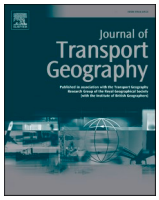
Journal of Transport Geography 109 (2023) 103590

Contents lists available at ScienceDirect

Journal of Transport Geography

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



Evaluating the impact of public transport travel time inaccuracy and variability on socio-spatial inequalities in accessibility

Carlos Kaue V. Braga [a,*], Carlos Felipe Grangeiro Loureiro [b], Rafael H.M. Pereira [c]

[a] Federal University of Ceara, Department of Transport Engineering, Campus do Pici - Bloco 703, CEP 60.440-900 Fortaleza, CE, Brazil
[b] Federal University of Ceara, Brazil
[c] Institute for Applied Economic Research - Ipea, Brazil

ARTICLE INFO

Keywords:
Accessibility
Public transport
Variability
GTFS
GPS
Equity

ABSTRACT

Urban accessibility by public transport has been attracting increasing attention from researchers and transport agencies in recent years. Many of the studies rely on public transport scheduled timetables from GTFS data to calculate accessibility indicators, overlooking the ways in which inaccuracies in scheduled levels of service, as well as day-to-day travel time variability, might impact different socioeconomic groups. This might generate unrealistic or biased results when analyzing accessibility socio-spatial inequalities and assessing transport projects. In this study, we consolidate a method to correct timetables of GTFS feeds based on historical GPS data, and use the city of Fortaleza, Brazil, to show how accessibility to work opportunities based on these two accounts can influence the results of accessibility analyses due to two issues: data inaccuracy and day-to-day travel time variabilities. We use 1-month archived GPS data to create new GTFS timetables that represent both a median level of service and a variability-state level of service; then we use these estimated GTFS to examine the impact of travel time inaccuracy and day-to-day variability on accessibility levels. Results show that, due to the problem of data inaccuracy, the scheduled GTFS underestimates accessibility by 1.5% on average, but in some areas accessibility estimates can be over or underestimated by more than 40%, with significant impact in low income regions. We also find that the variability of travel times have a significant impact of 50% on average on accessibility estimates. This impact is unequally distributed both spatially and across income groups, raising accessibility inequality by 30%. The underlying causes of these impacts are related to several factors, including the GTFS feed's quality, the high concentration of jobs in the city center, and higher travel time variability in the corridors that connect lower-income areas to the city center. These results highlight the importance of considering both inaccuracy and day-to-day variability issues in public transport travel times when estimating accessibility levels and evaluating transport projects, particularly from an equity perspective.

1. Introduction

Urban accessibility has been increasingly recognized as a valuable metric for assessing the benefits associated with transportation and land use systems (Geurs and van Wee, 2004; Levinson and King, 2020). As a result, accessibility has been receiving growing attention from researchers and transportation agencies, particularly in the context of public transportation planning (Farber and Fu, 2017; Mayaud et al., 2019; Pereira, 2019). A common practice in these studies is that they rely on transit scheduled timetables from GTFS data to calculate accessibility estimates. By doing so, they overlook inherent uncertainties in public transport (PT) travel times, and the ways in which 1) scheduled levels of service might differ from what is delivered to the population (*inaccuracy*), and 2) delivered levels of service might vary across different days (*variability*), as well as how these factors might impact different socioeconomic groups and neighborhoods. These two issues might generate unrealistic or biased results when analyzing accessibility socio-spatial inequalities or assessing transport projects.

In this paper we demonstrate how to correct timetables of GTFS feeds based on historical GPS data and show how accessibility estimates based on these two accounts can influence the results of accessibility analyses concerned with transport planning and equity while accounting for day-to-day travel time variability. Using the city of Fortaleza (Brazil) as a case study, we examine how social and spatial inequalities in access to

* Corresponding author.
E-mail addresses: kaue@det.ufc.br (C.K.V. Braga), felipe@det.ufc.br (C.F.G. Loureiro), rafael.pereira@ipea.gov.br (R.H.M. Pereira).

https://doi.org/10.1016/j.jtrangeo.2023.103590
Received 7 September 2022; Received in revised form 22 February 2023; Accepted 14 April 2023
Available online 5 May 2023

employment opportunities can differ when real speed data is considered. We use a method that transforms raw GPS data into two real-time GTFS feeds using two approaches: one based on real-time median speeds to serve as comparison to the scheduled GTFS, and one based on a dispersion value of speeds to serve as comparison to the median one. These comparisons are used to evaluate the impact of inaccuracy and variability on accessibility estimates and socio-spatial inequalities.

Recent studies have developed methodologies to correct scheduled GTFS timetables with GPS data and improve accessibility accuracy (Wessel et al., 2017; Wessel and Farber, 2019; Liu et al., 2022). The authors have found that, on average, using scheduled GTFS data leads to overestimated employment accessibility levels in North American cities, and that these biased estimates show strong spatial patterns. However, these studies generally considered average travel times and did not take day-to-day travel time variability into consideration. The use of average measures of travel times in the network can lead to overestimation of the opportunities that are accessible, especially in public transport systems that present high variability in levels of service. Moreover, even though there are systematic deviations between scheduled and delivered PT services, there is still little understanding about how these disparities affect different income groups, raising concerns about socio-spatial inequalities estimates and subsequent equity-based policy evaluations.

This paper advances the literature in different ways. First, it demonstrates that both issues of inaccuracy and day-to-day travel time variability in GTFS data can importantly influence the estimates of accessibility levels and inequalities. Second, this paper consolidates a method that allows researchers and transport agencies to use GPS data to generate accessibility estimates that account for both inaccuracy and day-to-day biases commonly present in GTFS data. Moreover, this paper shows how using measures of dispersion of observed travel times can lead to greater robustness in transport accessibility analyses. Finally, another contribution of this paper is that the method we proposed to generate GTFS files based on GPS data can be applied to other transportation contexts. This is largely because the method uses the GTFS standard format, which is widely used by transport agencies worldwide, and GPS records, which tend to have similar structure worldwide.

The remainder of this paper is organized as follows. In Section 2, we review the importance of urban accessibility and the current methodologies that incorporate travel time uncertainty in accessibility estimates. Section 3 provides a brief contextualization of our study area, Fortaleza. Section 4 describes the procedure for the reconstruction of the GTFS timetables and the method for analyzing accessibility indicators used in this paper. In Section 5, we present the results of our analyses for the city of Fortaleza. Finally, Section 6 presents the main findings from the case study's results and the broader implications from the methodological contributions of this paper.

2. Literature review

This section provides a literature review on how the problems of travel time *inaccuracy* and *variability* may impact accessibility and inequalities estimates. We start by discussing the methods found in the literature that use vehicle location data (GPS) to improve accessibility accuracy. Then, we proceed to evaluate the state of the art of day-to-day travel time variability in accessibility calculation. Lastly, we summarize the main research gaps identified on both topics.

2.1. Improving accessibility accuracy with GPS data

Over the last two decades, transport accessibility models have become more sophisticated and easier to use, especially after the creation of the GTFS (*General Transit Feed Specification*) data format (Farber and Fu, 2017; Pereira, 2019). The GTFS standard has allowed the emergence of several transport routing models and accessibility tools that account for door-to-door travel time estimates in complex multimodal transport networks (Pereira et al., 2021; Higgins et al., 2022). A key piece of information used by these routing and accessibility models is the scheduled timetables for PT routes, which are presented in the *stop_times.txt* table of GTFS feeds. However, planned travel times in the GTFS can significantly differ from the actual travel times of PT vehicles for several reasons, such as mixed traffic, weather conditions, service interruptions, and bus bunching (Mandelzys and Hellinga, 2010; Elgeneidy et al., 2011; Palm et al., 2020; Park et al., 2020). Furthermore, it is not clear how transit agencies take these factors into consideration when building the timetables represented in their GTFS feeds. Wessel et al. (2017) noted that agencies' schedules may be conservative to guarantee a higher adherence, as they have observed for their case study in Toronto, but it is not clear how such practice could impact travel time and accessibility estimates. These factors can lead to *inaccurate* travel times, which can produce biased accessibility measurements.

