Article:
Tang, T., Gu, Z., Yang, Y. et al. (2024) A data-driven framework for natural feature profile of public transport ridership: Insights from Suzhou and Lianyungang, China. Transportation Research Part A: Policy and Practice, 183. 104049. ISSN: 0965-8564

https://doi.org/10.1016/j.tra.2024.104049

Accepted by Transportation Research Part A: Policy and Practice

A data-driven framework for natural feature profiling in public transport ridership: Insights from Suzhou and Lianyungang, China

Tianli Tang[1], Ziyuan Gu[1*], Yuanxuan Yang[2], Haobo Sun[1], Siyuan Chen[1] and Yuting Chen[3]

[1] Jiangsu Key Laboratory of Urban ITS, Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, School of Transportation, Southeast University, Nanjing 211189, China

[2] Institute for Transport Studies, University of Leeds, Leeds LS2 9JT, UK

[3] Department of Civil Engineering, University of Wisconsin-Madison, Madison WI 53706, USA

Abstract

Urban public transport systems, characterised by their complexity, generate vast data sets that pose challenges to traditional analytical methods. To address this issue, our research introduces an innovative natural feature profile framework, leveraging a comprehensive, data-driven approach that incorporates big data, data mining, machine learning, and correlation analysis. This approach provides detailed insights essential for transport planning and policy development. The framework's core is its three-layered structure: the data layer, the feature layer, and the application layer, complemented by a unique four-level feature tagging system. This system investigates correlations, significance, and sensitivities amongst feature tags. It facilitates the extraction of natural feature profiles from voluminous data sets, rendering the framework highly applicable in practical scenarios. The implementation of this framework in Suzhou and Lianyungang demonstrated its adaptability and effectiveness. The findings underscored distinct city-specific transport patterns, highlighting the necessity for customised transport strategies. Furthermore, our framework excels at capturing spatial-temporal dynamics, offering essential insights grounded in evidence. Overall, this paper introduces a methodical, adaptable, and data-oriented framework, signalling a promising future for the development of intelligent and sustainable urban public transport systems.

Keywords: Natural features, big data analytics, public transport operation, policy-making support, green transport mode

* Corresponding author

1. Introduction

The agenda of *Transit Metropolis* has emerged as a pivotal focal point in the pursuit of sustainable urban transport systems (Cervero, 1998). One primary solution, embraced by major cities globally to address their transport system challenges, is the promotion of public transport-based trips (Mo et al., 2021; Wang, 2022; Wang and Tang, 2023).

Public transport systems not only offer a safer transit for passengers, as evidenced by reduced accident-related fatalities (Johnson, 2021), but also contribute to the reduction of carbon emissions and air pollutants, particularly through bus electrification (McGrath et al., 2022; Xylia et al., 2019). However, numerous regions have experienced bottlenecks in bus patronage, exemplified by declines of 52% in Shenzhen, 31% in Shanghai, and 49% in England (Johnson, 2021; STC, 2015; Xue, 2021). Scholars like Shiftan et al. (2015) argue that passengers' perceptions influenced their choice of trip modes, and Tao et al. (2017) posit that the level-of-service (LoS) affected passengers' loyalty towards bus travel. Thus, understanding the inherent features of ridership patterns, as our study proposes, can provide the granular insights necessary for refining LoS, making bus travel more appealing and improving patronage (Arana et al., 2014; Erhardt et al., 2022).

Efforts to enhance bus LoS have led to initiatives aimed at providing reliable bus services, reducing passenger waiting times, and expending bus network coverage (Liu and Sinha, 2007; Wu et al., 2017). Nevertheless, earlier research has largely depended on oversimplified models of bus networks or restricted survey data, which may lead to a disparity between theoretical models and the realities of real-world scenarios (Saberi et al., 2020). The advent of multi-source data has transformed public transport systems, offering a broader range of data, from smart card transactions and surveillance footage for ridership analysis, to road traffic data for estimating arrival times and even weather data for studying changes in passenger boarding and alighting patterns (Tang et al., 2023, 2021, 2020). However, the rapid expansion and variety of data introduce a new challenge: pinpointing and deriving meaningful insights from pertinent data (Gu et al., 2022, 2018; Qin et al., 2022).

In this scenario, the concept of *Natural Features* in public transport ridership becomes particularly insightful. Natural features in the realm of data analysis are defined as the inherent characteristics and patterns that are found within datasets, which remain unaffected by any external influence or alteration. Essentially, these features present an unadulterated, organic representation of the data at hand. When applied to public transport ridership, these features elucidate the unscripted behaviours, preferences, and trends apparent among passengers. Recognising and understanding these natural features is foundational, as it offers a genuine snapshot of ridership dynamics, thereby enabling more aligned and effective transport planning and policy-making.

The natural features of public transport ridership encapsulate the inherent characteristics, behavioural tendencies, and patterns exhibited within a public transport system. The crux of natural features, lies in an in-depth understanding of the essential characteristics and behavioural patterns

embedded within the public transport ridership, illuminating the foundational rules and the norms of its operation. These nature features typically emerge from the intricate dynamics of public transport ridership, influenced by a host of factors including, but not limited to, passenger preferences, temporal and spatial variability, urban morphology, service reliability, and network design (Lyu et al., 2022).

Deciphering these natural features provides crucial understanding of the operational patterns and interactions within the public transport network. This understanding is pivotal to advancing transport services and decision-making, thereby fostering sustainable urban transport development. By offering a true representation of the public transport system, these natural features facilitate informed decision-making for policymakers, urban planners, and transport operators, which in turn enhances the efficiency of urban transport planning and management (He et al., 2022; Peled et al., 2021).

To address the challenges posed by the diverse, large-scale, and complex data in public transport systems, this paper proposes a problem-oriented, data-driven framework for the in-depth analysis of public transport ridership's natural features. Our goal is to extract valuable information that enhances public transport services, improves control mechanisms, and refines decision-making processes. We aim to advance a precise understanding of public transport dynamics.

Our objectives include: (i) developing a scenario-based label system for bus ridership's natural feature profile, (ii) exploring effective methods for extracting and interpreting insights from multi-source data, (iii) examining multi-source data processing techniques considering the relationships between various ridership features, and (iv) suggesting standardised methods for profiling natural features and comprehensive evaluation across different data collection platforms and technological contexts

The remainder of this paper is structured as follows. Section 2 reviews the analysis characteristics of bus ridership and their impacts on individual boarding behaviour. Section 3 introduces the framework of natural feature profiling. Two case studies of the public transport systems in the cities of Suzhou and Lianyungang, China, are presented in Section 4. Section 5 discusses the implications of the natural feature, and Section 6 offers the conclusion of this study and outlines future work.

2. Literature review

2.1. Characteristics analysis on bus ridership

A deep understanding of bus ridership characteristics is vital for improving bus services and fostering sustainable urban transport. Numerous methods and models have been employed to analyse bus ridership, which can be generally classified into three categories: regression models, machine learning models, and travel demand models (Taylor and Fink, 2013).

Regression models, encompassing multiple linear regression (MLR) and Poisson regression,

have been widely utilised to determine the relationship between ridership and an array of factors, including socio-economic, land use, and service quality characteristics (Hu and Chen, 2021; Kim et al., 2016; Yang et al., 2021). For example, Boisjoly and El-Geneidy (2017) employed MLR to investigate the correlation between bus stop amenities and ridership, revealing that shelters and benches positively impact ridership. Vergel-Tovar and Rodriguez (2018) discovered that built environment characteristics, such as the mixture of land uses surrounding Bus Rapid Transit (BRT) stations and their integration into the urban fabric, are crucial determinants of BRT ridership and the necessity for sustainable mass transit systems. Zhou et al. (2017) emphasised that metro stations in urban areas are more susceptible to outdoor weather concerning ridership, while regular transit users exhibit resilience to weather changes. However, these models often suffer from small sample sizes or aggregate-level data limitations, which may not fully capture the complexity and dynamics of urban travel behaviour (Gutiérrez et al., 2011).

Machine learning models, comprising decision trees, support vector machines, and artificial neural networks, have gained popularity in recent years for predicting bus ridership (Chen et al., 2022; Tang et al., 2021; Ullah et al., 2022). For instance, Tang et al. (2020) utilised a tree-based model, gradient boosting decision tree (GBDT), to estimate potential alighting stops for individual bus trips. Their study demonstrated that time-dependent variables were of greater importance than others, and point-of-interest (POI)-related variables exhibited weaker correlations with alighting choice behaviour. Wu et al. (2021) proposed a novel scaled stacking GBDT model to predict bus passenger flow using multi-source datasets, which effectively addressed the multicollinearity issue with multi-source data and prioritised influential factors for passenger flow prediction. Although these models provide enhanced accuracy and adaptability for handling non-linear relationships and large-scale data compared to traditional regression models, some approaches may not adequately address the interactions and spatial-temporal dependencies among factors influencing bus ridership (Sivakumar Nair et al., 2023; Yousefzadeh Barri et al., 2022).

Travel demand models, such as four-step and activity-based models, have been applied to forecast bus ridership by simulating individual travel behaviour and capturing intricate interactions between travellers, land use, and transport systems (Chen et al., 2023; McNally, 2007; Pinjari and Bhat, 2021). For example, Deepa et al. (2022) developed a direct demand model for bus transit ridership in Bengaluru, India, and examined the impact of service frequency and inter-route relationships on ridership. Berrebi and Watkins (2020) analysed the decline in bus ridership in four US cities between 2012 and 2018, finding that the principal cause of the decline predominantly affected white bus riders, while neighbourhoods with non-white, carless, and high-school-educated residents were more likely to exhibit high ridership. Nonetheless, activity-based models may not fully capture the complexity of human travel behaviour and decision-making processes, which can be influenced by factors such as social norms, perceptions of safety, and individual preferences (Bradley et al., 2010; Malayath and Verma, 2013).

2.2. Impacts on bus ridership prediction

Accurate prediction of bus ridership is essential for improving public transport planning and policy-making. The factors influencing individual choices for bus travel cover a broad spectrum, including mode choice, travel purpose, boarding and alighting stops, travel distance, and travel time, etc. Researchers have employed an assortment of models and methods to examine these factors and their implications on overall travel demand (Carpio-Pinedo, 2014; Wei et al., 2019; Zhou et al., 2013).

Discrete choice models have been extensively employed to study mode choice behaviour, emphasising the effects of service quality attributes, socio-economic characteristics, and individual preferences (Boisjoly and El-Geneidy, 2017; Xue et al., 2015). For instance, Shiftan et al. (2015) applied a nested logit model to investigate the influence of service quality attributes and socio-economic characteristics on mode choice. While these models have yielded valuable insights, potential biases in data sources and model assumptions may affect the generalisability and robustness of the findings.

Structural equation modelling has been utilised to explore the impact of the level of service on passengers' loyalty and satisfaction with bus travel (Currie and Delbosc, 2011; Tao et al., 2017). These studies have underscored the importance of service quality, reliability, and accessibility in shaping travellers' perceptions and choices. However, these models may not capture the full range of factors influencing bus ridership, including the interactions and spatial-temporal dependencies among various determinants.

Machine learning techniques, e.g., decision trees, support vector machines, and artificial neural networks, have emerged as promising approaches for predicting bus ridership due to their capacity to handle large-scale data, non-linear relationships, and complex interactions (Khalil et al., 2021; Toqúe et al., 2016). For example, Liu et al. (2019) and Zhang and Cheng (2020) applied deep learning models to forecast bus ridership, demonstrating improved prediction accuracy compared to traditional regression models. However, few studies have effectively integrated multi-source data, such as smart card data, road traffic data, and weather conditions, to provide a more comprehensive understanding of bus ridership and its influencing factors (M. Zhang et al., 2022).

In summary, despite the use of various methods to predict bus ridership and inverstigate factors affecting travel choices, there remains a need for further research to address challenges associated with data sources, model assumptions, and contextual factors (Kuo et al., 2023). Incorporating a data-driven approach using large-scale, multi-source data and advanced modelling techniques can potentially reveal new insights into the determinants of bus ridership, including the identification of natural features that influence travel behaviour. By uncovering these natural features, researchers and policymakers can devise more effective strategies to improve the accuracy of ridership predictions and enhance the overall quality of bus services.

3. Methodology: Natural feature profile framework

The methodological framework presented herein aims to analyse the natural feature profile of public transport ridership. Utilising a bottom-up approach, it is firmly grounded in the real-world data, to thoroughly represent every facet and situation of the actual public transport system. This approach is particularly designed to support applications in public transport systems, including policy-making and strategic planning. The framework is structured systematically into three distinct layers, comprising the data layer, the feature layer, and the application layer, as depicted in Figure 1. The Data layer pertains to data acquisition and preprocessing; the feature layer centres on feature construction and analysis, and the application layer guides the problem analysis of public transport ridership.

