Journal of Internet Services and Applications, 2026, 17:1, doi: 10.5753/jisa.2026.6668

Analysis of the Quality of Service of Public Urban Buses Based on GPS Monitoring

Tiago Borzino Rocha [**Universidade Federal do Rio de Janeiro** | ***borzino@gta.ufrj.br***]
Fernando Dias de Mello Silva [**Universidade Federal do Rio de Janeiro** | ***fernandodias@gta.ufrj.br***]
Aline Carneiro Viana [**INRIA** | ***aline.viana@inria.fr***]
Luís Henrique M. K. Costa [**Universidade Federal do Rio de Janeiro** | ***luish@gta.ufrj.br***]

Centro de Tecnologia - 972, Av. Horácio Macedo, 2030 - Sala H-301 - Cidade Universitária da Universidade Federal do Rio de Janeiro, Rio de Janeiro - RJ, 21941-598, Brazil.

Received: 19 October 2025 • **Accepted:** 11 February 2026 • **Published:** 23 May 2026

Abstract

The introduction of GPS-equipped IoT devices onboard of urban buses allows the collection of data to monitor the fleet and assess the quality of service of this public transportation modality. The present work analyzes the GPS data from Rio de Janeiro buses, which includes executed trajectories, bus routes and identifiers to estimate per-route performance metrics with the collected data. In particular, the time between two buses, the number of bunched buses, the time spent on the inner fraction of the route, and the entropy are investigated as indicators of quality of service. The results reveal the unpredictability degree of each analyzed route. Lastly, we investigate the correlations between different metrics as a means of discovering relationships between metric performances. Our analysis shows that some routes exhibit greater regularity than others when individual metrics are considered. In addition, we observe a correlation between bus interval and bus bunching, with longer intervals reducing the likelihood of bunching. Also, we observed a correlation between entropy and time spent in the inner fraction of the route. Those results provide useful information in fleet planning activities for public transport.

Keywords: Internet of Things, intelligent transportation systems, smart cities

1 Introduction

Vehicle connectivity with Internet of Things (IoT) devices, alongside a centralized server that aggregates data regarding these vehicles, allows for a comprehensive view of every connected vehicle and an assessment of numerous aspects related to mobility in the urban environment. In connected cities, integrated sensors collect information from the public transport, enabling fleet monitoring. It is possible to use collected data to analyze the quality of service of different bus routes, offering insights that might assist in urban planning and the development of new bus itineraries. In this paper, we take the city of Rio de Janeiro as a use case: Rio's urban buses are equipped with sensors that gather information, including GPS (Global Positioning System) location, speed, date and time, bus identifier, and the route currently being served. This data is then transmitted to a server using the cellular network and made available by Rio de Janeiro's city hall through the DataRio website DATA.RIO [2025].

With data collected from such a smart fleet, it is possible to estimate different metrics to evaluate the quality of the service provided by bus companies – private in the case of Rio. Some metrics permit the analysis of different route features, like bus interval, bunching, traversal time and entropy. Those metrics are useful to determine which lines require attention or intervention, and can indicate whenever anomalous activities happened in the city traffic. Nonetheless, the device used to discover the bus position, like any GPS device, is susceptible to imprecision in the estimated coordinates. Therefore, some of the information that is transmitted may be located outside the street limits, even inside buildings or rivers, and lakes. Thus, a trajectory correction algorithm is necessary to mitigate this problem and eliminate incompatible data.

This paper[1] analyzes different metrics to evaluate the public transportation system's quality of service. This analysis consists of first implementing a bus trajectory correction algorithm based on the informed route, the different quality of service metrics computation, and metric comparison. Those metrics are evaluated over each trip for multiple days, and demonstrate the behavior of different lines in the system.

Our analysis demonstrates how different characteristics of each metric manifest for each line, based on the observation of individual metrics. Those observations can indicate how regular two distinct lines are. On the other hand, the correlation analysis only indicates a significant correlation between bus bunching and bus interval metrics, with a higher interval leading to less-likely bus bunching, and a correlation between entropy and time spent in the inner section of the trajectory. Other metric combinations seem uncorrelated.

We summarize the contributions of this paper as follows:

- We evaluate the quality of service of Rio de Janeiro's bus routes using different metrics, with each metric analyzing a different aspect of the service.
- We compare the metrics, aiming to find correlations between them. The results show that bus bunching and

[1] Part of this work is based on our paper published in Portuguese in the Proceedings of the IX Workshop de Computação Urbana (CoUrb) available at `https://sol.sbc.org.br/index.php/courb/article/view/35266`. This utilization is permitted by the Brazilian publisher, as seen in `https://sol.sbc.org.br/index.php/indice/conduta`.

Author	Data Type	Data Source	Metrics
Li *et al*. [2022]	Speed and position	Loop Detectors	Speed, position
Nguyen *et al*. [2023]	GPS	Bus	Travel time, bus bunching
Du and Dublanche [2018]	Smart card	People	Bus bunching
Song *et al*. [2010]	CDR	People	Entropy
He [2015]	Arrival times	Bus	Bus bunching
Wang *et al*. [2021]	Crowding information	Simulated	Bus bunching
Zhou *et al*. [2022]	Crowding information	Bus	Bus bunching
Teixeira *et al*. [2019]	GPS and CDR	People	Entropy
Cuttone *et al*. [2018]	Position and timestamp	People	Predictability
Huang *et al*. [2024]	Check-in data	People	Entropy
Teixeira *et al*. [2021]	GPS and CDR	People	Entropy
Ikanovic and Mollgaard [2017]	GPS	People	Entropy
Zhao *et al*. [2016]	Taxi trips	Taxi	Entropy
Zheng *et al*. [2022]	Survey	People	Speed, reliability, etc
Hosseini *et al*. [2025]	Survey	People	Commute time, wait time, etc
dos Santos and Lima [2021]	Questionnaire	People	Travel time, frequency, etc
Smith *et al*. [2014]	GPS	People	Entropy
Lu *et al*. [2013]	CDR	People	Entropy
Ours	GPS	Bus	Bus interval, bus bunching, entropy, time spent in the inner section

Table 1. Summary of data types, sources, and metrics used in related studies.

bus interval, and entropy and time spent in the inner section of the trajectory have a significant correlation.

This paper is organized as follows. Section 2 reviews related works on using entropy to evaluate regularity and bus quality of service. Section 3 describes the data used and the data processing made. Section 4 details how the different metrics used are calculated. Section 5 describes the choices for the period analyzed and the cell size selection. Section 6 presents and discusses the obtained results. Finally, Section 7 concludes this work and discusses future research directions.

2 Related Work

Urban mobility is a concern for medium to large cities worldwide. Various works in the literature consider the use of technology to monitor the transportation systems. Li *et al*. [2022] verified how traffic varies during the day on a road in the Netherlands, trying to predict how foreseeable the spatial position and vehicle speed are along the road, using data from loop detectors that record average speed and vehicle flow. Nguyen *et al*. [2023] studied the Los Angeles bus system, examining different aspects related to the quality of service using GPS data originated in the buses provided by LA Metro, such as bus bunching and regularity in the time a transport arrives at some of the stops in its itinerary.

