Intelligent Transportation Infrastructure, 2024, liae019
https://doi.org/10.1093/iti/liae019
Data Note

UrbanBus: a fine-grained dataset for bus ridership analysis

Liwen Ke[1], Xiangyu Guo[2], Tianrui Li[2], Chongshou Li[2,*]

[1]SWJTU-Leeds Joint School, Southwest Jiaotong University, Xi'an Road, Pidu District, 611756, Sichuan, China
[2]School of Computing and Artificial Intelligence, Southwest Jiaotong University, Xi'an Road, Pidu District, 611756, Sichuan, China
[*]Corresponding author. lics@swjtu.edu.cn

Bus ridership analysis plays a critical role in intelligent transportation systems. Numerous methods have been proposed for this purpose. However, compared to other public transportation modes (e.g. metro, taxi), there is a lack of public datasets for bus ridership analysis, including forecasting, scheduling, and ~~clustering~~. To address this issue, this note presents a novel bus ridership dataset, UrbanBus, consisting of approximately 727 million automatic fare collection transactions collected over six months from the intelligent transportation system in a major Asian city. The unique characteristic of the dataset is its fine granularity at the transaction level. The time granularity is at the second level, allowing for various aggregation operations for high-level analysis. Additionally, the bus routes and pairwise distances between stops are provided. This dataset is expected to benefit the intelligent transportation research community and is publicly available at: https://github.com/ableyyyx/UrbanBus.

Keywords: dataset; bus ridership analysis; intelligent transportation.

1. Introduction

As global urbanization accelerates, city populations are expanding, resulting in increased demand for urban transportation. A well functioning urban public transportation system is crucial for enhancing the quality of life for city residents, mitigating traffic congestion, and reducing environmental pollution. Buses, in particular, offer advantages over subways due to their lower construction costs and broader coverage area, which better meets the travel needs of a larger number of residents. They also offer greater flexibility in responding to urban expansion, new developments, and residential districts' demands, providing significant potential for future urban growth (Gao and Zhu, 2022). Conversely, transportation issues such as traffic congestion are becoming increasingly severe in metropolitan areas (Kumar and Raubal, 2021). Analyzing bus ridership is essential for improving the efficiency and effectiveness of urban public transportation systems (Vuchic, 2002). For example, accurate bus ridership forecasts can enable early interventions to enhance the operational efficiency of public transport and provide valuable insights for route planning and scheduling.

An essential foundation of bus ridership analysis is the dataset. Bus ridership analysis is a typical spatiotemporal data analysis problem and has been extensively studied in the existing literature (Li, 2018; Bowen, 2019; Han, 2019; Ye et al., 2019; Li, 2020; Luo, 2020; Liu, 2021; Luo, 2021; Zou, 2022). Various methods have been applied, including traditional time series forecasting methods (Ye et al., 2019; Liu, 2021) and deep-learning-based methods (Li, 2018; Bowen, 2019; Han, 2019), such as recurrent neural networks (RNNs), gated recurrent units (GRUs), and long short-term memory (LSTM). Datasets are crucial for these data-intensive methods. This note introduces a stop-level bus ridership dataset, UrbanBus, consisting of approximately 727 million automatic fare collection (AFC) transactions collected from around six million smart cardholders over six months from the bus transportation systems in a major city in Asia. (To maintain privacy, the city name has been withheld. This does not affect the nature of the data or the insights and results derived from this study.)

Although intelligent transportation datasets are popular and widely available in the existing literature (Wang et al., 2023), publicly accessible datasets specifically for bus ridership analysis are insufficient. Due to various issues such as privacy concerns, only analytic results and insights are typically provided (Van Oort et al., 2015; Shoman et al., 2020; Farahmand et al., 2023). To address this gap, this note introduces a fine-grained bus ridership dataset with several key advantages:

1. **Large-scale and fine-grained data**: The dataset includes approximately 727 million AFC transactions collected from around six million smart cardholders over the period from 1 August 2017 to 31 March 2018. The time granularity of the AFC transactions is at the second level. The dataset covers 5,144 bus stops, allowing for various aggregation operations for high-level analysis.
2. **Comprehensive route and stop data**: The dataset includes bus routes and pairwise distances between stops, enabling stop-level analysis. Descriptive analysis and insights are provided to facilitate the development of various analytical models.
3. **Privacy protection**: To address privacy concerns, encryption rules have been discussed and applied. This ensures that the

Received: July 28, 2024. **Revised:** November 19, 2024. **Accepted:** December 4, 2024

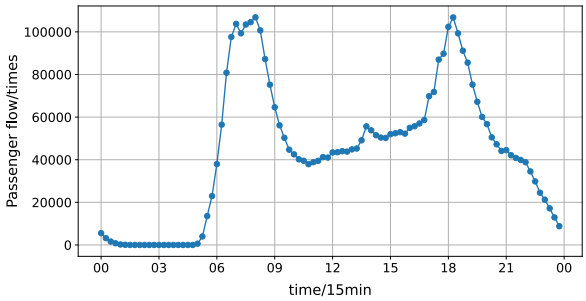

Figure 1 Passenger flow statistics every 15 minutes.

dataset is released legally while maintaining the integrity of the data patterns.

4. **Supporting resources**: In addition to the dataset, basic input and output Python source codes are provided. These resources are publicly available at https://github.com/ableyyyx/UrbanBus.

2. ~~Data cleaning and processing~~

The UrbanBus dataset is collected from a real-world system, and the raw data may contain various issues and inconsistencies that could negatively affect subsequent analysis, modeling, and visualization. Therefore, data cleaning and processing are essential steps. We have undertaken several cleaning operations to ensure the quality and reliability of the dataset.

- Handling missing values: We checked for missing values in critical fields, such as timestamps and bus stop information, to ensure data completeness.
- Handling outliers: Outliers, including unreasonable passenger flow values or anomalous timestamps, were identified and addressed. For example, instances where the alighting time was earlier than the boarding time were removed from the dataset.
- Removing duplicates: Duplicate records were detected and eliminated to avoid skewing the analysis results.
- Data type conversion: We verified that data types were accurate, particularly for timestamps and passenger flow fields, to ensure proper data interpretation.
- Data consistency: The dataset was checked for adherence to expected standards, ensuring that passenger flow values were positive integers and timestamps were formatted correctly.
- Data standardization: To ensure comparability across different data sources, the data were normalized to a consistent unit or scale. In the UrbanBus dataset, bus operations occur throughout the entire 24-hour period. However, passenger flow data before 5:30 AM was significantly lower compared to other times, as shown in Fig. 1. To maintain consistency and data stability, a time window from 05:30 AM to 11:30 PM was selected for standardization. AFC transaction data were aggregated into 15-minute intervals during this period for model training, ensuring that the model was trained on data from relatively stable time periods.
- Saving cleaned data: The cleaned data were saved into a new file for subsequent analysis.

To ensure the anonymity of bus stops and their associated access points while maintaining the uniqueness and distinguishability of the data, we designed a systematic encryption rule. The specific rules are as follows:

1. **Counting and sorting of bus stops and access points**: Each bus stop may have multiple access points, such as different boarding and alighting locations. We first count the number of access points for each bus stop and then sort the bus stops in descending order based on the number of access points.
2. **Encryption representation rules**: According to the sorting results, we assign unique identifiers with numerical indices to each access point of the bus stops. The identifiers are generated in increasing alphabetical order, starting with single letters from A to Z, then double-letter combinations (e.g. AA to AZ), followed by triple-letter combinations, and so on. For access points within the same bus stop, consecutive natural numbers starting from 1 are used as indices (e.g. A_1, A_2, etc.), ensuring that all access points within the same bus stop have ordered and unique identifiers. For example, if the top-ranked bus stop has seven access points, their encrypted identifiers would be A_1, A_2, A_3, A_4, A_5, A_6, and A_7.