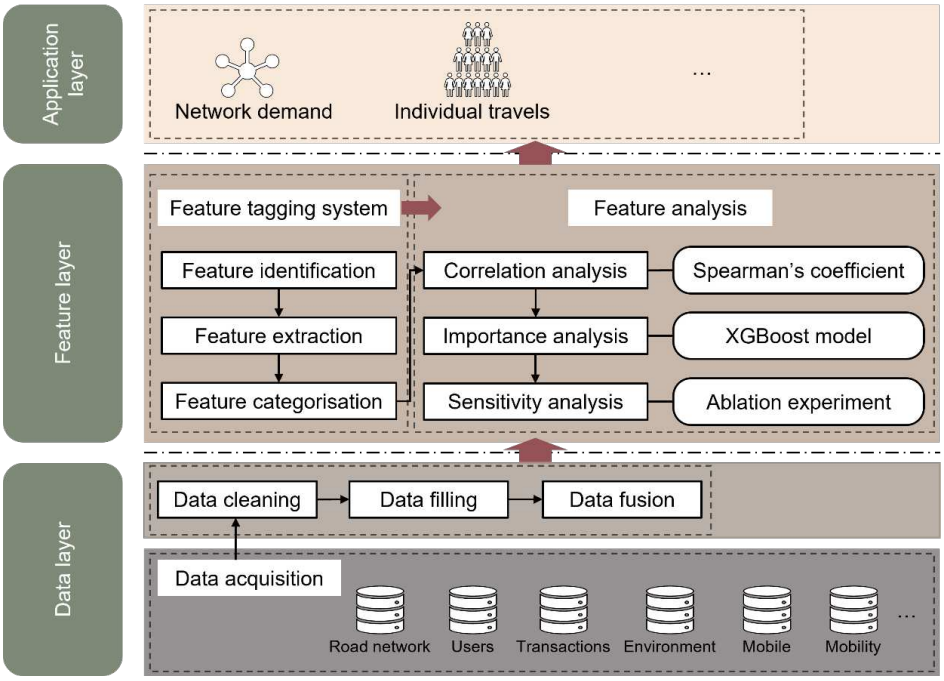


Figure 1 A framework of the natural feature profiling in the public transport ridership.

3.1. Data layer

With the advancement of communication and sensor technologies, an increasing number of devices are being equipped in vehicles, stations, and on roads, thereby facilitating comprehensive monitoring of public transport systems. In this section, we introduce five, but not limited, principal data sources ubiquitously accessible and pertinent to most public transport systems (Iliashenko et al., 2021; Welch and Widita, 2019):

- **Network Information**: This encompasses fundamental data of public transport networks, documenting the attributes of stations (e.g., position, length, type, etc.) and bus lines (e.g., length, total number of stations, average travel time, etc.). Additionally, the network topology ought to be discernible, for example, the subordination and sequence of stations to service lines and the interconnections amongst stations.

- **Scheduling Information**: This dataset contains information on schedules, including planned arrival and departure times at each station, headway between buses, and other timetable-related data.
- **Transaction Records**: Fares are primarily paid in three main ways: cash, pre-paid tickets (monthly paid tickets), and smart cards. Cash payments and pre-paid tickets yield limited useful information, while smart-card data, recording intricate boarding information and sometimes alighting information, is invaluable for bus ridership analyses.
- **Mobile Data**: This data, sourced from smartphones and other mobile devices, provides real-time insights into public transport passenger movements, enabling a more detailed understanding of users' travel activities and patterns (Wang et al., 2022; J. Zhang et al., 2022).
- **Mobility Data**: This dataset combines the intricacies of urban transit, encapsulating micro-mobility's role in bridging public transport gaps, shared mobility's rise via rideshares and carpooling, and the personal travel choices reflecting urban transport patterns (Huo et al., 2020; Ma and Zhang, 2022; Zhong and Sun, 2022).
- **Other Environmental Factors**: This dataset typically contains point-of-interest (POI) information and weather factors, both significantly influencing passengers' travel behaviour and the running status of bus vehicles.

Public transport system databases typically utilise incremental backups for data storage, which can lead to the presence of duplicate and expired data. Additionally, these databases may contain noise data, which are irrelevant or incorrect information that can skew analysis results. It is crucial to eliminate such invalid data before proceeding with any analysis. Besides these issues, other data challenges such as missing information and errors often arise. Addressing these requires tailored strategies for either supplementing or removing data, depending on the specific conditions of the dataset (Wang et al., 2022). Finally, it is necessary to integrate data from various sources. This integration is achieved by leveraging the relationships between corresponding fields across different tables within the database, ensuring a cohesive and comprehensive data set for analysis.

3.2. Feature layer

3.2.1. Feature tagging system

In this section, we craft an elaborate feature tagging system for the public transport system. As illustrated in the middle stage of Figure 1, we initially identify an extensive range of features based on expert experience and findings from previous studies. Each individual feature or a combination thereof delineates a specific aspect of the transport system. For instance, 'weather events' may potentially instigate alterations in travel choice, whereas'departure time' and 'arrival time' exemplify network reliability. Subsequently, we extract the values of these features. The selected features from the prior step should be easily measurable, implying that they can be directly measured

from the data or through a straightforward data fusion method. Cities adopt different architectural approaches for the digitalisation and informatisation of their public transport systems. Considering the differences in data field structures and information content among cities, this section proposes a public transport system feature tagging system that is compatible with data and accommodates differences. The feature tagging system primarily encompasses four levels of structure, namely system elements, feature objects, feature tags, and tag entities. This feature tagging system serves as the cornerstone for our analysis and modelling of the public transport system.

- *System elements*

For urban public transport systems, the principal system elements to be considered comprise network structure, passenger behaviour, operation plans, and other factors. Network structure delineates the spatial topology of infrastructure, such as bus lines, stations, and depots, forming the foundational framework of the transit system. Passenger behaviour characterises the temporal, spatial, and volumetric traits of passenger flow at individual, station, line, and regional levels, serving as the demand source for the transit system. Operation plans encompass the scheduling plans and actual operating conditions of bus routes and vehicles, representing the service supply of the transit system. Other factors include external factors beyond the transit system that may impact the supply and demand of transit, such as changes in land use and weather conditions.

- *Feature objects*

Feature objects are the intricate aspects within each system element, which specifically identify different aspects of information. For example, when describing ‘network structure’, the feature tags should cover objects such as bus stops and routes in the transit network. When describing "service supply," feature tags can be divided into planning schemes and actual situations. When describing "travel demand," feature tags can be divided into individual travel, station demand, regional demand, route passenger flow, etc., according to different top-level design requirements. "External environment" can be classified and summarised based on data foundation and application objectives.

- *Feature tags*

Feature tags are specific descriptions of physical or logical variables in the transit system, and they also constitute the basic information obtained directly from the data. For instance, the previous layer of "station information" merely summarises and categorises variables that describe relevant information of a station. To define a station in the network, specific information such as station name, range, location, sequence number, and relationship with routes needs to be relied upon. In addition, some feature tags are obtained through simple statistics or inference. For example, when calculating station passenger flow, individual travel needs to be aggregated and summed up.

- *Tag entities*

Tag entities are the specific representations of feature tags. Different forms of presentation may exist for the same feature tag, so it is necessary to choose a representation that is easy to obtain,

convenient for analysis, and accurately expresses the information. For example, when describing station location, it can be described based on reference objects (such as "near the east gate of People's Park"), or street guidance (such as "120 meters north of the intersection of Wuyi Road and Labour Road"). However, the most precise representation is through latitude and longitude positioning.

3.2.2. Feature analysis

Within the context of big data, public transport systems exhibit a wide array of feature tags. The selection of these labels is informed by both available data and expert experience. Nevertheless, the relationships amongst these labels, as well as their connection to the application objectives, often remain ambiguous. Thus, it is crucial to examine and assess these feature tags within their specific application context. Subsequently, three evaluation methods and indicators are proposed to derive a natural feature profile from the intricate system of feature tags.

- *Correlation among feature tags*

An examination of the correlation amongst feature tags necessitates the consideration of the correlation between feature tags and application objectives, in addition to the correlation among disparate feature tags. The correlation between feature tags is primarily ascertained through correlation coefficients. Depending on the data types of the feature tags, they can be categorised as ordered or unordered variables. Ordered variables encompass continuous variables and ordered categorical variables, whilst unordered variables predominantly consist of nominal variables. The methods for calculating the correlation coefficients between feature tags and application objectives, as well as among different feature tags, depend on the data types.

Spearman's rank correlation coefficient is suitable for the analysis of correlation between ordered variables, such as the day of the week and passenger flow at a station. This coefficient facilitates linear correlation analysis utilising the rank order of two variables and does not necessitate any assumptions regarding the distribution of the original variables. It constitutes a non-parametric statistical method. The formula for calculating Spearman's rank correlation coefficient is as follows:

$$\rho_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{1}$$

where n represents the sample size, and d_i symbolises the rank difference between X_i and Y_i, i.e., the position of this number in the column after sorting from smallest to largest.

- *Importance of feature tags*

In the context of the corresponding application, the significance of feature tags is indicative of their importance. Feature tags with higher significance should be accorded greater attention in model construction and decision-making processes. The extreme gradient boosting (XGBoost) model calculates the importance of feature tags, employing the gain of split scores to determine the

feature transition at the splitting point (Tang et al., 2020). The importance of a specific feature is ascertained by its role across all trees, meaning that the more frequently an attribute is utilised in constructing decision trees within the model, the higher its importance. The xgboost feature importance index typically evaluates the average reduction in loss when a feature is employed as a splitting attribute, that is, the information gain of the feature. The calculation of its importance can be expressed as:

$$V(k) = \frac{1}{2} \frac{\sum_{t=1}^{T} \sum_{i=1}^{N(t)} I(\beta(t,i) = k) \left(\frac{G_{\gamma(t,i,L)}^2}{H_{\gamma(t,i,L)} + \lambda} + \frac{G_{\gamma(t,i,R)}^2}{H_{\gamma(t,i,R)} + \lambda} - \frac{G_{\gamma(t,i)}^2}{H_{\gamma(t,i)} + \lambda} \right)}{\sum_{t=1}^{T} \sum_{i=1}^{N(t)} I(\beta(t,i) = k)} \tag{3}$$

where k represents a node, T denotes the total number of trees, $N(t)$ signifies the number of non-leaf nodes in the t-th tree, $\beta(t,i)$ represents the splitting feature of the i-th non-leaf node in the t-th tree, i is the indicator function, and λ is the regularisation term hyperparameter. $G_{\gamma(t,i)}$ and $H_{\gamma(t,i)}$ represent the sum of first- and second-order derivatives of all samples that fall into the i-th non-leaf node in the t-th tree, respectively. $G_{\gamma(t,i,L)}$ and $G_{\gamma(t,i,R)}$ signify the sum of first-order derivatives of samples that fall into the left and right child nodes of the i-th non-leaf node in the t-th tree, respectively. Similarly, $H_{\gamma(t,i,L)}$ and $H_{\gamma(t,i,R)}$ represent the sum of second-order derivatives of samples that fall into the left and right child nodes of the i-th non-leaf node in the t-th tree, respectively. Therefore, the following equations hold:

$$G_{\gamma(t,\mathrm{i})} = G_{\gamma(t,\mathrm{i},L)} + G_{\gamma(t,\mathrm{i},R)} \tag{4}$$

$$H_{\gamma(t,\mathrm{i})} = H_{\gamma(t,\mathrm{i},L)} + H_{\gamma(t,\mathrm{i},R)} \tag{5}$$

Information gain is employed in Gain-method, which can readily identify the most direct features. In practice, the values at the beginning and end of the ranking of Gain-method often exhibit significant differences, as the optimisation of subsequent features may not occur on the same scale. This discrepancy is similar to the variation of magnitudes in neural networks when optimising loss functions, which may differ by tens or hundreds of times.

- *The Sensitivity of feature tags*

Feature tags' sensitivity analysis entails examining the effectiveness of each feature tag within the application context, identifying feature tags with negligible or no impact on the decision-making process for the application target. This approach can considerably reduce model complexity, diminish the workload of data analysis and processing, and significantly enhance model accuracy. Furthermore, the differences in sensitivity of feature tags among various city bus systems are vital considerations in intrinsic profiling, as they can inform the development of city-specific planning and management strategies.

The proposed framework employs feature ablation experiments to demonstrate the role and impact of each feature tag within the application context. These experiments aid in understanding causality within the system and provide a direct means of generating reliable knowledge about the target application. During feature ablation experiments, each feature tag is systematically removed,

and the first-order indicators of the underlying passenger flow prediction model's performance are analysed, measuring the contribution of each feature tag variable to the output. This analysis measures the contribution of each feature tag to the model's output, thereby elucidating the role and relationship of each feature tag variable.

Conducting feature ablation experiments enables the determination of each feature tag's contribution to model performance and the assessment of their sensitivity. If the ablation of a particular feature tag has little significant impact on the model performance, it indicates that this feature tag has negligible or no influence on the decision-making for the application target, and it may be considered for removal from the model to reduce complexity. Simultaneously, by comparing the effects of ablation of different feature tags, the differences in their roles within the application context can be evaluated, which can help gain insights into the sensitivity differences of various feature tags and inform the development of city-specific planning and management strategies.