Recent studies have developed new methods to use GPS data to overcome this limitation of GTFS feeds in accessibility analyses. Wessel et al. (2017) developed a methodology to use Automated Vehicle Location data (AVL, from GPS) to correct scheduled GTFS timetables for Toronto, improving the accuracy of cumulative accessibility estimates. The authors transformed pre-processed GPS points into a *stop_times* timetable, with one timetable per day of service. In a subsequent paper, Wessel and Farber (2019) applied their methodology to four North American cities and calculated two accessibility indicators for jobs, comparing access after and before the GTFS correction. They used AVL data with the location of all vehicles in each of the cities analyzed, for five days, apart from high-capacity lines for which they used scheduled data. Then, they built a GTFS for each day and calculated the accessibility levels resulting from the service observed on the corresponding day; finally computing an average accessibility for this 5-day sample. They found that schedule-based accessibility tends to *overestimate* accessibility levels by 5% to 15% on average. The authors also found that these differences in accessibility estimates show consistent spatial patterns, with detectable clusters of under and over estimations across the cities analyzed. Regions where accessibility is overestimated by GTFS data are usually located in peripheral areas, where there is generally more service variation, while regions with higher underestimation are in central areas.

More recently, Liu et al. (2022) used GPS real-time data to evaluate the impact of considering scheduled GTFS for the city of Columbus, Ohio. The authors used two methodologies to calculate travel time between origin-destination (OD) points: the first one used the method proposed by Wessel et al. (2017), where travel time estimation is based on the fastest route found in the retrospective GTFS; the second calculated the route based on the scheduled GTFS, and then used the retrospective GTFS travel time for the same route to represent the OD travel time. The authors argued that the second method better represents the users' route decision as they usually plan their routes beforehand and cannot predict future delays or interruptions. They found very similar accessibility estimates when considering Wessel et al. (2017) method and scheduled GTFS data; nevertheless, they found significantly lower accessibility levels when using their proposed method.

2.2. Travel time variability and accessibility

While the advances discussed above are significant, these studies ended up using travel time average measures from multi-day samples of GPS data to calculate accessibility. Using GPS data for 5 different days, Wessel and Farber (2019) calculated travel time matrices at every minute for each day and then measured accessibility by averaging travel times for each origin/h. By only considering the average of travel times, the method used by the authors does not account for travel time variability across different days. However, PT performance can significantly vary between days due to demand fluctuations, driving behavior variations, weather conditions, services disruptions, and non-recurrent disruptions due to road maintenance or crashes. These sources of variability can importantly affect people's ability to consistently reach

opportunities, which requires that users budget extra time when commuting and reduces public trust in the transit system (Mazloumi et al., 2010; Chung, 2019).

According to Mazloumi et al. (2010), the variability of travel times of a given OD pair can be analyzed from three perspectives: 1) *vehicle-to-vehicle* variability, which is the travel time differences between vehicles traveling on the same route at the same time; 2) variability *within the day*, which is the variability of travel time according to the time of day; and 3) *day-to-day variability*, which is the variability of the same trip, made at the same time, on different days. A few studies, such as the work of Farber and Fu (2017) and Conway et al. (2018), address the variability of travel times within the day in their accessibility analyses, despite still using scheduled data. Meanwhile, the work of Wessel and Farber (2019) addresses both vehicle-to-vehicle and within-the-day variabilities in travel time. However, by taking the average accessibility over multiple days, the authors overlook how accessibility is impacted by the dispersion from the day-to-day variability of the public transport system.

Recent studies have proposed methods to incorporate day-to-day variability into accessibility estimates. Focusing on accessibility by car, Chen et al. (2012) made a significant contribution by addressing the impact of travel time variability. Firstly, they devised a method that accounted for this variability by creating a reliable shortest path algorithm that considered the distribution of link travel times. This algorithm considers an on-time arrival probability, providing a more rigorous representation of the uncertainty involved in travel times. Their method has been applied to GPS data from cars in subsequent studies in various accessibility contexts, such as in space-time prisms (Chen et al., 2013), access to food services (Chen et al., 2017), and shopping services (Chen et al., 2019). More recently, Chen et al. (2020) leveraged similar automobile travel time data to analyze how travel time uncertainty can impact healthcare accessibility while accounting for competition for opportunities with a two-step floating catchment area accessibility metric.

For public transport, Arbex and Cunha (2020) used AVL and smartcard data to calculate observed travel times between OD pairs for multiple days in the city of São Paulo. In order to account for day-to-day variability in travel times, the authors estimated a travel time buffer for each OD pair, with this buffer being calculated as the difference between the 95th and the 50th percentiles of all travel times recorded between each OD pair over the course of 20 weekdays.

Despite these recent efforts, there is still little understanding of how both travel time inaccuracy and variability may have different impacts on the accessibility estimates for different socioeconomic groups, and

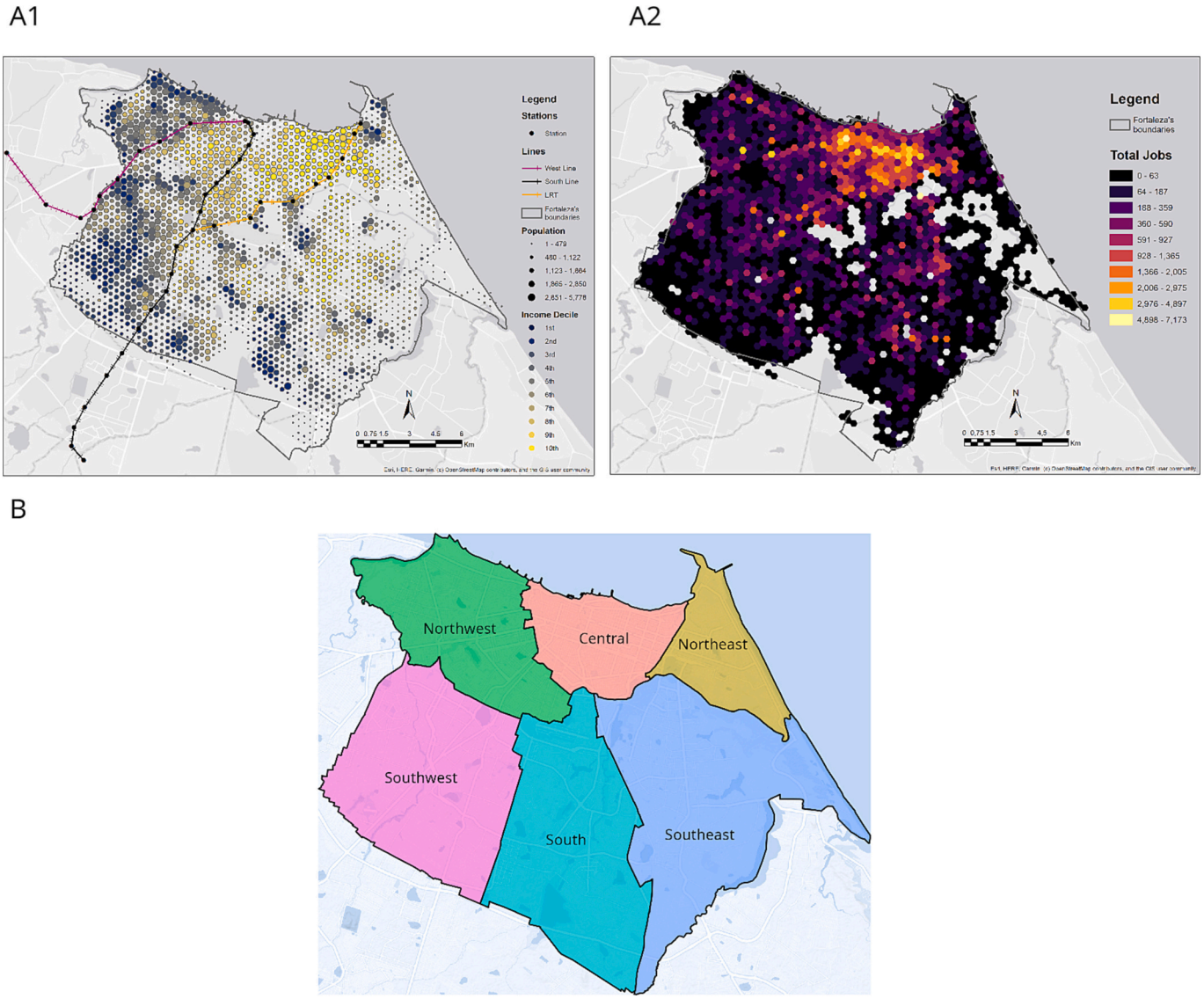


Fig. 1. Fortaleza's population and income distribution from 2010 (A1) and formal jobs distribution from 2019 (A2); Fortaleza's analysis regions (B).

