. https://doi.org/10.1016/j.trc.2010.12.008.

Cui, M., Levinson, D., 2018. Full cost analysis of accessibility. J. Transp. Land Use 11, 661–679. https://doi.org/10.5198/jtlu.2018.1042.

El-Geneidy, A.M., Levinson, D.M., 2006. Access to destinations: development of accessibility measures. In: Technical Report, Minnesota.

El-Geneidy, A., Levinson, D., Diab, E., Boisjoly, G., Verbich, D., Loong, C., 2016. The cost of equity: assessing transit accessibility and social disparity using total travel cost. Transp. Res. Part A Policy Pract. 91, 302–316. https://doi.org/10.1016/j.tra.2016.07.003.

Farber, S., Fu, L., 2017. Dynamic public transit accessibility using travel time cubes: comparing the effects of infrastructure (dis)investments over time. Comput. Environ. Urban. Syst. 62, 30–40. https://doi.org/10.1016/j.compenvurbsys.2016.10.005.

Fayyaz, S.K., Liu, X.C., Porter, R.J., 2017. Dynamic transit accessibility and transit gap causality analysis. J. Transp. Geogr. 59, 27–39. https://doi.org/10.1016/j.jtrangeo.2017.01.006.

FHWA (Federal Highway Administration), 2010. Travel Time Reliability: Making it There on Time, all the Time. U.S. Department of Transportation, Federal Highway Administration. [WWW document]. URL. https://ops.fhwa.dot.gov/publications/tt_reliability/TTR_Report.htm.

Geurs, K.T., van Wee, B., 2004. Accessibility evaluation of land-use and transport strategies: review and research directions. J. Transp. Geogr. 12, 127–140. https://doi.org/10.1016/j.jtrangeo.2003.10.005.

Google, 2019. Google Transit - O que é GTFS? [WWW Document]. URL. https://developers.google.com/transit/gtfs/?hl = pt-br (accessed 4.3.19).

Gordon, J., Koutsopoulos, H., Wilson, N., Attanucci, J., 2013. Automated inference of linked transit journeys in London using fare-transaction and vehicle location data. Transp. Res. Rec. J. Transp. Res. Board 2343, 17–24. https://doi.org/10.3141/2343-03.

Hansen, W.G., 1959. How accessibility shapes land use. J. Am. Inst. Plann. 25, 73–76. https://doi.org/10.1080/01944365908978307.

Haywood, L., Koning, M., Monchambert, G., 2017. Crowding in public transport: who cares and why? Transp. Res. Part A Policy Pract. 100, 215–227. https://doi.org/10.1016/j.tra.2017.04.022.

He, L., Nassir, N., Trépanier, M., Hickman, M., 2015. Validating and Calibrating a Destination Estimation Algorithm for Public Transport Smart Card Fare Collection Systems.

Hernandez, D., 2018. Uneven mobilities , uneven opportunities : Social distribution of public transport accessibility to jobs and education in Montevideo. J. Transp. Geogr. 67, 119–125. https://doi.org/10.1016/j.jtrangeo.2017.08.017.

Hickman, M.D., 2016. Transit origin-destination estimation. In: Public Transport Planning with Smart Card Data. CRC Press, pp. 15–35.

Hörcher, D., Graham, D.J., Anderson, R.J., 2017. Crowding cost estimation with large scale smart card and vehicle location data. Transp. Res. Part B Methodol. 95, 105–125. https://doi.org/10.1016/j.trb.2016.10.015.

Iacono, M., Krizek, K.J., El-Geneidy, A., 2010. Measuring non-motorized accessibility: issues, alternatives, and execution. J. Transp. Geogr. 18, 133–140. https://doi.org/10.1016/j.jtrangeo.2009.02.002.

IBGE (Instituto Brasileiro de Geografia e Estatística), 2018. Instituto Brasileiro de Geografia e Estatística: Censo Demográfico 2010 [WWW Document]. URL. https://censo2010.ibge.gov.br/.

Kieu, L.-M., Bhaskar, A., Chung, E., 2015. Public transport travel-time variability definitions and monitoring. J. Transp. Eng. 141, 1–9. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000724.

Kittelson & Associates, I, Brinkenhoff, P., Group, K, Institute, T.A.T., Arup, 2013. Transit Capacity and Quality of Service Manual, Third Edition. Transit Cooperative Highway Research Program (TCRP) Report 165. Third Edit. ed. Transportation Research Board, Washington, D.C. https://doi.org/10.17226/24766.