By employing these methods, a comprehensive understanding of the natural features in public transport systems and their relationships can be developed. This knowledge will prove essential for further analysis and application in various scenarios and objectives, such as demand estimation, travel choice analysis, reliability analysis, and more.

Whilst we present certain models as part of our general framework, we recognise that the landscape of analytical techniques is broad and multifaceted. For instance, when determining feature importance, aside from the models mentioned, other algorithms based on decision trees, like LightGBM, could also be considered. Depending on the type of data, there might be a need to choose appropriate correlation measures; for continuous data, Pearson's coefficient could be suitable, while for non-linear relationships, the Maximal Information Coefficient (MIC) might be more pertinent (Mao et al., 2022). We urge researchers to evaluate and select techniques that best align with the characteristics of their data and the specific research questions at hand.

3.3. Application layer

The application layer serves as the final component in the natural feature acquirement framework. This layer is responsible for integrating and processing data and features from the data layer and feature layer to generate valuable insights that support informed decision-making.

To provide a more grounded understanding, reference can be made, but not limited, to several applications in the current environment.

- **Real-time Urban Planning Platforms**: Applications that leverage real-time data to recommend adjustments in urban transit schedules, ensuring optimal bus timings in accordance with dynamic urban patterns (Kwon et al., 2023; Lian et al., 2023).
- **Predictive Analytics Tools**: Solutions that employ advanced algorithms to forecast passenger flow during peak and off-peak times. By anticipating surges, transit authorities can better allocate resources and manage crowd control (Jiang et al., 2023).

- **Integrated Data Visualisation**: Applications that amalgamate diverse data streams to visually represent potential bottlenecks, traffic patterns, or passenger preferences. These visual tools are essential for decision-makers to identify challenges swiftly.
- **User Experience Enhancers**: Tools that focus on passenger experience, collecting feedback in real-time, and suggesting immediate remedial measures. This ensures the transport system remains responsive to user needs.

The application layer encompasses more than raw analysis; it also orchestrates the deployment of various analytical techniques and machine learning algorithms. By exploring relationships, pinpointing trends, and crafting predictive models using the assimilated data and cherry-picked features, one can glean insights into the bus network's functioning, discern passenger flow trajectories, and spotlight potential enhancement zones.

The applications birthed from this layer are invaluable. They shed light on the bus network's efficacy, paint a picture of passenger mobility patterns, and highlight areas ripe for improvement. The advice that emanates from these insights can tackle myriad facets of the bus network. This might span from dissecting demand patterns to enhancing the overall user experience of the public transport ecosystem.

In essence, the application layer is pivotal. It transforms data and features from its preceding layers into tangible insights and advice. This is achieved through a symphony of data amalgamation, feature sifting, intricate analysis, adept modelling, vivid visualisation, and astute decision support. By doing so, the natural feature acquisition framework promises to be a catalyst in refining and streamlining the bus network

4. Case study

In this section, we present two distinct case studies of bus networks in the cities of Suzhou and Lianyungang, China. These cities differ significantly in city size, population, bus network structure, and public transport ridership, as illustrated in Table 1.

Table 1 Statistical indicators of urban and public transport network development in the cities of Suzhou and Lianyungang[1].

Descriptions	Suzhou	Lianyungang
City size	4,652.84 km^2	3,032.42 km^2
Population	5.2 million	2.3 million
Disposable income per capita	6,8191 RMB	39,862 RMB
Transport expenditure	5,946 RMB	1,390 RMB
Number of bus lines	252	116
Number of bus stop	4,233	2,121
Daily ridership	≈300,000	≈50,000

[1] Open Data Source: Suzhou City Statistics Bureau (http://tjj.suzhou.gov.cn/sztjj/tjnj/2022/zk/indexce.htm) and Lianyungang City Statistics Bureau (http://tjj.lyg.gov.cn//tjxxw/upload/ad0e70de-a800-4141-88f8-3e9c244de7ad.pdf), 2021.

Suzhou, located in eastern China, is a rapidly developing city renowned for its rich history, cultural heritage, and picturesque gardens. In recent years, Suzhou has experienced substantial economic growth, resulting in an increase in population and urbanisation. As the city expands, transport and public transit have become indispensable components of Suzhou's development.

Lianyungang, one of the most important port cities on the eastern coast, connects major railway lines such as the China-Europe Railway Express and Longhai Railway, as well as the Port of Lianyungang, which operates numerous international shipping routes. Lianyungang serves as the eastern bridgehead of the Second Eurasian Land Bridge, responsible for transporting over 90% of transit containers. Lianyungang is the junction of multimodal transport for rail and shipping. However, Lianyungang is not a large city in terms of size and population.

4.1. Bus network in Suzhou and Lianyungang

Before exploring the detailed descriptions of the bus networks in Suzhou and Lianyungang, we first compare the primary features of both systems in a consolidated Table 2:

Table 2 Comparative Overview of Bus Networks in Suzhou and Lianyungang.

Descriptions	**Suzhou**	**Lianyungang**
Daily Passengers	Over 300,000	Approximately 40,000
Total Bus Routes	Over 200	6 BRT lines + Over 100 regular
Bus Stops	Over 4,000	300 BRT stations + 1,500 regular
Specialised Buses	Night, tourist, customised lines	BRT
Sustainability Initiatives	Electrification, ITS	Modern BRT system
Popular Payment Modes	Cash, smart cards	Over 80% mobile payments
Special Discounts	Students, seniors, frequent riders	Not mentioned
Heatmap Reference	Figure 2	Figure 3

Transport in Suzhou is well-developed, boasting a comprehensive network of roads, railways, and waterways that connect the city to its surrounding areas. The city possesses an efficient public transit system, crucial in meeting the transport demands of its residents and visitors. Suzhou's bus network serves as a vital mode of transport, accommodating an impressive volume of passengers daily. With over 300,000 passengers per day, Suzhou's bus system addresses the mobility requirements of a significant portion of the city's population, as mapped in the heatmap of Figure 2. The network comprises over 200 bus routes, including night buses, tourist lines, and customised buses, encompassing a wide range of destinations within and beyond the city. These routes serve over 4,000 bus stops, offering extensive coverage across Suzhou and its surrounding areas. The buses are modern, well-maintained, and equipped with amenities such as air conditioning and real-time information displays. Affordable fares, payment options including cash and smart cards, and special discounts for students, seniors, and frequent riders render the bus system convenient and accessible. Suzhou's bus system also emphasises sustainability, with initiatives like route optimisation, upgrading to electrification buses, and implementing intelligent transport systems. Overall, Suzhou's bus network provides a reliable, convenient, and sustainable mode of transport for residents and visitors alike, accommodating a substantial number of passengers daily.

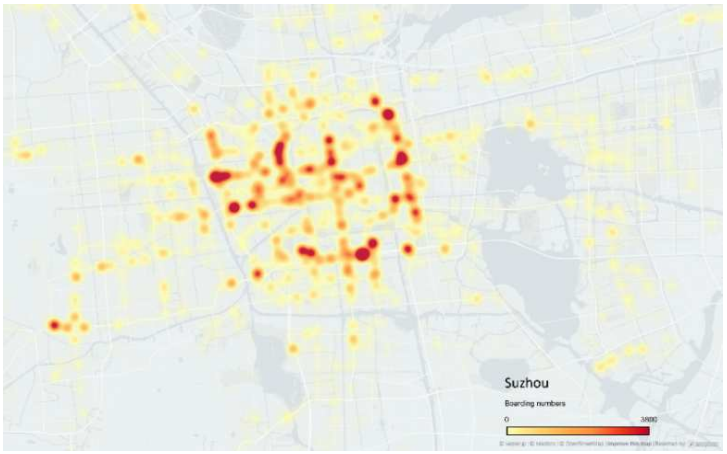


Figure 2 The average daily boardings at bus stations in the city centre of Suzhou. Map produced by © kepler.gl, Map tiles by © Mapbox, Data by © OpenStreetMap.

Lianyungang's bus system is an integral component of the city's transport infrastructure, offering reliable and convenient services to approximately 40,000 passengers daily. The system encompasses a diverse fleet of buses, including six BRT lines with 300 BRT bus stations and over 100 regular bus lines with over 1,500 conventional bus stops. The BRT system in Lianyungang is a modern and efficient mode of transport that delivers fast and reliable services through dedicated bus lanes, advanced fare collection systems, and well-designed stations. The BRT lines cover key routes in the city, connecting major commercial, residential, and transport hubs, making it a popular choice for commuters and travellers alike. In addition to the BRT lines, Lianyungang's bus system also incorporates a comprehensive network of regular bus lines that serve various destinations within the city. These conventional bus lines provide extensive coverage across different neighbourhoods, linking residential areas, business districts, educational institutions, and other essential locations, catering to the diverse mobility needs of the local population. Lianyungang's bus system has witnessed significant adoption of mobile payment, with over 80% of passengers opting for this convenient mode of fare payment. Figure 3 shows the heatmap of the average daily boarding among a month at bus stations in the city centre of Lianyungang.

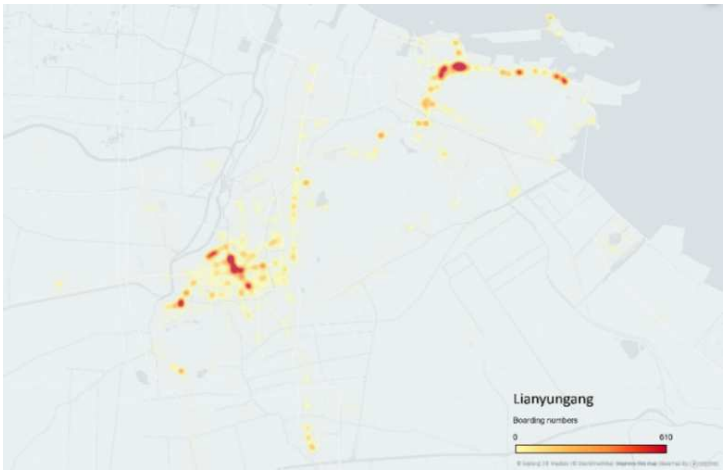
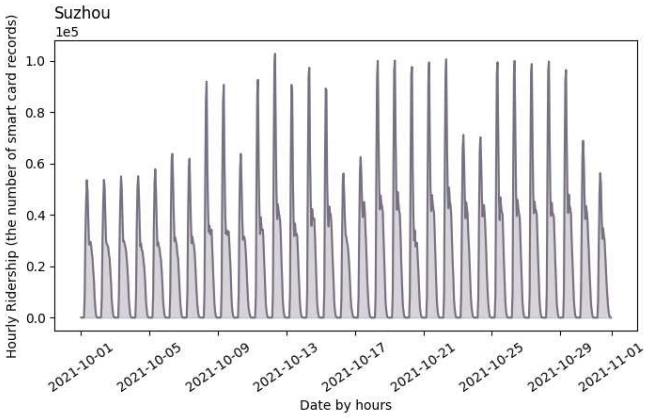


Figure 3 The average daily boardings at bus stations in the city centre of Lianyungang. Map produced by © kepler.gl, Map tiles by © Mapbox, Data by © OpenStreetMap.

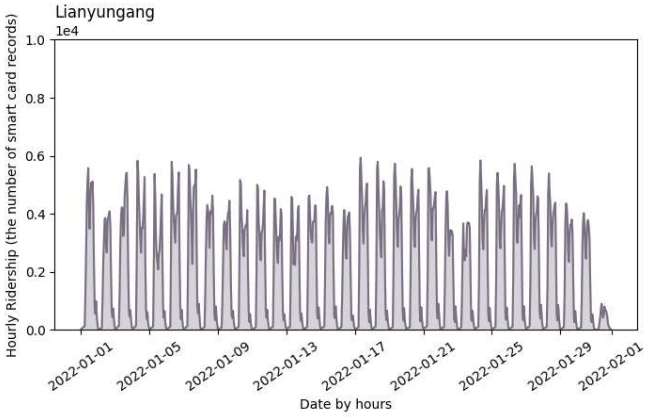
Suzhou and Lianyungang, each possessing unique urban dynamics, have been chosen to validate our natural feature profile framework. Suzhou represents an example of urban sophistication, boasting an advanced public transport system. In contrast, Lianyungang, as a developing city, is actively working on improving its transportation infrastructure. This deliberate contrast serves two purposes: firstly, to highlight the wide applicability of our framework in different urban contexts, and secondly, to enhance understanding of ridership patterns through the comparative analysis of insights drawn from these two distinct cities.

4.2. Data description

This study employs a diverse range of data sources to craft a multi-dimensional picture of public transport ridership. For the purposes of this investigation, the acquired data has been compartmentalised into three critical categories: smart-card data, POI data, and network data. The following details each type of data utilised, the manner in which they were procured, and the subsequent processing and analysis steps that have been undertaken. Notably, the data considered are derived from Suzhou in October 2021 and Lianyungang in January 2022, offering us an opportunity to examine public transport dynamics in two distinct urban settings. Figure 4 provides a summarised snapshot of the smart card data utilised in this study in Suzhou and Lianyungang.