hence affect inequalities in access to opportunities. This is particularly important in cities with marked socio-spatial segregation and inequalities issues, as commonly found in Global South metropolises (UN-HABITAT, 2011). Particularly in Brazil's largest cities, the low-income population is mostly located in the peripheral regions and suffers from poor accessibility conditions and strong inequality levels (Pereira et al., 2019; Bittencourt et al., 2020; Boisjoly et al., 2020; Pinto et al., 2023). As they often live in urban peripheries far from the city center, with poor access to private vehicles, they need longer and more complex transit commutes to reach opportunities, what might make them more exposed to accuracy and variability issues.

3. Study area: Fortaleza, Brazil

Fortaleza is the fifth most populous city in Brazil, with an estimated population of 2.7 million inhabitants (IBGE, 2021). Its public transport system is composed of 318 bus lines and 3 low-to-medium frequency high-capacity train routes that transport approximately one million daily passengers. According to its 2019 Household Travel Survey, Fortaleza's average PT commuting travel time is 52 min. Fig. 1 shows the spatial distribution of people by income (A1) and jobs (A2), which will be considered in the calculation of accessibility indicators. The figure also shows Fortaleza's regions that will be considered in our analysis (B).

Recent studies have shown significant patterns of socio-spatial inequalities in access to opportunities in Brazilian metropolises. Pinto et al. (2023) revealed how Fortaleza's Central region (where the richest live) have much higher accessibility levels than its periphery (where the poorest live), and how these patterns are consistent across multiple accessibility measures. Bittencourt et al. (2020) found that black communities are disadvantaged in their access to jobs opportunities in São Paulo. Results from the Access to Opportunities project (2019) (Pereira et al., 2019) estimate that the richest have on average twice as much accessibility as the poor by public transport in Brazilian largest cities.

4. Data and methods

As stated in Section 1, this research effort addresses two main questions: 1) how using scheduled PT data may impact the accuracy of accessibility estimates compared to real-time PT data; and 2) what is the impact of day-to-day travel time variability on accessibility? We also examine the extent to which both impacts vary across space and among income groups. To answer these questions, we compare accessibility levels calculated using different PT data inputs. To address the first question, we compare the ***scheduled accessibility*** (based on scheduled GTFS) against the ***median-corrected accessibility***, which uses historical GPS data to fix travel times in the GTFS. To address the second question, we compare the ***median-corrected accessibility*** with a ***dispersion-corrected accessibility*** (based on a version of GTFS corrected by a dispersion measure of real speeds). The method to correct GTFS time tables with historical GPS data is described in the following subsections, and is illustrated in the diagram presented on Fig. 2.

4.1. GPS and GTFS data

GPS and GTFS data from the bus system were provided by Fortaleza's Urban Transport Company (ETUFOR). GTFS data from the train and metro systems were obtained from the state transport agency, METROFOR. Both databases referred to the month of September 2018. According to the GTFS data, a typical business day in the Fortaleza's PT system had 35,000 vehicle trips distributed across 301 routes.

GPS records registered the timestamp and the spatial coordinates of each vehicle every 30 s in most cases, covering only for the bus system. On average, there were approximately 4 million data points per day during the 19 business days of September 2018. Approximately, 85% of the bus fleet was covered in the GPS dataset. The other 15% missing from the GPS dataset consisted of smaller vehicles running on feeder routes with low capacity and passenger demand.

4.2. Creating timetables from GPS data

The method proposed here follows the methodology developed in the work of Braga et al. (2020). This methodology was selected because, in contrast to the method developed by Wessel et al. (2017), it yields a travel time distribution for each public transport segment between consecutive bus stops and time interval across multiple days, allowing us to calculate and analyze day-to-day travel time variability. Moreover, such an approach also has the advantage to cope with incomplete GPS data, which is a common problem likely to happen in most cities of the Global South. The first step of the proposed method was to convert the raw GPS data to a timetable format, like the format used on a GTFS *stop_times.txt* file. This process was only possible because there is equivalence between the route number from the GPS database with the *route_*id from the GTFS *shapes.txt* file. The following steps were repeated for each vehicle and for each day:

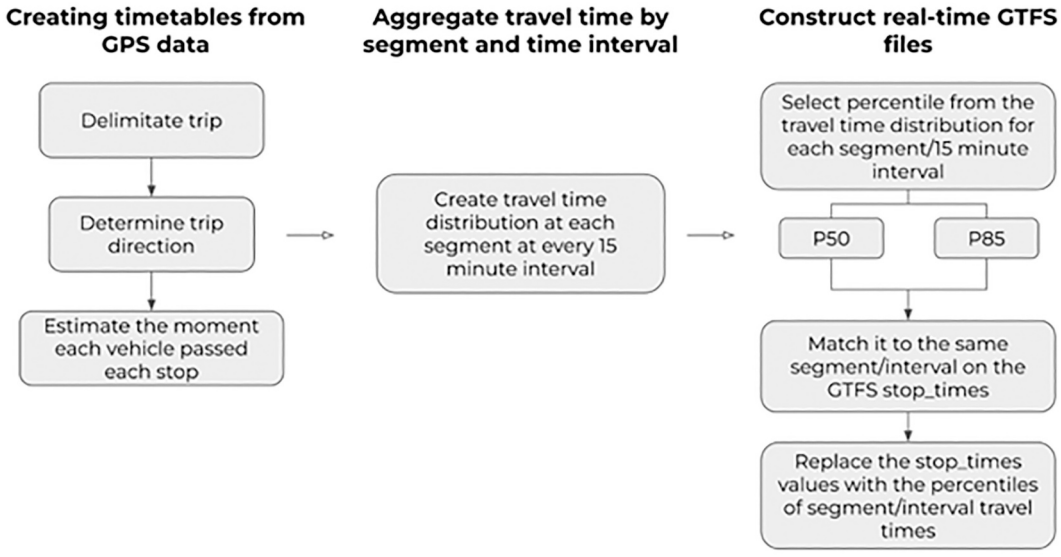


Fig. 2. Adopted method to calculate real-time GTFS feeds from raw GPS data.

- **Delimitate each trip**: we create a buffer of 100 m around the start/endpoints of each route (obtained from the GTFS *shapes.txt* file) and evaluate: if a vehicle consistently (for more than 10 GPS points) enters the route's endpoint, the trip finishes there. The next trip starts when the vehicle consistently leaves the endpoint;
- **Determine the direction of the trip (in or outbound)**: we snap each vehicle's GPS point to the stop sequence for the corresponding *inbound* route in the GTFS file. This will produce either an ascending sequence (point 1 was snapped to stop 1, point 10 was snapped to stop 2, …) or a descending sequence (point 1 was snapped to stop 20, point 10 was snapped to stop 19, …). If the correspondent sequence of each stop is ascending, it means that the trip is inbound. Otherwise, outbound;
- **Estimate the moment when each vehicle passed through each stop**: we snapped both the GPS points and the stops to the route geometry and calculated the cumulative distance along the route itinerary for each one. We linearly interpolated the timestamp of each stop based on the distances and the timestamp of each GPS point. The resulting table is similar to the one found in the GTFS *stop_times.txt* file.