Other works in the literature further analyze the occurrence of bus bunching as a surrogate metric of the quality of service. Du and Dublanche [2018] researched the influence of the weekday on bus bunching. Other studies focused on proposing new strategies to handle bunching, as in He [2015] and Wang *et al*. [2021]. Bus bunching is a problem since buses depart at regular intervals, and thus, if two or more buses end up grouped alongside one another, this is an indication that some unexpected situation occurred and directly affected the proper operation of the route. Finally, Zhou *et al*. [2022] verified the effect that making available information regarding how crowded a bus is affected the behavior of the possible passengers, whether they would choose to wait for the next one or not.

In order to estimate the amount of uncertainty in someone's movement pattern, different works explored the usage of entropy. Song *et al*. [2010] uses entropy as a method to estimate the amount of uncertainty which exists in the routine movement of a person. The introduction of contextual information (such as weekday and weather) and the exclusion of impossible next locations (i.e., locations not reachable within one sampling interval) have been explored as approaches to reduce uncertainty by Teixeira *et al*. [2019]; Cuttone *et al*. [2018]; Huang *et al*. [2024]. Separating one's locomotion into routine (places that were previously visited) and novelty (first-time visits) was also tried in order to assess their influence on the entropy (Teixeira *et al*. [2021]) and, thus, predictability. Ikanovic and Mollgaard [2017] and Smith *et al*. [2014] varied the time and spatial resolutions with the intention of evaluating their influence on entropy and its impact on the uncertainty of a person's position. Lu *et al*. [2013] investigated the limit of predictability in human mobility with Markov Chains and observed a correlation between higher entropy and higher uncertainty. Moreover, entropy was further utilized to predict the taxi demand in a city in China Zhao *et al*. [2016].

Aspects of the service that affect the user's sense of quality of service were also studied by Zheng *et al*. [2022]; Hosseini *et al*. [2025]; dos Santos and Lima [2021], revealing how the different factors played a role in how the bus public transport system is perceived. Characteristics such as reliability in wait time and commute time were frequently associated with good public transport.

Different from previous work that uses human data, being both Call Detail Records (CDR) and GPS, we use data from vehicles, more specifically, public buses. We investigate the

use of entropy as a quality of service metric for metropolitan bus service. Moreover, we compare traditional quality of service metrics, aiming at discovering whether the variation of a metric has any reflection in another metric. Our goal is to investigate how different metrics obtained from bus movement traces can effectively model the quality of the public service. A comparison between our work and the literature is summarized in Table 1, highlighting the different data types, sources, and metrics used.

3 Bus Monitoring and Data Processing

In this section, we present the main characteristics of the dataset employed and the processing methodology needed to prepare the data and use it to produce the different quality of service metrics.

3.1 Datasets

The information used in this work is extracted from two distinct datasets, both obtained from the DataRio website DATA.RIO [2025]. The first contains raw GPS data, and the second has information about the different trips made.

The first dataset contains GPS records, with each entry comprising information about the position of the vehicle (latitude and longitude), hour and day, route, and bus identifier. In this work, we consider the period ranging from December 10, 2022, to January 10, 2023. An example of some of the data in this dataset is exemplified in Table 2. Each entry consists of a timestamp, the bus identifier, the route it is serving, and the GPS position.

The second dataset consists of records of complete trips. Each trip is defined by a bus identifier, departure and arrival time, route, and direction. A fragment of this dataset is in Table 3. Each entry consists of the bus identifier, the route it serves, the direction ("I" or "V", going or coming back on a round trip), departure from origin and arrival at destination timestamps.

Table 2. Samples from the GPS records dataset.

Timestamp	Vehicle ID	Route	Latitude	Longitude
2022-12-16 08:39:18	A48176	415	-22.92401	-43.22901
2022-12-16 08:39:19	A48109	415	-22.93160	-43.23905
2022-12-16 08:39:22	A48071	415	-22.92086	-43.21778
2022-12-16 08:39:24	A48008	415	-22.93816	-43.24676
2022-12-16 08:39:27	A48040	415	-22.90577	-43.19274
2022-12-16 08:39:33	A48008	415	-22.93819	-43.24678
2022-12-16 08:39:36	A48071	415	-22.92060	-43.21702
2022-12-16 08:39:41	A48007	415	-22.90437	-43.18863

3.2 Trajectory Selection

In order to compute the metrics, each trip of each route is considered independently. For this purpose, the information from both datasets, trips and GPS records, is combined. Additionally, records in which the bus is parked and still transmitting or sending incorrect data are eliminated.

Figure 1 shows a flowchart of the process whereas Figure 2 shows how the union is performed. The trip corresponding

Table 3. Samples from the trips dataset.

Vehicle ID	Route	Direction	Departure timestamp	Arrival timestamp
A27515	433	I	2022-12-15 11:41:01	2022-12-15 12:59:01
A27528	548	V	2022-12-15 06:28:00	2022-12-15 07:26:30
D87394	870	V	2022-12-15 22:33:04	2022-12-15 23:08:34
D87397	771	I	2022-12-15 07:54:35	2022-12-15 09:09:28
A29173	435	I	2022-12-15 20:16:52	2022-12-15 21:31:22

Figure 1. Flowchart of the trajectory selection.

to the first entry in the trips table is chosen. Then, the GPS records table is searched, given the constraints of the trip regarding vehicle identifier, route, and timestamp. Only the records that match the criteria are selected and then grouped to form the corresponding trip. If a record from the GPS records table does not match any trip defined in the complete trips table, the GPS record is discarded.

3.3 Trajectory Correction

The positions collected by the GPS device carried in the buses can be imprecise due to localization errors, transmission failures, and device configuration mistakes. These uncertainties can lead to a bus informing a position corresponding to other streets and, in worst cases, inside buildings or rivers. For this reason, it is necessary to pre-process the data sent by the buses, according to the process presented in Figure 3.

Given that the trajectory of each bus route is known, it is possible to apply a coordinate correction algorithm based on the line information. First, the information sent by the vehicles is separated according to the line and, following this, the process described in Section 3.2 is used to identify the data that belongs to each trip.

Subsequently, each trip is corrected using a script that automates the process. To achieve this, each coordinate of a trip is compared to the ideal trajectory of the line, available at DataRio (DATA.RIO [2025]), and then the value of the real position will be substituted by the coordinates of the ideal position that had the smallest distance to the real one. If the distance to every pair of latitude and longitude of the ideal trajectory is greater than a threshold of 500 meters, the real record that originated this distance will be discarded. This ensures that the analyzed trips will not only stay within the street limits, but also in the correct direction. We used the value of 500 meters since it comprehends a larger area than the spacing provided by the ideal trajectory, accepting small deviations to the predetermined route while still being able to discard any data that deviated from the expected course.

