scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Salvador Urban Network Transportation (SUNT): A Landmark Spatiotemporal Dataset for Public Transportation

Marcos V. Ferreira[1], Matheus Souza[2], Tatiane N. Rios[1,3], Islame F. C. Fernandes[1], Jorge Nery[1,2], João Gama[4], Albert Bifet[3] & Ricardo A. Rios[1,3] ✉

Efficient public transportation management is essential for the development of large urban centers, providing several benefits such as comprehensive coverage of population mobility, reduction of transport costs, better control of traffic congestion, and significant reduction of environmental impact limiting gas emissions and pollution. Realizing these benefits requires a deeply understanding the population and transit patterns and the adoption of approaches to model multiple relations and characteristics efficiently. This work addresses these challenges by providing a novel dataset that includes various public transportation components from three different systems: regular buses, subway, and BRT (Bus Rapid Transit). Our dataset comprises daily information from about 700,000 passengers in Salvador, one of Brazil's largest cities, and local public transportation data with approximately 2,000 vehicles operating across nearly 400 lines, connecting almost 3,000 stops and stations. With data collected from March 2024 to March 2025 at a frequency lower than one minute, SUNT stands as one of the largest, most comprehensive, and openly available urban datasets in the literature.

Background & Summary

In this work, we focus our investigation on efficient urban mobility by modeling data from public transportation systems due to its importance to the population. Any decision regarding this system directly impacts urban mobility, especially in developing countries, where it is often the only means of transport available to low-income populations. When poorly planned, it delivers low-quality services with delayed and overloaded vehicles, concentrates traffic in specific regions while leaving others unattended, and aggravates pollution with higher gas emission rates.

Efficient urban mobility depends on a comprehensive set of strategies to optimize traffic management, yielding benefits such as improved safety, reduced travel time, lower costs, and enhanced environmental sustainability. These strategies have led researchers to explore various solutions, including vehicle-to-vehicle communication, route optimization, the integration of the Internet of Things for connected transportation systems, and more effective public transit scheduling[1–5].

According to Ceder[6], the success of intelligent transportation systems (ITS) relies on collecting and analyzing accurate data, which has led several studies to focus on data-driven approaches. In this sense, Wang *et al*. and Gordon *et al.*[7,8] have used passenger fare data and vehicle location tracking to develop heuristics to better estimate boarding and alighting times and locations in London (England). Similarly, researchers collected passenger and vehicle data from Harbin (China), which was later modeled using unsupervised machine learning methods to understand public transit riders' travel patterns better[9]. Researchers from Seoul (the Republic of Korea) also developed a methodology for estimating non-tagged alighting stop information gradually, by considering the characteristics of trip types and utilizing transportation card data[10]. In New York (USA), researchers analyzed data from the transit system, where riders swipe a fare card only when entering a station or boarding a bus. They used this information to estimate alighting stops based on the bus boarding locations[11]. A similar problem was

[1]Federal University of Bahia, Institute of Computing, Salvador, Brazil. [2]Integra (Association of Transport Companies of Salvador), Salvador, Brazil. [3]Univeristy of Waikato, Hamilton, New Zealand. [4]University of Porto, Porto, Portugal. ✉e-mail: ricardoar@ufba.br

Dataset	#Nodes	#Edges	Period	Shortest time interval
METR-LA[22–25]	207	2,369	March 1, 2012 to June 30, 2012	5 minutes
PeMS-BAY[23–28]	325	1,515	January 1, 2017 to May 31, 2017	5 minutes
TaxiBJ[29,30]	*NA	NA	July 1, 2013 to October 30, 2013	30 minutes
			March 1, 2014 to June 30, 2014	
			March 1, 2015 to June 30, 2015	
			November 1, 2015 to April 10, 2016	
BikeNYC[29,30]	50	NA	April 1, 2014 to September 30, 2014	1 hour
Shanghai Metro[31]	288	958	July 1, 2016 to September 30, 2016	15 minutes
Hangzhou Metro[31–33]	80	248	January 1, 2019 to January 31, 2019	15 minutes
Beijing Metro[31,34]	276	NA	February 29, 2016 to April 3, 2016	NA
Chongqing Metro[31]	170	NA	March 1, 2019 to March 31, 2019	15 minutes
Stockholm County[35]	NA	NA	NA	NA
UVDS[2,36]	104	NA	Three months	5 minutes
SUNT	**2,871**	**4,526**	**March 01, 2024 to March 31, 2025**	**$<$ 1 minute**

Table 1. Key Characteristics of SUNT in Comparison with Existing Datasets from the Literature.

addressed in Southeast Queensland (Australia) by using a Deep Neural Network to predict unknown alighting locations after being trained in a dataset with a combination of transactional and public transit network attributes[12]. Although these aforementioned citations are more related to our work, further research on the problem of inferring boarding-alighting locals in public transportation systems is detailed in a review published by Mohammed and Oke (2023)[13].

As extensively discussed in the literature, understanding and improving public transportation directly impact urban mobility, particularly in developing countries, where it is often the primary means of transport for low-income populations. Poorly planned systems provide low-quality services, leading to delays, overcrowded vehicles, traffic congestion in specific areas while neglecting others, and increased pollution due to higher gas emissions. After an in-depth investigation of published manuscripts focused on public transportation, we noticed a limitation in the availability of a totally public dataset containing comprehensive quantitative, spatial, and temporal information about passengers, vehicles, lines, stops, and stations. Moreover, despite the increasing advancements in Machine Learning (ML) methodologies, particularly for intelligent transportation systems (ITS), there remains a significant lack of datasets with detailed information about public transportation with their respective passengers.

To address this limitation, we present the Salvador Urban Network Transportation (SUNT) dataset, the most comprehensive public transportation dataset currently available in the literature. Collected in Salvador, Brazil, between March 2024 and March 2025, SUNT covers an area of approximately 694 km^2 and serves nearly 3 million residents. The transportation used by the local population in Salvador comprises three systems: regular buses, subway, and BRT (Bus Rapid Transit). The regular bus system is the most extensive transportation in Salvador, serving most of the population. Currently, there are about 1,900 buses distributed on approximately 400 lines with around 3,000 stops and stations, supporting roughly 470,000 passengers daily. The subway system spans about 35 km across 2 lines with 20 stations. Approximately 210,000 passengers use this system daily. The BRT (Bus Rapid Transit) system was recently inaugurated, further enhancing urban mobility and serving about 30,000 passengers daily. Currently, about 40 buses are operating on 3 lines and 20 stations.

In addition to publicly available vehicle information, which is commonly shared by several cities worldwide, SUNT stands out for its innovative inclusion of passenger data, such as boarding and alighting details, and its diverse data formats. These include graph representations with over 2,000 nodes and 4,000 edges, as well as temporal data streams with a granularity of less than one minute.

In Table 1, we summarize important related works, which models different urban datasets. The missing information ("NA") in this table reflects the fact that several datasets commonly used in research articles are partially described in the publications and are not freely shared in public repositories with the same level of detail as ours. For example, we have noticed that information about the number of nodes, edges, or specific temporal intervals is often unavailable. As a result, researchers face challenges in reproducing experiments or fully understanding the scope and limitations of the datasets referenced in these studies.

On the other hand, we offer the Salvador Urban Network Transportation (SUNT) dataset, which stands out as an exception, offering 2,871 nodes, 4,526 edges, and a temporal granularity of less than one minute, with an in-depth dataset construction, which are pivotal for addressing key deficiencies identified in recent studies on learning benchmarks. SUNT offers a robust foundation for developing models that can learn complex spatiotemporal patterns and adapt to rapidly changing conditions in ITS scenarios. Additionally, being recently collected, it reflects an updated urban configuration, in contrast to the most recent previously available dataset, which dates back to 2019.

The significance of sharing the SUNT dataset lies in its dual impact: advancing scientific research and informing public policy. For researchers, SUNT provides a comprehensive and high-quality resource to develop and evaluate a wide range of data-driven methods, such as supervised and unsupervised learning, concept drift and anomaly detection, time series analysis, graph-based optimization, and high-performance computing techniques for large-scale transit data. For policymakers and transit agencies, SUNT enables the simulation of

route_short_name	pt_sequence	direction_id	longitude	latitude	stop_id	route_long_name	service_code
0116	1	1	− 38.51123	− 12.983389	43768720	Avenida Vale Do Tororo	53786
0116	2	1	− 38.511097	− 12.986428	45832898	Avenida Vale do Tororo, 291	53786
0116	3	1	− 38.511448	− 12.990091	44782328	Praça Dr. João Mangabeira	53786
0116	4	1	− 38.504387	− 12.990533	44784448	Av. Vaco da Gama, S/N -	5378
0116	5	1	− 38.501972	− 12.992005	44784449	Av. Vasco da Gama, 271 -	53786
0116	6	1	− 38.499004	− 12.993324	45833116	Av. Vasco da Gama, S/N -	53786

Table 2. AVL-lines features: information regarding bus lines.

vehicle	route_short_name	direction_id	gps_datetime	longitude	latitude	stop_id	service code
20001	0310	0	2024-03-01 05:53:20	− 38.512428	− 12.978642	45834426	45546
20001	0310	0	2024-03-01 05:53:53	− 38.509964	− 12.975935	45834425	45546
20001	0310	0	2024-03-01 05:53:57	− 38.509964	− 12.975935	45834425	45546
20001	0310	0	2024-03-01 05:54:02	− 38.508957	− 12.975689	44782954	45546
20001	0310	0	2024-03-01 05:54:47	− 38.508957	− 12.975689	44782954	45546
20001	0310	0	2024-03-01 05:55:58	− 38.507446	− 12.97867	44428471	45546

Table 3. AVL-vehicle features: information concerning vehicles' routes and bus schedules.

real-world transportation scenarios, supporting evidence-based decisions and reducing risks when planning or modifying public transportation systems. This combination of scientific utility and practical application underscores the dataset's relevance and potential to contribute meaningfully to the fields of urban mobility and smart city development.

Methods

This section describes the steps taken to create the SUNT dataset. First, we present in detail the data collected from four distinct public transportation sources. We then explain the use of the Trip Chaining approach to integrate these data sources, resulting in a complete origin-destination matrix. Finally, we describe the process of modeling this matrix as a graph that connects stops, enriched with several attributes related to passenger boarding and alighting.

Raw Datasets. In this study, we utilized an Automated Data Collection System (ADCS) to gather data from multiple sources[13]. The first source was the Automatic Vehicle Location (AVL) system, which monitors all regular and BRT buses, providing details about their geospatial positions over time.

In summary, AVL records real-time vehicles' geographical locations, which are important to estimate several relevant information, such as passengers boarding and alighting, public transportation network planning, and monitoring and controlling traffic operations. The daily AVL information shared in our repository[14] contains two different set of features: AVL-lines and AVL-vehicles. AVL-lines comprise static information regarding the lines, whose features are shown in Table 2. These columns provides different information about the lines: `route_short_name` – identification; `pt_sequence` – stop sequence order; `direction_id` – direction, where 1 stands for one-way and 0 for return trip; `longitude` and `latitude` – geographical coordinates for the bus stop identified in column `stop_id;route_long_name` – stop names; `service_code` – the trip along the line.