4.3. Aggregate travel time by segment and time interval

The second step of the proposed method is to aggregate the travel times of all vehicles between each pair of consecutive stops (each PT network link) over 15-min time intervals. Thus, each link will have a travel time distribution for each time interval, with the number of observations depending on the number of routes and frequencies of buses running on each link. To increase the reliability of these measurements, travel times of segment and interval combinations that have less than 10 observations are replaced by the travel time already available in the scheduled timetable. The aggregation generated around 120 thousand combinations of segments and 15-min intervals (5400 segments × 22 time intervals in the morning). About 17% of the combinations had a sample of less than 10 travel times. Most of these combinations are from segments that are in peripheral areas and with little bus traffic, which to a certain extent makes the data trimming less problematic as these areas tend to present less variability in travel times.

This methodological step helps us overcome the problem of missing data in the GPS records. Although Fortaleza's GPS data didn't cover minibuses and vans, we were also able to generate reliable timetables for the services running with low-capacity vehicles by using the travel times estimates aggregated between PT stops.

4.4. Construct real-time GTFS files

Using the distribution of travel times between consecutive pairs of PT stops, we proceed to build the real-time GTFS feeds. To do this, we replace the scheduled travel times between stops (present in *stop_times.txt* from the Scheduled GTFS) with the observed travel times collected from GPS, and use this information to recalculate the departure and arrival times from/at each PT stop for every trip. In order to do so, we assume that the first departure time for each vehicle recorded in the scheduled GTFS feed is correct. This was only possible because we could individually identify each vehicle on the *trip_id* column of the *stop_times.txt* file, and this particularity from Fortaleza's GTFS allowed us to incorporate observed travel times in a way that each vehicle's trip timetable is affected by previous trips from the same vehicle.

To construct the GPS-based median real-time GTFS, we extract the median value (50th percentile) from each travel time distribution between PT stops, match it to the same link and interval in the scheduled *stop_times*, and replace the scheduled travel time. We call this the *Realtime P50 GTFS*. Next, in order to represent the dispersion of travel times in the GPS data, we choose the percentile 85 from the travel time distributions and follow the same method to create a new GTFS. We call this the *Realtime P85 GTFS*. We choose a high percentile as a way to calculate more reliable accessibility estimates given the day-to-day variability in service levels. We particularly recommend the percentile 85 because it represents around 1 standard deviation above the mean – assuming travel times are distributed according to a Normal distribution, as studies about travel time variability have found to be adequate (Abkowitz et al., 1987).

After the Realtime P50 GTFS and the Realtime P85 GTFS are generated, we calculate employment accessibility with these GTFS feeds (along with the Scheduled GTFS) and start drawing comparisons between accessibility estimates to address our research questions. Comparing accessibility estimates based on Scheduled GTFS against Realtime P50 GTFS help us examine how the accuracy problem in scheduled PT data can affect accessibility analyses. On the other hand, comparing accessibility estimates based on Realtime P50 GTFS against Realtime P85 GTFS allows us to capture the impact of day-to-day travel time variability on accessibility analyses. In Section 4.5, we describe the method used to calculate and compare accessibility estimates.

A shortcoming of the method adopted in this paper is that it only corrects for the speeds between PT stops. That means that it doesn't account for how PT frequencies might differ between planned and delivered services. So, if there is a large difference between scheduled and observed number of trips, this method should be used with caution, as it assumes that frequencies don't change. According to Fortaleza's agency, this should not be a problem for this case. Moreover, it is worth noting that the approach used in this study considers an implicit assumption that the entire public transport network is operating at the 50th or 85th percentile condition, while estimating central tendency and dispersion of accessibility levels. Therefore, by considering the percentiles of the travel time distribution on every link, our method leads to an overestimation of the levels-of-service experienced by public transport users. This is because link travel times are highly stochastic due to momentary fluctuations in traffic conditions (Chen et al., 2019). Nonetheless, we mitigate this issue by aggregating for every link the travel times of multiple vehicles over a period of 20 days, which may increase estimates robustness.

Despite its limitations, the approach used in this paper has two advantages compared to similar methods adopted in previous studies. First, our method is computationally more efficient, since it requires computing only two synthetic GTFS feeds and their corresponding accessibility levels, one for the central tendency (P50) and another for the dispersion (P85). Moreover, the method can be easily adapted to periods of analysis of different durations, which in the case of this study is a 20-day period. In contrast, the method used by Wessel et al. (2017) requires generating multiple historical GTFS feeds (one feed for every day of GPS records) and calculating accessibility estimates for each day. This can be computationally costly for large periods of analysis, and restricting sample size can potentially compromise the robustness of travel-time and variability estimates.

Another important advantage of our method is that it compensates for eventual limitations of missing data in the GPS records. These records can often be incomplete due to signal losses, faulty equipment or simply because not all vehicles have an onboard GPS device functioning correctly all the time. We believe our method mitigates this problem by calculating link travel times from aggregate data of all vehicles over a relatively long period of time, combined with planned service information from GTFS feeds. The method used by Wessel et al., (2017), on the other hand, leaves minimal room for missing GPS data because it requires that the GPS dataset captures every vehicle in the system, otherwise the accessibility calculations may be incomplete.

4.5. Analyzing accessibility differences

In this paper, we measure access to employment using a cumulative opportunities metric. We chose a cumulative measure because it is the most used accessibility metric (Manaugh et al., 2015; Papa and Bertolini, 2015; Wessel and Farber, 2019), and has all of its limitations well

defined and discussed in the literature (Geurs and van Wee, 2004).

The first step to calculate accessibility was to estimate the travel times between origin-destination (OD) pairs by public transport. To achieve this, the R5 routing engine was used through the r5r package in R (Pereira et al., 2021). The application provides detailed travel time estimates considering door-to-door PT journey: walk time to stop, waiting time, time in the vehicle, and waiting time for possible integrations. The OD pairs are created considering the centroids of the H3 hexagonal spatial index developed by Uber. The chosen aggregation is a hexagon with a diagonal of 357 m and area of 0.1 km^2, which allows capturing the variability of accessibility at a high spatial resolution level.

For each of the GTFS feeds, we calculated the median travel time between 6 am and 8 am, considering multiple departure times every minute in that time window. The calculation of a matrix per minute is made because even small differences in the start time of the trip can imply large differences in the total trip time, mainly due to the loss of integrations and long waits (Conway et al., 2018; Stępniak et al., 2019).