4 Quality of Service Metrics

In this section, we discuss the metrics used to evaluate the quality of service of the bus public transportation service. The

Trips				
vehicle_id	route	direction	departure_time	arrival_time
A	123	I	12/12/2022 13:30:39	12/12/2022 14:27:45
B	456	I	12/12/2022 14:50:57	12/12/2022 16:10:11
C	789	I	12/12/2022 15:25:00	12/12/2022 16:00:00

GPS records				
vehicle_id	route	timestamp	latitude	longitude
A	123	12/12/2022 13:57:51	-22.89843	-43.29701
B	456	12/12/2022 15:30:11	-22.88424	-43.29754
A	123	12/12/2022 13:49:45	-22.89961	-43.29739
A	123	12/12/2022 14:12:43	-22.90104	-43.29783
A	123	12/12/2022 16:27:45	-22.90196	-43.29816
C	789	12/12/2022 15:45:00	-22.85807	-43.38753

Result				
vehicle_id	route	timestamp	latitude	longitude
A	123	12/12/2022 13:57:51	-22.89843	-43.29701
A	123	12/12/2022 13:49:45	-22.89961	-43.29739
A	123	12/12/2022 14:12:43	-22.90104	-43.29783

Figure 2. Example of combination of datasets.

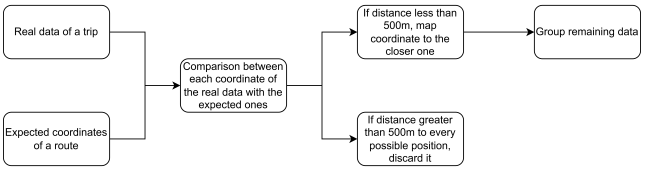

Figure 3. Flowchart of the trajectory correction.

main metrics considered are bus interval, bus bunching, time spent in the inner section of a route and the entropy. Before any metric is calculated, the algorithm in 3.3 is used to correct the bus' trajectories. A flowchart of the data can be seen in Figure 4.

4.1 Bus Interval

The bus interval is calculated by measuring the time difference between the arrivals of two buses from the same route at the same location. To the passengers, it measures the amount of time they will likely have to wait if they have just missed the bus passing through a particular stop.

In order to compute the bus interval, a point along the line trajectory is chosen as a reference and a radius of 500 meters is defined around it. If the distance of the bus to the chosen location is smaller than the radius, it will be considered that the bus passed in the delimited area of reference. To check whether a bus passed through the chosen location, the coordinates of each record of the trips are compared to the reference location. If the distance of any record to the coordinates is smaller than the threshold, the timestamp at which it occurred for the first time is recorded. With the aim of avoiding duplicate detections of the same bus, for each vehicle, the timestamp of the last detection is stored and a new one is only considered if the time difference between the last detection and the current detection is at least 20 minutes. We chose this value since Rio de Janeiro has a great number of routes, with varied lengths, and 20 minutes represented a balance in which both smaller and greater routes could have the same bus traveling the same direction but in another trip, since choosing a value individually for each line is impracticable. Repeating this process for every trip, and organizing the detection temporally, the bus interval among vehicles that operate in the same line is computed. To avoid the influence of regular departing times, we use as the reference location one of the last possible positions the bus can send, namely the penultimate position of a route. This choice also allows the usage of the same parameter in the different analyzed routes instead of selecting a specific position for each line.

4.2 Bus Bunching

The distance among buses that travel the same line in the same direction, or bus bunching, is another metric that can be used to evaluate the quality of the service offered (Nguyen *et al*. [2023]). A higher number of buses that travel in groups is usually a symptom of irregularity in the service.

To compute the number of buses close to one another, the distances between pairs of buses that are traveling on the same route and in the same direction are computed. To account for different sampling intervals between buses, the smallest distance between buses in a time window of 180 seconds is kept. In case this distance is lower than a threshold of 200 meters at any moment of the time window, both buses are classified as bunched and part of the same group. Some buses can stay still in their first and last stops, because of that, the data sent in these regions is discarded. We chose a time window of 3 minutes and 200 meters for the distance, since those are values that would allow a user to perceive that two or more buses are too close to each other because they would see more than a single bus in a narrow time window.

As seen in Figure 5, two buses can be part of the same bunching group even if they are more distant than the threshold. Although buses B and D are not close to each other, they belong to the same group because both are grouped with C. Bus A is distant from buses B, C and D, and therefore, it is not part of any group. If two or more groups have buses in common, those groups are merged and the total number of buses for a group increases. As there can exist more than one group at the same time with different sizes, the value of bus bunching will be the sum of the sizes of each group of buses.

4.3 Time Spent in the Inner Route Section

The buses have schedules they are supposed to comply with, with a well-defined arrival time at the last stop. The drivers have access to this information and, to comply with the given time, they often try to compensate for the delay at the end of the route. To isolate the effect of this compensation on the performance of the travel, we define a metric which is the time spent in the inner part of the trajectory. As this measure only evaluates the intermediary part of the trip, it

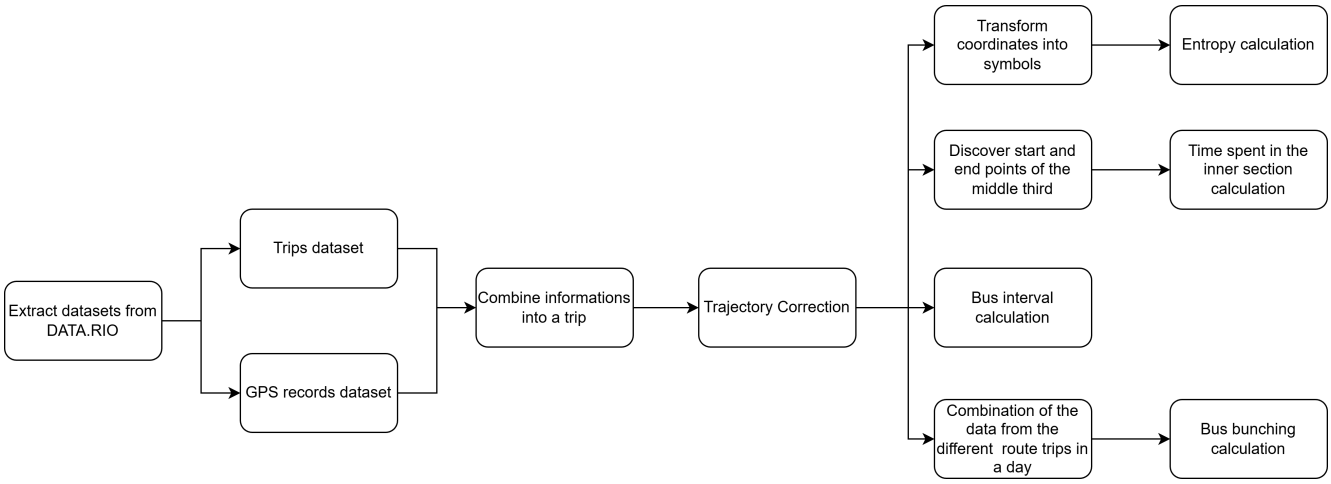

Figure 4. Flowchart of the metrics.