Table 3 presents AVL-vehicles features, which comprise information concerning vehicles' routes and bus schedules. One of the most important columns is `gps_datetime`, which provides the vehicle's arrival date and time at the stop identified by the `stop_id` column. If `gps_datetime` contains two values, the earlier timestamp corresponds to the bus arrival time at the stop, while the later one represents the departure time. The stop sequence of the bus line must be consistent with the values in `gps_datetime`; that is, for each stop, the arrival time must be earlier than the departure time, and the departure time must be earlier than the arrival time at the next stop. The remaining columns are similar to those described in Table 2.

The second collecting source is the Automatic Fare Collection (AFC) system, which contains information from the ticketing systems, recording the time when users' contactless cards are used for payments. In addition to the exact time of card usage, it also includes details on the vehicles and their respective lines. In our scenario, AFC is used to gather data from passengers using regular buses, subway, and BRT. For buses, data collection occurs at two points: when passengers validate their tickets either at the vehicle's built-in turnstile or at a mobile turnstile. In the case of the subway, the AFC system records entries through turnstiles located at station entrances. For the BRT system, AFC combines both methods—collecting data through turnstiles inside vehicles as well as those installed at station entrances.

A subsample of AFC, shown in Table 4, illustrates the available attributes: `cod_card` is the number of passenger's card, randomly generated to avoid recovering any user identification, `afc_datetime` represents the time when the passenger registers the payment, `integration` indicates the possibility of a connection between vehicles, `route_short_name` is the route identification, `direction_id` shows the bus direction

cod_card	afc_datetime	integration	route_short_name	direction_id	value	vehicle
02310034266847	2024-03-01 06:22:03	False	1386	I	0.0	20390
02310034266847	2024-03-01 06:22:10	False	1386	I	0.0	20390
02310033002113	2024-03-01 06:22:57	False	1386	I	0.0	20390
02310032345960	2024-03-01 08:12:25	False	1386	I	0.0	20390
02320033736512	2024-03-01 06:04:08	False	1386	I	0.0	20390
03620033306428	2024-03-01 06:10:17	False	1386	I	5.2	20390

Table 4. Illustrative example of the AFC dataset.

agency_id	agency_name	agency_url	agency_timezone	agency_lang	agency_phone
1	company_I	www .	America/Sao_Paulo	pt	
2	company_II	www .	America/Sao_Paulo	pt	

Table 5. GTFS Agency: information about the bus companies.

route_id	agency_id	route_short_name	route_long_name	route_type
4089	1	1230	Sussuarana x Barra R1.	3
4450	1	1321	São Marcos x Barroquinha	3
4518	1	1103	Alto do Cruzeiro/Pernambués x Shop.Bela Vista/Term Ac.Norte	3
4523	1	1405	Estação Pirajá x Cajazeiras 8	3
4524	1	1137	Pernambués x Barra	3

Table 6. GTFS Routes: information about bus lines.

route_id	service_id	trip_id	direction_id	block_id	shape_id
4089	26082_D_1046761	1046761_D_1_0	0	4089_001M	26082_I
4089	26082_D_1046761	1046761_D_1_1	1	4089_001M	26082_V
4089	26082_D_1046761	1046761_D_2_0	0	4089_002M	26082_I
4089	26082_D_1046761	1046761_D_2_1	1	4089_002M	26082_V
4089	26082_D_1046761	1046761_D_3_0	0	4089_002T	26082_I

Table 7. GTFS Trips: information about the trips and the paths followed by the bus.

trip_id	arrival_time	departure_time	stop_id	stop_sequence	pickup_type	drop_off_type
1046761_D_1_0	08:30:00	08:30:00	43968810	1	0	0
1046761_D_1_0	08:31:41	08:31:41	47566106	2	0	0
1046761_D_1_0	08:33:49	08:33:49	44782337	3	0	0
1046761_D_1_0	08:34:55	08:34:55	44784470	4	0	0
1046761_D_1_0	08:35:44	08:35:44	44784471	5	0	0

Table 8. GTFS Stops Times: the chronological order of bus stops where each trip paused.

(I – one way or V – return) considering its initial and final stops, `value` is the trip cost, and `vehicle` is the code used to identify the vehicle. This dataset contains approximately 35 million records per month.

Additionally, we used static data based on the General Transit Feed Specification (GTFS) format, which defines a standard format for public transportation schedules associated with geographic information (http://gtfs.org/). Using this format, we provided geospatial details about stations and stops along with their sequential order, lines, and directions. In the SUNT dataset, the GTFS provides 5 files (GTFS Agency, GTFS Routes, GTFS Trips, GTFS Stops Times, and GTFS Stops) that describe the entire network and services of public transportation related to local companies. GTFS Agency, illustrated in Table 5, contains information about the bus companies, which are associated with GTFS Routes (Table 6) by the attribute `agency_id`. GTFS Routes contains information about bus lines and is associated with GTFS Trips (Table 7) by the attribute `route_id`. GTFS Trips shows all the trips and the paths followed by the bus and is directly associated with GTFS Stops Times (Table 8), which maps the chronological order of bus stops where each trip paused. Finally, GTFS Stops (Table 9) contains information about each bus stop and is associated with the GTFS Stops Times by the attribute `stop_id`.

stop_id	stop_name	latitude	longitude	location_type	parent_station
43968810_S	R. São Cristóvão 2	− 12.931565284729	− 38.444393157959	1	
43968810	R. São Cristóvão 2	− 12.931565284729	− 38.444393157959	0	43968810_S
47566106_S	Av. Ulysses Guimarães 4067	− 12.93385887146	− 38.4467735290527	1	
47566106	Av. Ulysses Guimarães 4067	− 12.93385887146	− 38.4467735290527	0	47566106_S
44782337	Av. Ulysses Guimarães 4314-4322	− 12.9351501464844	− 38.4405784606934	0	

Table 9. GTFS Stops: information about each bus stop.

route_short_name	service_code	direction_id	vehicle	start_trip	end_trip	activity
T014	74335	I	20401	01/03/2024 17:03:49	01/03/2024 17:10:45	Leaving the garage
T014	74335	I	20516	01/03/2024 05:37:16	01/03/2024 05:40:36	Leaving the garage
T014	74335	I	20516	01/03/2024 17:11:40	01/03/2024 17:20:58	Normal
T014	74335	I	20086	01/03/2024 05:39:27	01/03/2024 05:46:38	Leaving the garage
T014	74335	I	20401	01/03/2024 12:37:47	01/03/2024 12:42:04	Returning to the garage

Table 10. LTI: the start of each trip made by a vehicle on a specific route.

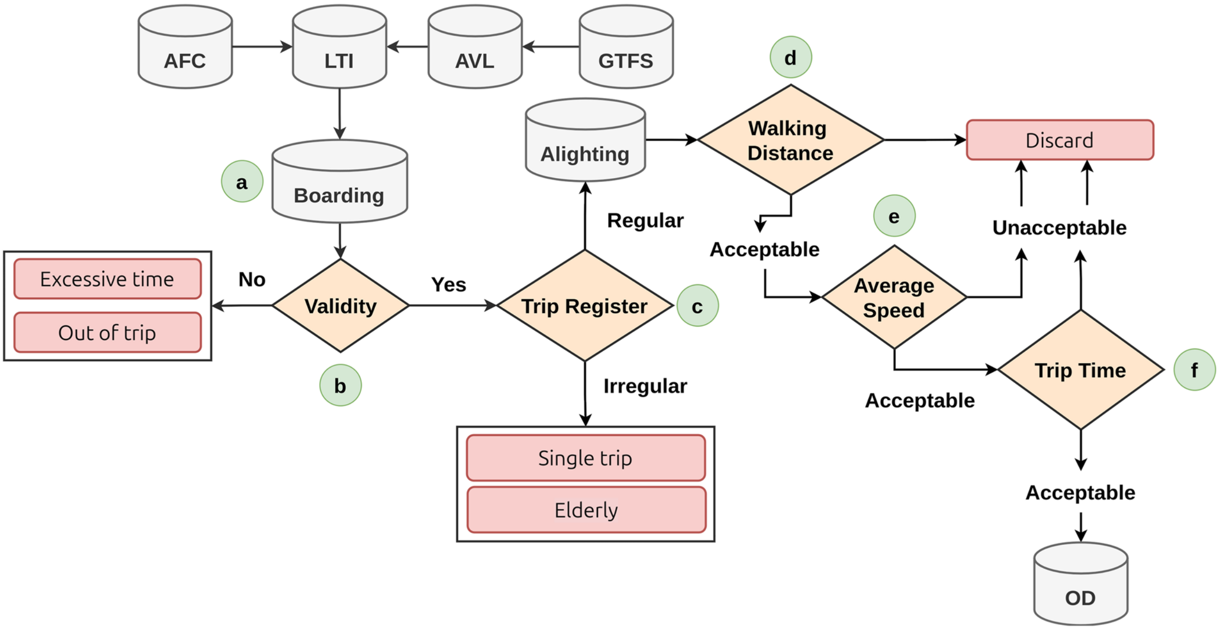


Fig. 1 Steps used to create our origin-destination dataset. Red boxes represent boarding data with no alighting correspondence.

Finally, we also provide a dataset containing Local Trip Information (LTI), which includes details about the expected and actual departure and arrival times for all vehicles on every line and in each direction. Due to the dynamic nature of data collected from the AVL system, missing data may occur, resulting in random loss of information about vehicle activities. This issue can be easily addressed by combining redundant vehicle information from GTFS and LTI. Table 10 summarizes the attributes of the trip mapping dataset. The most important attributes are `start_trip` and `end_trip`, which indicate the start and end times of each trip, respectively. The `activity` attribute categorizes the trip as either a regular service, a departure from the garage, or a return to the garage. This dataset complements the AVL dataset by providing trip-level information, which is not included in the AVL records. On average, it contains approximately 700,000 records per month.

Trip Chaining. After organizing these four datasets (AVL, AFC, GTFS, and LTI), the first challenge is to find out the boarding locations for all users. As illustrated in Fig. 1(a), this information is computed by integrating AVL and AFC, and retrieving the exact latitude and longitude positions when the users' cards performed the payments. Using these positions, we can estimate the closest stop or station that indicates the boarding local. Next, we merge multiple boarding locations to classify the users' trips as initial, intermediate, and final. Such a classification is relevant to map all possible connections that compose a complete user's trip. Finally, all boarding positions with their respective time instants are used to organize trip chains that describe passengers' behavior, as

Column	Sample Value	Dtype
tripuserid	02300033357538_20240301184830	object
type_bus	bus	object
user_type	driver	object
set	company_i	object
registers	2	int64
trip_id	20097_0310_7	object
start_trip	2024-03-01 17:56:43	datetime64[ns]
end_trip	2024-03-01 20:08:27	datetime64[ns]
tolerance	NaT	datetime64[ns]
integration	False	bool
cod_card	2300033357538	object
stop_time	2024-03-01 19:36:35	datetime64[ns]
register_time	2024-03-01 18:48:30	datetime64[ns]
service_code	45546	object
route_short_name	0310	object
vehicle_afc	20097	object
vehicle	20097	object
stop_id	44782849	object
order	1	float64
direction_id	I	object
trip_em	7.0	float64
dif_boarding	48.083	float64
trip	Inside	object
classification	irregular	object
motive	excessive time	object
trip	firt_trip	object
set_nb	company_i	object
stop_time_nb	2024-03-01 20:04:39	datetime64[ns]
route_short_name_nb	1067	object
vehicle_nb	20446	object
stop_id_nb	44164980	object
diff_nb	0.53	float64
motive_pe	regular	object
target_boarding	irregular	object

Table 11. Description of Columns in the Dataset Boarding.

summarized in Table 11. As shown, most attributes in this table result from the integration process, with bus type being a particularly important attribute in the dataset.