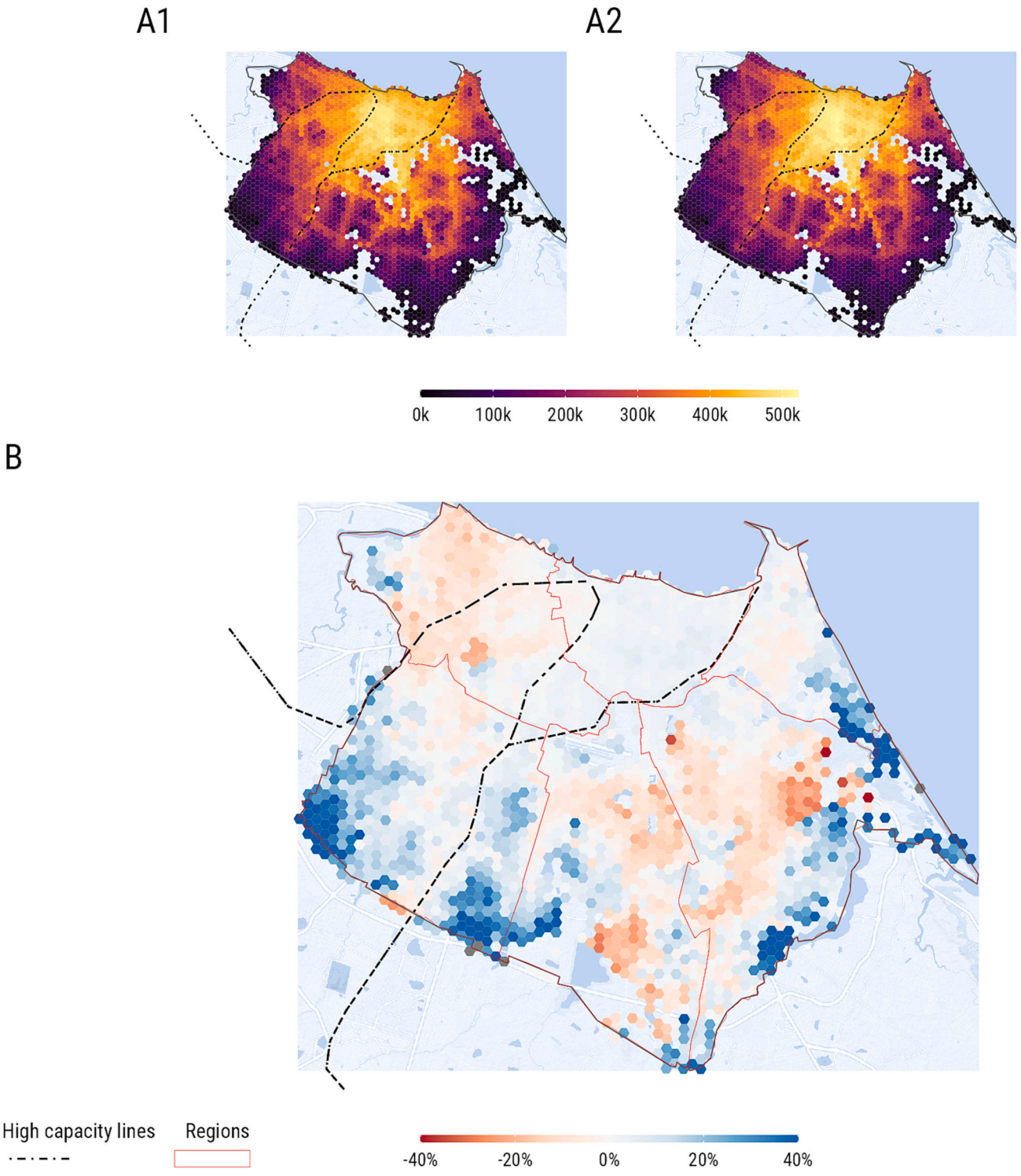


Fig. 3. Accessibility distribution estimated by Scheduled GTFS and Realtime P50 GTFS (A1 and A2); and the relative difference between the accounts (B).

With the median travel time matrices at hand, we calculated cumulative access to jobs following Eq. 1, where Dj is the total opportunities at location j.

$$A_i = \sum_{i=1}^{j} D_j f(c_{ij}) \quad (1)$$

Where

$$f(c_{ij}) = 1, if\ t_{ij} \leq t_{max}$$
$$f(c_{ij}) = 0, if\ t_{ij} > t_{max}$$

Cumulative indicators are calculated for each GTFS for employment opportunities with a time threshold of 60 min, which roughly represents Fortaleza's average PT commute travel time of 58 min (Braga et al., 2022). Employment data was obtained from the Ministry of Labor's Annual List of Social Information (*Relação Anual de Informações Sociais* - RAIS) in 2018. The dataset of geolocated jobs at hexagons was downloaded from the R package *aopdata* (Pereira et al., 2019).

To investigate the impact of the type of public transport input data on accessibility inequalities, we 1) examine how accessibility levels are distributed across income deciles and 2) use the Palma Ratio as a summary inequality measure. Palma Ratio is an inequality indicator that calculates the ratio between the average accessibility of the richest (decile 10) by the average accessibility of the poor (deciles 1 to 4), and which has been advocated and used in the literature to represent inequalities in access to urban activities (Pritchard et al., 2019; Geurs, 2020; Herszenhut et al., 2022).

5. Results

This section is divided into two main topics: the first one evaluates how accessibility analyses can be affected by the inaccuracy problem in scheduled GTFS when compared to real-time data. In the second part, we analyze the impact of day-to-day travel time variability on accessibility estimates.

5.1. Inaccuracy problem in scheduled GTFS

Fig. 3 shows the distribution of accessibility levels calculated based on Scheduled GTFS (A1) and for the Realtime P50 GTFS (A2), as well as the relative difference in accessibility estimates from both input data (B). To compare the accessibility levels from different GTFS feeds, we calculated the relative change $\frac{y-x}{x}$ multiplied by 100 to communicate the results as percentage (%).

The maps A1 and A2 show how Fortaleza's Central region has better accessibility conditions in comparison to its peripheral areas. We can also see the effect of the transport corridors on employment accessibility. Upon visual inspection of the two maps, no discernible pattern of differences in accessibility levels between the scheduled GTFS and the Realtime P50 GTFS can be identified. Map B helps us highlight the difference between the two accounts. Negative values indicate areas where Scheduled GTFS overestimates accessibility compared to the Realtime P50 GTFS. The median difference between the two approaches is just +1.5%, which indicates that Scheduled GTFS slightly underestimates accessibility levels when compared to Realtime P50 GTFS. However, there is a certain balance between over and underestimation by the Scheduled GTFS, with the interquartile interval ranging from −5% to 7%.

The map shows that the Central region, which concentrates most of the job opportunities, is barely affected by the type of data input. Locations in this region already access most of the opportunities in in the city in less than 60 min, and the difference between observed and scheduled times has minimal impact on them. Meanwhile, the type of data input has much larger impacts on accessibility levels in peripheral regions, where there are contrasting patterns. While scheduled GTFS tends to overestimate accessibility in the Northwest, South and Southeast regions, it tends to underestimate accessibility in the Southwest region, where most of the low-income population lives. Besides, the areas especially along the South metro line are barely affected by the type of data input because we had to assume that the metro's scheduled timetables wouldn't change in the scenarios.

There are two main factors that may be causing these distinct patterns. First, areas where Scheduled GTFS overestimates accessibility levels (South and Southeast regions, especially) concentrate more congested roads, which can lead to greater impacts on travel times. The Southwest region, on the other hand, has better infrastructure of dedicated bus lanes that help protect PT bus trips from traffic congestion. Second, there could be issues regarding the quality of scheduled GTFS timetables. As discussed by Wessel et al. (2017), GTFS timetables can be built conservatively, perhaps to avoid system operators being punished for possible delays. Furthermore, the GTFS of Fortaleza, as well as of other Brazilian cities, only informs the time of departure from the 1st stop and the arrival time at the last stop of each trip. In these cases, routing engines such as R5 or OpenTripPlanner assume that the trip has a constant average speed and linearly interpolates the arrival times at the stops. This means that accessibility estimates with this type of Scheduled GTFS ignore speed variations along the route, which may disproportionately affect certain areas of the city, causing divergent patterns.

Because the inaccuracy of scheduled GTFS data is heterogeneously distributed in space, this raises the question about whether the choice of using Scheduled GTFS or Realtime P50 GTFS could impact the results of our accessibility inequality analysis. The boxplots in Fig. 4 present how differences in accessibility from both accounts vary by income groups. The peripheral regions, which are the ones with the most distinct differences, concentrate most of the low-income population in the city. This explains why low-income groups (deciles 1 to 4) have a greater dispersion in the difference values, showing larger variation between positive and negative values. Meanwhile, the wealthiest population lives mostly in the Central region, where there is only a small variation in accessibility estimates based on scheduled and Realtime P50 GTFS.

These results for Fortaleza suggest that the accessibility differences between Scheduled GTFS and Realtime P50 GTFS are fairly symmetrical (with positive and negative values equally distributed) for most income groups. Consequently, we find no change in accessibility inequality with Scheduled or Realtime P50 GTFS. Using both GTFS inputs, the Palma ratio stays at approximately 2, meaning that the richest can access two times more employment opportunities than the poor.

In summary, the evidence here is slightly different from previous studies. The work of Wessel and Farber (2019) found that scheduled GTFS, on average, overestimated accessibility by 5% to 15% in the North American cities analyzed, with a consistent pattern of overestimation in their peripheral regions. The difference between our results and the results from Wessel and Farber (2019) may occur because of 1) different methodologies, 2) different spatial aggregation units, 3) differences in PT systems' reliability, and 4) differences in GTFS quality. Based on their results, we also expected for Fortaleza a more defined pattern of *overestimation* of accessibility in peripheral locations, which would result in an underestimation of income inequalities in accessibility. The distinct patterns found in this paper reflect the complexity and uncertainty that lies behind the creation of scheduled timetables in the Global South.