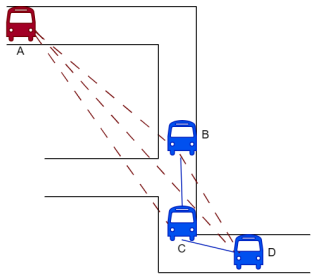

Figure 5. Bus bunching example.

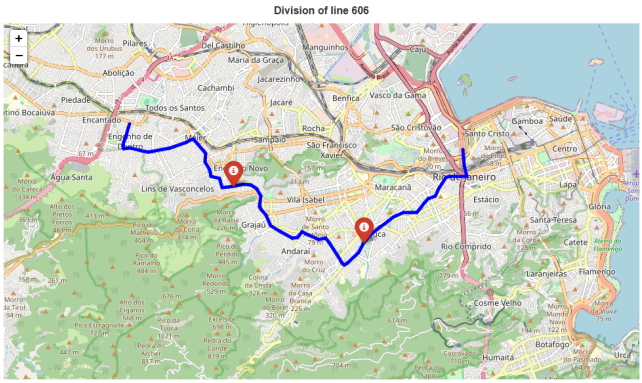

Figure 6. Example of the division of one bus route in to three parts, showing the inner section of the trajectory.

tries to capture the influence of the traffic conditions on the trip.

Since the trajectory of a route is known, it is possible to divide it into equally spaced parts. One of the parts is then chosen to compute the time a bus takes to travel its length. In this work, we divided the trajectory into three equal parts and analyzed the middle one (from 1/3 to 2/3 of the trajectory length). This division is represented by the two red markers in Figure 6. To discover when a bus started traveling the chosen segment, the positions that correspond to the start and end of the segment are recorded. Then, each position of a trip is compared to the ones at the start and end of the segment. Since a position can be skipped, either because of speed and lost data or another factor, some positions ahead of the ones that delimit the part are also considered if they are needed. The metric is then calculated by subtracting the first known timestamp inside this trajectory interval from the last timestamp inside the interval.

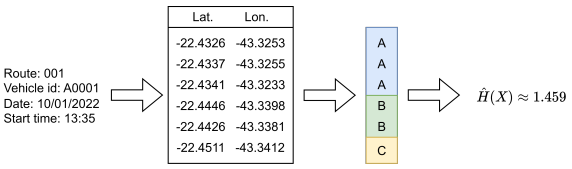

Figure 7. Entropy calculation example.

4.4 Entropy

The entropy H of a random variable X can be interpreted as a measure of how much uncertainty exists in this random variable. According to Shannon [1948], the entropy can be calculated as:

$$H(X) = -E[\ln(p(x))] = -\sum_{x} p(x)\ln(p(x)), \quad (1)$$

where $p(x)$ is the probability of each output of the random variable X.

In the context of urban buses, the random variable represents the set of possible locations a bus can take while traveling its trajectory. To estimate the entropy, the sequence of positions that comprise a trip is transformed into symbols. Following this, the entropy of a trip is estimated using Equation 1.

Since the entropy will be computed over the buses' latitude and longitude pairs, which are two different floating-point values, it is necessary to make a conversion of the two-dimensional coordinates into a one-dimensional symbol from a finite set that is able to represent the position of the vehicle. The process is illustrated in Figures 7 and 8. The former illustrates an overview of the whole process, from the trip selection, to the conversion and finally to the entropy calculation, the latter focuses on the conversion itself, showing how different coordinates originate different symbols: We multiply both latitude and longitude by a multiplication factor γ and, after that, the resulting values are truncated, and then the modified values of latitude and longitude are concatenated. The resulting single value is the symbol. Thus, each trip is represented by a sequence of symbols that represent the positions the bus sent while traveling. Finally, symbol sequences representing each trip are used for entropy computation. The multiplication factor γ, along with the spacing in the positions used in the trajectory correction, are used to define the area

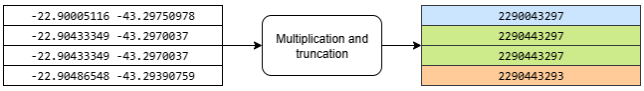

Figure 8. Transformation of coordinates into symbols.

one symbol represents, producing a tesselation of the map. Since the original position is replaced by the ideal position that is spaced by Ω meters, there will be a position every Ω meters. Subsequently, each position is multiplied by γ. The tesselation occurs as a result of the number of decimal digits in the GPS coordinates that are used in symbol formation. A larger number of decimal digits used designates a narrower area, and thus more symbols, due to increased precision in latitude and longitude values. In contrast, fewer digits encompass a bigger area, leading to less certainty in the real original position. The chosen cell size has an impact on symbol usability. Smaller cell sizes add a numerous amount of symbols per route, and buses are susceptible to skipping symbols. This would result in a symbol distribution between two different alphabets, which is not a fair comparison. Larger cell sizes do not capture more nuanced traffic conditions due to the many possible events happening in the covered area.

Figure 8 shows an example of the conversion, where the colors highlight the different symbols produced. The different positions lead to different symbols, whilst the same or close positions result in the same symbol.

5 Methodology

In this section, the period chosen for the analysis is defined with its time and day constraints, with the motives behind the restrictions being explained in Section 5.1. After that, Section 5.2 presents the cell size choice.

5.1 Analyzed Period

The period chosen to analyze the bus routes starts on December 10th, 2022 ending January 10th, 2023, adding up to 32 days. Due to different operational schedules defined for outside working hours and weekends, we only consider business days and working hours, since some routes in Rio de Janeiro do not operate, have their trajectories modified, or reduced fleet size outside this interval. Another motivation is the different traffic patterns observed between weekends and weekdays, which are more intensive for the latter. Therefore, the observed time is set between 6:00 am and 10:00 pm.

The routes analyzed are selected based on the number of trips made during the considered period. The threshold was set at a minimum of 340 trips during the observed interval. A graphical representation of the choice is shown in Figure 9, where the threshold is represented by the change of color of the samples in the graphic. The blue color represents the routes used in the analysis, whilst the red dots represent the lines that had less than 340 trips and were therefore discarded.

5.2 Cell Size

The cell size, as explained in Section 4.4, depends on the multiplication factor chosen to convert the GPS coordinates into symbols. To choose this multiplication factor, an experiment was conducted to verify the frequency of the symbols present in different trips of the same route. After adding up the total number of occurrences of each symbol in the other trips, we divide the sum of the occurrences of each symbol by the total number of trips analyzed. With that, the mean number of occurrences of each symbol per trip was obtained.

The choice of the cell size is made by verifying which size is the most adequate for the majority of the routes. To determine the optimal size, various values are tested, and the selected one is the one that displays a balance between symbols that appear in every trip, cells that do not exhibit a high number of repetitions, and a few cells that do not appear in every trip.