In the next phase, Fig. 1(b), we assess the validity of the boarding registration by checking two specific conditions. Firstly, a user's boarding is discarded if the time difference between the AFC-recorded fare payment and the AVL-recorded bus arrival at the stop exceeds a certain threshold. This threshold has two possible values: (i) 20 minutes for bus stations; and (ii) 5 minutes for regular stops. This differentiation is necessary because buses typically remain longer at stations. Secondly, another discarding possibility happens when there is no direct connection between AVL and AFC records, which is considered in this figure as "out of trip". According to the literature[7], appropriate time intervals for integration depend on the specific dynamics of each city. For instance, in London, a 5-minute interval is considered suitable for integrating payment and boarding processes. In the case of Salvador, local studies conducted by the bus operating companies, taking into account factors such as delays, low-frequency routes, scheduled departures, and traffic conditions, led to the definition of time intervals considered most appropriate for modeling the dynamic interactions between passengers and buses.

In the subsequent phase, Fig. 1(c), we analyzed user types to determine the feasibility of estimating their alighting points. In Salvador, there is no device to validate the passengers' alighting; therefore, the main challenge is to estimate it by analyzing the following boarding. Moreover, in Salvador, passengers aged 65 or older are entitled to free public transportation and are not required to use any form of electronic ticket or identification card. They can board simply by presenting a personal document. As a result, their boarding and alighting events are not recorded in the system. However, this group represents only a small portion of the total passenger volume. To account for their presence in the dataset, we apply a probability distribution to allocate these passengers along the bus line within the analyzed time interval. Another particular case that prevents us from identifying users' alighting points occurs when there is only a single trip registration on a given day. In such cases, we can only determine the boarding point, with no information available about the alighting point. Therefore, we cannot consider such situations in our analyses. The alighting dataset, illustrated in Table 12, is one of the most

Column	Sample Value	Dtype
tripuserid	02300033520791_20240301104958	object
stop_time_ali	2024-03-01 10:55:27	datetime64[ns]
stop_id_ali	44165441	object
order_ali	6.0	float64
walk_target	excessive	object
trip_ali	8.0	float64
walk_dis	1.299	float64
walk_time	15.588	float64
walk_speed	5.5	float64
diff_de_pe	68.4	float64
wait_time	52.812	float64
trip_dis	1.884	float64
trip_time	5	float64
vel_media	22	float64
bridge	False	bool
bridge_type	no bridge	object
bridge_id	None	object
chain	bus-bus	object
target_ws	regular	object
target_avs	regular	object
target_tt	regular	object
target_td	regular	object
target_alighting	regular	object

Table 12. Dataset summary Alighting.

important dataset produced by our research, whose examples of relevant attributes are: `stop_time_ali` – the estimated alighting time, `stop_id_ali` – the stop ID where alighting is inferred; `walk_target` – classification of walking behavior post-alighting, `walk_dis` – estimated walking distance in kilometers, `walk_time` – estimated walking time in minutes, `wait_time` – estimated waiting time at the stop in minutes, `trip_dis` – distance covered during the trip in kilometers, and `trip_time` – duration of the trip in minutes. A complete description of all features is available in our repositories.

For all remaining cases, alighting points can be inferred by analyzing each user's sequence of boarding points, i.e., by applying the Trip Chaining strategy, which is widely adopted in the ITS literature as discussed in the previous section[6–8,10,11,13]. To better understand this inference, consider the three scenarios illustrated in Fig. 2. In Scenario I, we observe a passenger boarding at 8:00 AM (B1) at Stop (b) and then boarding again at 6:00 PM (B2) at Stop (f). Therefore, it can be inferred that the passenger boarded at Stop (b), disembarked at Stop (f) on the first trip (b → f), and then made the return journey at the end of the day (f → b).

In the second scenario, we observe a user trip with a connection. In this situation, there are two boarding points for each trip. Initially, the user boarded at Stops (b), at 8:00 AM (B1), and (d), 8:20 AM (B2), being the first alighting registered at Stop (d). At the end of the day, the user boarded at Stops (j), at 6:00 PM (B3), and (d), at 6:10 PM (B4), respectively. Therefore, we infer the first user's trip was b → j starting at 8:00 AM, and their return was j → b at 6:00 PM.

In our final scenario, we illustrate a situation when a user utilizes a connection between two different stops by walking a short distance between them. In this case, they register a first boarding at Stop (b), at 8:00 AM (B1), and the second one at Stop (x), 8:50 AM (B2). As one may notice, Stop (x) is in a different line. Hence, we look for its closest stop, respecting the maximum walking distance (Δ), Stop (f) in this case, to represent the first alighting. Considering they register another boarding at Stop (u), at 7:00 PM (B3), followed by boarding at Stop (f), at 7:30 PM (B4), we can map their full daily trip using the same rule previously considered. Therefore, we infer the first user's trip was $b \longrightarrow f\Delta x \longrightarrow u$ starting at 8:00 AM, and their return was $u \longrightarrow x\Delta f \longrightarrow b$ at 7:00 PM.

As shown in Fig. 1(d), a walking distance is deemed acceptable if it is limited to 1.1 km. Concerning the average velocity, Fig. 1(e), and the trip time, Fig. 1(f), all registers with values greater than 80 km/h and 2 hours are unconsidered. These values were estimated by local specialists based on the passengers' usage patterns and the transportation infrastructure in Salvador. Similar to the time interval, walking distance is also influenced by the specific infrastructure of each city. In London, for example, two different studies adopted thresholds of 1 km[8] and 750 meters[7]. In the case of Salvador, the walking distance was defined by the operational planning team of the city's bus consortium, based on the typical spacing between local stops and stations. It is important to note that both the time and distance thresholds can be adjusted by readers to suit their own scenarios, as we provide access to both raw and processed data.

In Fig. 1, all red boxes represent situations in which we cannot precisely use the passengers' occurrences in our analyses. Nevertheless, even in minority cases, it is essential to consider their general behavior to mitigate imprecision in further estimations, such as the load of passengers on the buses. In this case, we use the data distribution for each line to allocate these occurrences across different buses, as recommended by the literature.

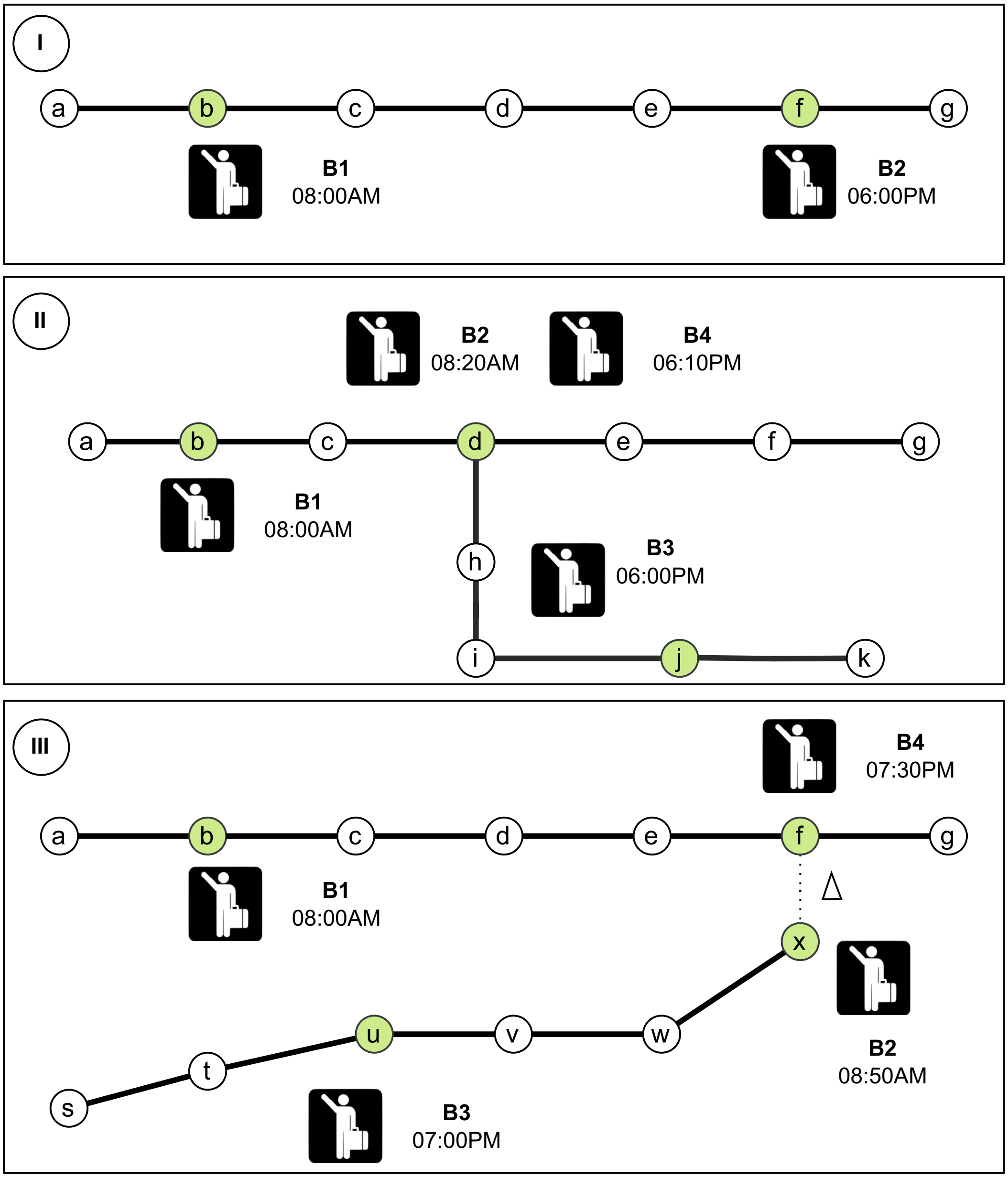


Fig. 2 Scenarios illustrating three different boarding-alighting situations: (I) a single line, (II) lines with a connection, and (III) two lines connected by walking distance.

After this correction, we have the processed Origin-Destination (OD) dataset, whose attributes are illustrated in Table 13. In this example, it is important to clarify that, at the first stop or station of each trip, the same timestamp is recorded for both the "start_trip" and "stop_time" attributes. This occurs because the bus is initiating its route, as indicated by the attribute "pt_sequence = 1". At this initial point, the bus is empty, as expected at the beginning of all trips. For instance, in such cases, the number of boardings ("n_boardings") reflects the total number of passengers entering the bus, while no alighting occurs ("n_alighting" = 0, "lag_loading" = 0, and "balance" = 0). Consequently, the loading after this first stop corresponds exactly to the number of boarded passengers.