5.2. Impact of day-to-day travel time variability

The impact of day-to-day variability on accessibility estimates is shown in Fig. 5, where we compare accessibility levels calculated using the Realtime P50 GTFS and with the Realtime P85 GTFS. The top maps (A1 and A2) show the spatial distribution of accessibility levels with the Realtime P50 GTFS and the Realtime P85 GTFS, respectively. These maps already give a sense of the impact of travel time variability on accessibility levels, where areas outside the Central region tend to be the most affected, showing much lower levels of employment accessibility

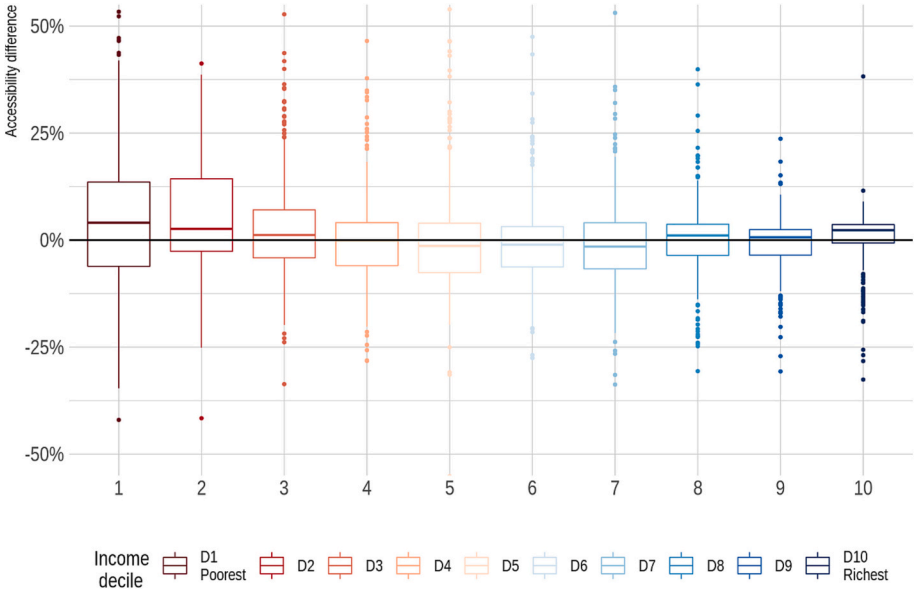


Fig. 4. Distribution of the relative difference between the Scheduled GTFS and the Realtime P50 GTFS by income deciles.

when accounting for day-to-day service variability (A2).

Map B on Fig. 5 shows the relative difference between P50 GTFS and P85 GTFS. There are only negative values in the distribution because, as expected, the day-to-day variability of services captured with P85 GTFS generates systematically lower levels of accessibility than the median real times from P50 GTFS. The median difference between the two estimates is −50%, indicating a significant impact of travel time variability on overall levels of employment accessibility. In absolute values, locations have, on average, access to 80,000 fewer jobs when we account for day-to-day PT variability. The distribution shows an interquartile interval ranging from −34% to −65%. In the most extreme cases, the day-to-day variability in service levels are so significant in some areas (colored in darker red) that residents could only access 80% fewer jobs than what would be assumed considering median travel times from P50 GTFS.

As expected, the Central region is the least affected by variability, as its proximity to activities makes it less sensitive to day-to-day variations in travel times. The peripheral regions present a consistent pattern where variability negatively affects accessibility, except for the areas surrounding the South metro line, which has no variability in its schedules due to quick access to rail services. Despite presenting no variability in their operation, the remaining rail routes didn't impact the difference as much, especially because they are low frequency lines (45 min in peak hour).

The travel time variability affected most significantly the South region. This region has important radial transport corridors which are regularly affected by congestion and disruptions that can cause great service variability. Meanwhile in the Northwest region, shades of lower variability impact emerge mainly due to the presence of a segregated bus corridor that goes toward the Central region of the city.

Finally, how does day-to-day travel time variability impact social inequality in access to jobs? When the impact of day-to-day variability is analyzed by income group (Fig. 6), we observe that the poorest are the most impacted, and there is a clear progression with decreasing effects for higher income groups. The poorest (decile 1) and the richest (decile 10) also present lower levels of within-group variance, indicating that the poorest consistently suffer more from variation while the richest suffer less. The Palma Ratio measure of inequality rises from 2 when considering P50 GTFS to 2.7 with the P85 GTFS. In other words, ignoring day-to-day travel time underestimates accessibility inequalities by 35% in our case study.

When compared to other studies in a Global South context, we found much larger impacts of day-to-day travel time variability on accessibility. The work of Arbex and Cunha (2020) found for São Paulo an average reduction of 6.2% in accessibility estimates when considering travel time variability - against 50% found in our study. This difference in magnitude may be due to methodological choices (different time thresholds, different trip departure times) and GTFS quality. Compared to the hypotheses that were raised in our literature review, the evidence confirms that accessibility measures which do not consider variability are underestimating inequality conditions.

6. Conclusions

In this study we analyzed how employment accessibility estimates by public transport can be affected by two common data problems: the fact that scheduled GTFS feeds are based on inaccurate travel times that do not necessarily reflect real-time speeds of PT systems; and the fact that GTFS feeds do not adequately capture day-to-day variability in PT services. Fortaleza, as one of Brazil's largest cities, was selected as a case study. Regarding the inaccuracy data problem, our results show overall that scheduled GTFS slightly underestimate accessibility when compared to real-time GTFS with median travel times (GTFS P50). Nonetheless, we find much larger accessibility differences across neighborhoods and socioeconomic groups. For the case of Fortaleza, using scheduled GTFS tends to overestimate employment accessibility in certain areas (Northwest and Southeast regions) and underestimate in others (Southwest region) while there is very little difference in the Central region. Moreover, in our case study, we found that the inaccuracy of GTFS feeds has no significant effect on aggregated estimates of accessibility inequality.

Furthermore, when we analyze the impact of day-to-day travel time variability on accessibility by comparing Realtime P50 GTFS to Realtime P85 GTFS. We found that, on average, ignoring day-to-day travel time variability can lead to an overestimation of accessibility by up to 50%, with an observed range of 34–67%. We also found that this impact is unequally distributed both spatially and across income groups. These results show that peripheral locations – where most low-income population live – are consistently more affected by day-to-day service variability than the Central region or high-income neighborhoods. Inequality levels, represented by the Palma Ratio, show a 35% increase when day-to-day travel time variability is considered.