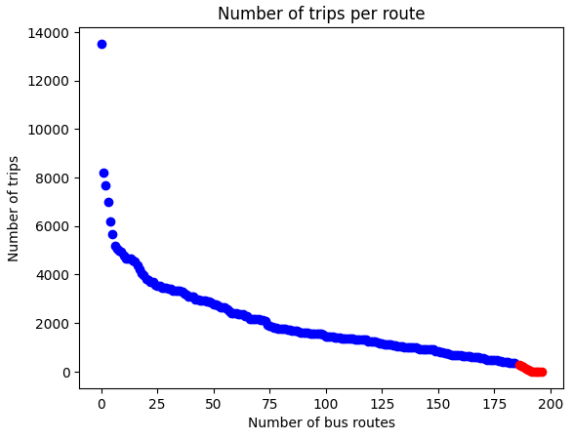

Figure 9. Threshold of choice for analysis. Blue dots represent the chosen and red dots represent the discarded.

Table 4. Parameter list

Metric	Name	Value	Unit
Bus Interval	Radius size	500	m
	Location	Penultimate position	-
Bus Bunching	Radius size	200	m
	Time window	180	s
Time to travel inner section	Location	Position of the start of the middle third	-
	Location	Position of the end of the middle third	-
Entropy	Cell size	250	m

6 Results

This section presents the results obtained for the different metrics over the bus position dataset. Section 6.1 shows the results used for cell size estimation and the choice for map tessellation. Section 6.2 presents the individual results of quality of service metric. Finally, Section 6.3 investigates the correlation between metrics.

Table 4 summarizes the relevant parameters defined for each metric. To compute the bus interval we use the penultimate position as the reference position with a radius of 500 meters around it. For bus bunching, we use a time window of

180 seconds and consider buses grouped if they are within 200 meters from one another. The time interval to travel the inner section of the trajectory uses as reference positions the one corresponding to the start and end points of the middle third (point that marks 1/3 for the start and 2/3 for the end). Finally, for the entropy we consider a cell size of 250 meter radius. The motivation for the parameter choices are explained in their respective sections, with the exception of the cell size, which is explained in Section 6.1.

6.1 Cell Size Selection

Table 5 displays the cell sizes (Ω) tested and the average entropy value for the analyzed trips using Ω as the cell size, where the third column contains the mean number of symbols that appear at least once in a trip, the fourth column represents the average number of GPS coordinates sent by buses in each trip made, the fifth column represents the mean of the average number of times each symbol appeared in the trips and the last column represents the mean of the number of empty cells per trip. We verify that as the cell size reduces, the mean number of unique symbols per trip grows, as well as the entropy value. This happens because the number of repetitions of a symbol decreases as the symbol corresponds to a smaller area. For a cell size of 100 meters, there are many cells that are not used in every trip. For 333 meters, tested since it represents a bus traveling with a uniform speed of 20 km/h, most cells are used more than once every trip, since most of the buses of the routes have a sampling interval between 30 and 60 seconds. Since with cell sizes bigger than 333 meters or smaller than 100 meters there are many cells used more than once or skipped multiple times, γ was set to 1,000, as a smaller value would produce an area of a 100 square meters or less and a bigger would represent an area larger than 1,000 square meters. Thus, the chosen value is 250 meters, because it provides the best balance between the number of cells that appeared on every trip and only a few cells with a single occurrence or multiple cells with many occurrences.

6.2 Individual Metrics

In the figures of this section, the box plots are ordered according to the number of trips considered in each route (identified in the X axis). Leftmost routes are the ones with largest trip numbers, as seen in Table 6, which also shows their respective length. After filtering, there are results for 185 routes in total. We choose to display the routes that are representative of the observed behavior of the rest of the routes.

Figure 10 shows the result of the bus interval metric. Some routes show a small median and interquartile range (IQR) for the bus interval, which coincides with routes that have a greater number of trips in the analyzed period, from routes 864 up to 415. The others show more irregularity in the bus interval and, at the same time, have fewer trips in the period analyzed, as routes 435 and 238.

Figure 11 shows the results of the number of bunched buses. It reveals that the routes that had more trips in the 32 days presented a higher number of bunched buses than the ones that had fewer trips. An increase in this metric suggests that this

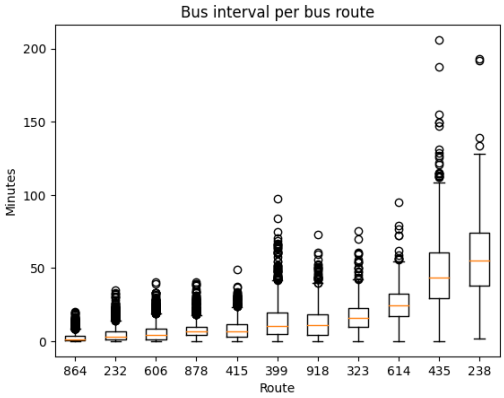

Figure 10. Box plot of the bus interval.

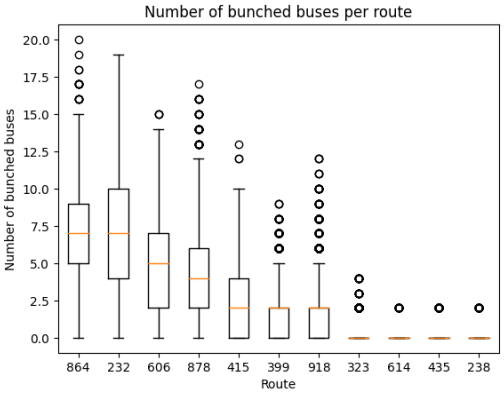

Figure 11. Box plot of bus bunching.

route could have a reduced fleet size to avoid this unnecessary grouping. Nonetheless, this decision must consider vehicle occupancy. On the other hand, the number of active-service buses directly influences the number of bunched vehicles. Given two routes of the same length but different number of active buses, the one that has a higher number of circulating vehicles will be more prone to bus bunching.

Figure 12 shows the result of the time interval to travel the inner part of the trajectory. The variance in the same route reflects the dynamic traffic conditions throughout the day, whilst the variation in the median values between the routes is due to the different speed limits on the trajectory roads, their length, and the geographical characteristics of the places through which the route passes, such as hills and turnings. The routes that show more regularity are 864 and 614, with most of the other lines presenting a variation of about five minutes.

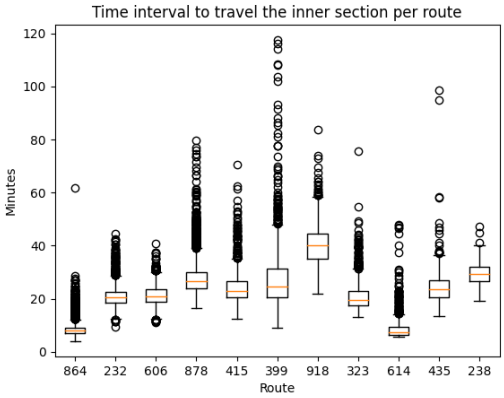

Figure 12. Box plot with the time interval to travel the inner section of the trajectory.