Bus route information is crucial for model accuracy. However, the relevant bus route information was not recorded during the collection of the UrbanBus dataset. To address this challenge, two methods for obtaining bus route information were considered: one is web scraping from official websites, and the other is constructing bus routes from the existing data. Although web scraping from official websites is a straightforward approach, the UrbanBus dataset was collected in 2017. The seven-year time span may have resulted in significant changes in the current bus route information, making it unsuitable for the model's needs. To obtain more accurate bus route information, this note proposes an innovative bus route construction algorithm by analyzing features in the existing dataset. This approach aims to compensate for the missing route information and provide more accurate inputs for the model. The bus route construction algorithm and its implementation steps are as follows:

1. **Step 1: Construct bus operation trajectory subsets**: Begin by extracting the bus route number feature to compile statistics on the set of bus route sets. Divide the dataset into different subsets according to these bus route sets. Further, categorize each bus route subset into bus operation trajectory subsets based on the **direction** of travel, **bus_reg_num**, and **bus_trip_num**. Finally, add a trajectory time feature to each subset, representing the travel time between consecutive stops within each trajectory.
2. **Step 2: Construct the bus route dataset**: For each bus route's operation trajectory subset, use the number of stops as a feature to sort the trajectories in descending order. Select the first trajectory in the sorted list as the baseline route. Then, complete the operation trajectory using the time difference algorithm.

There are two key challenges in constructing the algorithm. First, the UrbanBus dataset comprises card transaction data, lacking direct information on the exact arrival and departure times at bus stops. To address this gap, we examined the busiest bus route in the dataset over the course of a day, analyzing the time differences between the first and last card transactions at the same stop, as well as between the first card transactions at adjacent stops. As shown in Figs 2 and 3, the average dwell time of a bus at a stop is approximately 20 seconds, while the average travel time between stops is around 180 seconds, significantly longer than the dwell time. Therefore, the first card transaction at each stop can be used to approximate the bus arrival time, and the time difference between the first card transactions at consecutive stops can be used to estimate the travel time between stops. This

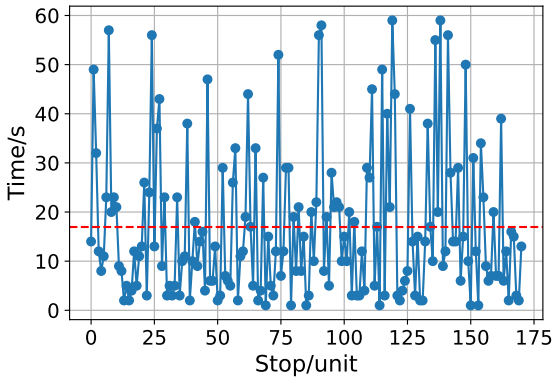

Figure 2 Stop dwell time statistics.

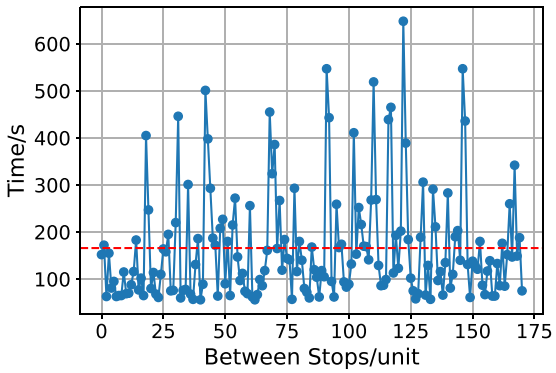

Figure 3 Travel time statistics between stops.

method allows us to approximate travel time information along the bus route.

Secondly, multiple scenarios arise when completing the operation trajectory for the baseline stops. As illustrated in Fig. 4, three trajectories travel from stop A to stop F, with black dots indicating stops where the bus halts and passengers board or alight. The baseline trajectory, *tr0*, has the most stops, namely {A, C, E, F}. Sub-trajectory *tr1* includes stops {A, C, D}, and sub-trajectory *tr2* includes stops {B, D, E}. The complete operation trajectory, *tr*, is {A, B, C, D, E, F}, which requires adding stops {B, D} to the baseline trajectory. Two situations arise in this context:

1) As seen in sub-trajectory *tr1*, the difference set between trajectories *tr0* and *tr1* is *trdiff* = {*D*}. Since the preceding stop exists in the baseline trajectory, the time difference between stops C and D in *tr1* is compared with the corresponding time differences in *tr0*. If the time difference is greater, stop D is inserted into the baseline trajectory *tr0*.
2) In sub-trajectory *tr2*, the difference set between trajectories *tr0* and *tr2* is *trdiff* = {*B*}. As there is no preceding stop for B in *tr2*, the time difference after B is compared with the preceding time differences for stop D in *tr0*. If this time difference is greater, stop B is inserted into the baseline trajectory *tr0*.

The route construction algorithm, detailed in Algorithm 1, demonstrates its effectiveness by generating bus routes that are more complete and better aligned with the current dataset compared to those obtained from web scraping. This advantage highlights the algorithm's capability to accurately reconstruct the actual bus routes within the dataset, thereby enhancing the precision and reliability of the bus route information.

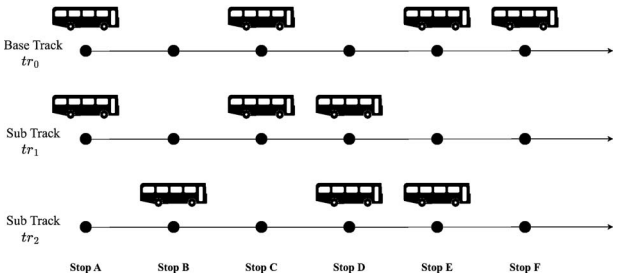

Figure 4 Bus sub-route diagram.

Algorithm 1 Public Transport Route Construction Algorithm

Input: Dataset $L = \{l_0, l_1, \ldots, l_n\}$
Output: Constructed public transport routes L
$i \leftarrow 0$
while $i \leq n$ **do**
  Extract all trajectories $TR = \{tr_0, tr_1, \ldots, tr_m\}$ of bus line L_i
  $j \leftarrow 0$
  while $j \leq m$ **do**
    Merge the boarding and alighting trajectories tr_j and sort by time feature
    Remove duplicate stops, keeping the first occurrence in tr_j
    Use the initial stop time as the time reference, calculate time differences for subsequent stops
    $j \leftarrow j + 1$
  end while
  Compute the base trajectory tr_0 using time difference method
  $i \leftarrow i + 1$
end while
Obtain the constructed public transport routes L

The above algorithm leverages raw trajectory data to reconstruct bus routes in a manner that is more comprehensive and accurate compared to relying on static data sources. Its systematic approach to merging, sorting, and duplicating trajectory data ensures a high-quality representation of public transport routes. By explicitly considering temporal features and multiple trajectories, the algorithm is able to produce a robust base trajectory that reflects real-world bus movements. Incorporating additional enhancements such as noise handling and dynamic adaptation could make the algorithm even more powerful in practical applications for planning and analysis of public transport.

3. Data visualization

3.1 Spatiotemporal distribution characteristics analysis

3.1.1 *Overall characteristic analysis*

The line chart of UrbanBus boarding passenger flow from 1 January 2018 to 28 February 2018, shown in Fig. 5, reveals clear periodic variations, excluding specific holidays. The decline in passenger flow on 1 and 2 January and the significant fluctuations from 28–30 January can be attributed to the New Year's Day and Spring Festival holidays. During these holidays, the substantial changes in passenger numbers are likely due to alterations in daily travel patterns, as people chose to stay home or travel for holiday activities rather than commute. Excluding the anomalous data from these holidays, the passenger flow on weekdays (Monday

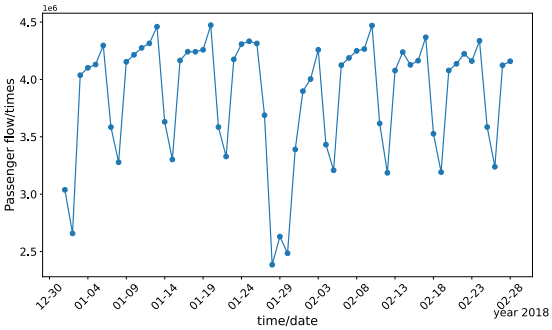

Figure 5 Boarding passenger flow.

to Friday) shows an increasing trend, often peaking on Fridays. This peak could be explained by increased social and leisure activities on Friday evenings as the workweek concludes, resulting in higher utilization of the bus system. Conversely, passenger flow significantly drops on weekends, especially on Sundays, which typically record the lowest ridership of the week. This pattern may reflect a reduction in routine activities over the weekend or a preference for alternative modes of travel during this period.