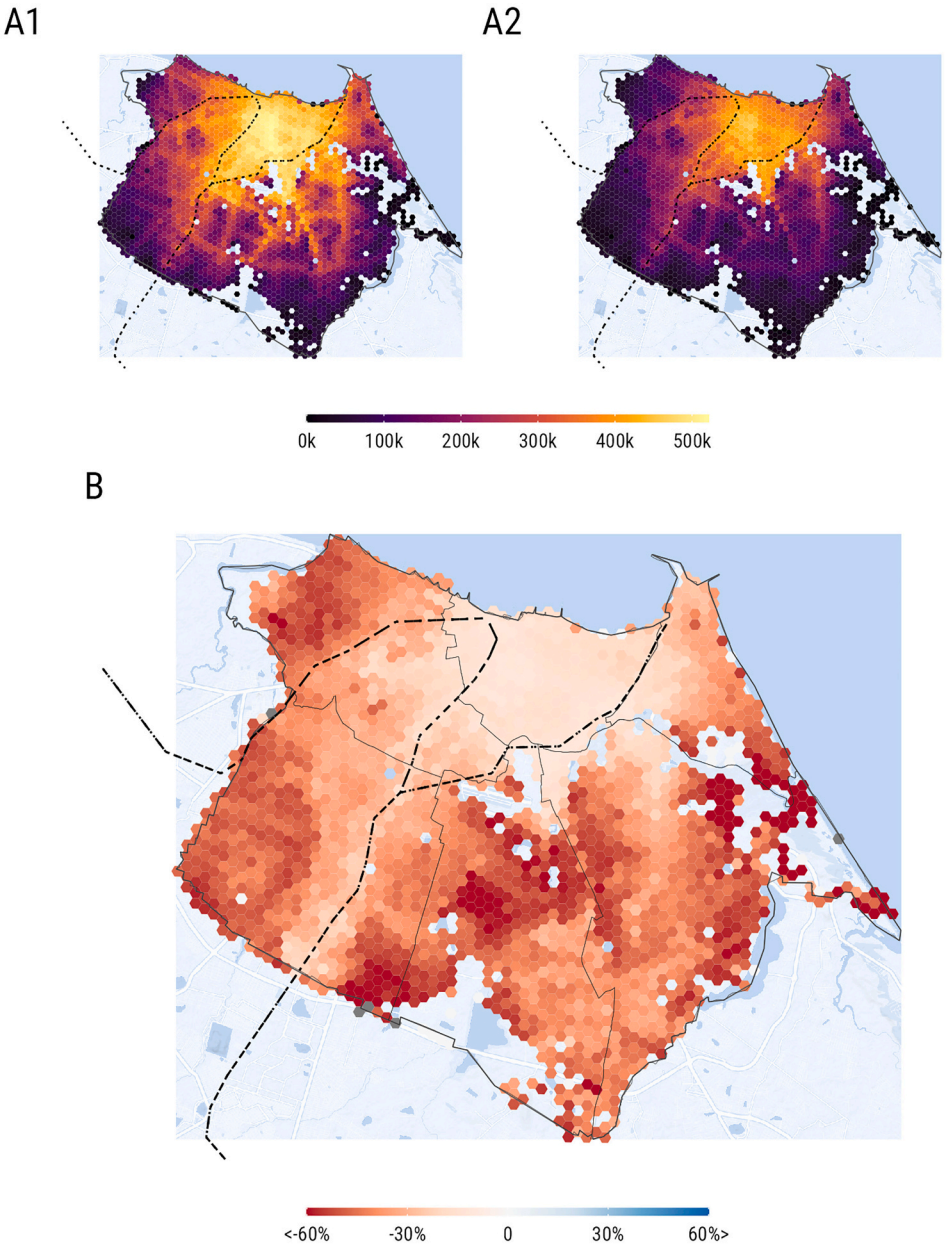


Fig. 5. Accessibility distribution calculated by Realtime P50 GTFS and the Realtime P85 GTFS (A1 and A2) and the relative difference between the accounts (B).

The problem of GTFS inaccuracy emerges for various reasons, including the reliability of the public transport system and the methodology each city uses to build its own scheduled timetables. In the case of Fortaleza, scheduled GTFS overestimates accessibility in some regions where bus services are more exposed to traffic congestion. This overestimation is somewhat lower than expected, which suggests the transport agency could be adopting a conservative approach to build its GTFS timetables. These factors raise questions about whether the findings of this case study could be valid to other cities given the issues commonly found in scheduled GTFS data. The strategies used by transport agencies to build their GTFS and the quality of their feeds play a crucial role in determining how reliable the results of accessibility analyses can be. As such, the magnitude of the impact of GTFS inaccuracy and variability issues should be individually analyzed for each city.

In the case of day-to-day variability impact, cities that show similar demographics, land use and travel patterns characteristics to Fortaleza - lower income population living on the outskirts, high concentration of jobs in the Central region, long and complex public transport commutes - should expect similar results. Dedicated infrastructure to public transport may play an important role in mitigating the impact of variability. In this aspect, fully segregated right-of-way (such as BRT and metro corridors) are more effective than bus lanes to guarantee more reliable services, since the latter are more likely to experience to delays due to traffic lights, vehicle interference, and other incidents.

Some limitations of this study should be highlighted. Employment data only refer to formal jobs, which may be problematic for Brazilian cities (Pinto et al., 2023). We used a single cumulative opportunities measure with one time threshold, and previous studies have shown how the time threshold choice can impact the results of accessibility analyses (Pereira, 2019). Further studies are necessary to analyze the extent to which the conclusions found in this paper hold when considering different accessibility indicators, impedance functions, as well as types of activities other than employment.

There are a few broad lessons that can be drawn from the findings in this paper. From a research perspective, the evidence found for this case study suggests that future work should whenever possible incorporate

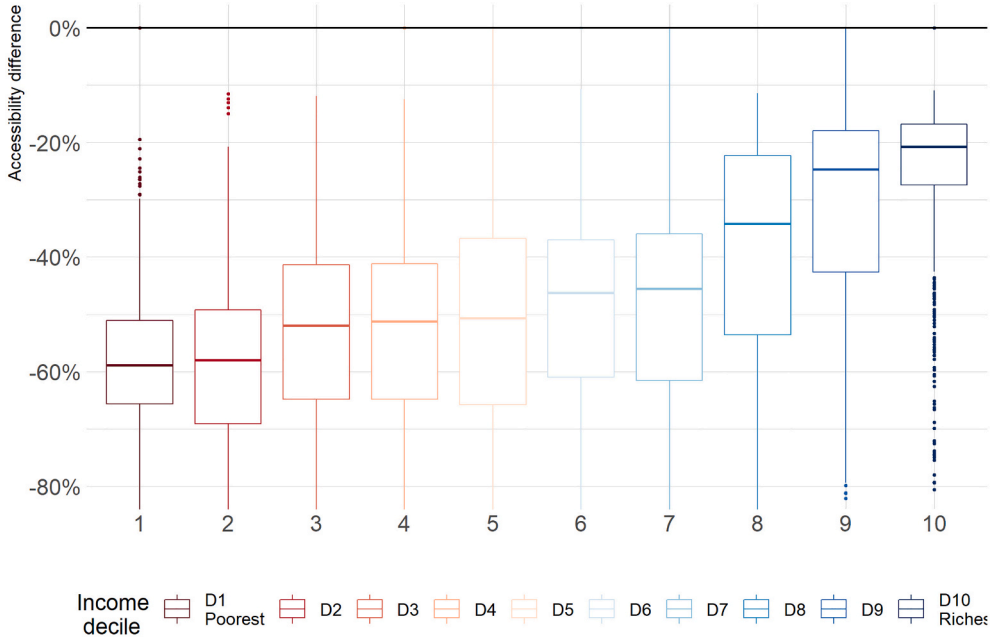


Fig. 6. Distribution of the relative difference between the Realtime P50 GTFS and the Realtime P85 GTFS by income deciles.

GPS data to generate more realistic accessibility estimates. Analysis solely based on scheduled GTFS are likely to generate spatially biased results, and particularly overestimate accessibility levels in areas commonly affected by congestion. Our findings also suggest that studies that ignore the day-to-day variability of public transport services are likely to overestimate accessibility levels and underestimate inequalities in access to opportunities, raising important concerns for transportation equity analyses. These biases are particularly important in the impact assessment of public transport projects, especially those that include mixed-traffic interventions, which are more subjected to delays and variability.

Declaration of Competing Interest

None.

Data availability

Data will be made available on request.

Acknowledgements

The authors would like to thank two anonymous referees for their very useful and constructive comments and suggestions. This work was supported by the National Council for Scientific and Technological Development (CNPq) Proc. 306922/2019-3 and the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) Proc. 88881.337362/2019-1.

References

Abkowitz, M., et al., 1987. Operational feasibility of timed transfer in transit systems. J. Transp. Eng. 113 (2), 168–177. https://doi.org/10.1061/(ASCE)0733-947X(1987)113:2(168).

Arbex, R., Cunha, C.B., 2020. Estimating the influence of crowding and travel time variability on accessibility to jobs in a large public transport network using smart card big data. J. Transp. Geogr. 85 (February), 102671 https://doi.org/10.1016/j.jtrangeo.2020.102671.

Bittencourt, T.A., Giannotti, M., Marques, E., 2020. Cumulative (and self-reinforcing) spatial inequalities: interactions between accessibility and segregation in four Brazilian metropolises. Environ. Plan. B: Urban Anal. City Sci. 0 (0), 1–17. https://doi.org/10.1177/2399808320958426.