Table 5. Example of cell sizes tested for route 871.

Ω	Mean entropy	Mean number of unique symbols	Total number of symbols	Mean of average symbol frequency	Mean number of empty cells
100	6.214	76.214	81.0	0.473	14
200	5.936	65.357	81.0	0.946	7
250	5.715	57.643	81.0	1.179	6
300	5.520	50.214	81.0	1.427	4
333	5.380	45.571	81.0	1.595	4

Table 6. Number of trips per bus route presented and their length.

Bus route	864	232	606	878	415	399	918	323	614	435	238
# of trips	13514	6195	5190	4537	3970	2399	2353	1403	1084	630	354
Length (km)	10.801	16.106	14.553	19.863	25.798	33.943	36.925	27.318	21.102	23.854	20.062

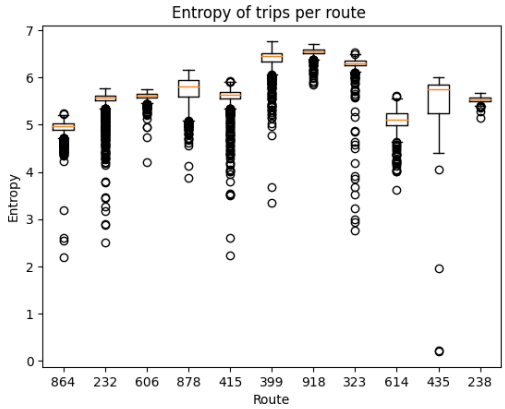

Figure 13. Box plot of the entropy.

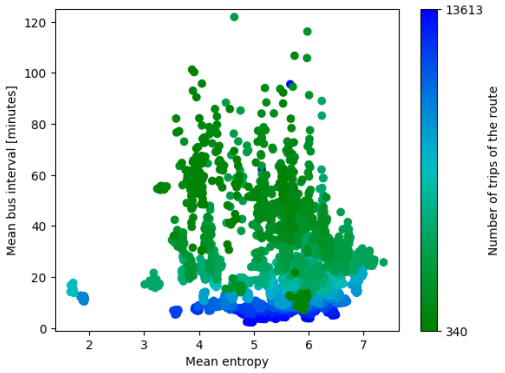

Figure 14. Comparison between mean entropy and mean bus interval.

Despite that, some of the routes have a higher variation, like lines 399 and 918, which have variations of about ten minutes between the time two vehicles take to travel the inner section of three equally spaced parts, indicating those routes have a higher irregularity in the amount of time their trips consume. Since routes 399 and 918 have a higher irregularity in the time taken, the users might opt to choose a more regular line, such as 864 and 232.

Figure 13 shows the result of entropy for the routes. A higher number of trips does not necessarily reflect a higher entropy variance, as routes 864 and 232 have a greater number of trips and less variation in their entropy than routes 878 and 435. Moreover, we observe that most lines followed the behavior of route 864, characterized by a small variance, indicating that the trip conditions of the same route are similar, as greater variations would reflect on the entropy value. Even though most measurements present a small IQR, there is still a significant number of outliers. This suggests that something anomalous occurred in those trips, such as partially complete trajectories, invalid GPS measurements, or even some drivers that drive the buses exceedingly faster or extremely slower than average.

Route 864, the shortest among those represented, exhibits the lowest entropy, whereas lines 399 and 918, the longest, show the highest. However, despite route 614 being longer than, for instance, lines 606 and 232, its entropy remains lower. This indicates that while route length influences entropy, travel time also contributes to the entropy value (as illustrated in Figure 12, where the inner section of route 614 is traversed more quickly than those of lines 232 and 606).

6.3 Metric Correlation Analysis

This section presents the results of the comparisons between the individual metrics. Different from Section 6.2, the plots consider all of the 185 bus routes obtained after the initial filtering. In the following graphs, each sample corresponds to the mean of one of the metrics of a route over one day, excluding one figure, in which the difference is explained. The color of each sample represents the number of trips the route made during the analyzed period, where dark blue is a higher number of trips and dark green represents the lines with fewer trips.

The first correlation investigated is between entropy and bus interval, shown in Figure 14. Both metrics show a great variation in their mean values. Most routes have mean bus intervals lower than 40 minutes, the vast majority, below 80 minutes. The mean daily entropy varies between ≈4 and 7 for the same bus interval mean, showing no direct relation between the two metrics ($\rho = -0.12$ and $p \ll 0.01$). Moreover, routes with a larger number of trips tend to have shorter intervals, as indicated by the blue color representing the routes with the highest trip counts. We also compare the mean of the entropy to the standard deviation of the bus interval, expecting that a higher deviation, and thus uncertainty, would lead to higher entropy, however, we found that the metrics are not related ($\rho = -0.09$ e $p \ll 0.01$).

The relation between entropy and the formation of bus bunching is shown in Figure 15. It reveals that even though the mean bus bunch sizes can vary from 0 to 12 buses, they all would share a mean entropy between 5 and 6.5, not exhibiting a clear relation ($\rho = 0.13$ and $p \ll 0.01$). A similar behavior is observed with the standard deviation of bus bunching and mean entropy ($\rho = 0.22$ and $p \ll 0.01$). Even though one

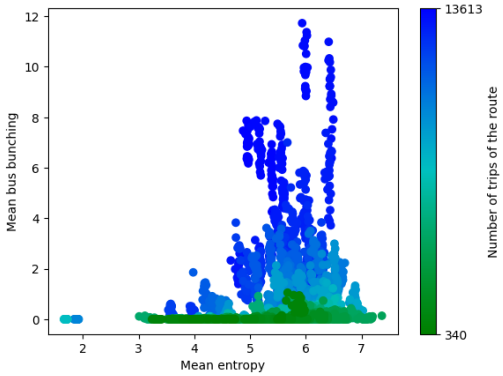

Figure 15. Comparison between mean entropy and mean bus bunching.

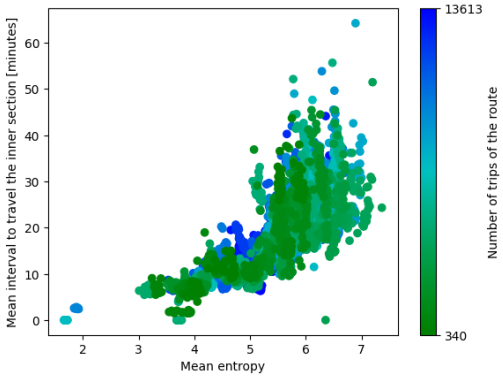

Figure 16. Comparison between mean entropy and mean time interval to travel the inner section of the trajectory.

could expect a correlation between large bus bunches and trajectory irregularity, we observe no joint variation in the entropy with the higher values of bunching, both between the metric means and between the mean of the entropy and bus bunching standard deviation.