(a)



(b)

Figure 4 Summary of the hourly ridership (i.e., the number of smart card records) of the public transport systems in (a) Suzhou and (b) Lianyungang.

Smart card data, an crucial data source in modern public transport studies, offers unparalleled insight into passenger travel behaviour. This data records critical variables such as the time of transactions, boarding time and locations, and vehicle and line identification in our case cities. Table 3 and Table 4 present the structures of (masked) smart card data of Suzhou and Lianyungang. After data cleaning to ensure validity and reliability of the information, these granular details allow us to generate a detailed passenger flow model across the public transport network.

Table 3 The data structure of masked smart card record used in Suzhou.

Card ID	Station	Boarding time	Bus line	Line name	Station No.	Direction	Vehicle
21**850	78c**b80	2021-10-01 06:41:52	c31**cf7	34	20	1	d39**b77
21**938	793**d9c	2021-10-01 06:37:40	AB9**F58	63	4	0	a64**d8a
21**130	540**e35	2021-10-01 17:46:08	6AC**034	629	8	0	f4d**3d0
21**153	6ae**f37	2021-10-01 10:26:54	50d**5b6	308	3	1	bd2**ab9
21**956	fd2**308	2021-10-01 16:46:02	2b9**bef	302	37	1	7D0**D57

Table 4 The data structure of masked smart card record used in Lianyungang[2].

Card ID	Transaction time	Bus line	Subline	Station	Vehicle	Operation trip
22****240	2022/1/7 08:35	930	93001	10****072	17****47	930****17
22****930	2022/1/7 08:35	930	93001	10****072	15****57	930****17
22****930	2022/1/7 09:45	236	23611	21****325	18****74	236****23
22****924	2022/1/7 10:31	27	02702	14****626	14****31	027****26
13****400	2022/1/7 13:22	183	18302	23****705	17****47	183****43

4.3. Results on the natural feature profiles

4.3.1. Feature tagging systems of Suzhou and Lianyungang

With a comprehensive and detailed understanding of the varied data utilised in our case studies of Suzhou and Lianyungang, we subsequently construct the corresponding feature tagging systems for each city following the approach delineated in Section 3.2.1. Drawing from the variety of data sources mentioned, the feature tagging system depicted in Figure 5 and Figure 6 comprehensively capture the distinct characteristics and dynamics of the public transport systems in each of these cities.

[2] For convenience of understanding, the names of the variables do not use the names of their field names within the database.

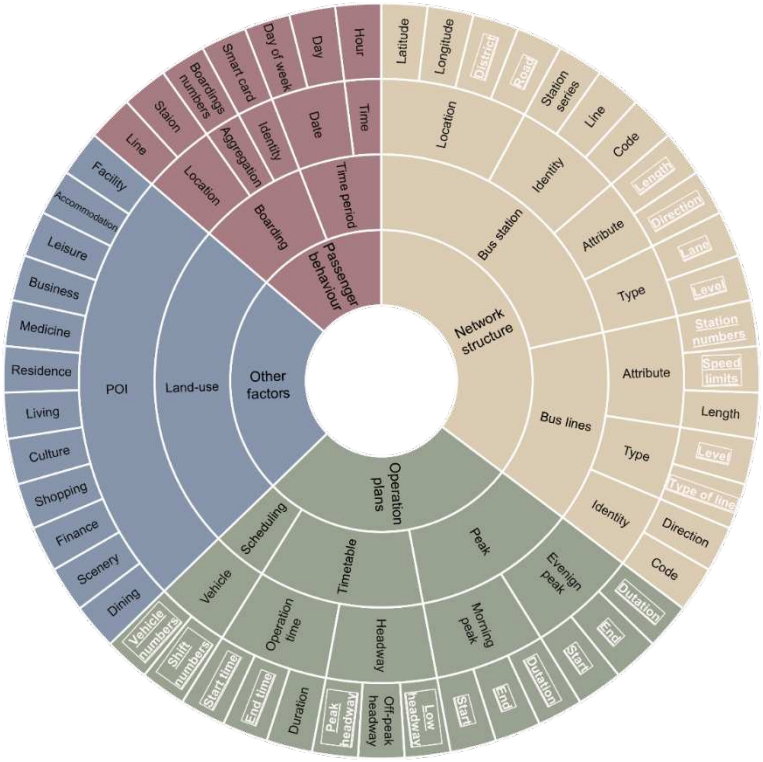


Figure 5 The four-level feature tagging systems utilised for Suzhou's bus system. The different feature tags are highlighted and outlined in white.

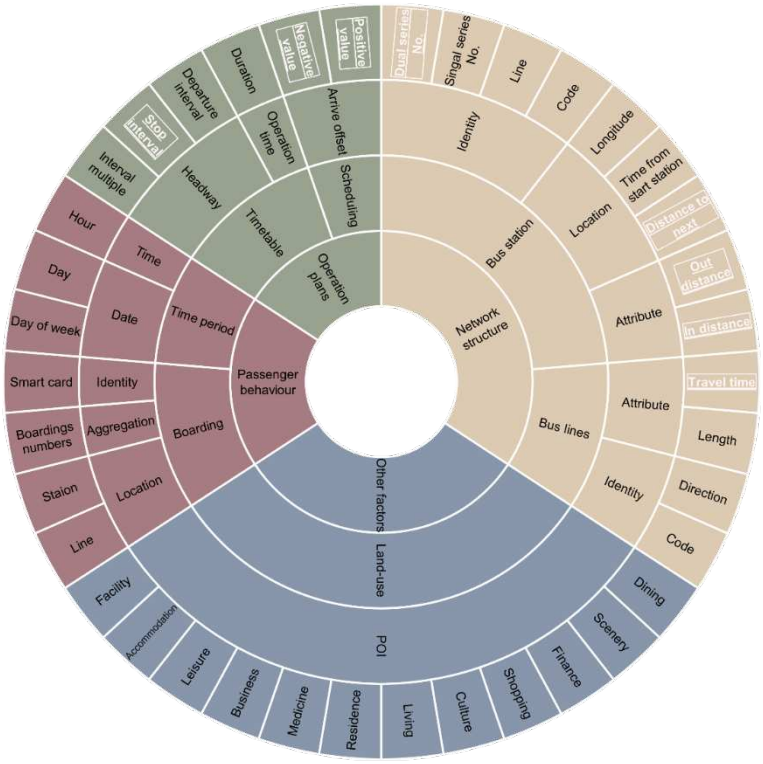


Figure 6 The four-level feature tagging systems utilised for Lianyungang's bus system. The different feature tags are highlighted and outlined in white.

In consideration of the constraints of our available data, certain individual information (e.g., gender, age), and data on other modes of travel (e.g., subway, ride-hailing services) remain unknown in this study. Hence, our analysis in this case primarily revolves around four system elements: network structure, passenger behaviour, operation plans, and other factors. Notably that in other urban contexts or scenarios, there might be access to a richer dataset. In such cases, additional potential feature labels can be procured based on requirements and seamlessly integrated into this four-level feature tagging system.

Our innovative four-level feature tagging system provides an efficacious and robust platform that effectively covers a broad spectrum of factors influencing public bus ridership, facilitating a rational categorisation and amalgamation of these factors. It is noteworthy that, within this overarching system, there are subtle discrepancies between the tag entities specific to Suzhou and Lianyungang. For instance, Suzhou incorporates grade divisions for features such as ‘bus line’ and ‘bus station’ a categorisation absent in the Lianyungang system. Such discrepancies underscore the versatility and adaptability of our four-tiered feature tagging system, demonstrating its capability to accommodate diversity and variance inherent in different cities and their corresponding datasets. The system, therefore, affirms its flexibility and adaptability, mirroring the unique aspects of distinct public transport systems while maintaining a uniform structure. This balanced methodology has been both thorough and precise, successfully capturing the subtleties of individual public transport ridership scenarios in the cities under study.

Having established these robust feature tagging systems, we proceed to leverage the techniques proposed in Section 3.2.2. Guided by the problem-oriented approach of the Application Layer outlined in Section 3.3, our focus shifts to an exhaustive feature analysis, centring on passenger flow at stations. In the forthcoming sections, we utilise these methods to draw conclusions about the natural feature profiles of public transport ridership in both Suzhou and Lianyungang. This cogent, data-driven approach allows us to surface insightful conclusions about the unique aspects of public transport ridership in each city, thereby providing a foundation for informed policy recommendations.

4.3.2. Feature analysis in Suzhou’s bus system

In this section, we explore the natural features of Suzhou’s bus system by examining the correlations, importance, and sensitivity of feature tags. These insights into bus ridership will enable city planners and policymakers to make informed decisions regarding public transport network improvements and adjustments, highlighting key indicators that impact the performance and efficiency of Suzhou's bus network.

- *Correlation analysis by Spearman's rank correlation coefficient*

Figure 7 illustrates a comprehensive correlation analysis amongst various feature tags utilised in Suzhou's bus system. This correlation matrix is a crucial element of the feature importance

evaluation in the framework for acquiring natural features of bus ridership in Suzhou. It provides an in-depth insight into the interactions between different feature tags and their collective impact on ridership patterns.

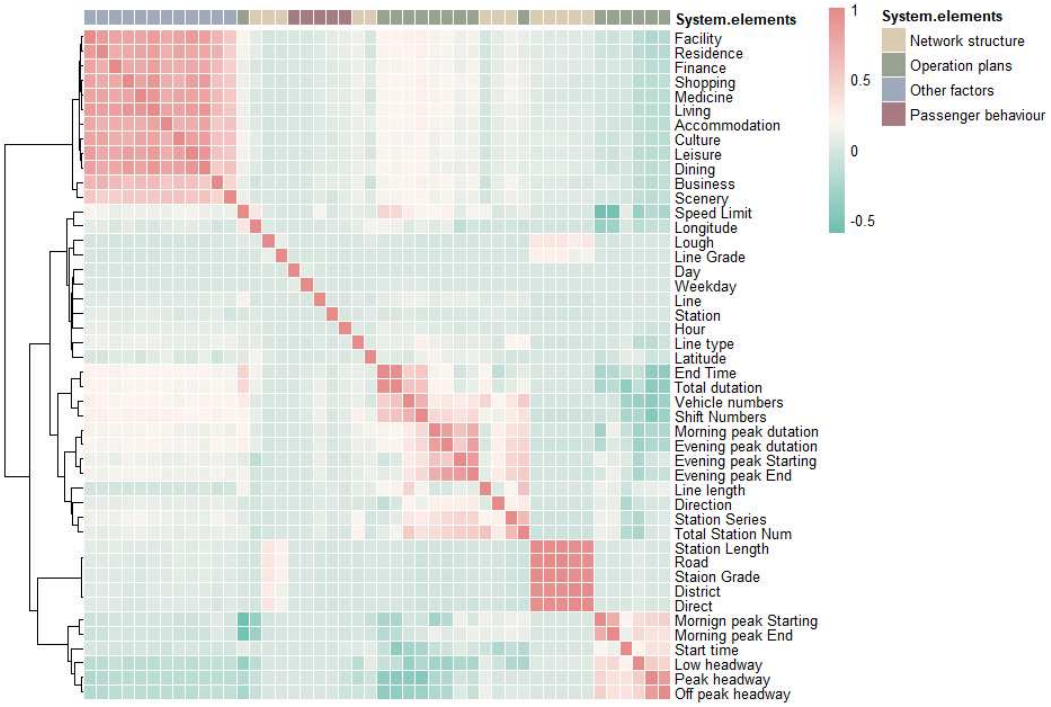


Figure 7 Correlation matrix illustrating the relationships among feature tags in Suzhou's bus system.