The analysis of the overall boarding passenger flow data indicates a clear periodic pattern. As shown in Figs 6 and 7, which present line charts of boarding passenger flow for two randomly selected bus stops and bus routes respectively, the periodic nature is evident. However, the trends in boarding passenger flow at the route level align more closely with the overall boarding passenger flow trend, whereas the periodicity observed at different bus stops appears more random. This suggests that route-level boarding passenger flow reflects a collective behavioral trend, while the boarding passenger flow at individual stops may be influenced by more localized factors, resulting in more random periodicity.

3.1.2 *Weekly variation characteristic analysis*

An in-depth analysis of the overall variation characteristics of passenger flow reveals that the total boarding passenger flow generally follows a weekly cycle. Thus, conducting a detailed analysis of the variations within a week becomes an effective strategy. Understanding the characteristics of data within a cycle allows for more accurate fitting and optimization of periodic models, enhancing the model's predictive performance. To ensure the representativeness of the weekly variation analysis, the period from 16–22 January was selected as a typical cycle for analysis. This choice helps avoid the impact of special occasions, providing more accurate and comprehensive insights into weekly passenger flow variations.

Figure 8 shows line charts of boarding passenger flow at two randomly selected stops, recorded at 15-minute intervals during the chosen week. On weekdays, the passenger flow trend typically exhibits a bimodal distribution, with two distinct peaks occurring during morning and evening rush hours. These periods likely correspond to commuting times when people are traveling to and from work. In urban traffic flow, these morning and evening peaks are commonly observed as people start and finish their workdays. However, the peak hours differ significantly between the two stops, indicating that different locations have distinct passenger flow patterns. Certain stops may experience higher passenger flow at specific times due to their proximity to commercial areas, schools, or residential neighborhoods. During off-peak hours, particularly from late night to early morning, the passenger flow significantly decreases, which aligns with typical daily life patterns, as most people are not traveling during these hours.

Any non-periodic spikes or drops may indicate special events, such as nearby activities, traffic accidents, or weather changes, which could temporarily affect passenger flow. On weekends, the two stops do not exhibit distinct peaks and troughs, suggesting that passenger flow patterns on weekends differ from those on weekdays. Overall passenger flow may be lower, reflecting reduced travel demand on weekends. If an increase is observed, it could be attributed to leisure activities, such as shopping, entertainment, or attendance at specific events, leading to higher passenger flow at certain times (e.g. afternoon or evening).

Figure 9 shows line charts of boarding passenger flow for two randomly selected routes, recorded at 15-minute intervals over the chosen week. Similar to the points made for Fig. 8, both routes follow a daily time cycle, reflecting typical passenger flow patterns, such as morning and evening rush hours. However, while passenger flow at individual stops may exhibit spikes associated with specific boarding or alighting locations, the peak at a particular stop might be related to its location, such as proximity to commercial areas, office zones, or schools, resulting in higher passenger flow in the morning and afternoon. In contrast, route-level passenger flow represents the cumulative passenger flow across all stops along the route, which may not show the same sharp peaks as individual stops. Instead, route-level passenger flow tends to be smoother, as it aggregates the flow from many stops, unless there is a major event or special occurrence along the route. Route flow fluctuations may better reflect overall travel patterns rather than the localized variations seen at individual stops.

To more accurately examine the characteristics of weekdays and weekends within a week, Fig. 10 displays the boarding passenger flow at 15-minute intervals for two randomly selected routes from Monday to Sunday. The figure illustrates that, on each weekday, passenger flow patterns follow a consistent trend, indicating relatively stable commuting behavior. In contrast, on weekends, passenger flow patterns become more dispersed, reflecting the diversity of travel purposes. This confirms that weekdays have a distinct periodicity, with a daily cycle. Due to the more pronounced fluctuations in passenger flow at individual stops, a detailed analysis of weekday and weekend characteristics was not conducted for random stops.

Analyzing passenger trip duration is crucial for passenger flow forecasting, as it provides key insights into passenger behavior and public transportation usage patterns, enhancing the accuracy and reliability of predictive models. Specifically, analyzing trip duration can reveal travel habits, such as when passengers begin their journeys, the duration of their trips, and when trips end. This helps identify peak periods at specific times and locations. Understanding trip duration can inform predictions of passenger density on particular routes or in specific areas, thereby forecasting future ridership demand trends.

As shown in Fig. 11, the distribution of passenger trip durations over the course of a week clearly indicates that the majority of trips are completed within 15 minutes. A detailed statistical analysis reveals that trips lasting less than 15 minutes account for approximately 70% of all trips. More specifically, the average trip duration over the week is 13.8 minutes, with a median of 9.9 minutes and a mode of 5.2 minutes. These statistics highlight a clear trend: people tend to make shorter trips, particularly within the 15-minute range. The values of the mean, median, and mode further emphasize this tendency.

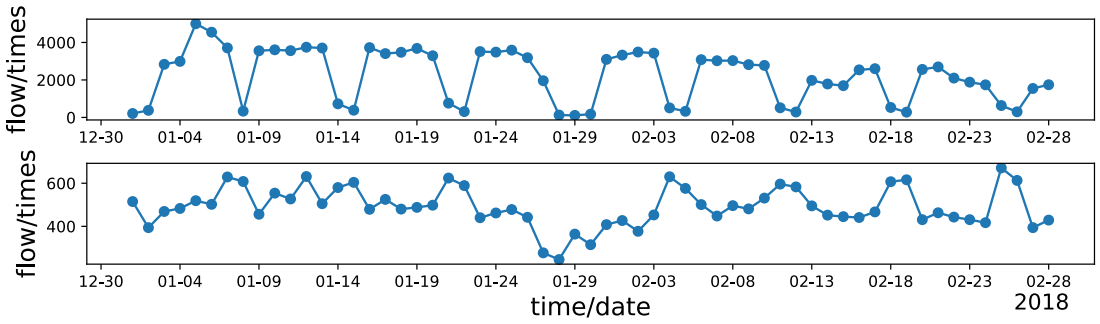

Figure 6 Stop-level boarding passenger flow.

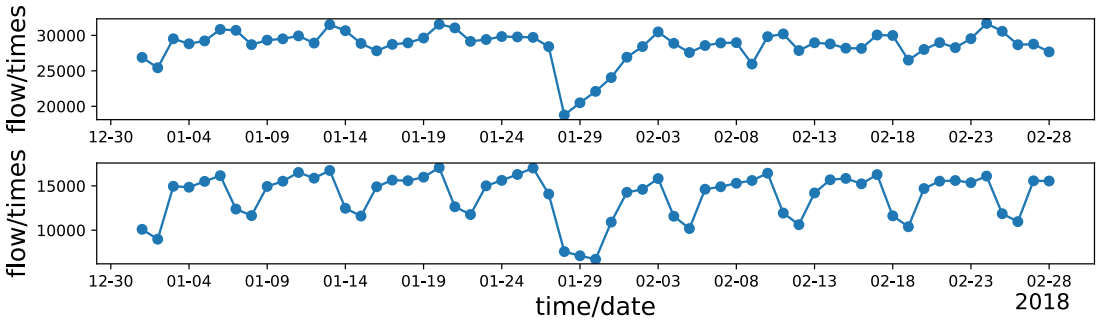

Figure 7 Route-level boarding passenger flow.