Graph Modeling. The organization of the OD dataset with passengers' boarding and alighting allowed us to create the SUNT dataset, embedding a set of quantitative, temporal, and geospatial variables as a complex network. Formally, we have used information on latitude, longitude, and time to create a spatial-temporal graph $G = \{G_1, G_2, ..., G_T\}$. For all $t = 1, ..., T$, $G_t = (V, E)$ stands for an attributed and directed graph at time t, where $V = \{v_1, v_2, ..., v_N\}$ is the set of N vertices corresponding to the bus stops and stations, and E is the set of edges corresponding to feasible routes. A directed edge $(v_i, v_j) \in E$ connects vertices $v_i, v_j \in V$ if, and only if, there is a feasible route for the bus traffic from the corresponding station v_i to v_j in the network. G_t is a fixed graph structure since sets V and E do not change over time.

Attribute	Description	Sample Values	Data Type
route_short_name	Route identification code	1521	object
register_code	Unique identifier in the database	55037	int64
direction_id	Vehicle direction: Inbound (I) or Outbound (V)	I	object
pt_sequence	Stop sequence number	1	int64
stop_id	Unique stop identifier	46021891	int64
vehicle	Vehicle identifier	30661	int64
trip_number	Trip sequence number	1	int64
trip_id	Unique trip identifier	30661_1521_1266	object
start_trip	Timestamp when the trip started	2024-03-01 06:59:11	datetime64[ns]
end_trip	Timestamp when the trip ended	2024-03-01 07:15:22	datetime64[ns]
stop_time	Timestamp when the vehicle arrived at the stop (e.g. 46021891)	2024-03-01 06:59:11	datetime64[ns]
n_boardings	Number of passengers boarding at the stop (e.g. 46021891)	42.0	float64
n_alighting	Number of passengers alighting at the stop (e.g. 46021891)	0	float64
lag_loading	Passenger load before the vehicle arrives at the stop (e.g. 46021891)	0	int64
balance	Passenger load after alighting (previous load minus alightings) at the stop (e.g. 46021891)	0	int64
loading	Passenger load after boarding (balance plus boardings) at the stop (e.g. 46021891)	42	int64

Table 13. Summary of the OD dataset.

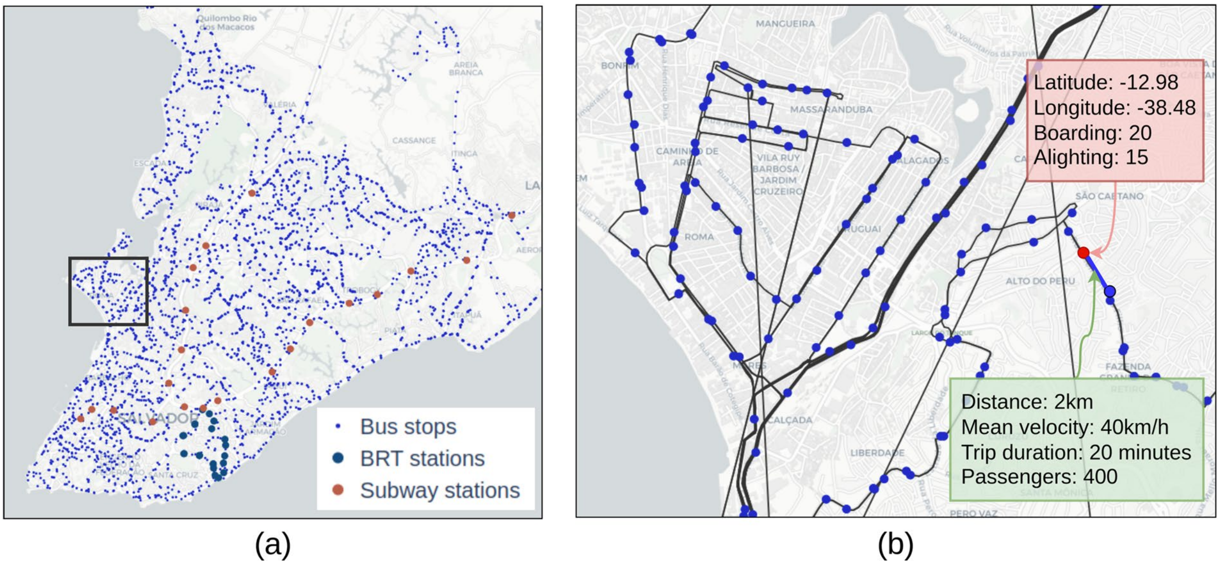


Fig. 3 (**a**) Salvador map with all stops and stations used by regular buses, BRT, and subway; (**b**) a sample of stops and stations (nodes) represented by blue dots and their respective lines (edges).

Figure 3(a) shows the map of Salvador with all vertices stored in our SUNT dataset, i.e., stops and stations used by regular and BRT buses, as well as subways. The geospatial information allows us to place them on the map, respecting their actual geographic position and the distances connecting them by the physical streets.

In our context, spatial data do not depend on time t, i.e., their information is time-invariant. Specifically, in every vertex $v_i \in V$, we store the following features: geographical position, number of boarding and alighting per vehicle, and passenger load. The features specifically concerning edges $(v_i, v_j) \in E$ include the distance between stops and stations, the trip duration, the mean velocity, and the Renovation Factor (RF). RF is a well-known metric used in transportation research to assess the total demand in a line, i.e., it is computed on a set of edges that belong to the line[15]. Formally, this metric is the ratio of the total demand of a line to the load on its critical link. Higher renovation factors occur when there are many short trips along the line. Corridors with very high renovation factor rates are more profitable because they handle the same number of paying customers with fewer vehicles[15]. Besides the individual features, there is relevant information shared by both vertices and edges, such as the number of passengers per vehicle, lines and directions, vehicle characteristics, altitude, and trips.

The black bounding box in Fig. 3(a) represents an essential region of the city, which gathers different lines and connections. Figure 3(b) zooms in this region with a portion of the full graph, illustrating some bus stops as vertices and lines connecting them as edges. The red explaining box contains some features related to that bus stop (vertex) such as its latitude and longitude position, and the amount of boarding and alighting passengers. In the green explaining box,

src	dst	distance	src_lat	dst_lat	src_lon	dst_lon	average_speed	trip_time	loading
100009577	345936831	0.254	−12.902	−12.902	−38.42	−38.417	25.6	4	78
100722777	100722778	0.362	−12.899	−12.897	−38.408	−38.408	11.3	8	20
100722777	44782645	1.062	−12.899	−12.899	−38.408	−38.413	40.2	5	45
100722777	45833440	0.417	−12.899	−12.897	−38.408	−38.409	50.5	10	90
100722777	66771046	0.934	−12.899	−12.897	−38.408	−38.413	26.2	6	30

Table 14. Sample of edge features.

node	loading	n-alighting	n-routes	n-boarding	n-trips	n-vehicles	average_speed
100009577	2.77	0.0	1.08	0.23	1.1	1.1	6.31
100722777	28.54	4.43	1.54	4.49	1.56	1.56	22.86
100722778	36.72	1.39	1.83	0.1	2.04	2.04	16.06
101214305	12.53	3.97	1.0	1.66	1.0	1.0	19.95
101269104	125.57	3.55	4.57	9.48	5.28	5.28	38.25

Table 15. Examples of node features.

we illustrate some features related to a line (edge), such as the distance connecting two stops, the mean velocity and trip duration among the buses in that section, and the total amount of traveling passengers.

In Table 14, we illustrate how edge information is shared: `src` – origin stop/station; `dst` – the destination; `distance` – the distance between them; `src_lat`, `dst_lat`, `src_lon`, and `dst_lon` – their geospatial locations; `average_speed` – the average speed; `trip_time` – the total time trip; and `loading` – the passenger load in a given edge (street or avenue).

In Table 15, we share information about nodes, i.e., details related to stops and stations. Some relevant information containing average values considering vehicles are: `loading` – passenger loading that crossed a given node; `n-boarding` and `n-alighting` – amount of boarding and alighting; `n-routes`, `n-trips`, and `n-vehicles` contain the number of routes, trips, and vehicles; and `average_speed` is the average of speed for each vehicle during their last trip up to the destination node. A complete discussion of all datasets is documented at https://github.com/LabIA-UFBA/SUNT/blob/main/docs/datasets.md, and the corresponding source code is available at https://github.com/LabIA-UFBA/SUNT/blob/main/docs/dataloader_sample.ipynb.

Positive Impacts and Future Works.

This paper introduced SUNT, a novel dataset collected from public transportation in Salvador, Brazil. This dataset is notably relevant to the scientific community for supporting investigations in several domains, such as planning public transportation, designing computational approaches, and managing environmental impact. As previously mentioned, other researchers have published related datasets, ratifying the importance of this subject. However, our dataset stands out due to its massive information and complete availability. Unlike manuscripts that only share outcomes, we have fully shared collections of raw and graph-based details of vehicles, passengers, stations, time, and geographic properties.

In summary, SUNT paves new ways to provide positive social impacts, such as better planning the allocation of buses to lines, reasonably defining regular and express trips, thus reducing traffic jams and carbon emissions, and offering better trip experiences. By sharing SUNT, we expect to provide a robust dataset for the community, supporting the advancement of several investigation possibilities like time-based models, graph algorithms, spatial approaches, deep neural networks, routing simulations, and search heuristics. To illustrate such possibilities, we have listed future work that is worth investigating from our perspective: (i) graph-based learning approaches designed to pass messages using both temporal and spatial information; (ii) Multi-objective optimization approaches to find the shortest path based on edges weighted by distance and time considering traffic jam; (iii) Multimodal ML models that combine different features (e.g., temporal, spatial, numerical, and categorical data) with varying encoding approaches as message passing; (iv) Queue theory to address the problem of attending passengers from a stop A to B; (v) Concept Drift methods designed to identify when passengers' pattern changes in real-world automatically; (vi) in multi-agent evolutionary algorithms, each agent handles a part of the search and, in each generation, spatial-temporal model could help to select the most suitable agent at each step of the evolutionary process; and (v) SUNT can be used to fine-tune time series foundation models, enabling similar transportation analyses in cities that lack equally complete and detailed datasets.

Ethical Declarations.

Our datasets has no human ethical concern. The identification of users' cards in the AFC data does not correspond to the actual card numbers, but rather an internal code that cannot be used to retrieve any personal information from external access. Although such recovery is highly unlikely, we implemented a hash-based solution (collision-free) to convert all internal identifications, adding a layer of privacy. It is important to note that no other attribute links individual users to their public transportation usage.

Variable	Mean	Std	Min	25%	50%	75%	Max
n_boardings	1.37	4.09	0	0.00	0.00	1.00	264
lag_loading	20.28	18.44	0	6.00	15.63	29.48	264
n_alighting	0.98	2.50	0	0.00	0.00	1.21	85
balance	19.30	17.84	0	5.54	14.67	28.00	264
loading	20.67	18.29	0	6.68	16.00	29.75	264

Table 16. Basic descriptive statistics.