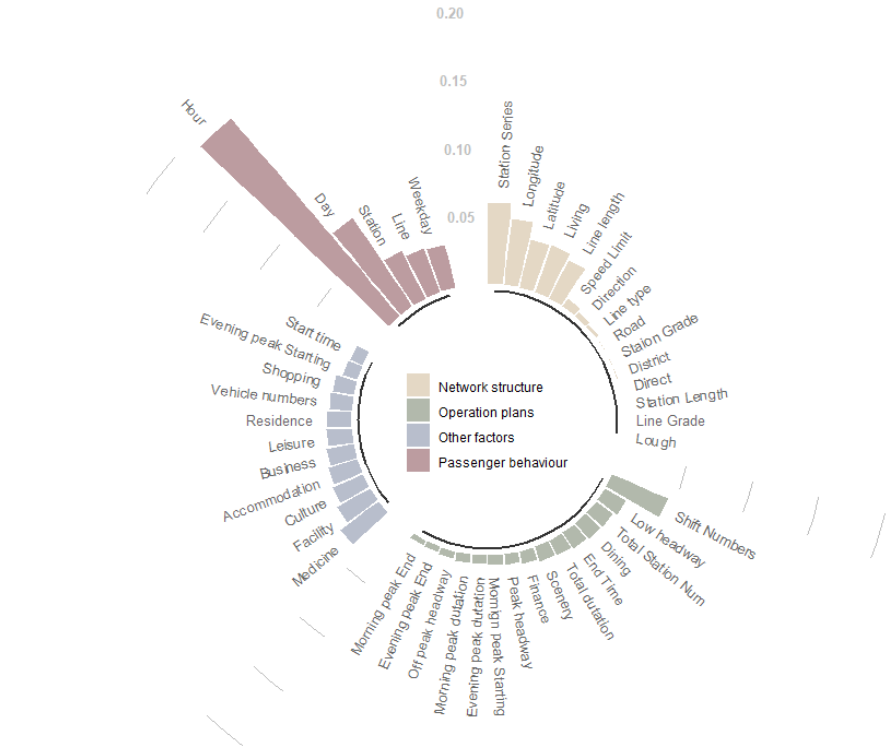
Upon the application of hierarchical clustering, it is observed that each cluster predominantly contains feature tags pertaining to a singular system element within the feature tagging systems. For example, tag entities associated with POI coalesce into a distinct cluster (visible in the top left corner of the matrix), with a similar congregation observed among tag entities relating to ‘passenger behaviour’. While feature tags from ‘network structure’ and ‘operation plans’ exhibit some overlap in clustering, feature tags of a more granular nature, such as those related to 'bus station', nonetheless preserve their affinity and are clustered together.

A notable observation from our correlation matrix is the prevailing trend of positive correlation among tag entities within the same cluster or system element. In contrast, a mild negative correlation characterises the relationship between tag entities from disparate clusters. This pattern provides a discernible indication of the specific interplay of features within individual system elements and their potential independence or divergent influence in relation to features from other system elements. Thus, these findings further underline the specificity and structure inherent in our feature tagging system, showcasing its capacity to effectively differentiate and categorise feature tags in relation to their source system elements.

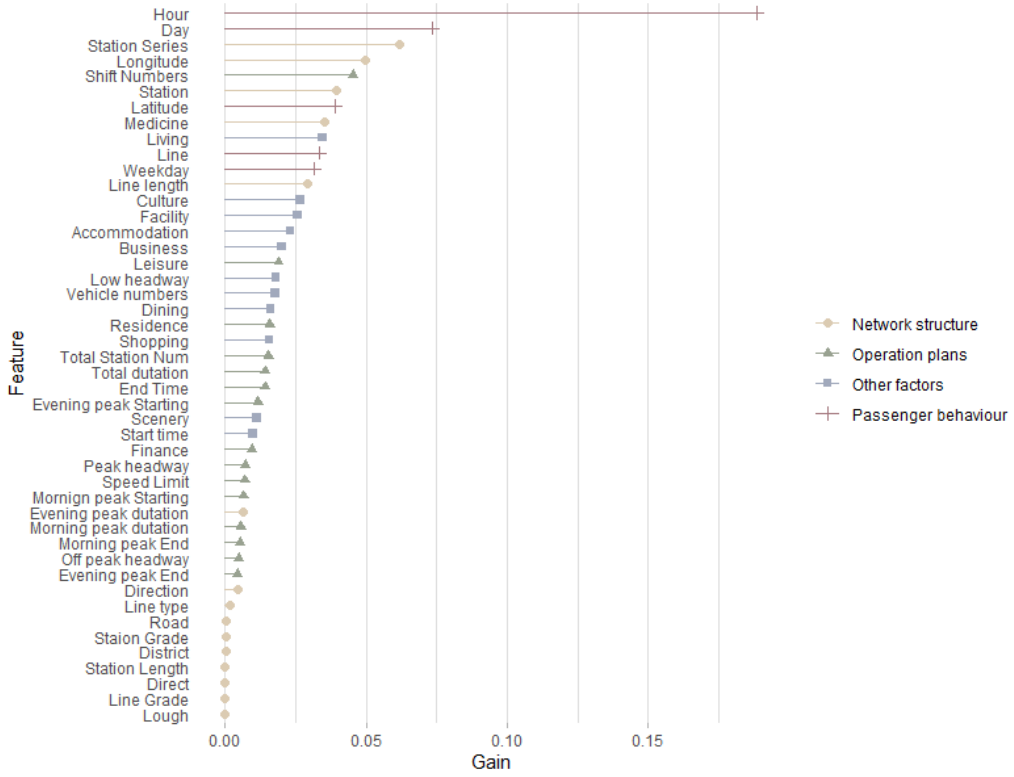
- *Importance analysis by Gain indicator in the XGBoost model*

Figure 8 presents an insightful ranking of the feature tags employed in the natural feature profile of Suzhou's bus system. This ranking, derived from the Gain indicator within the XGBoost

model, measures the relative importance of each feature tag in the context of public transport ridership in Suzhou.



(a)



(b)

Figure 8 Relative importance ranking of feature tags to profile the natural feature of Suzhou's bus ridership via the Gain indicator within the XGBoost model. (a) Intra-group rankings, and (b) overall rankings.

The ‘hour’ tag emerges as the most influential, accounting for nearly a fifth of the total relative importance. It's twice as important as the second most significant tag, ‘day’. Even though there are fewer feature tags under ‘passenger behaviour’, their overall relative importance is highest, accounting for 0.37 in total. Therefore, passenger travel time and location play a critical role in influencing the bus ridership in Suzhou. Subsequent tags of significance are network structure-related, e.g., ‘station series’, ‘longitude’, and ‘latitude’. These location-associated tags reflect that bus ridership is closely tied to geographical variables. Contrarily, some feature tags describing station attributes, like ‘lough’, ‘station length’, ‘line grade’, carry a low relative importance. This suggests that while the station infrastructure may add to the convenience of waiting passengers, they do not significantly impact the bus ridership in Suzhou. Within ‘operation plans’, ‘shift numbers’ is the most significant feature tag, ranking fifth and surpassing other tags in this category by a wide margin—it's 2.4 times more important than ‘low headway’.

The feature tags related to ‘operation plans’ in Suzhou’s bus ridership system generally have lower overall rankings. Within this system, there are 14 distinct feature tags, but collectively, they account for only 0.17 of the relative importance. This indicates that operational factors are not major influencers of bus ridership patterns in the city. On the other hand, feature tags related to POIs usually hold a moderate level of importance. Among these, ‘living’ and ‘medicine’ carry greater importance, ranking ninth and tenth, respectively. Conversely, ‘scenery’ and ‘finance’ are less influential, occupying the 27th and 28th positions (within the top two-thirds). The overall relative importance of POI-related tags exceeds that of the other two system element categories, ‘operation plans’ and ‘network structure’. This finding highlights the significant role that POIs have in influencing public bus ridership trends in Suzhou.

- *Sensitivity analysis by feature ablation experiment*

Our analysis scrutinises the sensitivity of feature tags in contributing to the understanding of Suzhou's bus passenger flow. Figure 9 represents the root mean square error (RMSE) derived from the XGBoost model used to predict hourly passenger flow at bus stops based on different feature tags. In this representation, the x-axis sequentially introduces the feature tags into the model, ordered in descending significance according to their relative importance, as determined by the rankings in Figure 8.

Despite ‘hour’ being the most significant feature, its individual contribution is insufficient for a comprehensive profiling of natural features in bus passenger flow. Following the integration of the fourth and fifth most important features, the RMSE value exhibits a substantial reduction by 33% and 41% respectively, compared to the initial value. Correspondingly, the cumulative relative importance indicators increase to 0.39 and 0.43. With the integration of the top 11 significant feature tags, the RMSE plunges from 7.13 to 3.96 (a reduction by 56%) as the cumulative relative importance indicator reaches 64%. The feature tags of high significance encompass those belonging to the ‘passenger behaviour’ category, however, none from the ‘operation plans’ category.

Furthermore, the POI-related feature tags of ‘living’ and ‘medicine’ are observed to exert a substantial influence on Suzhou’s bus passenger flow. Beyond this point, the addition of further feature tags does not enhance the model's performance. To summarise, the profiling of natural features in bus passenger flow necessitates a holistic incorporation of relevant and critical feature tags, rather than a reliance on singular, albeit significant, tags. The robustness and precision of such a model are optimised by the thoughtful inclusion of diverse and interrelated feature tags that capture the multifaceted nature of bus passenger flow.

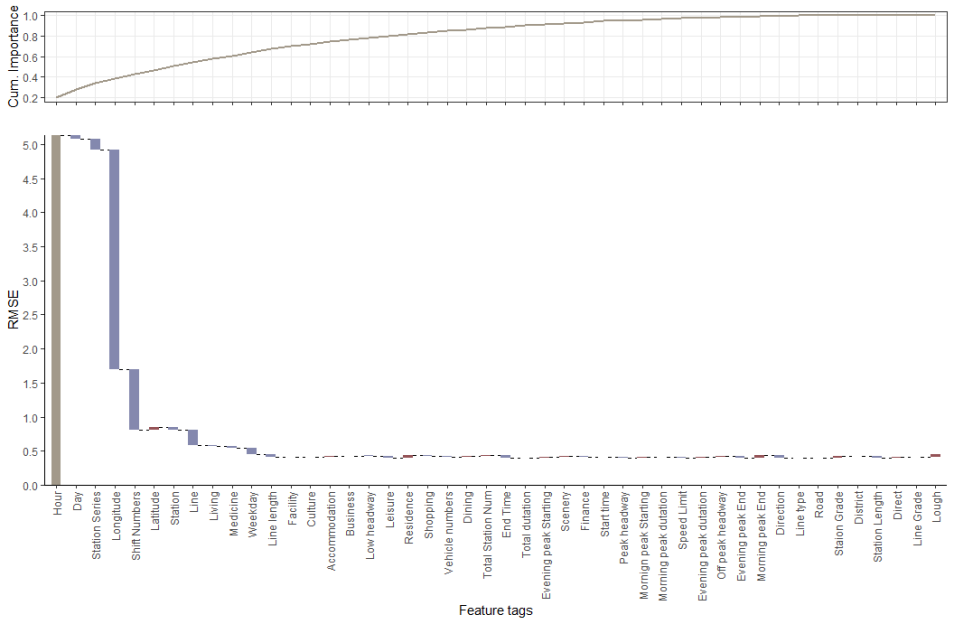


Figure 9 Sensitivity analysis of feature tags in Suzhou's bus passenger flow.

For Suzhou, the natural features of public transport demand highlight the significant influence of the surrounding POIs, especially medical facilities, on bus station passenger flows, with temporal fluctuations emphasizing intra-day variations in ridership.

4.3.3. Feature analysis in Lianyungang’s bus system

This section investigates the natural feature of Lianyungang's bus network, with a focus on the correlations, significance, and sensitivity of diverse feature tags in relation to the intrinsic passenger flow dynamics.

- *Correlation analysis by Spearman's rank correlation coefficient*

Figure 10 presents the correlations among different tag entities within the system elements utilised in the Lianyungang bus system. The fundamental characteristics of this figure bear a resemblance to those observed in Figure 7. Tag entities that fall within the same system elements are predominantly clustered together.

Within the context of Lianyungang's feature tags, those belonging to ‘network structure’ and ‘operation plans’ system elements tend to cluster together, rather than showing the interweaving pattern observed in the data from Suzhou. However, the associations between these tag entities do

not exhibit the same strength as in the case of Suzhou, with most showing a weak negative correlation. The vast majority of tag entities related to POI-related form a distinct cluster and exhibit a positive correlation amongst themselves. Yet, three POI-related tags (i.e., accommodation, facility, and scenery) exhibit minimal correlation with other POI-related tags. Furthermore, POI-related tag entities generally demonstrate a mild or weak negative correlation with tag entities belonging to other systems. This analysis illuminates the differential behavioural characteristics of feature tags between Lianyungang and Suzhou, thereby emphasising the flexibility and adaptability of the applied feature analysis in capturing public transport ridership variations across different urban contexts.