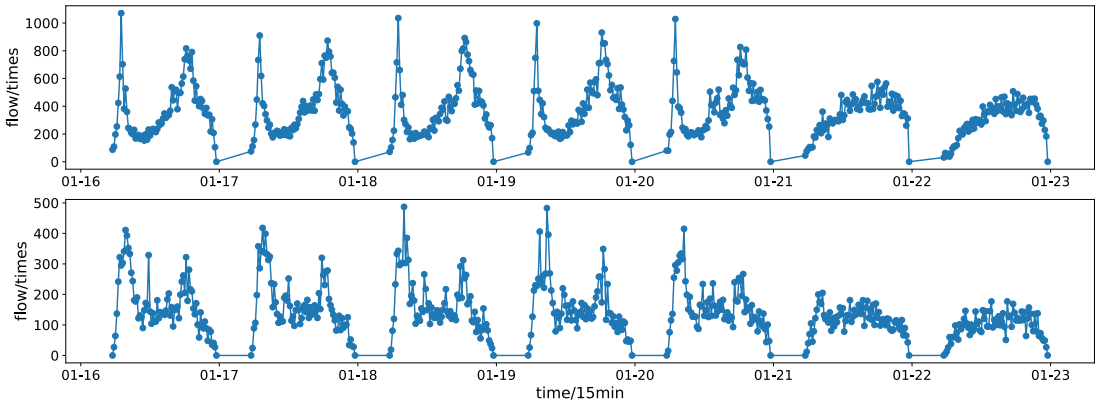

Figure 8 Stop-level boarding passenger flow in 15 minutes.

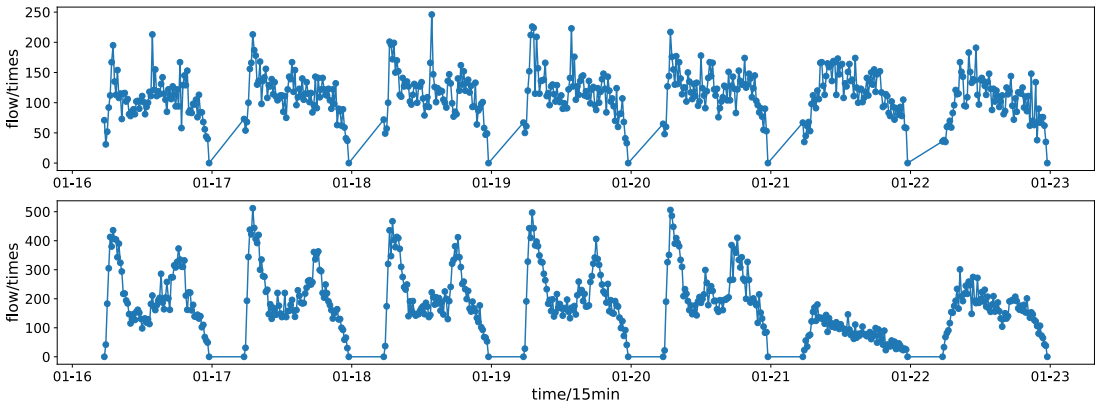

Figure 9 Route-level boarding passenger flow in 15 minutes.

3.1.3 *Daily variation characteristic analysis*

This section adopts an analytical method that categorizes weekday passenger flow distribution patterns into single-peak, double-peak, and non-peak types (Wang, 2016). The boarding passenger flow of buses on Tuesday 17 January was analyzed accordingly.

- **Single-peak type** Figure 12 illustrates passenger flow graphs for routes and stops that exhibit a single-peak pattern. In

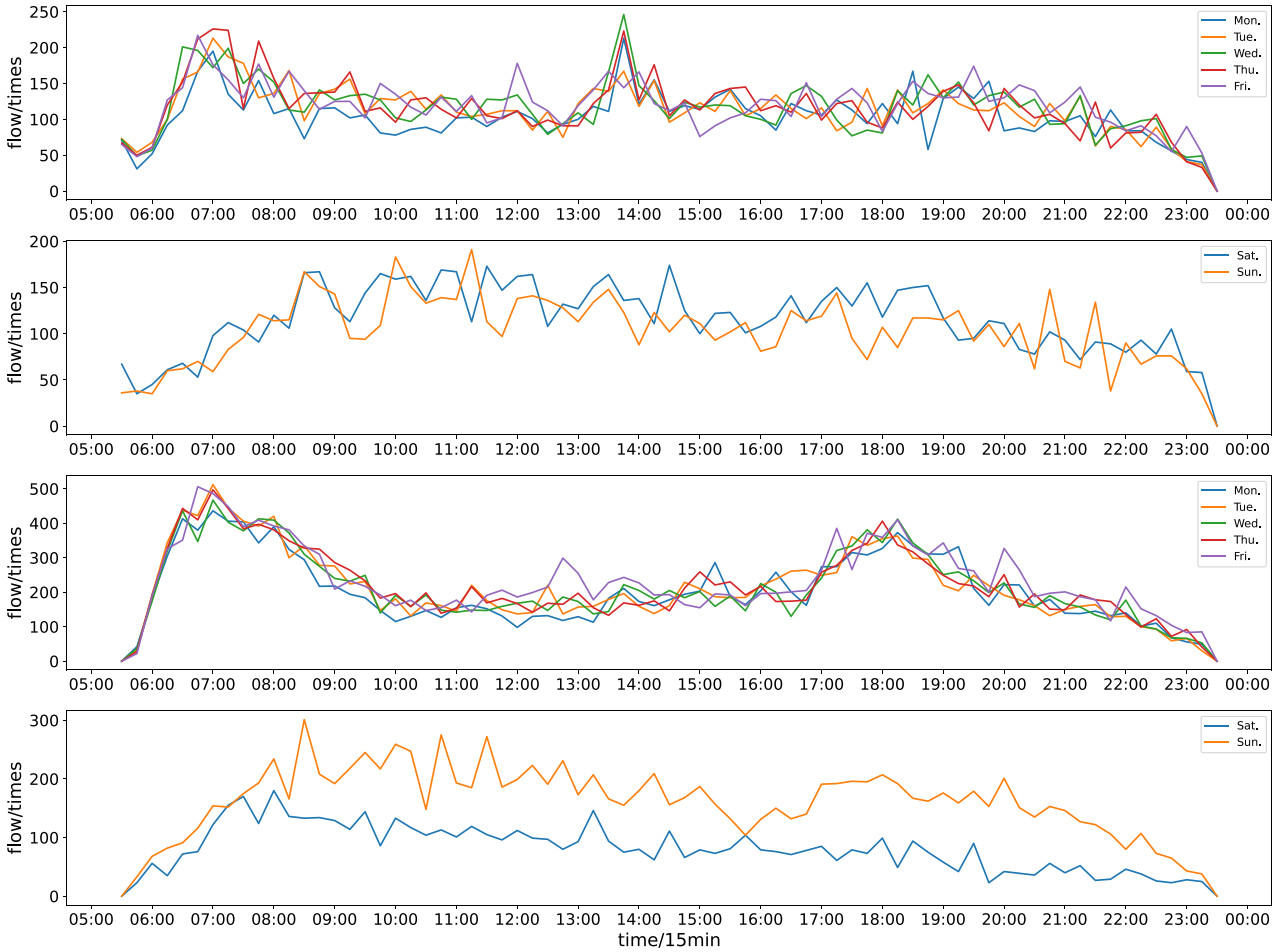

Figure 10 Route-level boarding passenger flow in 15 minutes within one week.

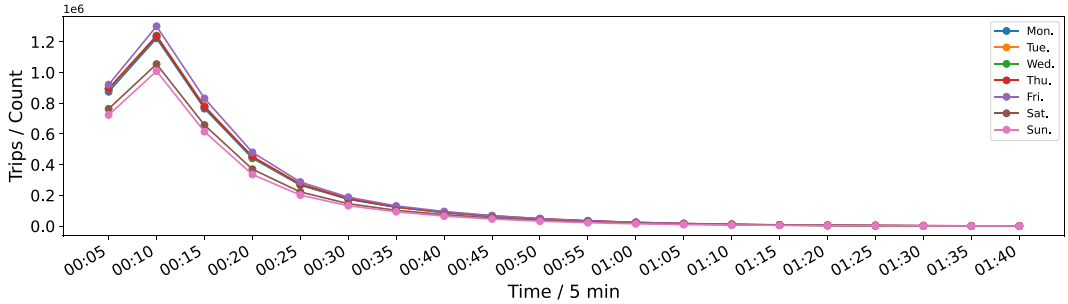

Figure 11 Passenger trip duration statistics.