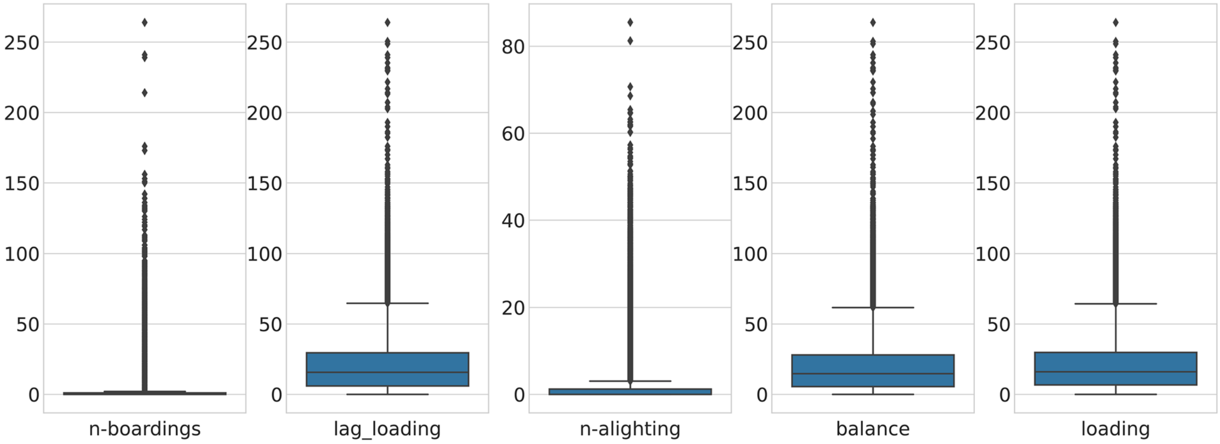


Fig. 4 Box plots summarizing descriptive statistics focused on passengers' behaviors.

Data Records

The dataset is available at Mendeley Data[14]. The raw data were categorized into the following folds: AFC (Automatic Fare Collection), AVL (Automatic Vehicle Location), GTFS (General Transit Feed Specification), and LTI (Local Trip Information). The processed data were then organized into three distinct folds: Alighting, containing information about where passengers began their trips; Boarding, including all estimated stops where passengers ended their trips; and OD, which provides complete origin-destination information, enabling the data to be modeled as a graph. These folds contain data from March 2024 to March 2025. For updates and new data, we recommend accessing the GitHub repository: https://github.com/LabIA-UFBA/SUNT.

Technical Validation

This section aims to demonstrate the quality, consistency, and technical validity of SUNT by detailing the procedures implemented to ensure data integrity. It includes statistical and temporal visualizations that confirm the dataset contains accurate and practically useful information. It is important to emphasize that this section focuses on evaluating the dataset itself (its structure, reliability, and coherence) rather than performing extensive machine learning experiments or domain-specific analyses.

Statistical Validation. This section presents a set of descriptive statistics used to demonstrate whether our datasets contains accurate and useful information about public transportation in Salvador. Table 16 summarizes the statistics of the five most relevant attributes in the OD dataset: "n_boardings", "lag_loading", "n_alighting", "balance", and "loading". A detailed examination of these attributes provides valuable insight into the overall dynamics of public transportation in Salvador. Moreover, these statistics provide a foundational understanding of data distribution, variability, and potential anomalies, which is essential for designing experiments, selecting models, and interpreting results using the SUNT dataset.

In addition to the basic statistics, Fig. 4 presents box plots for all attributes, highlighting the challenges of modeling transport data. These challenges arise from sensitivity to rush hours, unexpected transit events, and various seasonal and daily patterns. An important detail in both Table 16 and Fig. 4 is the presence of outliers with values exceeding 250. For instance, a recorded boarding count (n_boardings) of 264 for a single vehicle far surpasses its actual capacity. This occurs in rare cases when a mobile turnstile registers multiple passengers at a location, even though they board different buses. Although this is part of the local public transportation dynamic, such events are rare, occurring in only 0.5% of cases. The presence of a mobile turnstile also impacts other variables, but with the same low probability.

Figure 5 illustrates the data distribution for these attributes. While the box plots reveal the presence of outliers, the overall distribution follows a skewed distribution, as expected in public transportation, with a majority of values concentrated near the lower stops and occasional extreme values indicating stops with significantly higher activity. The "n-boardings" and "n-alightings" histograms confirm that the previously discussed outliers are rare, low-frequency events.

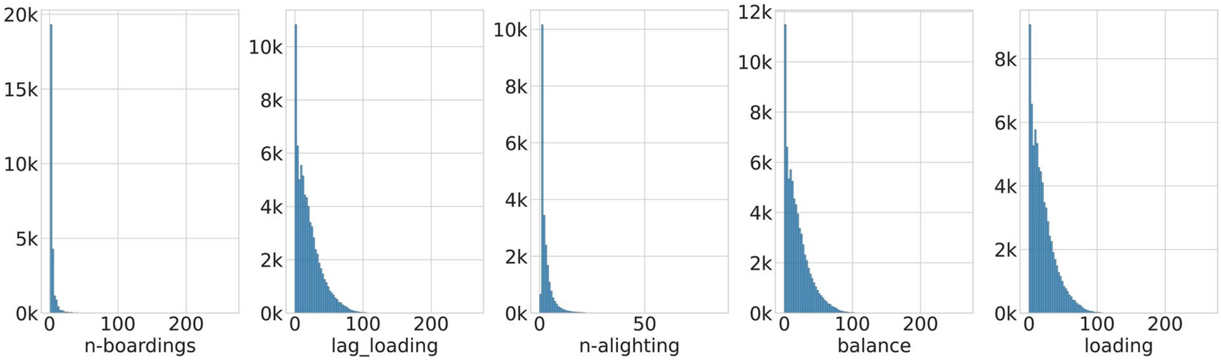


Fig. 5 Histogram describing different attributes related to passengers' behaviors.

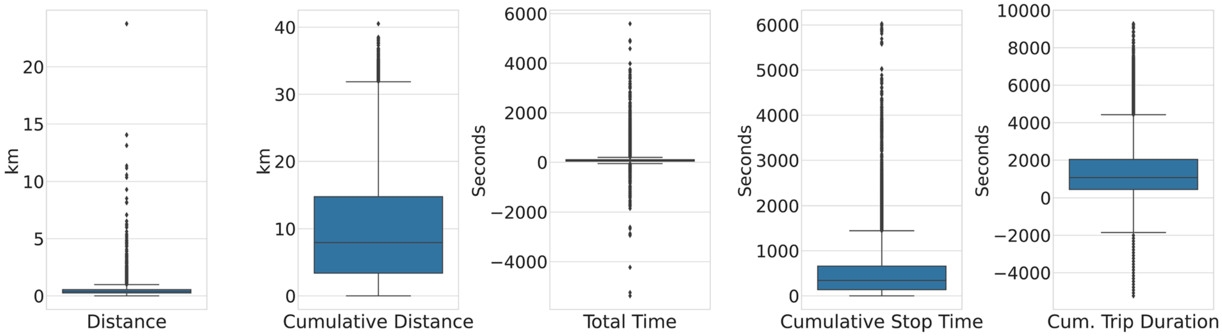


Fig. 6 Box plots summarizing descriptive statistics focused on trips' behaviors.

	Mean	Std	Min	25%	50%	75%	Max
Distance	0.458	0.472	0.000	0.251	0.367	0.546	23.758
Cumulative Distance	9.902	8.032	0.000	3.382	7.946	14.781	40.506
Total Trip Time	85.453	89.147	0.000	41.000	66.000	104.000	5,590.000
Cumulative Stop Time	448.724	400.953	0.000	137.000	344.000	660.000	6,027.000
Cumulative Trip Duration	1,379.671	1,172.054	0.000	447.000	1,083.000	2,038.000	9,281.999

Table 17. Descriptive statistics of the variables related to vehicle trips.

The exponential distribution pattern suggests that while most bus stops experience relatively low passenger movement, a few stops handle significantly higher volumes. Together, these figures emphasize the high variability in passenger demand, the presence of peak usage at specific stops, and the challenges associated with modeling and optimizing transit operations.

The next statistics (Table 17) extracted from SUNT summarize key aspects of vehicle trips, including individual and cumulative distances between stops, total trip time, cumulative stop time, and cumulative trip duration. In addition to cumulative data, SUNT also provides individual trip details. We compiled the accumulated values to highlight the volume of processed data.

Figure 6 presents box plots for the variables presented in Table 17, illustrating both general trends and the presence of outliers. In these cases, outlier values remain within the expected range. However, an analysis of "Total Trip Time" (Total Time) and "Cumulative Trip Duration" (Cum. Trip Duration) reveals negative values, which are inconsistent with time-based attributes. These anomalies arise due to real-world monitoring challenges, where delays in the internal clocks of devices collecting GTFS data can cause discrepancies. The apparent magnitude of these negative values, exceeding 4, 000, is a result of the large dataset. In practice, the actual time differences between stops amount to only a few seconds. To mitigate this issue, transportation companies aggregate data into one-hour intervals, and delayed clocks can be corrected through interpolation.

Likewise, the histograms in Fig. 7 highlight that the overall trip behavior follows the expected pattern for public transportation. The outliers with negative time values are rare events that can be excluded from analyses without impacting modeling performance. However, we chose to retain these values in SUNT to preserve the dataset's real-world nature and provide researchers with the flexibility to address them through alternative methods, such as estimating corrected values.

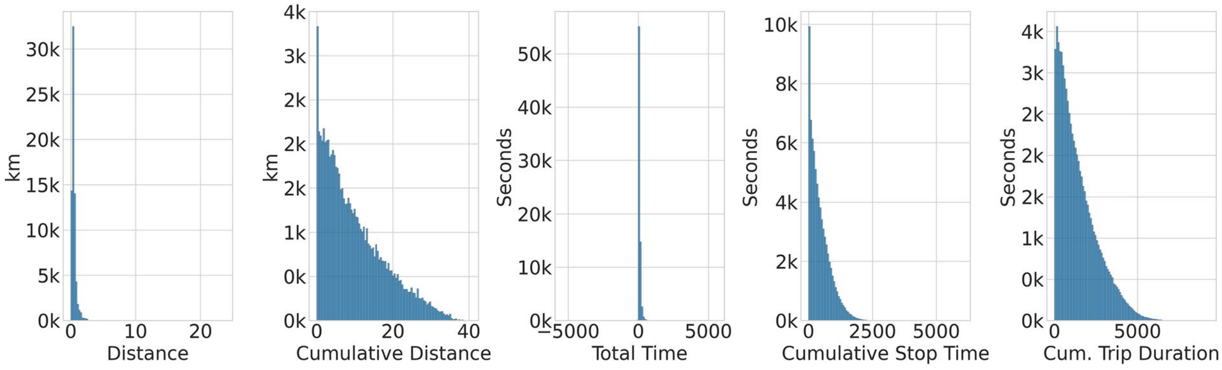


Fig. 7 Histogram describing different attributes related to trips' behaviors.