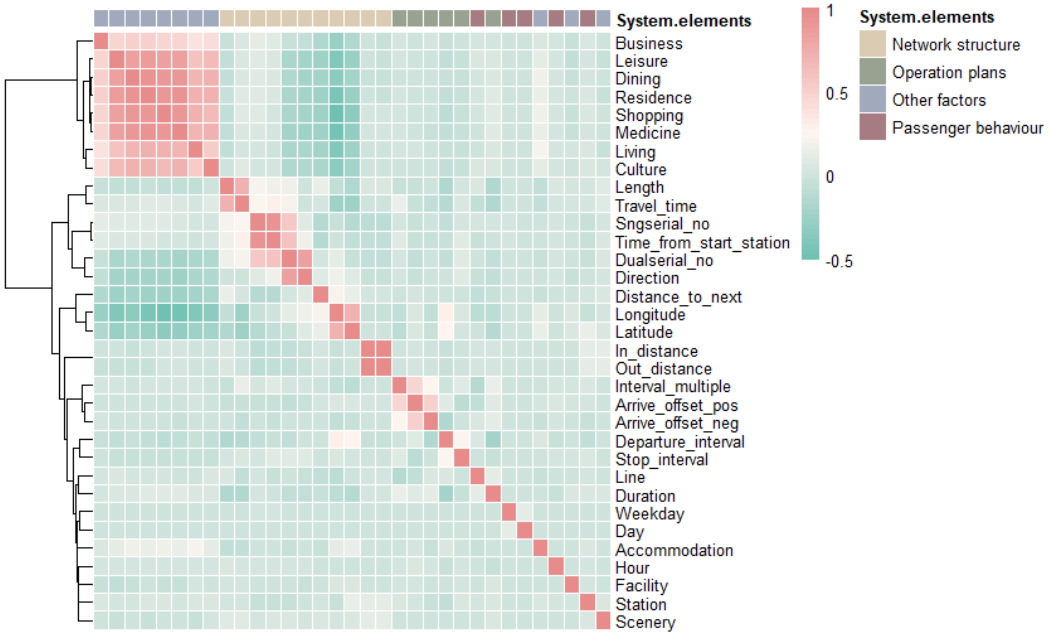


Figure 10 Correlations among different tag entities within the systems elements utilised in Lianyungang's bus system'

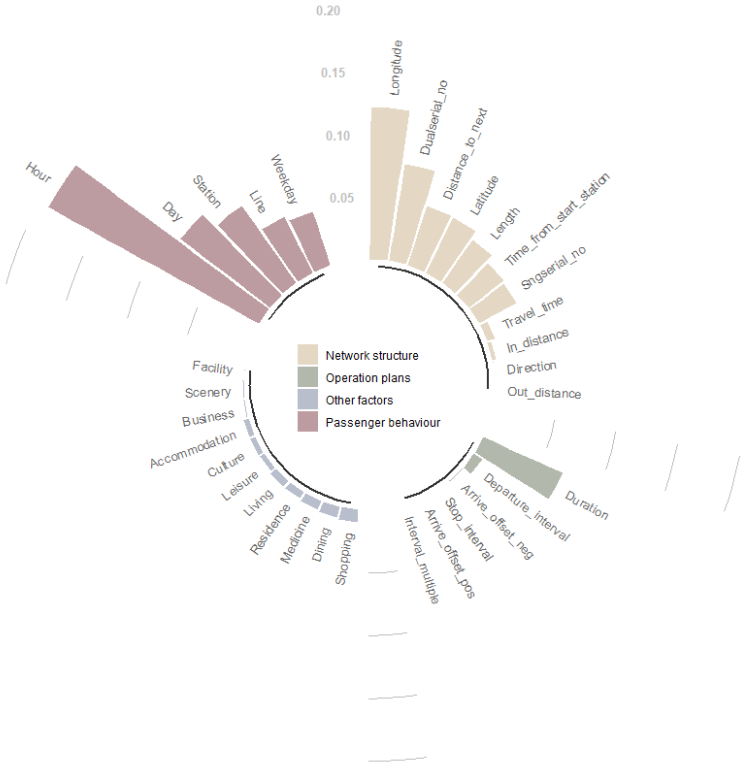
- *Importance analysis by Gain indicator in the XGBoost model*

Figure 11 showcases the importance of various factors in determining the intrinsic passenger flow patterns within Lianyungang's bus network, based on the Gain indicator of the XGBoost model.

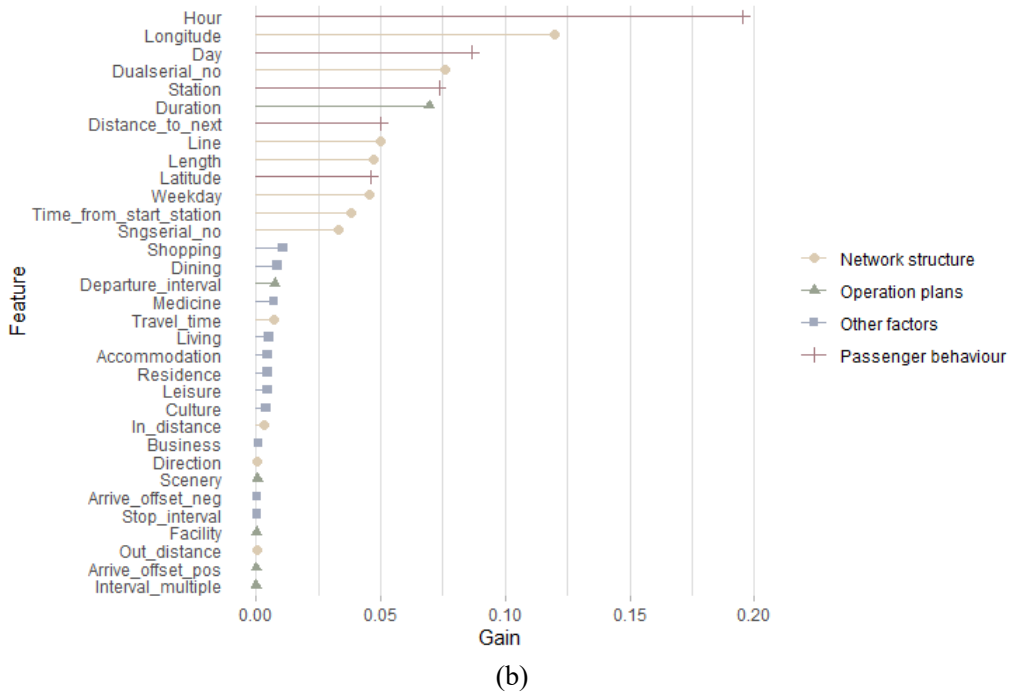
Paralleling the patterns identified in Suzhou, 'hour' emerges as the most salient factor, denoting those fluctuations in Lianyungang's bus ridership over the course of the day are distinctly pronounced. Tag entities within the 'passenger behaviour' system maintain a significant presence, amassing a collective relative importance of 0.48. Notably diverging from the case of Suzhou, however, are the tag entities associated with the 'network structure', which exhibit heightened importance in Lianyungang. Longitude of the bus stations, for instance, ranks second among all tags, and the total relative importance for the entire 'network structure' system stands at 0.45.

In terms of the 'operation plans', while the operation duration of bus routes secures the sixth position in terms of individual importance, the remaining tag entities within this system are of substantially diminished importance, appearing mostly in the lower rankings. The importance of

POI-related tag entities in Lianyungang also deviates from those observed in Suzhou. In Lianyungang, POI-related tags generally do not significantly influence bus ridership patterns. These results, taken together, indicate the distinctive nature of public transport ridership dynamics in Lianyungang.



(a)



(b)

Figure 11 Relative importance of various feature tags in the natural feature profile of Lianyungang's bus ridership based on the Gain indicator in the XGBoost model. (a) Intra-group rankings, and (b) overall rankings.

- *Sensitivity analysis by feature ablation experiment*

The sensitivity analysis illustrated in Figure 12 shows a consistent marginal impact of high-dimensional feature tags on bus ridership in Lianyungang. The RMSE, a measure of prediction accuracy, undergoes a stair-step change as feature tags are incrementally included into the model.

A notable reduction of 31% in RMSE is witnessed following the integration of the tag entity of 'longitude,' at this point, the cumulative relative importance index stands at 39%. Upon the inclusion of two additional significant tag entities, the RMSE experiences a further decline of 64% from its initial value, leading the cumulative relative importance index to reach 48%. When the ten most crucial tag entities are considered, the cumulative relative importance index ascends to 81%, and the RMSE significantly drops from its initial 4.61 to 0.65, representing a decrease of 86%. Mirroring the trends observed in Suzhou, the addition of further tag entities does not confer any substantial improvements to the model's predictive accuracy. This suggests that the variance of some subsequent tag entities does not provide a reliable reflection of their impact on fluctuations in bus ridership.

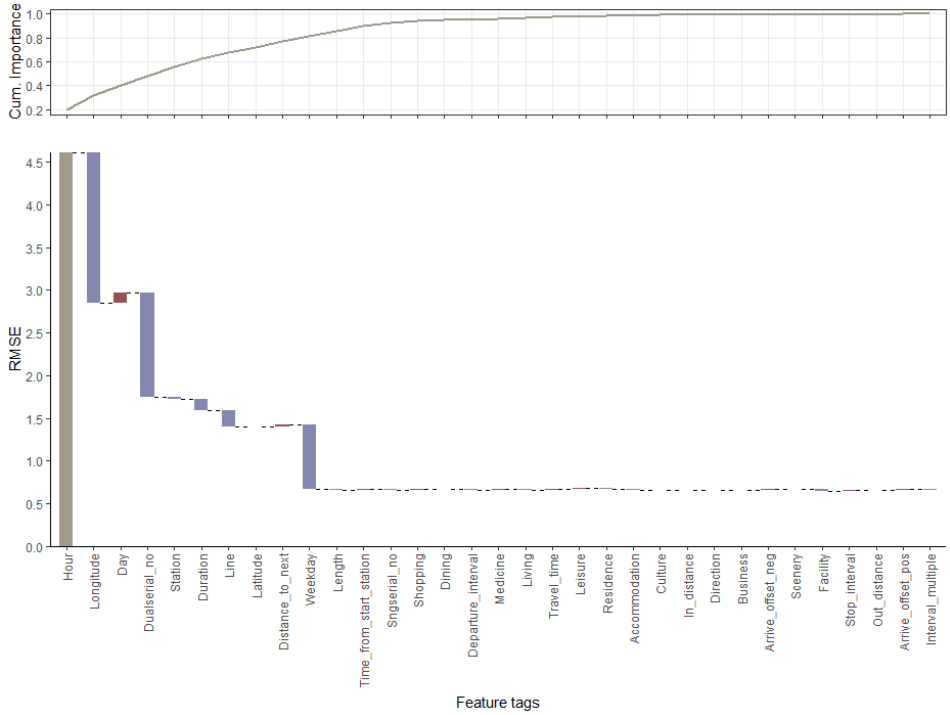


Figure 12 Sensitivity analysis of high-dimensional feature tags in Lianyungang's bus network.

The results of the sensitivity analysis underscore the influence of tag entities such as 'hour', 'longitude', 'day', 'dual serial no', 'station', 'duration', 'distance to next', 'weekday', 'latitude', and 'line' on the ridership of Lianyungang's buses. These factors characterise the natural features of bus ridership in Lianyungang, providing vital insights into the city's public transport dynamics.

Drawing a comparative analysis with the findings from Suzhou, we can discern that, despite employing the same feature analysis methods, the resultant conclusions bear both commonalities and differences. This highlights the distinct characteristics that define bus ridership natural features

in different cities. Our framework, therefore, demonstrates remarkable versatility and adaptability, suitable for application in varied urban settings. Importantly, it showcases a refined capacity to capture the subtle variations and distinct characteristics that differentiate one city from another in terms of public transport ridership dynamics. This emphasises the need for a context-specific approach when analysing and interpreting public transport data, reinforcing the validity and applicability of our methodology across a range of scenarios.

In summary, the dominant natural features for Lianyungang emphasise the importance of route design and bus station placement, with ridership also showing pronounced intra-day fluctuations but being more influenced by the spatial distribution of bus routes and stations.

4.3.4. Diversity of natural features between cities of Suzhou and Lianyungang

Comparing the diversity of natural features between Suzhou and Lianyungang reveals significant differences in the scale and layout of these two cities, as well as the distribution of POIs and bus station locations. Suzhou's bus network covers a large portion of residents' travel routes, facilitating commuting, leisure activities, and errands through public transport. In contrast, Lianyungang's bus network focuses more on daily activities such as commuting and shopping for residents. Moreover, while Suzhou's bus network is larger and more complex overall, the correlations between different feature tags are more apparent, allowing for consolidation and simplification in data analysis, model building, and policy-making.

Further explorations of the natural features of bus passenger flows between Suzhou and Lianyungang yield striking similarities. Both cities' bus ridership prominently demonstrates intra-day fluctuations, reflecting daily oscillations in ridership volume. Although the influence of specific days and weekdays is discernible, its impact is not as potent. Such understanding of bus ridership patterns has significant implications for the operational strategies of public transport systems in both cities. It is suggested that both Suzhou and Lianyungang transit authorities could consider implementing measures to adapt to the prominent intra-day fluctuations. This may include dynamic scheduling and frequency adjustments, ensuring that the supply of bus services aligns with the demand throughout the day. Furthermore, even though the influence of specific days and weekdays is less substantial, transport authorities could still enhance service efficiency by making modest adjustments to accommodate these minor variations in demand.

In terms of spatial distribution, Suzhou exhibits a stronger relationship between bus station passenger flows and surrounding POI attributes and quantities, especially medicine and facilities. This suggests that the location of essential services, such as healthcare facilities, exerts a notable influence on passenger flow, potentially driving demand in their vicinity. Conversely, the spatial distribution of bus passenger flows in Lianyungang is more dependent on the relationship between bus stations and routes, indicating that the route design and bus station placement may have a more prominent influence on passenger flows in Lianyungang. Therefore, public transit planning in Lianyungang might be more effectively optimised by focusing on the strategic allocation of bus

stations and the rational design of bus routes, to better correspond with passenger demand patterns.