Fig. 12(a), passenger flow rises rapidly in the morning, reaching a peak, and then gradually decreases, typically reflecting the commuting pattern during the morning rush hour. The peak is likely to occur between 7:00 AM and 8:00 AM, coinciding with the time that many people begin their workday. After the peak, the passenger flow drops sharply but remains relatively steady throughout the rest of the day, likely due to ongoing but less concentrated passenger movement. The absence of a similar evening peak suggests that passengers return home at more dispersed times or that the route experiences significant passenger turnover. In Fig. 12(b), a particular stop shows a sharp peak in the morning, followed by a rapid decline, which aligns with a pattern of people leaving the stop within a concentrated time window. After the morning peak, passenger flow remains low throughout the day, suggesting that the stop is primarily used for morning commutes rather than continuous daily usage. Unlike the route-level flow, the stop does not exhibit a notable evening increase, possibly indicating that passengers do not heavily use this stop in the evening or that there are more popular alighting points nearby.

- **Double-peak type** Figure 13 depicts passenger flow graphs for routes and stops with a double-peak pattern. Both graphs show a similar trend: passenger flow begins to increase gradually from 5:00 AM, reaching the first peak between 7:00 AM and 8:00 AM, reflecting the start of the morning commute. Subsequently, passenger flow decreases and remains low around midday, as most commuters have reached their workplaces. In the late afternoon, passenger flow rises again, peaking during the evening rush hour (approximately 5:00 PM to 6:00 PM), reflecting the demand for public transit as people commute home. The double-peak distribution indicates that

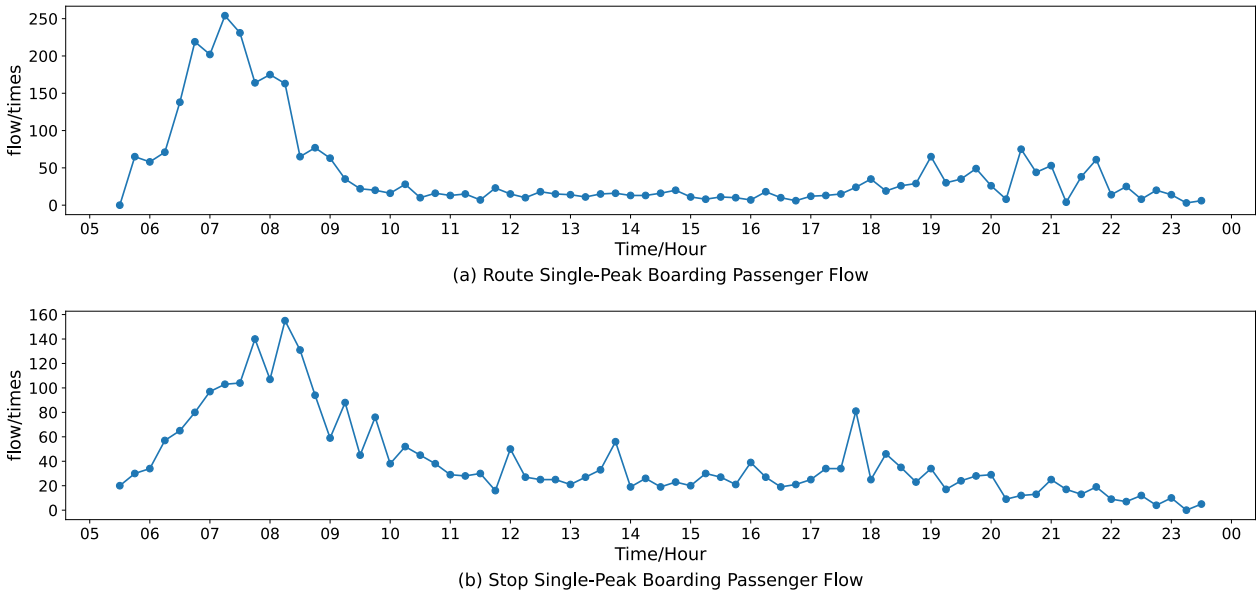

Figure 12 Single-peak passenger flow.

the stop and route mainly serve commuters who rely on public transportation during the morning and evening rush hours.

- **Non-peak type** Figure 14 shows passenger flow graphs for routes and stops without distinct peaks. For these routes, there are no pronounced morning or evening rush peaks. This may suggest that the route serves multiple stops where passenger boarding and alighting are relatively evenly distributed, or the areas serviced by the route are not strictly commercial or office zones, resulting in less commuting pressure. For stops, the absence of distinct peaks may imply that the stop is not a major commuter hub or that it serves a more diverse community, such as residential or mixed-use areas, where travel demand is spread throughout the day.
- Statistical analysis of daily passenger flow across stops and routes reveals that route-level flow often exhibits a double-peak pattern, while single-peak and non-peak patterns are relatively less common. In contrast, stop-level flow tends to show either a single-peak or a more stable non-peak pattern. This phenomenon could be explained as follows: for routes, the double-peak pattern reflects the typical daily commuting behavior, with concentrated peaks in the morning and evening corresponding to people's commute times. This is because routes typically span multiple stops and regions, catering to the travel needs of a large number of commuters, resulting in two distinct peaks. On the other hand, the flow patterns at individual stops are more influenced by their geographical location, surrounding environment, and the functionality of nearby destinations. For instance, a stop near a commercial area might experience a sharp increase in passenger flow in the morning as people arrive but lack a significant evening peak because most people stay at work during the day. Additionally, a stop located in a residential area may have a more evenly distributed demand throughout the day, leading to no obvious peaks.

3.2 Spatiotemporal correlation analysis

In this section, Pearson's correlation coefficient (Goh, 2007) is used to assess the correlations between routes and stops on weekdays and weekends, as well as between weekdays and weekends, and between adjacent stops. As discussed in Section 3.1, boarding passenger flow on bus routes and at stops exhibits periodicity, with distinct characteristics on weekdays and weekends within each cycle. To gain a more intuitive understanding of the degree of correlation, passenger flow at 15-minute intervals from 16–22 January was analyzed. Figure 15 illustrates the correlation analysis of boarding passenger flow between randomly selected weekday stops, Fig. 16 shows the correlation between weekday and weekend stops, and Fig. 17 depicts the correlation between weekend stops. Because routes and stops exhibit similar characteristics during the week, only stop-level correlations are presented.

Passenger flow at adjacent bus stops is expected to show strong spatial correlation. An analysis of boarding passenger flow at two adjacent stops on the same bus route over specific time intervals across two months demonstrates this strong spatial correlation, as shown in Fig. 18.

Although bus routes do not have spatial adjacency relationships and cannot form a physical adjacency matrix, passengers' travel habits typically revolve around their places of residence, work, and other daily destinations. These habitual travel paths often lead to a high degree of correlation between neighboring or functionally complementary bus routes. For instance, passengers might board one route and then transfer to a nearby route to reach their destination. An analysis of boarding passenger flow over specific time intervals on two particular bus routes across two months also shows strong spatial correlation, as illustrated in Fig. 19.

3.3 Analysis of differences between passenger flow at bus stops and on routes

Passenger flow prediction at bus stops primarily focuses on the number of passengers boarding and alighting at each stop. The data source is card swipe records at the stops, which exhibit strong spatiotemporal periodicity and are significantly affected by factors such as morning and evening rush hours, holidays, and weather.