Kwon, J., Barkley, T., Hranac, R., Petty, K., Compin, N., 2011. Decomposition of travel time reliability into various sources. Transp. Res. Rec. J. Transp. Res. Board 2229, 28–33. https://doi.org/10.3141/2229-04.

Lee, J.-H., Cho, S.-H., Kim, D.-K., Lee, C., 2016. Valuation of travel time reliability accommodating heterogeneity of route choice behaviors. Transp. Res. Rec. J. Transp. Res. Board 2565, 86–93. https://doi.org/10.3141/2565-10.

Li, Z., Hensher, D.A., 2013. Crowding in public transport: a review of objective and subjective measures. J. Public Transp. 16, 107–134. https://doi.org/10.5038/2375-0901.16.2.6.

Li, T., Sun, D., Jing, P., Yang, K., 2018. Smart card data mining of public transport destination: a literature review. Information 9 (18), 1–21. https://doi.org/10.3390/info9010018.

Liu, X., Zhou, Y., Rau, A., 2019. Smart card data-centric replication of the multi-modal public transport system in Singapore. J. Transp. Geogr. 76, 254–264. https://doi.org/10.1016/j.jtrangeo.2018.02.004.

Lomax, T., Schrank, D., Turner, S., Margiotta, R., 2003. Selecting Travel Reliability Measures, Texas Transportation Institute. College Station, Texas.

Luo, D., Bonnetain, L., Cats, O., van Lint, H., 2018. Constructing spatiotemporal load profiles of transit vehicles with multiple data sources. Transp. Res. Rec. J. Transp. Res. Board 2672 (8), 175–186. https://doi.org/10.1177/0361198118781166.

Ma, Z.L., Ferreira, L., Mesbah, M., Hojati, A.T., 2015. Modelling bus travel time reliability using supply and demand data from automatic vehicle location and smart card systems. 94th Annu. Meet. Transp. Res. Board 14.

Metrô SP, 2018. Pesquisa Origem Destino 2017-50 anos [WWW Document]. URL. http://www.metro.sp.gov.br/pesquisa-od/.

Monteiro, J., Pons, I., Speicys, R., 2015. Big Data para análise de métricas de qualidade de transporte: metodologia e aplicação. ANTP, São Paulo.

MTE (Ministério do Trabalho), 2018. Ministério do Trabalho - Relação Anual de Informações Sociais (RAIS) [WWW Document]. URL. www.mte.gov.br/rais.

Munizaga, M., Palma, C., 2012. Estimation of a disaggregate multimodal public transport origin-destination matrix from passive smartcard data from Santiago. Chile. Transp. Res. Part C Emerg. Technol. 24, 9–18. https://doi.org/10.1016/j.trc.2012.01.007.

NumPy, 2018. NumPy - Package for Scientific Computing with Python [WWW Document]. URL. http://www.numpy.org/.

OTP (OpenTripPlanner), 2019. OpenTripPlanner - Multimodal Trip Planning [WWW Document]. URL. http://www.opentripplanner.org/ (accessed 4.12.19).

Owen, A., Levinson, D.M., 2015. Modeling the commute mode share of transit using continuous accessibility to jobs. Transp. Res. Part A Policy Pract. 74, 110–122. https://doi.org/10.1016/j.tra.2015.02.002.

Páez, A., Scott, D.M., Morency, C., 2012. Measuring accessibility: positive and normative implementations of various accessibility indicators. J. Transp. Geogr. 25, 141–153. https://doi.org/10.1016/j.jtrangeo.2012.03.016.

Pandas, 2018. Pandas - Python Data Analysis Library [WWW Document]. URL. https://pandas.pydata.org/.

Pelletier, M.-P., Trépanier, M., Morency, C., 2011. Smart card data use in public transit: a literature review. Transp. Res. Part C Emerg. Technol. 19, 557–568. https://doi.org/10.1016/j.trc.2010.12.003.

Pereira, R.H.M., 2019. Future accessibility impacts of transport policy scenarios: equity and sensitivity to travel time thresholds for bus rapid transit expansion in Rio de Janeiro. J. Transp. Geogr. 74, 321–332. https://doi.org/10.1016/j.jtrangeo.2018.12.005.