Suzhou, an economically developed city, offered us a window into how established urban infrastructures impact bus ridership. Our natural feature profile of Suzhou revealed some telling patterns. The dominance of the ‘hour’ feature underscores the critical role of time-specific factors, such as office hours and commercial activities. Interestingly, while the ‘passenger behaviour’ category had fewer tags, features like travel time and location stood out in their influence on ridership. This emphasis on temporal and behavioural attributes contrasts with the geographical variables, where tags such as ‘longitude’ and ‘latitude’ indicate the significance of the city’s geography on ridership. Operationally, certain tags like 'shift numbers' were notably crucial, but as a whole, operational factors didn't significantly sway Suzhou's bus ridership.

In contrast, Lianyungang, presented a different developmental backdrop, allowing us to discern varied influences on its public transport ridership. Here too, the ‘hour’ feature was pivotal, highlighting the universal significance of daily patterns in public transportation. However, the city’s natural feature profile diverged in other aspects, notably in the network structure. The heightened influence of tags like the longitude of bus stations paints a different picture compared to Suzhou. Moreover, the diminished role of Points of Interest (POIs) and the lower ranking of most operational plan tags underscore the distinctive nature of Lianyungang's public transport dynamics.

Comparing these insights side-by-side, a few observations stand out. Both cities reiterate the importance of time-based factors (e.g., ‘hour’) in influencing bus ridership, pointing towards universally significant daily patterns in urban areas. However, the increased significance of network structure attributes in Lianyungang suggests that geographical dynamics play a more influential role in cities undergoing rapid urbanisation. The contrasting importance of POIs between the two cities further illustrates this: while established urban centres like Suzhou exhibit considerable influence from surrounding amenities and services, cities like Lianyungang, in their growth phase, might not exhibit the same dependencies.

These findings highlight the importance of considering both similarities and differences in public transport planning and policy-making for different cities. While expert experience can still provide valuable insights, the era of big data presents opportunities for fine-grained analysis and discussion of natural features of bus passenger flows. The proposed basic technical framework in this section can facilitate the customisation of natural feature profiles for different cities and different feature data, allowing for more targeted and effective strategies in improving public transport systems.

Conclusively, the comparison of the diversity of natural features between Suzhou and Lianyungang reveals significant differences in their scale, layout, and the distribution of POIs and bus station locations. While both cities exhibit similarities in the correlations between bus passenger flows and time and space, there are differences in the details. Spatially, Suzhou shows a stronger relationship between bus station passenger flows and surrounding POI attributes and quantities, while Lianyungang’s bus passenger flows are more influenced by route design and bus station

placement. These findings underscore the importance of considering both similarities and differences in public transport planning and highlight the opportunities and challenges presented by big data in improving public transport systems.

5. Discussions and implications

The natural feature profile framework for data-driven urban transport system analysis presents a distinctive and invaluable instrument for discerning the inherent characteristics of transport systems across cities. Its adaptability to varying cities and time periods, dependence on core key technologies, and case-specific application render it a potent resource for guiding transport planning and policy-making. The ramifications of this framework for implementation and decision-making are substantial.

One primary implication of the natural feature profile framework lies in its capacity to capture the unique attributes of different cities. By analysing extensive datasets, the framework can extract natural feature profiles reflecting the particular aspects of a city's transport system, such as bus passenger flows, temporal and spatial correlations, and spatial distribution patterns. This city-specific understanding can inform bespoke strategies and policies for enhancing transport efficiency and effectiveness, tailored to the requirements and characteristics of each city.

Moreover, the framework's capability to analyse data from disparate time periods enables temporal analysis and comprehension of transport system dynamics. By comparing bus passenger flows during weekdays versus weekends, or across distinct seasons or holidays, the framework can unveil insights into shifting travel patterns and demand fluctuations. This temporal analysis can prove invaluable in guiding transport planning and policy-making, particularly in dynamic urban environments where transport needs evolve over time.

The core key technologies underpinning the natural feature profile framework, such as data mining, machine learning, and spatial analysis, provide the foundation for its ability to extract meaningful insights from large-scale datasets. These technologies facilitate the identification of patterns and correlations within the data, the creation of predictive models for transport system attributes, and the exploration of spatial dynamics of urban transport. This accentuates the importance of harnessing big data and advanced technologies in urban transport research and the implications for advancing the field.

Our proposed methodology serves as a foundational blueprint for understanding the natural feature profile of public transport ridership. We emphasise that every city, with its unique characteristics ranging from population density to transportation infrastructures, demands a tailored approach. Recognising this, the case-specific application of our framework offers significant implications. It can be adapted and customised to diverse urban settings, enabling not just a standalone analysis, but also facilitating comparisons across cities. Such comparative insights can reveal both similarities and differences in various transport systems, playing a pivotal role in identifying best practices and facilitating knowledge transfer among cities. Harnessing local data

and gaining an in-depth understanding of city-specific transport challenges are pivotal for the optimal application of our framework. Moreover, its flexibility extends to its applicability over varying time periods, offering a longitudinal perspective on transport system evolutions and trends. This inherent versatility and adaptability make our framework a robust tool for delving deep into the idiosyncrasies of different cities and their transport dynamics.

The natural feature profile framework serves as a pivotal tool for urban planners and policymakers. This framework allows us to dive deep into the intricacies of public transport ridership, offering a dual perspective by capturing both spatial and temporal dynamics. As these main points are revealed, we gain enhanced capability to pinpoint the underlying reasons for changes in passengers' preferences, the variations in ridership at different times of day or week, and the impact of external elements such as urban events or infrastructural modifications.

One of the profound revelations of our research is the potential areas of intervention in the LoS. As we understand more about ridership characteristics, it becomes evident where improvements in LoS can be most impactful. For instance, if a certain route consistently witnesses low ridership despite high potential demand, it might be an indicator of sub-optimal service frequency or quality. Alternatively, patterns of high patronage during certain times might point towards the need for more vehicles or increased frequency during peak hours.

The data-driven nature of the natural feature profile framework also bears implications for decision-making in transport planning and policy-making. By delivering evidence-based insights and recommendations grounded in actual data, the framework can facilitate more targeted and efficacious strategies for ameliorating urban transport systems. This focus on evidence-based decision-making can contribute to the development of intelligent and sustainable cities, where transport planning and policy-making are rooted in data and empirical analysis.

Furthermore, the findings from our natural feature profile framework offer pivotal insights for policy-making in the realm of Intelligent Transportation Systems (ITS). By integrating this framework into ITS development, we can significantly improve the efficiency and effectiveness of urban public transport systems. The framework's ability to analyse and interpret complex, multi-source data sets enables a deeper understanding of urban transport dynamics, which is crucial for ITS-related policy development. Policies concerning data management, system design, and service optimization can greatly benefit from the refined insights provided by our framework. Particularly, the framework's capacity to identify and analyse the intrinsic patterns of public transport ridership can aid policymakers and urban planners in crafting data-driven, evidence-based strategies for ITS implementation. These strategies could encompass advanced traffic management systems, dynamic route planning, and real-time passenger information systems, all tailored to the specific needs of individual cities. The integration of our framework into ITS policy-making and planning processes marks a significant stride towards intelligent, efficient, and user-centric urban transport systems.

As we transition towards sustainable urban ecosystems, the role of public transport becomes even more pronounced. Our research lays down a pathway not just for understanding but for acting

upon the challenges faced by public transport systems. Through evidence-backed interventions, we can reinvigorate bus patronage, making public transport the preferred choice for urban dwellers and advancing the vision of a sustainable transit metropolis.

6. Conclusion

Urban public transport systems, comprising elements such as passengers, stations, and routes, are influenced by a multitude of factors. The extensive data generated by these systems presents challenges due to diverse sources, storage methods, and heterogeneous data types. Addressing the challenges associated with multi-source, heterogeneous, massive, holistic, high-temporal, and high-spatial correlation data, this paper presents a problem-oriented framework to obtain comprehensive natural feature profiles of public transport ridership. This framework enables the extraction of valuable information pertinent to urban transport services, management, and decision-making, facilitating a precise perception of urban transport systems.

A key contribution of this research is the introduction of a novel natural feature profile framework for data-driven public transport ridership analysis. This innovative approach enables the identification of natural features of bus ridership across different cities. By utilising big data and advanced technologies such as data mining, machine learning, and spatial analysis, the framework proves effective in guiding transport planning and policy-making.

Furthermore, this research fills a crucial gap often overlooked in current studies. While many investigations delve into city-specific issues, such as the impact of weather on bus travel in cities like Suzhou, they tend to focus intensely on particular cases or the development of model details. Such studies rarely consider broader applicability or transferability to other contexts. Our natural feature profile framework, however, offers a solution to this limitation. It is designed to connect complex data, varied technological approaches, and diverse application scenarios within a universal framework. This holistic approach extends the utility of our research beyond specific case studies, making it adaptable to a wide array of urban contexts. This universality represents a key innovation of our framework, distinguishing it from narrower studies and enabling broader implementation across various urban settings.

Through case studies conducted in Suzhou and Lianyungang, key and interesting findings have emerged. For instance, while both cities exhibit similarities in the correlations between bus passenger flows and time and space, there are differences in the details. Spatially, Suzhou shows a stronger relationship between bus station passenger flows and surrounding POI attributes and quantities, while Lianyungang's bus passenger flows are more influenced by route design and bus station placement.

The research problem addressed herein necessitates a customisable and adaptable framework capable of capturing city-specific characteristics, analysing temporal dynamics, and providing evidence-based insights for transport planning and policy-making. The key findings and results of this research underscore the potential of the natural feature profile framework in fulfilling these

objectives.

A principal contribution and innovation of this research is the development of the natural feature profile framework itself, furnishing a systematic and data-driven method for comprehending urban transport systems. Although the tripartite division of our framework, encompassing the data layer, feature layer, and the application layer, might resonate with conventional structures, its essence encapsulates profound novelty. Most existing research ventures into addressing tailored problems, pivoting around unique case-specific conditions. However, the absence of a holistic, adaptable framework in existing literature underscores the pioneering nature of our approach. What sets our framework apart is its inherent versatility. It embraces a wide gamut of data conditions and feature variances, thereby offering a canvas expansive enough to cater to myriad applications. The framework's adaptability to varying cities and time periods, dependence on core key technologies, and case-specific application render it a distinctive and invaluable resource for transport analysis. By virtue of this adaptability, it transcends the limitations of specificity, rendering it an indispensable, all-encompassing tool that champions practicality in diverse real-world transport scenarios.

In addition to its adaptability, another aspect that bolsters the innovation is the intricate interlinking between its constituent layers. This facilitates a transformative journey, transmuting raw, often nebulous, data into structured insights and, subsequently, actionable strategies. Such a cohesive and fluid transition, bridging the chasm between data and real-world application, is seldom found in other frameworks.

Additionally, our approach's granularity, evident in the four-level feature tagging system, sets a new benchmark. It champions a depth of analysis that transcends merely scratching the surface. By delving into minute patterns and correlations, our framework brings to light often-overlooked intricacies in transportation data, enriching our understanding and enhancing our strategies' efficacy. Another pivotal innovation is our accentuated focus on spatial and temporal dynamics within the feature layer. While many frameworks might acknowledge these aspects, the profundity with which our structure incorporates them is unique. By capturing these fleeting yet impactful trends, we are better positioned to anticipate and cater to the ever-evolving demands of urban transportation, ensuring that our strategies remain not just relevant but pioneering in the face of rapid urban metamorphoses.

The ramifications of this research are considerable for implementation and decision-making in transport planning and policy-making. By harnessing our framework's adaptive nature, it becomes possible to accommodate the ever-evolving transportation needs of cities. Its distinctive capacity to capture city-specific characteristics and decipher temporal dynamics equips stakeholders with a detailed understanding, facilitating the design of bespoke strategies and policies. These are not generic; they are tailored to the unique attributes and requirements of each urban environment. The tangible, evidence-based insights put forth by our framework pave the way for not only more efficient transport planning but also more effective policy-making. Through this integration, cities

are empowered to progress toward being smarter, more responsive, and ultimately, more sustainable in their transport initiatives.

Nonetheless, this research exhibits certain limitations. One constraint is the reliance on large-scale datasets, which may not be obtainable or accessible for all cities or time periods. Another limitation is the dependence on the quality and accuracy of the data employed in the analysis, as well as the assumptions made during the modeling process. These limitations ought to be considered when applying the natural feature profile framework in practice.

For future research, several avenues can be explored. Firstly, further refinement and validation of the natural feature profile framework can be undertaken to enhance its accuracy and reliability. Secondly, the framework can be applied to additional cities and time periods to broaden its generalisability and applicability. Furthermore, integrating other pertinent factors, such as socio-economic and environmental variables, into the framework can yield a more comprehensive understanding of urban transport systems. While our framework provides a comprehensive overview of public transport ridership, we acknowledge that the breadth of potential analytical techniques exceeds our present scope. Subsequent research will delve deeper into these specific technical details, enriching the overall framework's substance and depth. Lastly, examining the potential of integrating the natural feature profile framework with other decision support tools or models can further augment its effectiveness in informing transport planning and policy-making.