In contrast, passenger flow prediction on bus routes emphasizes the movement of passengers across the entire route,

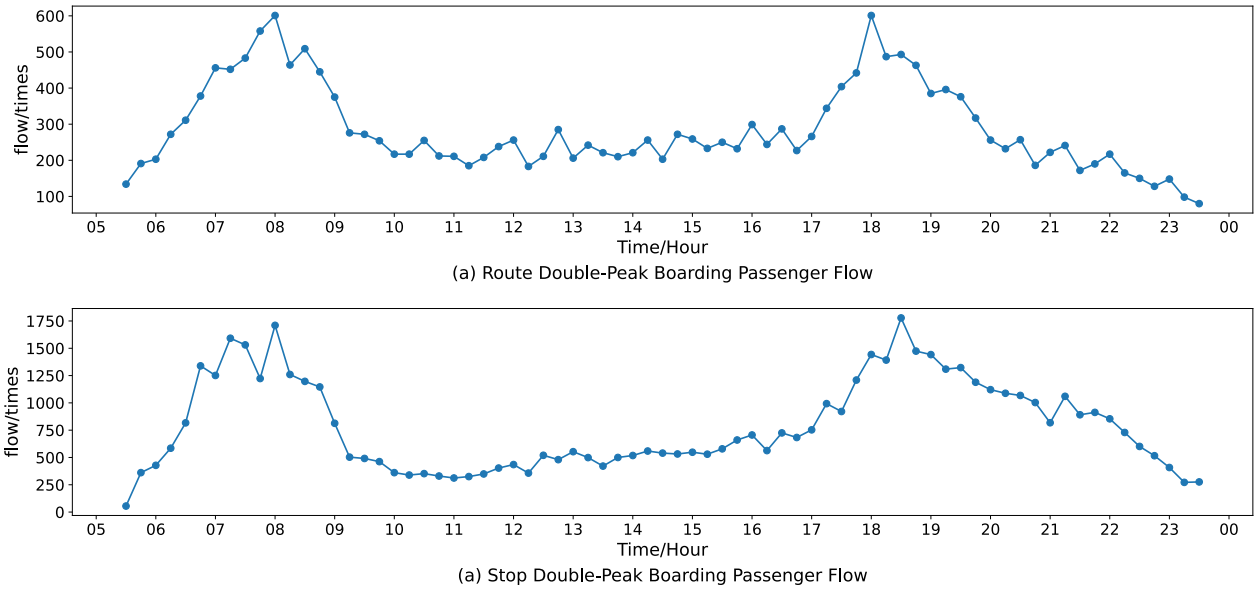

Figure 13 Double-peak passenger flow.

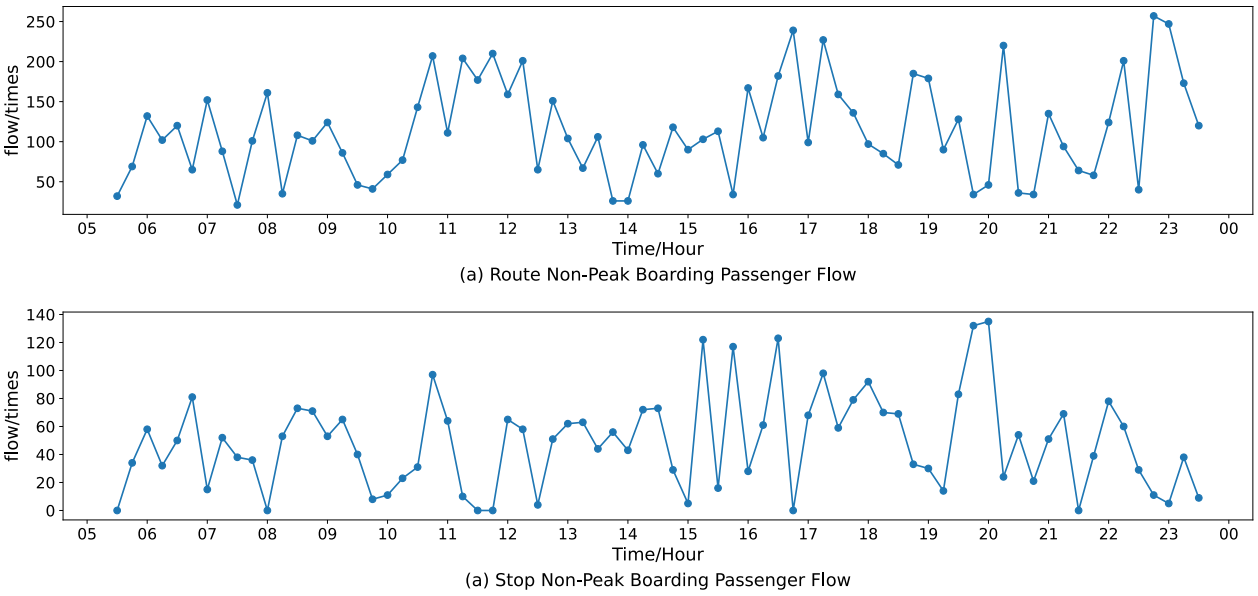

Figure 14 Non-peak passenger flow.

including the flow between stops. The data are primarily derived from card swipe records along the route, highlighting the overall trends of passenger movement and being influenced by factors such as route length, the stops along the way, and travel time.

As shown in Table 1, there are notable differences in data characteristics between bus stop and route passenger flow prediction in the UrbanBus dataset. Passenger flow prediction at stops involves a larger number of nodes, lower average passenger flow, moderate variance, and a high proportion of zero values, indicating high data sparsity and volatility. Thus, models capable of handling highly volatile and sparse data, such as time series analysis and deep-learning models, are necessary. In contrast, route passenger flow prediction involves fewer nodes, higher average passenger flow, greater variance, and a lower proportion of zero values, with more concentrated data. These types of data are suitable for models like graph neural networks and self-attention mechanisms, which focus on capturing complex dynamic relationships and overall flow trends along the route.

4. Data description

The dataset presented in this study offers a comprehensive collection of urban bus data, including a symmetric distance matrix (with distances in kilometers), bus routes, and smart card transaction records over a six-month period. Derived from the processed automatic fare collection (AFC) transaction data for an Asian city, the dataset spans 182 days from 1 October 2017 to 31 March 2018, covering over 727 million records, 123 weekdays, and 59 days of rest (weekends and holidays). An overview of this dataset is provided in Table 2. The dataset comprises two main types of data: 1) AFC transactions and 2) bus stop-related information,

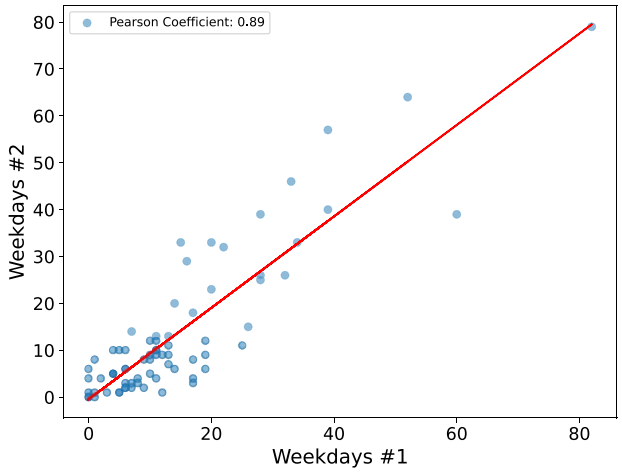

Figure 15 Relationship between boarding passenger flow at stops on weekdays.

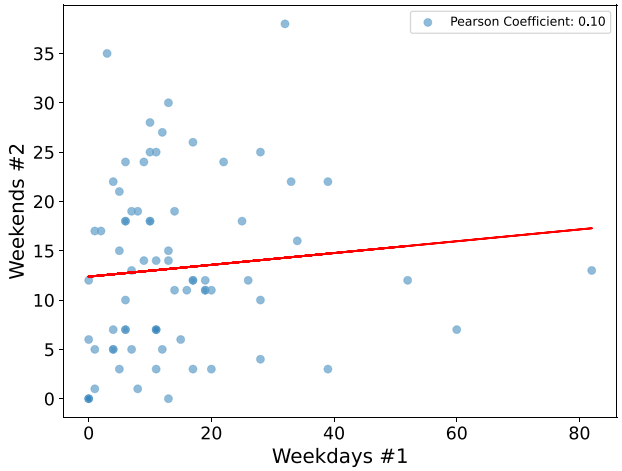

Figure 16 Relationship between boarding passenger flow at stops on weekdays and weekends.