Figure 16 compares the mean entropy with the mean time to travel the inner section of the trajectory. We expected to find a correlation between these two metrics, since days that have a higher entropy were expected to lead to an increase in the time taken to complete a trip, thus also a fraction of it. In general, this is the observed behavior, with higher mean entropy values leading to a longer time taken to travel the inner section of the trajectory, exhibiting a Pearson correlation of 0.76, with a negligible p value ($p \ll 0.01$). Moreover, the mean entropy and the standard deviation of the time to travel the inner part of the trajectory also exhibit a relation ($\rho = 0.43$ and $p \ll 0.01$), indicating that a higher uncertainty in the time to travel the inner section of the trajectory would also correspond to days on which the entropy has a higher mean value.

To further dig into the correlation between entropy and time to travel the inner section of the trajectory, Figure 17 presents a different plot where each dot represents the metrics of a single trip, different from the other figures where each represents the mean over a day. In this case, we find a correlation of 0.61, with a negligible p value, further indicating that both metrics are correlated, whether referring to the raw metrics or their means.

Figure 18 implies a correlation between the mean of bus bunching and the mean of bus interval, with a Pearson correlation of -0.39 and a negligible p value. We believe this is due

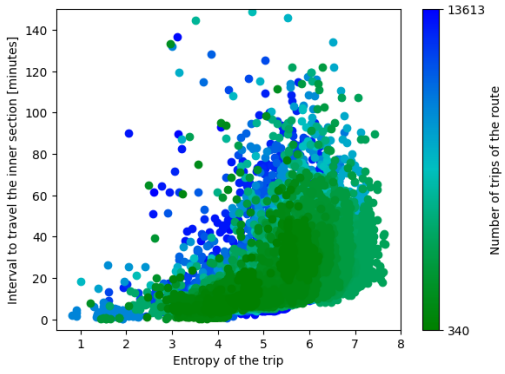

Figure 17. Comparison between entropy per trip and time interval to travel the inner section of the trajectory per trip.

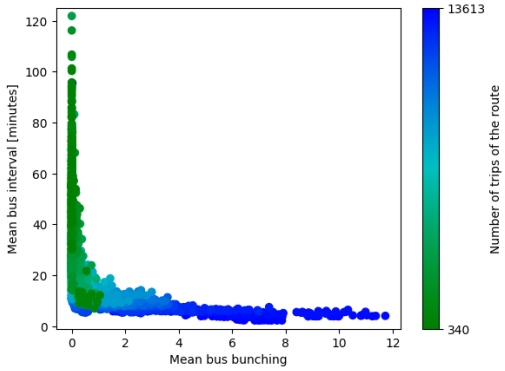

Figure 18. Comparison between mean bus bunching and mean bus interval.

to the fact that a lower bus interval of a route leads to closer buses and, the smaller the distance between the vehicles, the more likely they will be grouped and traveling together. However, the standard deviation of both metrics did not reveal a correlation between them ($\rho = -0.34$ and $p \ll 0.01$).

Figure 19 shows the comparison between the means of bus interval and time interval to travel part of the trajectory. We expected to find a relation between these two metrics, since a bus that takes longer than average to complete its trajectory is more likely to take longer to get to the next location. Nonetheless, we observe that the metrics varied independently ($\rho = -0.1$ and $p \ll 0.01$), possibly because when one bus takes longer than average to arrive at a location it will be closer to the one that started the trip after it, but it will also be further away from the one that left before it. Thus, the anomalous behavior will be compensated by the means of the values, where one will be smaller and the other higher. The standard deviation of both metrics also did not exhibit any correlation ($\rho = 0.08$ e $p \ll 0.01$).

Likewise, we expected that bus bunching and time to travel the inner section of the trajectory might be correlated, as a bus that travels slower than the average is more likely to be closer to other vehicles, thus being grouped together. Nevertheless, the metrics proved to be uncorrelated (Figure 20) ($\rho = 0.15$ and $p \ll 0.01$). Buses that travel slower than others are more likely to form groups with other vehicles that are faster than them. On the other hand, if the slower speed is caused by traffic jams, all the buses tend to travel slower.

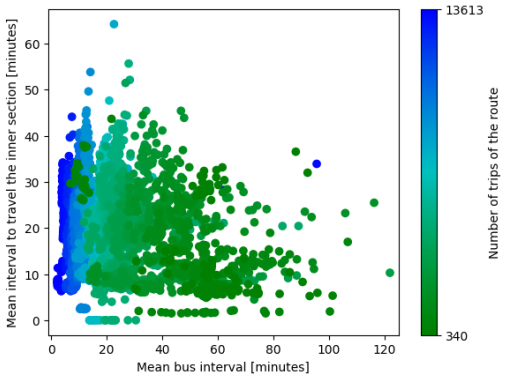

Figure 19. Comparison between mean bus interval and mean time interval to travel the inner section of the trajectory.

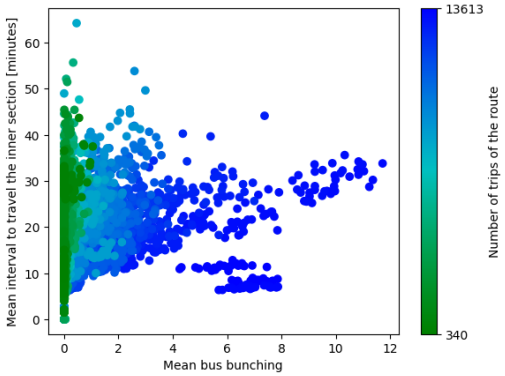

Figure 20. Comparison between mean bus bunching and mean time interval to travel the inner section of the trajectory.

6.4 Discussion

In Rio de Janeiro, most routes do not have a high entropy variation, given that the chosen lines for representation reflect the behavior of all the lines analyzed. Similarly, if a trip of any route has its entropy value discordant, there is an indication that something atypical happened in this specific trip, either because of a failure when the data is collected and transmitted or some traffic condition during the trajectory.

The high number of routes that have bunched buses, being either the ones in which it is a normal situation or the ones that are the exception, indicates that bus bunching is normal in Rio de Janeiro, mainly in routes that have more trips when compared to the others. However, it is paramount to highlight that it is still unwanted, as it can compromise the bus interval, inducing higher waiting times and also negatively affecting the users' perception of the quality of service.

From the planning perspective, routes that have a higher bus interval are logical candidates to receive more vehicles, because longer waiting times are likely to discourage their usage. Furthermore, less regular routes make the users feel less confident about the service, since it is more difficult to predict the time the next bus will traverse the determined location, given that the time of the last bus in that location is known.

The comparison between the metrics revealed that routes with high bus intervals are less likely candidates to have bunched buses because their distance would be higher. The other metrics, however, do not show any clear correlation among them.

7 Conclusion and Future Work

This work evaluated the quality of service of Rio de Janeiro's bus routes through geopositioning information sent periodically by the vehicles. Different metrics were used that reflect different aspects of the quality of service and were compared among them with the intention of evaluating if one influenced another. In order to avoid inconsistent data, a filtering algorithm and a trajectory correction algorithm were applied based on the trajectory the route informed and, following that, the quality of service metrics were calculated.