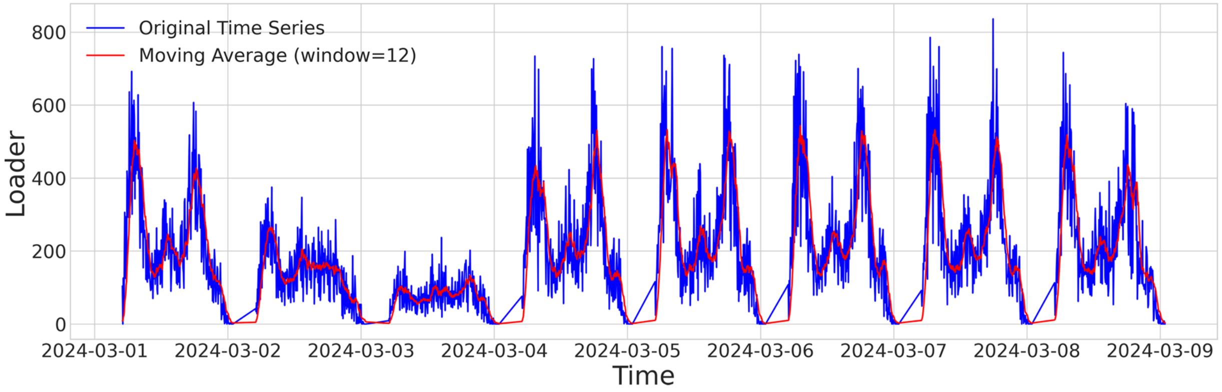


Fig. 8 In blue, the time series containing loading information in a bus station, collected every 5 minutes between March 1st, 2024, and March 9th, 2024. In red, we show the time series transformed by SMA using a window size of 12 observations.

The statistical patterns observed in the OD dataset are also reflected in the derived datasets, supporting further exploration and analysis. Building on the previous descriptive analysis, the SUNT dataset was examined as a time series to investigate temporal patterns, including cyclical and seasonal behaviors related to weekdays, weekends, and holidays. This analysis of temporal dynamics helps identify key characteristics, contributing to the validation of the dataset's consistency and overall quality.

Temporal Validation. To check the temporal characteristics of the SUNT dataset, we selected several stops and stations with high passenger flow and multiple connection options between lines and buses. Figure 8 shows a time series (in blue) whose observations represent the loading of passengers at a given station, collected every 5 minutes from March 1st, 2024 to March 9th, 2024. As one may notice, the time series is characterized by a significant frequency fluctuation as noise that may affect its modeling and prediction. To address this issue, a simple moving average (SMA) with a window size of 12 observations can be applied to smooth the time series, as illustrated by the red line. The key advantage of selecting this window size is its ability to capture seasonal and cyclical patterns. As expected, analyzing SUNT as a time series allows for modeling daily variations (higher frequencies during rush hours) and differences between business days and weekends (lower frequencies on weekends).

Similarly, to illustrate the seasonal relationships between stations and highlight the importance of the underlying graph structure, we selected the top five stations (nodes: 694, 2772, 1203, 592, and 561) with the highest passenger transit from a total of 2,871 possible stations represented in the SUNT graph. Figure 9 presents the smoothed time series for these stations, revealing how their patterns relate to one another while exhibiting distinct amplitude variations.

Continuing the focus on temporal relationships, Fig. 10 illustrates passenger volume over time across three transportation modes: BRT, subway, and regular bus. This analysis enables the investigation of how demand varies and interacts across different systems, supporting the development of forecasting models that incorporate multiple modes of transportation.

Spatial Validation. Leveraging the inherent spatial structure of SUNT represented as a graph, where nodes represent stop-time events and edges denote direct connections between them, we conducted two preliminary graph-based learning tasks, node classification and edge classification. These illustrative experiments highlight the dataset's spatio-temporal potential and its applicability to real-world scenarios, particularly in the context of route optimization.

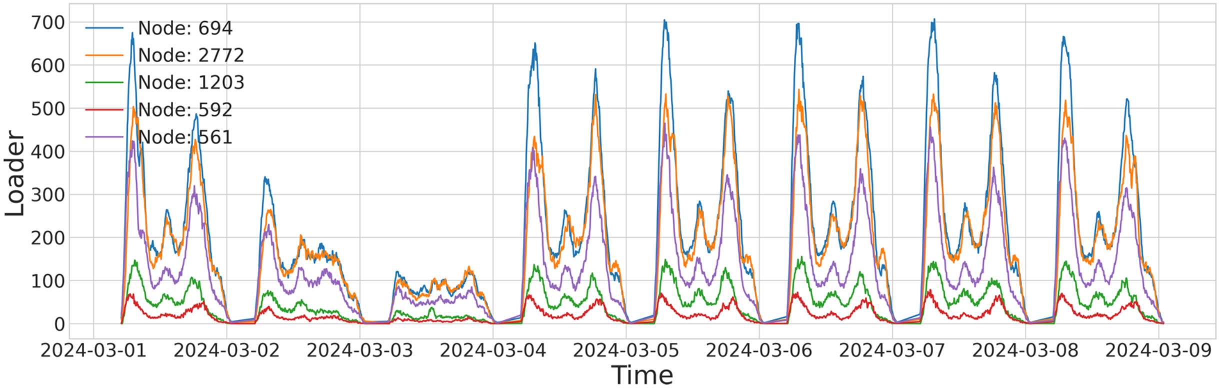


Fig. 9 Five time series with intense transit of passengers to illustrate the node regression task.

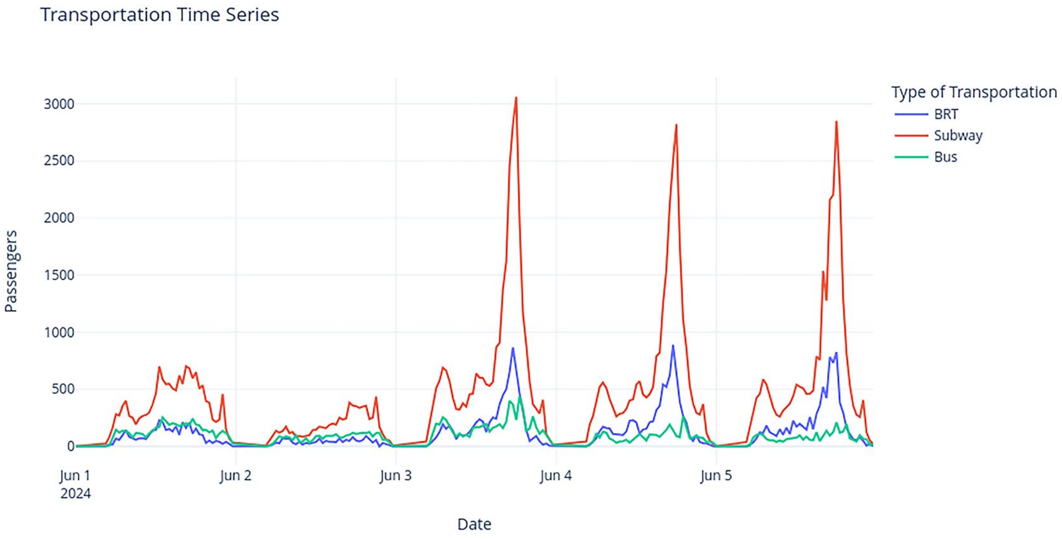


Fig. 10 Time series representing multiple transportation modes.

We selected eight features from the dataset, including passenger loading, mean velocity, distance between stops or stations, boarding and alighting counts, and the total number of lines, vehicles, and trips. To validate the spatial utility of SUNT, passenger loading was used as the target variable for node classification, while mean velocity served as the target for edge classification. These tasks are particularly relevant for route optimization, as they leverage the rich spatio-temporal information embedded in the graph structure to uncover patterns in passenger demand and traffic flow—key factors for improving transit planning, operational efficiency, and overall network performance.

For the node classification, passenger loading, a numerical variable, was discretized into four categories based on its quartiles: "maximum", "high", "medium", and "low". Similarly, for edge classification, mean velocity was binarized using the median value as a threshold. These class intervals were designed to illustrate local transportation demand. Alternative discretization schemes can be easily applied using our publicly available dataset. The spatial relationships between these attributes were then assessed using Graph Neural Networks (GNNs): Graph Convolutional Network (GCN)[16], Chebyshev spectral graph convolutional operator (CHEB)[17], SAmple and aggreGatE (SAGE)[18], and Graph Attention Networks (GAT)[19]. GCN employs a graph convolution operation to learn representations of nodes in a graph. A key characteristic of GCNs is weight sharing, meaning the same weight matrix is applied to all nodes. This is achieved through symmetric normalization of the adjacency matrix and the inclusion of self-loops to ensure each node incorporates its own features during the aggregation process. CHEB implements an efficient generalization of Convolutional Neural Networks (CNNs) to arbitrary graph structures by expressing graph convolutional filters as polynomials of the graph Laplacian $\mathbf{L}$ of a graph G. As discussed in[20], using a polynomial of degree M ensures that the output at each node is influenced by information from its M-hop neighborhood, enabling localized and scalable filtering on graphs. SAGE is a GNN architecture that, instead of learning individual embeddings for each node, learns a set of aggregation functions that operate over a node's local neighborhood to generate its embedding[18]. Each aggregator function combines information from neighbors at a specific distance, referred to as the number of hops or search depth, allowing the model to capture multi-scale structural and feature information. GAT is a type of GNN that incorporates

(a) Node Classification Results					
Model	**Accuracy**	**F1**	**MCC**	**Precision**	**Recall**
GCN	0.67 ± 0.06	0.64 ± 0.08	0.57 ± 0.08	0.70 ± 0.09	0.66 ± 0.07
CHEB	0.72 ± 0.06	0.68 ± 0.08	0.65 ± 0.07	**0.76 ± 0.06**	0.72 ± 0.06
SAGE	**0.76 ± 0.1**	**0.75 ± 0.1**	**0.68 ± 0.14**	0.75 ± 0.11	**0.76 ± 0.11**
GAT	0.67 ± 0.08	0.66 ± 0.09	0.56 ± 0.1	0.68 ± 0.09	0.67 ± 0.07
(b) Edge Classification Results					
Model	**Accuracy**	**F1**	**MCC**	**Precision**	**Recall**
GCN	**0.62 ± 0.03**	**0.62 ± 0.03**	**0.24 ± 0.06**	**0.62 ± 0.03**	**0.62 ± 0.03**
CHEB	0.61 ± 0.03	0.61 ± 0.03	0.23 ± 0.07	0.61 ± 0.03	0.61 ± 0.03
SAGE	0.61 ± 0.03	0.61 ± 0.03	0.23 ± 0.07	0.61 ± 0.04	0.61 ± 0.03
GAT	0.61 ± 0.03	0.61 ± 0.03	0.22 ± 0.06	0.61 ± 0.03	0.61 ± 0.03

Table 18. GNN results for node and edge classification.

attention mechanisms to learn node representations in a graph[19]. These mechanisms enable the model to assign different importance weights to each neighbor, allowing it to focus on the most relevant nodes during the aggregation process and thereby improving performance. GATs can also employ multiple attention *heads*, a concept closely related to the multi-head attention mechanism introduced in the transformer architecture by (Vaswani, Ashish, *et al*., 2017)[21]. Our validation of node and edge classification employed 10-fold cross-validation, a widely accepted method in machine learning to ensure robust and reliable evaluation. We used the same set of evaluation metrics for both tasks: Accuracy, F1-score, Matthews Correlation Coefficient (MCC), Precision, and Recall.

These comprehensive metrics are particularly important given the nature of transportation data, where certain categories (e.g., high passenger loads or low velocities) may be less frequent but critical for decision-making. The detailed evaluation framework thus ensures that the models' strengths and limitations are fully understood, guiding future improvements and practical applications of the SUNT dataset in transit system analysis.