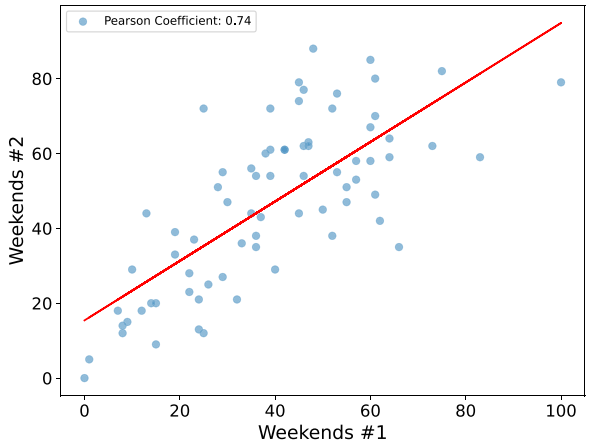

Figure 17 Relationship between boarding passenger flow at stops on weekends.

including: a) pairwise distance matrix of bus stops and b) bus routes.

For each AFC transaction, there are 13 attribute fields, detailed in Table 3. Most of these fields are self-explanatory. Notably, for card_type, the value set and explanations are provided in

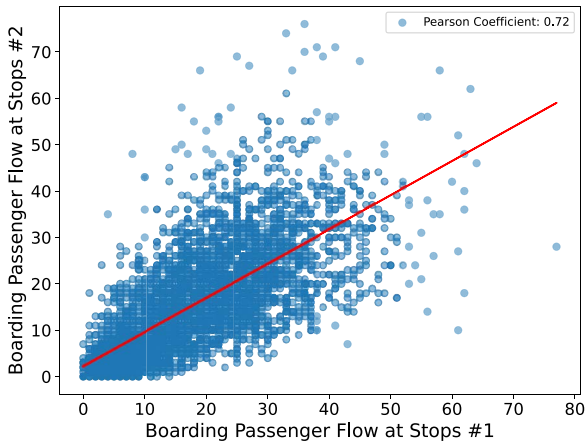

Figure 18 Boarding passenger flow at adjacent stops.

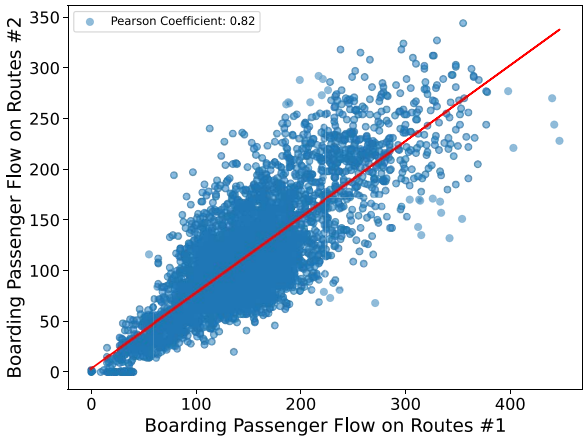

Figure 19 Route passenger flow relationship.

Table 4. Each bus stop has a unique name and ID, stored in BusStopList.CSV, as shown in Table 9. Table 6 displays attributes field of BusRoutes.pickle. Example data from DistanceMatric.CSV, BusRoutes.pickle, and BUS_DATA_MAR_2018.CSV are provided in Table 5, 7, and 8, respectively. Additionally, guidance on how to read these data files using Python is available on the GitHub page at https://github.com/ableyyyx/UrbanBus.

5. Benchmark results of ridership predictions

In the previous section, we discussed how UrbanBus can be utilized for clustering to identify passenger flow patterns and for scheduling optimization to improve public transit operations.

In this section, to comprehensively assess the dataset's applicability, UrbanBus is evaluated using six benchmark models for bus ridership prediction tasks. The benchmark models are: AGCRN (Bai, 2020), DCRNN (Li, 2017), MTGNN (Wu, 2020), Graph WaveNet (Wu, 2019), STGCN (Yu et al., 2018), and T-GCN (Zhao, 2020). The experimental results, presented in Table 10, analyze the performance of these models at intervals of 15 minutes, 30 minutes, 45 minutes, and 60 minutes, demonstrating the suitability of the dataset for time series forecasting. By presenting these case studies and potential use cases, we emphasize the extensive relevance of the dataset and its unique contribution to the advancement of research in public transportation.

Table 1. Stop and route feature data.

Data type	Number of nodes	Mean passenger flow	Passenger flow variance	Proportion of zero values	Time interval
Stop	5144	10.67	36.63	27.60%	15 min
Route	307	120.95	96.78	3.10%	15 min

Table 2. Dataset overview.

Type	File name	Description
AFC transaction	BusData.CSV	ACF transaction attribute field list
	BUS_DATA_MAR_2018.CSV	AFC transaction records of March 2018
	BUS_DATA_FEB_2018.CSV	AFC transaction records of February 2018
	BUS_DATA_JAN_2018.CSV	AFC transaction records of January 2018
	BUS_DATA_DEC_2017.CSV	AFC transaction records of December 2017
	BUS_DATA_NOV_2017.CSV	AFC transaction records of November 2017
	BUS_DATA_OCT_2017.CSV	AFC transaction records of October 2017
Bus stop-related information	BusStopList.CSV	Names and IDs of all bus stops
	DistanceMatrix.CSV	Linear distance matrix between any two stops
	BusRoutes.pickle	Dictionary data of bus routes

Table 3. ACF transaction attribute field list (BusData.CSV)

Field	Data type	Information
card_num	Int	The card number
card_type	String	The type of card
travel_mode	String	The mode of travel
bus_service:num	String	The unique service number
direction	String	The direction of travel
bus_trip_num	String	The unique trip number
bus_reg_num	String	The unique registration number
boarding_stop_stn	String	The name of the boarding stop
alighting_stop_stn	String	The name of the alighting stop
ride_start_date	String	The start date of the ride
ride_start_time	String	The start time of the ride
ride_end_date	String	The end date of the ride
ride_end_time	String	The end time of the ride

Table 4. Values of field card_type.

card_type value	Meaning	Explanation
A	Adult	The cardholder is an adult
S	Student	The cardholder is a student
C	Child	The cardholder is a child
SC	Aged people	The cardholder is aged over 65

Table 5. Example data for DistanceMatrix.CSV (symmetric matrix with linear distances in kilometers).

BUS_STOP	HS_1	FBW_1	HN_2	ETE_1	EVL_1	ESZ_1	ESY_1	...
HS_1	**0**			...				...
FBW_1	3.28	**0**			...			...
HN_2	21.77	23.89	**0**			...		...
ETE_1	12.29	12.16	16.08	**0**				...
EVL_1		...		...	**0**			...
ESZ_1	...		...		...	**0**		...
ESY_1	15.84	16.21	13.03	4.36	23.64	17.37	**0**	...
...		...		...		...		**0**

Table 6. Attribute field of BusRoutes.pickle.

Field	Data type	Information
key	String	The unique service number of bus
stop_stn	String [DataFrame]	The unique name of the bus stop
ride_time	Datetime [DataFrame]	The timestamp of the bus arrival
sub	Timedelta [DataFrame]	The time difference to the start stop

Table 7. Example data for BusRoutes.pickle.

Key	stop_stn	ride_time	Sub
SER_52f1	NR_1	13:58:02	0 days 00:00:00
	BMY_1	13:59:13	0 days 00:01:11
	DIR_1	14:00:54	0 days 00:02:52
	...	...	...
...	...	...	...

Table 8. Example data for AFC transaction (BUS_DATA_MAR_2018.CSV).

card_num	card_type	travel_mode	bus_service:num	direction	bus_trip_num	bus_reg_num	boarding_stop_stn	alighting_stop_stn	ride_start_date	ride_start_time	ride_end_date	ride_end_time
0	A	Bus	SER_52f1	Start	TRIP_6b86	REG_736e	JH_1	DGE_1	2018/3/10	9:38:48	2018/3/10	9:50:19
1	S	Bus	SER_52f1	Start	TRIP_7902	REG_c076	IU_2	CU_1	2018/3/9	16:00:50	2018/3/9	16:04:58
2	SC	Bus	SER_52f1	Start	TRIP_d0ff	REG_03ff	HB_1	CHU_1	2018/3/18	6:42:57	2018/3/18	6:48:24
...	...	...	...	...	...	...	...	...	...	...	...	...