Boisjoly, G., et al., 2020. Accessibility measurements in São Paulo, Rio de Janeiro, Curitiba and Recife, Brazil. J. Transp. Geogr. 82 (September 2019), 102551 https://doi.org/10.1016/j.jtrangeo.2019.102551.

Braga, C.K.V., Loureiro, C.F.G., Pereira, R.H.M., 2020. Analisando a variabilidade de estimativas de acessibilidade por transporte público a partir de dados de GPS. Transportes 28 (5), 169–184. https://doi.org/10.14295/transportes.v28i5.2175.

Braga, C.K.V., et al., 2022. TD 2767 - Impactos da expansão do metrô de Fortaleza sobre o acesso a oportunidades de emprego, saúde e educação. Texto para Discussão 2767, 1–50. https://doi.org/10.38116/td2767.

Chen, B.Y., et al., 2012. Reliable shortest path finding in stochastic networks with spatial correlated link travel times. Int. J. Geogr. Inf. Sci. 26 (2), 365–386. https://doi.org/10.1080/13658816.2011.598133.

Chen, B.Y., et al., 2013. Reliable space-time prisms under travel time uncertainty. Ann. Assoc. Am. Geogr. 103 (6), 1502–1521. https://doi.org/10.1080/00045608.2013.834236.

Chen, B.Y., et al., 2017. Measuring place-based accessibility under travel time uncertainty. Int. J. Geogr. Inf. Sci. 31 (4), 783–804. https://doi.org/10.1080/13658816.2016.1238919.

Chen, B.Y., et al., 2019. Understanding travel time uncertainty impacts on the equity of individual accessibility. Transp. Res. Part D: Transp. Environ. 75 (August), 156–169. https://doi.org/10.1016/j.trd.2019.08.027.

Chen, B.Y., et al., 2020. Evaluating spatial accessibility to healthcare services under travel time uncertainty: a reliability-based floating catchment area approach. J. Transp. Geogr. 87 (June), 102794 https://doi.org/10.1016/j.jtrangeo.2020.102794.

Chung, E., 2019. Public transport travel time variability definitions and monitoring. Grants Register 2020, 661–662. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000724.

Conway, M.W., Byrd, A., Van Eggermond, M., 2018. Accounting for uncertainty and variation in accessibility metrics for public transport sketch planning. J. Transpo. Land Use 11 (1), 541–558. https://doi.org/10.5198/jtlu.2018.1074.

El-geneidy, A.M., Building, M., Horning, J., 2011. Analyzing transit service reliability from automatic vehicular locator systems. J. Adv. Transp. 45 (2011), 66–79.

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.

Geurs, K., 2020. Accessibility and Transport Appraisal - Approaches and Limitations. International Transport Forum.

Geurs, K.T., van Wee, B., 2004. Accessibility evaluation of land-use and transport strategies: review and research directions. J. Transp. Geogr. 12 (2), 127–140. https://doi.org/10.1016/j.jtrangeo.2003.10.005.

Herszenhut, D., et al., 2022. The impact of transit monetary costs on transport inequality. J. Transp. Geogr. 99, 103309 https://doi.org/10.1016/j.jtrangeo.2022.103309.

Higgins, C.D., et al., 2022. Calculating place-based transit accessibility: methods, tools and algorithmic dependence. J. Transpo. Land Use 15 (1), 95–116. https://doi.org/10.5198/jtlu.2022.2012.

IBGE, 2021. Estimativas da População. Available at: https://www.ibge.gov.br/estatisticas/sociais/populacao/9103-estimativas-depopulacao.html?=&t=resultados.

Levinson, D., King, D., 2020. Transport Access Manual: A Guide For Measuring Connection Between People and Places. Available at: https://ses.library.usyd.edu.au/handle/2123/23733.

Liu, L., Porr, A., Miller, H.J., 2022. Realizable accessibility: evaluating the reliability of public transit accessibility using high-resolution real-time data. J. Geogr. Syst. https://doi.org/10.1007/s10109-022-00382-w.

Manaugh, K., Badami, M.G., El-Geneidy, A.M., 2015. Integrating social equity into urban transportation planning: a critical evaluation of equity objectives and measures in transportation plans in north america. Transp. Policy 37, 167–176. https://doi.org/10.1016/j.tranpol.2014.09.013.

Mandelzys, M., Hellinga, B., 2010. Identifying causes of performance issues in bus schedule adherence with automatic vehicle location and passenger count data. Transp. Res. Rec. J. Transp. Res. Board 2143 (1), 9–15. https://doi.org/10.3141/2143-02.

Mayaud, J.R., et al., 2019. Future access to essential services in a growing smart city: the case of Surrey, British Columbia. Comput. Environ. Urban. Syst. 73 (July), 1–15. https://doi.org/10.1016/j.compenvurbsys.2018.07.005.

Mazloumi, E., Currie, G., Rose, G., 2010. Using GPS data to gain insight into public transport travel time variability. J. Transp. Eng. 136 (7), 623–631. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000126.

Palm, M., Shalaby, A., Farber, S., 2020. Social equity and bus on-time performance in Canada's Largest City. Transp. Res. Rec. 2674 (11), 329–342. https://doi.org/10.1177/0361198120944923.

Papa, E., Bertolini, L., 2015. Accessibility and transit-oriented development in European metropolitan areas. J. Transp. Geogr. 47, 70–83. https://doi.org/10.1016/j.jtrangeo.2015.07.003.

Park, Y., et al., 2020. Assessing public transit performance using real-time data: spatiotemporal patterns of bus operation delays in Columbus, Ohio, USA. Int. J. Geogr. Inf. Sci. 34 (2), 367–392. https://doi.org/10.1080/13658816.2019.1608997.

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 (October 2018), 321–332. https://doi.org/10.1016/j.jtrangeo.2018.12.005.

Pereira, R.H.M., et al., 2019. 'Desigualdades socioespaciais de acesso a oportunidades nas cidades brasileiras, 2019', Texto para Discussão IPEA, 2535. Available at: http://www.ipea.gov.br/portal/images/stories/PDFs/TDs/td_2535.pdf.

Pereira, R.H.M., et al., 2021. r5r: rapid realistic routing on multimodal transport networks with R 5 in R. Findings 1–10. https://doi.org/10.32866/001c.21262.

Pinto, D.G.L., Loureiro, C.F.G., Sousa, F.F.L.D.M., Motte-Baumvol, B., 2023. The Effects of Informality on Socio-Spatial Inequalities in Accessibility to Job Opportunities. J. Transp. Geogr. 108, 103577. https://doi.org/10.1016/j.jtrangeo.2023.103577.

Pritchard, J.P., et al., 2019. An international comparison of equity in accessibility to jobs: London, São Paulo, and the Randstad. Transp. Find. 0–1. https://doi.org/10.32866/7412.

Stępniak, M., et al., 2019. The impact of temporal resolution on public transport accessibility measurement: review and case study in Poland. J. Transp. Geogr. 75 (January), 8–24. https://doi.org/10.1016/j.jtrangeo.2019.01.007.

UN-HABITAT, 2011. Cities for All: Bridging the Urban Divide – State of the World's Cities 2010/2011 by UN-HABITAT, First published by Earthscan in the UK and USA in 2008 for and on behalf of the United Nations Human Settlements Programme (UN-HABITAT). Available at: http://unhabitat.org/books/state-of-the-worlds-cities-20102011-cities-for-all-bridging-the-urban-divide/.

Wessel, N., Farber, S., 2019. On the accuracy of schedule-based GTFS for measuring accessibility. J. Transpo. Land Use 12 (1), 475–500. https://doi.org/10.5198/jtlu.2019.1502.

Wessel, N., Allen, J., Farber, S., 2017. Constructing a routable retrospective transit timetable from a real-time vehicle location feed and GTFS. J. Transp. Geogr. 62 (January), 92–97. https://doi.org/10.1016/j.jtrangeo.2017.04.012.