Pu, W., 2011. Analytic relationships between travel time reliability measures. Transp. Res. Rec. J. Transp. Res. Board 2254, 122–130. https://doi.org/10.3141/2254-13.

Robinson, S., Narayanan, B., Toh, N., Pereira, F., 2014. Methods for pre-processing smartcard data to improve data quality. Transp. Res. Part C Emerg. Technol. 49, 43–58. https://doi.org/10.1016/j.trc.2014.10.006.

Sarkar, P.K., Jain, A.K., 2017. Defining and assessing congestion inside metro trains and at station: case study of Delhi metro. India. Int. J. Traffic Transp. Eng. 7, 93–107.

Scipopulis, 2018. Scipopulis - Construindo Conhecimento com a sabedoria coletiva [WWW Document]. URL. https://www.scipopulis.com/.

Seaborn, C., Attanucci, J., Wilson, N.H.M., 2009. Using smart card fare payment data to analyze multi-modal public transport journeys in London. Transp. Res. Rec. J. Transp. Res. Board. https://doi.org/10.3141/2121-06. 2121.-1 55–62.

SPTrans, 2018. SPTrans - São Paulo Transporte [WWW Document]. URL. www.sptrans.com.br.

Stewart, A.F., 2017. Mapping transit accessibility: possibilities for public participation. Transp. Res. Part A Policy Pract. 104, 150–166. https://doi.org/10.1016/j.tra.2017.03.015.

Tirachini, A., Hensher, D.A., Rose, J.M., 2013. Crowding in public transport systems: effects on users, operation and implications for the estimation of demand. Transp. Res. Part A Policy Pract. 53, 36–52. https://doi.org/10.1016/j.tra.2013.06.005.

Tirachini, A., Sun, L., Erath, A., Chakirov, A., 2016. Valuation of sitting and standing in metro trains using revealed preferences. Transp. Policy 47, 94–104. https://doi.org/10.1016/j.tranpol.2015.12.004.

Tirachini, A., Hurtubia, R., Dekker, T., Daziano, R.A., 2017. Estimation of crowding discomfort in public transport: results from Santiago de Chile. Transp. Res. Part A Policy Pract. 103, 311–326. https://doi.org/10.1016/j.tra.2017.06.008.

Trépanier, M., Tranchant, N., Chapleau, R., 2007. Individual trip destination estimation in a transit smart card automated fare collection system. J. Intell. Transp. Syst. Technol. Planning, Oper. 11, 1–14. https://doi.org/10.1080/15472450601122256.

Van Oort, N., Drost, M., Brands, T., Yap, M., 2015. Data-driven public transport ridership prediction approach including comfort aspects. In: CASPT 2015: Conference on Advanced Systems in Public Transport. CASPT, Rotterdam.

Walker, J., 2018. To predict with confidence. Plan for Freedom. J. Public Transp. 21, 119–127. https://doi.org/10.5038/2375-0901.21.1.12.

Wang, W., Attanucci, J., Wilson, N., 2011. Bus passenger origin-destination estimation and related analyses using automated data collection systems. J. Public Transp. 14, 131–150. https://doi.org/10.5038/2375-0901.14.4.7.

Wessel, N., Farber, S., 2019. On the accuracy of schedule-based GTFS for measuring accessibility. J. Transp. Land Use 12, 475–500 (https://doi.org/DOI 10.17605/OSF.IO/S5K3B).

Whelan, G.A., Crockett, J., 2009. An Investigation of the Willingness to Pay to Reduce Rail Overcrowding. First International Conference on Choice Modelling. (https://doi.org/http://www.icmconference.org.uk/index.php/icmc/icmc2009/paper/download/31/51/).

Widener, M.J., Minaker, L., Farber, S., Allen, J., Vitali, B., Coleman, P.C., Cook, B., 2017. How do changes in the daily food and transportation environments affect grocery store accessibility? Appl. Geogr. 83, 46–62. https://doi.org/10.1016/j.apgeog.2017.03.018.

Wilson, N., Hemily, B., 2016. Opportunities, challenges and thoughts for the future. In: Public Transport Planning with Smart Card Data. CRC Press, pp. 245–261.

Yap, M., Cats, O., Yu, S., van Arem, B., 2017. Crowding valuation in urban tram and bus transportation based on smart card data. Thredbo 15 Compet. Owndersh. L. Passeng. Transp. 60, 4305–4325.