Table 9. Attribute fields of BusStopList.CSV.

Field	Data type	Information
fid	Int	Bus stop ID (unique)
bus_stop	String	Bus stop name (unique)

Table 10. Performance comparison of bus ridership prediction baseline methods on the UrbanBus dataset at 15-minute, 30-minute, 45-minute, and 60-minute intervals.

Model	15 min			30 min			45 min			60 min		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
AGCRN (Bai, 2020)	4.22	8.86	0.54	4.24	8.9	0.53	4.26	9.02	0.54	4.31	9.2	0.54
DCRNN (Li, 2017)	4.78	9.91	0.61	4.91	10.37	0.62	5.14	11.06	0.64	5.36	11.77	0.66
MTGNN (Wu, 2020)	4.17	8.7	0.53	4.18	8.8	0.53	4.19	8.85	0.52	4.25	9	0.52
Graph WaveNet (Wu, 2019)	4.48	9.38	0.56	4.53	9.53	0.57	4.59	9.73	0.58	4.69	10.02	0.6
STGCN (Yu et al., 2018)	4.55	9.22	0.63	4.64	9.62	0.62	4.8	10.17	0.62	4.95	10.75	0.63
T-GCN (Zhao, 2020).	9.2	16.2	1.6	9.41	17.83	1.63	9.66	19.03	1.68	9.93	20.24	1.81

6. Conclusion

In this note, we have presented the UrbanBus dataset, a comprehensive resource for bus ridership analysis, comprising approximately 727 million AFC transactions over six months from around six million smart cardholders in a major Asian city. This dataset, with its fine granularity at the second level and detailed bus routes, fills a critical gap in publicly accessible data for intelligent transportation research. It supports advanced analysis, including forecasting, scheduling, and clustering, while ensuring privacy through robust encryption measures. By making the dataset and accompanying Python source codes publicly available, we aim to facilitate research and innovation in urban public transportation systems. The UrbanBus dataset is a valuable tool to improve our understanding and management of urban transportation, paving the way for more efficient and effective transportation solutions.

Acknowledgments

The authors thank the associate editor and anonymous reviewers for their valuable suggestions. This work was supported by the National Science Foundation of China (Grant Nos. 62202395,

62176221), and the Sichuan Science and Technology Program (Grant Nos. 2024NSFTD0036, 2024ZHCG0166).

Author contributions

Liwen Ke (Conceptualization, Methodology, Visualization), Xiangyu Guo (Conceptualization, Methodology, Resources, Visualization), Tianrui Li (Funding acquisition, Project administration), and Chongshou Li (Funding acquisition, Methodology, Resources, Validation, Visualization).

References

Bai, Lei et al. (2020) 'Adaptive Graph Convolutional Recurrent Network for Traffic Forecasting', arXiv, https://doi.org/10.48550/arXiv.2007.02842, 22 October 2020, preprint: not peer reviewed.

Bowen, Du et al. (2019) 'Deep Irregular Convolutional Residual Lstm for Urban Traffic Passenger Flows Prediction', *IEEE Transactions on Intelligent Transportation Systems*, **21**: 972–85.

Farahmand, Zakir H., Gkiotsalitis, Konstantinos and Geurs, Karst T. (2023) 'Predicting Bus Ridership Based on the Weather Conditions Using Deep Learning Algorithms', *Transportation Research Interdisciplinary Perspectives*, **19**: 100833.

Gao, Yuan, and Zhu, Jiaxing (2022) 'Characteristics, Impacts and Trends of Urban Transportation', *Encyclopedia*, **2**: 1168–82.

Goh, Kwang-Il et al. (2007) 'The Human Disease Network', *Proceedings of the National Academy of Sciences*, **104**: 8685–90.

Han, Yong et al. (2019) 'Short-Term Prediction of Bus Passenger Flow Based on a Hybrid Optimized LSTM Network', *ISPRS International Journal of Geo-Information*, **8**: 366.

Kumar, Nishant, and Raubal, Martin (2021) 'Applications of Deep Learning in Congestion Detection, Prediction and Alleviation: A Survey', *Transportation Research Part C: Emerging Technologies*, **133**: 103432.

Li, Chuan et al. (2020) 'Forecasting Bus Passenger Flows by Using a Clustering-Based Support Vector Regression Approach', *IEEE Access*, **8**: 19717–25.

Li, X., Chen, Z., Zhu, F., Chang, W., Tan, C., Xiong, G. (2018) "Short-term Bus Passenger Flow Forecast Based On Deep Learning," *International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, Jinan, China, pp. 372–376. https://doi.org/10.1109/SPAC46244.2018.8965619.

Li, Yaguang et al. (2017) 'Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting, arXiv, https://doi.org/10.48550/arXiv.1707.01926, 6 July 2017, preprint: not peer reviewed.

Liu, Shu Ying et al. (2021) 'Research on Forecast of Rail Traffic Flow Based on Arima Model', *Journal of Physics: Conference Series*, **1792**: 012065.

Luo, Dan et al. (2020) 'Fine-Grained Service-Level Passenger Flow Prediction for Bus Transit Systems Based on Multitask Deep Learning', *IEEE Transactions on Intelligent Transportation Systems*, **22**: 7184–99.

Luo, Dan et al. (2021) 'Spatiotemporal Hashing Multigraph Convolutional Network for Service-Level Passenger Flow Forecasting in Bus Transit Systems', *IEEE Internet of Things Journal*, **9**: 6803–15.

Shoman, Maged, Aboah, Armstrong and Adu-Gyamfi, Yaw (2020) 'Deep Learning Framework for Predicting Bus Delays on Multiple Routes Using Heterogenous Datasets', *Journal of Big Data Analytics in Transportation*, **2**: 275–90.

Van Oort, Niels, Brands, Ties and de Romph, Erik (2015) 'Short Term Ridership Prediction in Public Transport by Processing Smart Card Data', *Transportation Research Record*, **2535**: 105–11.

Vuchic, Vukan R. (2002) *Urban Public Transportation Systems*, Vol. **5**, pp. 2532–58. Philadelphia, PA: University of Pennsylvania Press.

Wang, Jingyuan, Jiang, Wenjun and Jiang, Jiawei (2023) 'Libcity-Dataset: A Standardized and Comprehensive Dataset for Urban Spatial–Temporal Data Mining', *Intelligent Transportation Infrastructure*, **2**: liad021.

Wang, Xuesong et al. (2016) 'Speed Variation during Peak and Off-Peak Hours on Urban Arterials in Shanghai', *Transportation Research Part C: Emerging Technologies*, **67**: 84–94.

Wu, Z., Pan, S., Long, G., Jiang, J., Zhang, C. (2019) Graph wavenet for deep spatial-temporal graph modeling. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence* (pp. 1907–1913).

Wu, Z., Pan, S., Long, G., Jiang, J., Chang, X., Zhang, C. (2020, August). Connecting the dots: Multivariate time series forecasting with graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, (pp. 753–763).

Ye, Y., Chen, L., and Xue, F. (2019) "Passenger Flow Prediction in Bus Transportation System using ARIMA Models with Big Data," *International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, Guilin, China, pp. 436–443, https://doi.org/10.1109/CyberC.2019.00081.

Yu, B., Yin, H., Zhu, Z. (2018) Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *IJCAI International Joint Conference on Artificial Intelligence*, Stockholm, Sweden, 3634–3640.

Zhao, Ling et al. (2020) 'T-Gcn: A Temporal Graph Convolutional Network for Traffic Prediction', *IEEE Transactions on Intelligent Transportation Systems*, **21**: 3848–58.

Zou, Liang et al. (2022) 'Passenger Flow Prediction Using Smart Card Data from Connected Bus System Based on Interpretable XGBoost', *Wireless Communications and Mobile Computing*, **2022**: 1–13.

Intelligent Transportation Infrastructure, 2024, liae019
https://doi.org/10.1093/iti/liae019
Data Note