Table 18 summarizes all results obtained from our illustrative experiments. In Table 18(a), SAGE achieved the best performance across all metrics, except for precision, where CHEB performed slightly better. In Table 18(b), all models exhibited very similar behavior, with a slight advantage for GCN. The results demonstrate satisfactory performance, with values exceeding 60%, which is notable given the inherent complexity of predicting numerical values on edges. The balanced nature of the dataset ensures that the models are effectively learning meaningful patterns. Furthermore, these outcomes highlight opportunities for future research, encouraging the development of novel GNN architectures and preprocessing strategies to further enhance performance using our dataset as a benchmark.

Importantly, the results from both node and edge classification tasks underscore the value of representing urban mobility data as a graph. The SUNT dataset, by encoding spatio-temporal relationships in a graph structure, enables the application of graph-based learning methods that can capture complex patterns, such as variations in passenger flow and average velocity across different segments of the network, that are often lost in traditional flat or tabular representations. These preliminary experiments GNNs demonstrate the feasibility and potential of such models to extract meaningful insights from our dataset. This reinforces not only the relevance of the graph-based representation itself, but also the utility of the SUNT dataset as a foundation for future research on graph-based learning tasks.

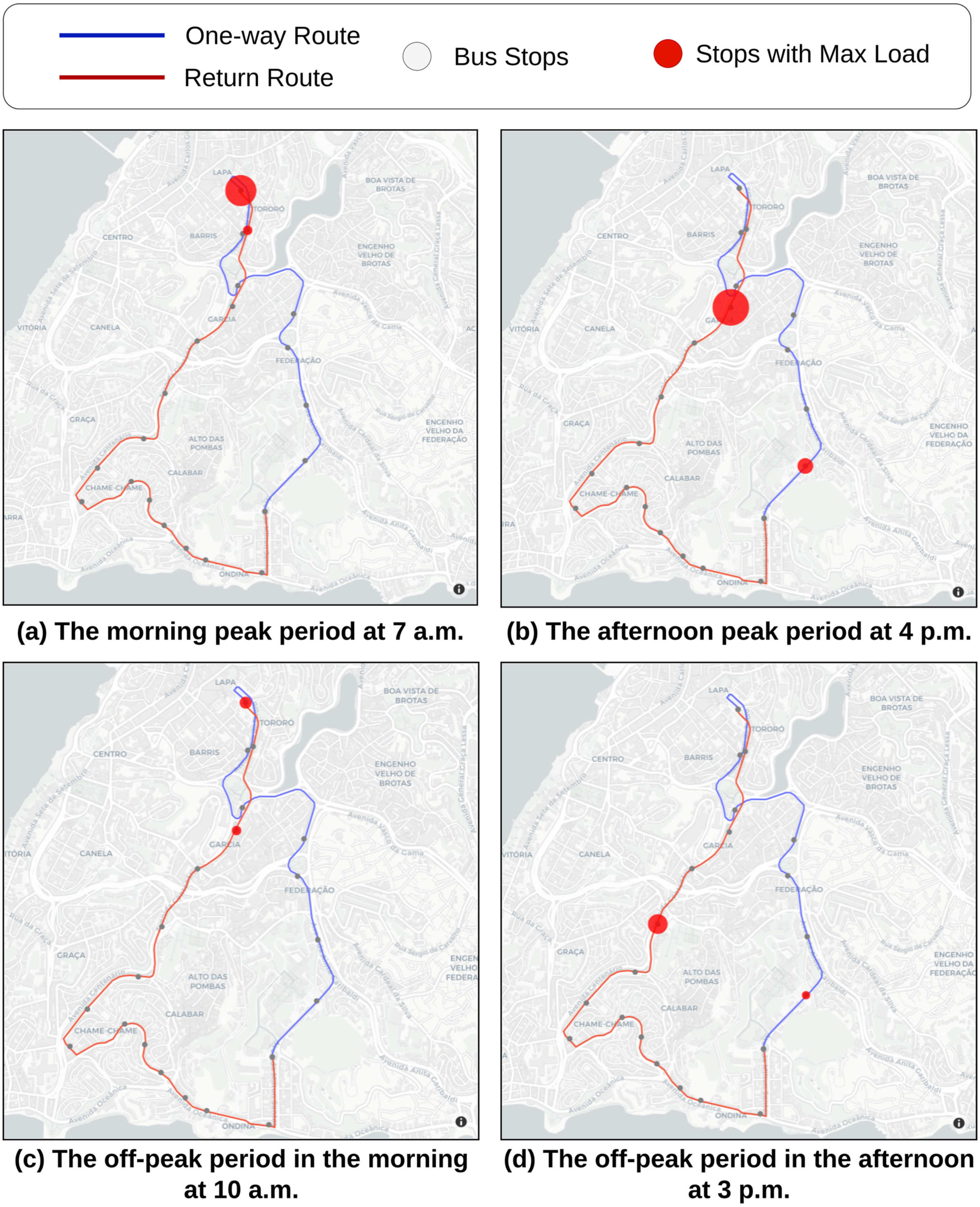
Transportation Planning Validation. By analyzing passenger loads through an Origin-Destination (OD) dataset and applying these insights to timetable planning, we demonstrate how SUNT is currently used to support transportation planning decisions. In the first example, the OD dataset was utilized to determine the maximum passenger load across different time intervals. According to (Ceder, 2016)[6], one of the fundamental objectives of transit service provision is to ensure sufficient capacity to accommodate the maximum number of passengers on board along the entire route within a given time period. Let us denote this time period (typically one hour) as j. Based on the peak-load factor concept, the required number of vehicles for period j is given by:

$$\mathscr{M}_j = \frac{\overline{P}_{mj}}{\gamma_j c} \tag{1}$$

In this equation, $\overline{P}_{mj}$ is the average maximum number of passengers (max load) observed on-board in period j, c denotes the vehicle's capacity (the total number of seats plus the maximum allowable standees), and γ_j is the load factor for period j, where $0 \leq \gamma_j \leq 1.0$.

To illustrate the importance of calculating the max load $\mathscr{M}_j$, we have selected a specific line and analyzed the max-load stops in Fig. 11 during four different time intervals: (a) 7 a.m. (morning rush hour), (b) 4 p.m. (afternoon rush hour), (c) 10 a.m. (morning off-peak), and (d) 3 p.m. (afternoon off-peak). By analyzing these maps, one can observe how the locations of maximum load stops vary across different time intervals. This information has been used to improve bus planning and allocation, enhancing service delivery to better meet the needs of the population. Notably, the highlighted maximum-load stops align well with the actual local transportation dynamics.

In our second example presented in Fig. 12, we illustrate how the information about max load can be used in practice to plan timetables, specifying which buses must be set as Express and Normal. Typically, when passenger



(a) The morning peak period at 7 a.m.

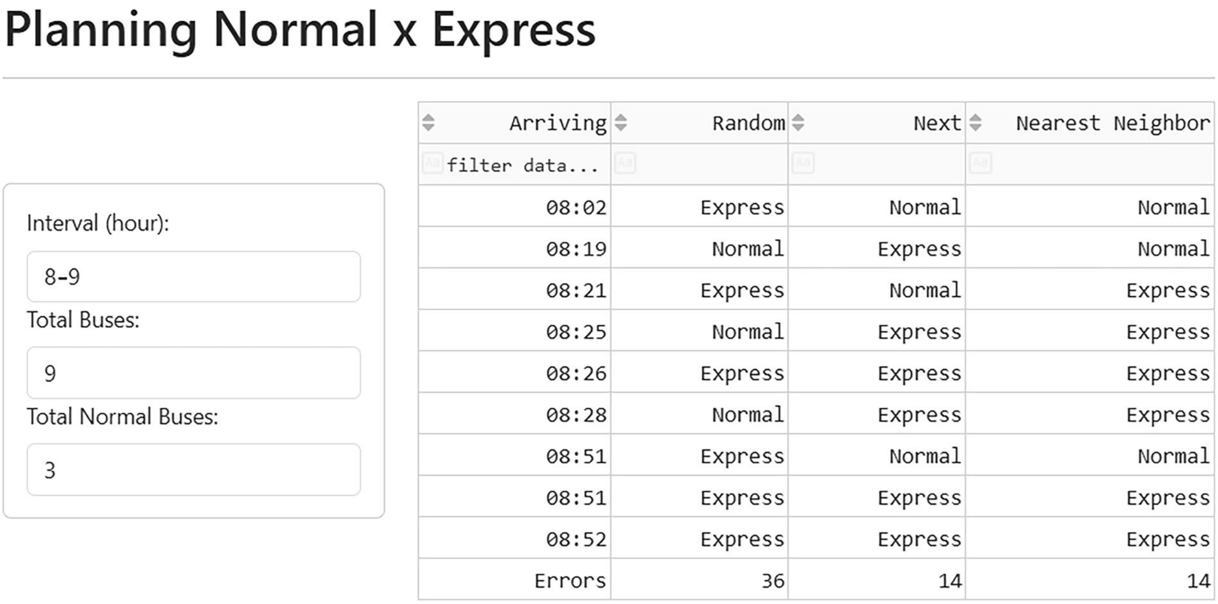
(b) The afternoon peak period at 4 p.m.

(c) The off-peak period in the morning at 10 a.m.

(d) The off-peak period in the afternoon at 3 p.m.

Fig. 11 Max load calculated for a specific line during different relevant time intervals.

data at stops/stations and along routes is unavailable, it becomes difficult to optimize bus services effectively. For example, during a rush hour interval (e.g., 8 AM-9 AM) with arrivals randomly defined, 9 buses (3 normal and 6 express) might be scheduled to serve passengers traveling from stop A to stop B, passing through intermediate stops. Without an estimate of maximum passenger load, all buses would need to return from stop B to stop A via the same route, stopping at all intermediate stations. Such a strategy has some problems: it wastes time and fuel, besides delaying the arriving time at A. Considering A is a neighborhood and B downtown, the amount of passengers from B to A is considerably lower during this rush time. Knowing the optimal number of buses required for the return trip and their appropriate schedules can significantly mitigate these issues.



Planning Normal x Express

Interval (hour):

8-9

Total Buses:

9

Total Normal Buses:

3

Arriving	Random	Next	Nearest Neighbor
filter data...			
08:02	Express	Normal	Normal
08:19	Normal	Express	Normal
08:21	Express	Normal	Express
08:25	Normal	Express	Express
08:26	Express	Express	Express
08:28	Normal	Express	Express
08:51	Express	Normal	Normal
08:51	Express	Express	Express
08:52	Express	Express	Express
Errors	36	14	14

Fig. 12 Planning timetables after calculating max load.