To compute the metrics, the data was separated according to the routes that comprise Rio de Janeiro's bus system and a temporal filter, selecting the period of the day in which the analysis was made alongside the exclusion of weekends. Those choices were made to analyze periods in which the expected service is the same.

The individual metrics revealed that there are routes that are more regular and others that are more irregular. Bus interval showed a variation from 5 minutes to 40 minutes in most routes. The occurrence of bunching is a common scenario in Rio de Janeiro, as in routes 864 and 232. Finally, the time interval to travel part of the trajectory was able to capture the variations in the traveling time, presenting differences of up to 15 minutes in the same route.

The comparison between entropy and the other metrics showed that entropy is correlated with the time interval to travel the inner section of a route and is not clearly related to bus interval and bus bunching. The bus bunching and bus interval varied jointly, when one diminished, the other grew. Lastly, the time interval to travel part of the trajectory did not show any correlation with either bus bunching or bus interval.

Future research includes trying new quality of service metrics and their relations with the ones used in this work and other methods to compute the entropy of the routes. Finally, it is possible to use these metrics to perform an online detection of anomalous buses, verifying oddities in the bus routes. From the planning perspective, these metrics can be paramount to identify routes that deserve more attention and adjustments in the service provided.

Declarations

Availability of data and materials

All the data utilized in this study originate from third-party datasets that are openly accessible online, as detailed in this article.

Competing interests

The authors declare no competing interests.

Funding

This work was partially sponsored by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Funding Code 001, CNPq (408255/2023-4), FAPERJ (E-26/204.122/2024), and FAPESP (2023/00673-7 and 2023/00811-0).

Authors' contributions

T.B.R. and F.D.M.S. collected the datasets and written the experiment codes; A.C.V. and L.H.M.K.C. have defined the methodology and reviewed the experimental results. All authors have contributed to writing and reviewing the manuscript.

References

Cuttone, A., Lehmann, S., and González, M. C. (2018). Understanding predictability and exploration in human mobility. *EPJ Data Science*, 7(1):2. DOI: 10.1140/epjds/s13688-017-0129-1.

DATA.RIO (2025). Data.rio - portal de dados abertos da cidade do rio de janeiro. Available at:`https://www.data.rio/`. Acesso: 14-Mar-2025.

dos Santos, J. B. and Lima, J. P. (2021). Quality of public transportation based on the multi-criteria approach and from the perspective of user's satisfaction level: A case study in a brazilian city. *Case Studies on Transport Policy*, 9(3):1233–1244. DOI: 10.1016/j.cstp.2021.05.015.

Du, B. and Dublanche, P.-A. (2018). Bus bunching identification using smart card data. In *2018 IEEE 24th International Conference on Parallel and Distributed Systems (ICPADS)*, pages 1087–1092. DOI: 10.1109/PADSW.2018.8644611.

He, S.-X. (2015). An anti-bunching strategy to improve bus schedule and headway reliability by making use of the available accurate information. *Computers & Industrial Engineering*, 85:17–32. DOI: 10.1016/j.cie.2015.03.004.

Hosseini, S. S., Ardabili, B. R., Azarbayjani, M., and Tabkhi, H. (2025). Demographic disparities, service efficiency, safety, and user satisfaction in public bus transit system: A survey-based case study in the city of charlotte, nc. *Transportation Research Interdisciplinary Perspectives*, 29:101296. DOI: 10.1016/j.trip.2024.101296.

Huang, Z., Xu, S., Wang, M., Wu, H., Xu, Y., and Jin, Y. (2024). Human mobility prediction with causal and spatial-constrained multi-task network. *EPJ Data Science*, 13(1):22. DOI: 10.1140/epjds/s13688-024-00460-7.

Ikanovic, E. L. and Mollgaard, A. (2017). An alternative approach to the limits of predictability in human mobility. *EPJ Data Science*, 6(1):12. DOI: 10.1140/epjds/s13688-017-0107-7.

Li, G., Knoop, V. L., and van Lint, H. (2022). Estimate the limit of predictability in short-term traffic forecasting: An entropy-based approach. *Transportation Research Part C: Emerging Technologies*, 138:103607. DOI: 10.1016/j.trc.2022.103607.

Lu, X., Wetter, E., Bharti, N., Tatem, A. J., and Bengtsson, L. (2013). Approaching the limit of predictability in human mobility. *Scientific Reports*, 3(1):2923. DOI: 10.1038/srep02923.

Nguyen, K., Yang, J., Lin, Y., Lin, J., Chiang, Y.-Y., and Shahabi, C. (2023). Los Angeles Metro Bus Data Analysis Using GPS Trajectory and Schedule Data. Available at:`https://rosap.ntl.bts.gov/view/dot/43622` Sponsored By California Department of Transportation.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

Smith, G., Wieser, R., Goulding, J., and Barrack, D. (2014). A refined limit on the predictability of human mobility. In *2014 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 88–94. DOI: 10.1109/PerCom.2014.6813948.

Song, C., Qu, Z., Blumm, N., and Barabási, A.-L. (2010). Limits of predictability in human mobility. *Science*, 327(5968):1018–1021. DOI: 10.1126/science.1177170.

Teixeira, D. d. C., Almeida, J. M., and Viana, A. C. (2021). On estimating the predictability of human mobility: the role of routine. *EPJ Data Science*, 10(1):49. DOI: 10.1140/epjds/s13688-021-00304-8.

Teixeira, D. d. C., Viana, A. C., Alvim, M. S., and Almeida, J. M. (2019). Deciphering predictability limits in human mobility. In *SIGSPATIAL 2019 - 7th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, Chicago, United States. DOI: 10.1145/3347146.3359093.

Wang, P., Chen, X., Zheng, Y., Cheng, L., Wang, Y., and Lei, D. (2021). Providing real-time bus crowding information for passengers: A novel policy to promote high-frequency transit performance. *Transportation Research Part A: Policy and Practice*, 148:316–329. DOI: 10.1016/j.tra.2021.04.007.

Zhao, K., Khryashchev, D., Freire, J., Silva, C., and Vo, H. (2016). Predicting taxi demand at high spatial resolution: Approaching the limit of predictability. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 833–842. DOI: 10.1109/BigData.2016.7840676.

Zheng, Y., Kong, H., Petzhold, G., Barcelos, M. M., Zegras, C. P., and Zhao, J. (2022). Gender differences in the user satisfaction and service quality improvement priority of public transit bus system in porto alegre and fortaleza, brazil. *Travel Behaviour and Society*, 28:22–37. DOI: 10.1016/j.tbs.2022.02.003.

Zhou, C., Tian, Q., and Wang, D. Z. (2022). A novel control strategy in mitigating bus bunching: Utilizing real-time information. *Transport Policy*, 123:1–13. DOI: 10.1016/j.tranpol.2022.04.022.