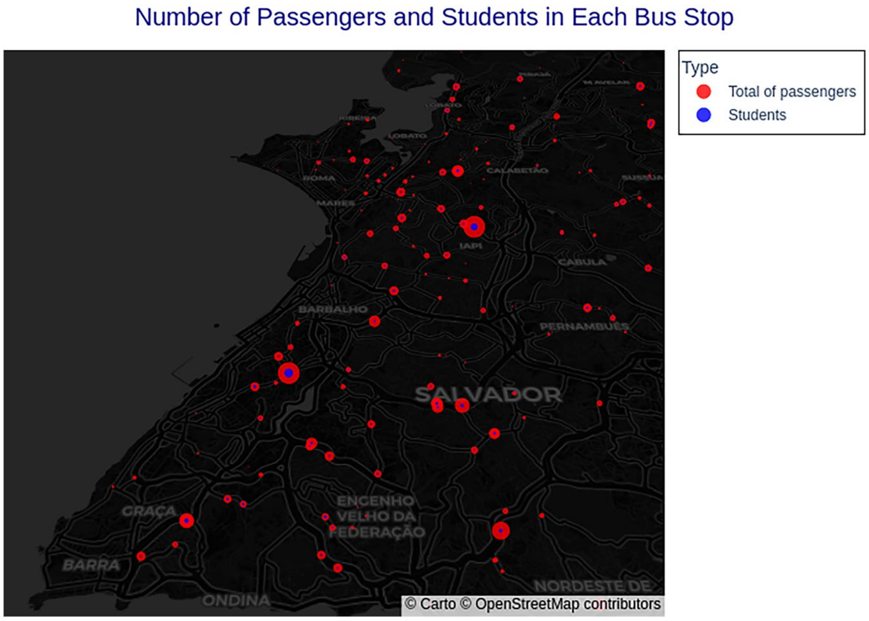


Fig. 13 Integrating SUNT with data from public schools. The total number of passenger per stop/station is shown in red. Students are represented by blue dots.

In Fig. 12, we illustrate three strategies for determining the bus type: (i) randomly selecting buses (Random information); (ii) dividing the hour into intervals based on the expected number of normal buses and designating the next bus within each interval as normal (Next information); and (iii) assigning normal buses based on the nearest neighbor approach, aiming to minimize the interval between them (Nearest information). The errors shown in this figure represent the differences, in minutes, between the estimated intervals – obtained using the random, next, and nearest neighbors strategies – and the optimal intervals. With this information, policymakers can effectively reduce passengers' waiting times at stops and stations while optimizing the management of bus transit within the city.

Cross-Dataset Validation. Another important validation was performed by integrating SUNT with other urban data sources. By combining it with a publicly available dataset on public schools in Salvador, made available by the Municipal Department of Education at https://dados.salvador.ba.gov.br/search?tags=educacao (in

Portuguese), it was possible to confirm the expected passenger load near schools during specific time periods. This dataset contains information on public schools, including their name, geographic coordinates (latitude and longitude), neighborhood, full address, and administrative details.

This example demonstrates the feasibility of integrating the SUNT dataset with external urban data sources, such as public school records, thereby expanding its range of applications beyond traffic analysis. As illustrated in Fig. 13, specialists analyze student passenger loads at bus stops or stations located near schools on specific days and times. In the visualization, the total number of passengers per stop or station is shown in red, while school locations are marked with blue dots. To enable this integration, each school is matched to its nearest stop or station based on geographic coordinates. Additionally, passengers in the SUNT dataset are identified as students through their transportation card classification.

This integrated analysis supports a variety of new research directions, such as evaluating school accessibility, understanding student mobility patterns, and informing policies for public transport planning in educational contexts.

Usage Notes

The dataset is licensed under Creative Commons (CC) BY 4.0. We encourage all interested researchers to download and use it to develop new AI-based methods and approaches aimed at enhancing urban mobility and public transportation management.

Code availability

The source code, models, and datasets, replicated from the Mendeley Repository, are freely available at https://github.com/LabIA-UFBA/SUNT. The repository is organized into a set of folders, each containing specific resources:

- **data** — raw data and graph-based representations;
- **data_design** — source code used to generate the datasets and train learning models;
- **docs** — dataset documentation;
- **images** — example of plots and visualizations summarizing dataset attributes;
- **integration** — examples of how to integrate other databases with the SUNT dataset;
- **models** — frozen AI models used to perform various prediction tasks;
- **outputs** — sample of model weights and prediction results;
- **stats** — Jupyter notebooks containing dataset statistics.

Received: 30 April 2025; Accepted: 23 July 2025;
Published online: 30 July 2025

References

1. Zhang, J. *et al*. Data-driven intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems* **12**, 1624–1639 (2011).
2. Rahmani, S., Baghbani, A., Bouguila, N. & Patterson, Z. Graph neural networks for intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems* **24** (2023).
3. Chattopadhyay, S. N. & Gupta, A. K. Unveiling critical transition in a transport network model: stochasticity and early warning signals. *Nonlinear Dynamics* 1–26 (2025).
4. An, Y. *et al*. Spatio-temporal multivariate probabilistic modeling for traffic prediction. *IEEE Transactions on Knowledge and Data Engineering* (2025).
5. Behura, A., Kumar, A. & Jain, P. K. A comparative performance analysis of vehicular routing protocols in intelligent transportation systems. *Telecommunication Systems* **88**, 26 (2025).
6. Ceder, A.*Public transit planning and operation: Modeling, practice and behavior* (CRC press, 2016).
7. Gordon, J. B., Koutsopoulos, H. N., Wilson, N. H. & Attanucci, J. P. Automated inference of linked transit journeys in London using fare-transaction and vehicle location data. *Transportation research record* **2343**, 17–24 (2013).
8. Wang, W., Attanucci, J. P. & Wilson, N. H. Bus passenger origin-destination estimation and related analyses using automated data collection systems. *Journal of Public Transportation* **14**, 131–150 (2011).
9. An, S., Wang, L., Yang, H., Wang, J. *et al*. Discovering public transit riders' travel pattern from GPS data: a case study in harbin. *Journal of Sensors* **2017** (2017).
10. Lee, S., Lee, J., Bae, B., Nam, D. & Cheon, S. Estimating destination of bus trips considering trip type characteristics. *Applied Sciences* **11**, 10415 (2021).
11. Barry, J. J., Freimer, R. & Slavin, H. Use of entry-only automatic fare collection data to estimate linked transit trips in New York city. *Transportation research record* **2112**, 53–61 (2009).
12. Assemi, B., Alsger, A., Moghaddam, M., Hickman, M. & Mesbah, M. Improving alighting stop inference accuracy in the trip chaining method using neural networks. *Public Transport* **12**, 89–121 (2020).
13. Mohammed, M. & Oke, J. Origin-destination inference in public transportation systems: A comprehensive review. *International Journal of Transportation Science and Technology* **12**, 315–328 (2023).
14. Ferreira, M. V. *et al*. SUNT: Salvador Urban Network Transportation. Mendeley Data, https://data.mendeley.com/datasets/85fdtx3kr5/1, https://doi.org/10.17632/85fdtx3kr5.1 (2025).
15. ITDP. The online brt planning guide Last Access: June 2024 (2016).
16. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. International Conference on Learning Representations (ICLR), https://arxiv.org/abs/1609.02907 (2016).
17. Defferrard, M., Bresson, X. & Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems (NeurIPS)***29** (2016).
18. Hamilton, W., Ying, Z. & Leskovec, J. Inductive representation learning on large graphs. *Advances in neural information processing systems (NeurIPS)***30** (2017).
19. Veličković, P. *et al*. Graph attention networks. International Conference on Learning Representations (ICLR), https://arxiv.org/abs/1710.10903 (2017).
20. Hamilton, W. L. Graph representation learning. *Synthesis Lectures on Artifical Intelligence and Machine Learning* **14**, 1–159 (2020).
21. Vaswani, A. *et al*. Attention is all you need. *Advances in neural information processing systems (NeurIPS)***30** (2017).

22. Jiang, R. *et al*. Spatio-temporal meta-graph learning for traffic forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, 8078–8086 (2023).
23. Wu, Z., Pan, S., Long, G., Jiang, J. & Zhang, C. Graph wavenet for deep spatial-temporal graph modeling. International Joint Conference on Artificial Intelligence (IJCAI), https://www.ijcai.org/proceedings/2019/0264.pdf (2019).
24. Cini, A., Marisca, I., Bianchi, F. M. & Alippi, C. Scalable spatiotemporal graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, 7218–7226 (2023).
25. Du, Y. *et al*. Graphgt: Machine learning datasets for graph generation and transformation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)* (2021).
26. Chen, W. *et al*. Multi-range attentive bicomponent graph convolutional network for traffic forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, 3529–3536 (2020).
27. Shao, Z., Zhang, Z., Wang, F., Wei, W. & Xu, Y. Spatial-temporal identity: A simple yet effective baseline for multivariate time series forecasting. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 4454–4458 (2022).
28. Oreshkin, B. N., Amini, A., Coyle, L. & Coates, M. Fc-gaga: Fully connected gated graph architecture for spatio-temporal traffic forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, 9233–9241 (2021).
29. Zhang, J., Zheng, Y. & Qi, D. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 31 (2017).
30. Bai, L. *et al*. Spatio-temporal graph convolutional and recurrent networks for citywide passenger demand prediction. In *Proceedings of the 28th ACM international conference on information and knowledge management*, 2293–2296 (2019).
31. Xie, P. *et al*. Spatio-temporal dynamic graph relation learning for urban metro flow prediction. *IEEE Transactions on Knowledge and Data Engineering* (2023).
32. Liu, L. *et al*. Physical-virtual collaboration modeling for intra- and inter-station metro ridership prediction. *Transactions on Intelligent Transportation System* **23**, 3377–3391, https://doi.org/10.1109/TITS.2020.3036057 (2022).
33. Ren, L., Chen, J., Liu, T. & Yu, H. Od-enhanced dynamic spatial-temporal graph convolutional network for metro passenger flow prediction. In *International Conference on Neural Information Processing*, 72–85 (Springer, 2023).
34. Zhang, J., Chen, F., Cui, Z., Guo, Y. & Zhu, Y. Deep learning architecture for short-term passenger flow forecasting in urban rail transit. *IEEE Transactions on Intelligent Transportation Systems* **22**, 7004–7014 (2020).
35. Klar, R. & Rubensson, I. Spatio-temporal investigation of public transport demand using smart card data. *Applied Spatial Analysis and Policy* **17**, 241–268 (2024).
36. Bui, K.-H. N., Yi, H. & Cho, J. Uvds: a new dataset for traffic forecasting with spatial-temporal correlation. In *Asian Conference on Intelligent Information and Database Systems*, 66–77 (Springer, 2021).

Acknowledgements

R.A.R. was supported by CNPq (Brazilian National Council for Scientific and Technological Development) grants to [404771/2024-6, 406354/2023-5, 312755/2023-6]. T.N.R. was supported by CNPq grant 313053/2023-5 and Maria Emilia Foundation grant to 01/2023. R.A.R. and T.N.R were supported by INCITE FAPESB grant to PIE0002/2022. I.F.C.F. was supported by CNPq grant to 402842/2023-5. M.V.F. was supported by FAPESB (Bahia Research Foundation) grant [1589/2021]. R.A.R. and J.N. were supported by UFBA/CNPq 68/2022 - MAI/DAI.

Author contributions

M.S., R.A.R., and M.V.F conceptualized the study. M.S., and M.V.F organized the data. M.V.F., M.S., R.A.R., T.N.R., and I.F.C.F. implemented the algorithms, conducted the experiments, and performed the analyses. M.V.F., M.S., R.A.R., T.N.R., I.F.C.F., J.N, J.G., and A.B. interpreted the results and wrote the manuscript. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to R.A.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.