The natural feature profile framework for data-driven public transport ridership analysis presents a promising approach to understanding the natural features of mobility within urban public transport systems in different cities. The framework's adaptability, reliance on core key technologies, case-specific application, and evidence-based decision-making hold significant implications for transport planning and policy-making. While this research has limitations, it contributes to the advancement of the field and provides a foundation for future studies in this area. Through the case studies conducted in Suzhou and Lianyungang and the interesting findings obtained, the natural feature profile framework demonstrates its potential to uncover valuable insights and facilitate informed decision-making in urban transport systems.

Acknowledgements

We would like to acknowledge the support from the National Natural Science Foundation of China (No. 52131203) and High-Level Personnel Project of Jiangsu Province, China (No. JSSCBS20220099). Tianli Tang is supported by the Jiangsu Funding Program for Excellent Postdoctoral Talent (No. 2022ZB114), the Natural Science Foundation of Jiangsu Province (No. BK20230852), and the 'Chunhui Jihua' of the Ministry of Education, China (No. HZKY20220156).

References

Arana, P., Cabezudo, S., Peñalba, M., 2014. Influence of weather conditions on transit ridership: A statistical study using data from Smartcards. Transp Res Part A Policy Pract 59. https://doi.org/10.1016/j.tra.2013.10.019

Berrebi, S.J., Watkins, K.E., 2020. Who's ditching the bus? Transp Res Part A Policy Pract 136, 21–34. https://doi.org/10.1016/J.TRA.2020.02.016

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

Bradley, M., Bowman, J.L., Griesenbeck, B., 2010. SACSIM: An applied activity-based model system with fine-level spatial and temporal resolution. Journal of Choice Modelling 3, 5–31. https://doi.org/10.1016/S1755-5345(13)70027-7

Carpio-Pinedo, J., 2014. Urban Bus Demand Forecast at Stop Level: Space Syntax and Other Built Environment Factors. Evidence from Madrid. Procedia Soc Behav Sci 160, 205–214. https://doi.org/10.1016/J.SBSPRO.2014.12.132

Cervero, Robert., 1998. The Transit Metropolis: A Global Inquiry. Island Press.

Chen, S., Liu, X., Lyu, C., Vlacic, L., Tang, T., Liu, Z., 2023. A holistic data-driven framework for developing a complete profile of bus passengers. Transp Res Part A Policy Pract 173, 103692. https://doi.org/10.1016/J.TRA.2023.103692

Chen, Z., Liu, K., Wang, J., Yamamoto, T., 2022. H-ConvLSTM-based bagging learning approach for ride-hailing demand prediction considering imbalance problems and sparse uncertainty. Transp Res Part C Emerg Technol 140, 103709. https://doi.org/10.1016/j.trc.2022.103709

Currie, G., Delbosc, A., 2011. Understanding bus rapid transit route ridership drivers: An empirical study of Australian BRT systems. Transp Policy (Oxf) 18, 755–764. https://doi.org/10.1016/j.tranpol.2011.03.003

Deepa, L., Rawoof Pinjari, A., Krishna Nirmale, S., Srinivasan, K.K., Rambha, T., 2022. A direct demand model for bus transit ridership in Bengaluru, India. Transp Res Part A Policy Pract 163, 126–147. https://doi.org/10.1016/J.TRA.2022.07.004

Erhardt, G.D., Hoque, J.M., Goyal, V., Berrebi, S., Brakewood, C., Watkins, K.E., 2022. Why has public transit ridership declined in the United States? Transp Res Part A Policy Pract 161, 68–87. https://doi.org/10.1016/J.TRA.2022.04.006

Gu, Z., Saberi, M., Sarvi, M., Liu, Z., 2018. A big data approach for clustering and calibration of link fundamental diagrams for large-scale network simulation applications. Transp Res Part C Emerg Technol 94, 151–171. https://doi.org/10.1016/j.trc.2017.08.012

Gu, Z., Wang, Z., Liu, Z., Saberi, M., 2022. Network traffic instability with automated driving and cooperative merging. Transp Res Part C Emerg Technol 138, 103626. https://doi.org/10.1016/j.trc.2022.103626

Gutiérrez, J., Cardozo, O.D., García-Palomares, J.C., 2011. Transit ridership forecasting at station level: an approach based on distance-decay weighted regression. J Transp Geogr 19, 1081–1092. https://doi.org/10.1016/j.jtrangeo.2011.05.004

He, J., Yan, N., Zhang, J., Yu, Y., Wang, T., 2022. Battery electric buses charging schedule optimization considering time-of-use electricity price. Journal of Intelligent and Connected Vehicles 5, 138–145. https://doi.org/10.1108/JICV-03-2022-0006

Hu, S., Chen, P., 2021. Who left riding transit? Examining socioeconomic disparities in the impact of COVID-19 on ridership. Transp Res D Transp Environ 90, 102654. https://doi.org/10.1016/J.TRD.2020.102654

Huo, X., Wu, X., Li, M., Zheng, N., Yu, G., 2020. The allocation problem of electric car-sharing system: A data-driven approach. Transp Res D Transp Environ 78, 102192. https://doi.org/10.1016/j.trd.2019.11.021

Iliashenko, O., Iliashenko, V., Lukyanchenko, E., 2021. Big Data in Transport Modelling and Planning. Transportation Research Procedia 54, 900–908. https://doi.org/10.1016/j.trpro.2021.02.145

Jiang, W., Zheng, N., Kim, I., 2023. Missing data imputation for transfer passenger flow identified from in-station WiFi systems. Transportmetrica B: Transport Dynamics 11, 325–342. https://doi.org/10.1080/21680566.2022.2064935

Johnson, M., 2021. Annual Bus Statistics: Year Ending March 2021. London.

Khalil, S., Amrit, C., Koch, T., Dugundji, E., 2021. Forecasting Public Transport Ridership: Management of Information Systems using CNN and LSTM Architectures. Procedia Comput Sci 184, 283–290. https://doi.org/10.1016/J.PROCS.2021.03.037

Kim, D., Ahn, Y., Choi, S., Kim, K., 2016. Sustainable Mobility: Longitudinal Analysis of Built Environment on Transit Ridership. Sustainability 8, 1016. https://doi.org/10.3390/SU8101016

Kuo, Y.-H., Leung, J.M.Y., Yan, Y., 2023. Public transport for smart cities: Recent innovations and future challenges. Eur J Oper Res 306, 1001–1026. https://doi.org/10.1016/J.EJOR.2022.06.057

Kwon, D., Lee, C., Kang, H., Kim, I., 2023. Large-Scale Network Imputation and Prediction of Traffic Volume Based on Multi-Source Data Collection System. Transp Res Rec 2677, 30–42. https://doi.org/10.1177/03611981231158324

Lian, Y., Lucas, F., Sörensen, K., 2023. The on-demand bus routing problem with real-time traffic information. Multimodal Transportation 2, 100093. https://doi.org/10.1016/j.multra.2023.100093

Liu, R., Sinha, S., 2007. Modelling urban bus service and passenger reliability. International Symposium on Transportation Network Reliability.

Liu, Y., Liu, Z., Jia, R., 2019. DeepPF: A deep learning based architecture for metro passenger flow prediction. Transp Res Part C Emerg Technol 101. https://doi.org/10.1016/j.trc.2019.01.027

Lyu, N., Wang, Y., Wu, C., Peng, L., Thomas, A.F., 2022. Using naturalistic driving data to identify driving style based on longitudinal driving operation conditions. Journal of Intelligent and Connected Vehicles 5, 17–35. https://doi.org/10.1108/JICV-07-2021-0008

Ma, Z., Zhang, P., 2022. Individual mobility prediction review: Data, problem, method and application. Multimodal Transportation 1, 100002. https://doi.org/10.1016/j.multra.2022.100002

Malayath, M., Verma, A., 2013. Activity based travel demand models as a tool for evaluating sustainable transportation policies. Research in Transportation Economics 38, 45–66. https://doi.org/10.1016/J.RETREC.2012.05.010

Mao, J., Huang, H., Lu, W., Chen, Y., Liu, L., 2022. Multi-precision traffic speed predictions via modified sequence to sequence model and spatial dependency evaluation method. Appl Soft Comput 130, 109700. https://doi.org/10.1016/j.asoc.2022.109700

McGrath, T., Blades, L., Early, J., Harris, A., 2022. UK battery electric bus operation: Examining battery degradation, carbon emissions and cost. Transp Res D Transp Environ 109, 103373. https://doi.org/10.1016/J.TRD.2022.103373

McNally, M.G., 2007. The Four-Step Model, in: Hensher, D.A., Button, K.J. (Eds.), Handbook of Transport Modelling. Emerald Group Publishing Limited, pp. 35–53. https://doi.org/10.1108/9780857245670-003

Mo, P., D'Ariano, A., Yang, L., Veelenturf, L.P., Gao, Z., 2021. An exact method for the integrated optimization of subway lines operation strategies with asymmetric passenger demand and operating costs. Transportation Research Part B: Methodological 149, 283–321. https://doi.org/10.1016/j.trb.2021.05.009

Peled, I., Lee, K., Jiang, Y., Dauwels, J., Pereira, F.C., 2021. On the quality requirements of demand prediction for dynamic public transport. Communications in Transportation Research 1, 100008. https://doi.org/10.1016/j.commtr.2021.100008

Pinjari, A.R., Bhat, C.R., 2021. Activity-based Travel Demand Analysis, in: de Palma, A., Lindsey, R., Quinet, E., Vickerman, R. (Eds.), A Handbook of Transport Economics. Edward Elgar Publishing, pp. 213–248. https://doi.org/10.4337/9780857930873.00017

Qin, X., Ke, J., Wang, X., Tang, Y., Yang, H., 2022. Demand management for smart transportation: A review. Multimodal Transportation 1, 100038. https://doi.org/10.1016/j.multra.2022.100038

Saberi, M., Hamedmoghadam, H., Ashfaq, M., Hosseini, S.A., Gu, Z., Shafiei, S., Nair, D.J., Dixit, V., Gardner, L., Waller, S.T., González, M.C., 2020. A simple contagion process describes spreading of traffic jams in urban networks. Nat Commun 11, 1616. https://doi.org/10.1038/s41467-020-15353-2

Shenzhen Statistics Bureau, 2015. Shenzhen Statistics and Information Yearbook. China Statistics Press, Shenzhen.

Shiftan, Y., Barlach, Y., Shefer, D., 2015. Measuring Passenger Loyalty to Public Transport Modes. J Public Trans 18, 1–16. https://doi.org/10.5038/2375-0901.18.1.7

Sivakumar Nair, G., Mirzaei, A., Ruiz-Juri, N., 2023. Investigating the Use of Machine Learning Methods in Direct Ridership Models for Bus Transit. Transp Res Rec 2677, 768–781. https://doi.org/10.1177/03611981221117540

Tang, T., Fonzone, A., Liu, R., Choudhury, C., 2021. Multi-stage deep learning approaches to predict boarding behaviour of bus passengers. Sustain Cities Soc 73, 103111. https://doi.org/10.1016/j.scs.2021.103111

Tang, T., Liu, R., Choudhury, C., 2020. Incorporating weather conditions and travel history in estimating the alighting bus stops from smart card data. Sustain Cities Soc 53, 101927. https://doi.org/10.1016/j.scs.2019.101927

Tang, T., Liu, R., Choudhury, C., Fonzone, A., Wang, Y., 2023. Predicting hourly boarding demand of bus passengers using imbalanced records from smart-cards: A deep learning approach. IEEE Transactions on Intelligent Transportation Systems 24, 5105–5119. https://doi.org/10.1109/TITS.2023.3237134

Tao, S., Corcoran, J., Mateo-Babiano, I., 2017. Modelling loyalty and behavioural change intentions of busway passengers: A case study of Brisbane, Australia. IATSS Research 41, 113–122. https://doi.org/10.1016/j.iatssr.2016.10.001

Taylor, B.D., Fink, C.N.Y., 2013. Explaining transit ridership: What has the evidence shown? Transportation Letters 5, 15–26. https://doi.org/10.1179/1942786712Z.0000000003

Toqúe, F., Côme, E., Mahrsi, M.K. El, Oukhellou, L., 2016. Forecasting dynamic public transport origin-destination matrices with long-short term memory recurrent neural networks. IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC 1071–1076. https://doi.org/10.1109/ITSC.2016.7795689

Ullah, I., Liu, K., Yamamoto, T., Al Mamlook, R.E., Jamal, A., 2022. A comparative performance of machine learning algorithm to predict electric vehicles energy consumption: A path towards sustainability. Energy & Environment 33, 1583–1612. https://doi.org/10.1177/0958305X211044998

Vergel-Tovar, C.E., Rodriguez, D.A., 2018. The ridership performance of the built environment for BRT systems: Evidence from Latin America. J Transp Geogr 73, 172–184. https://doi.org/10.1016/J.JTRANGEO.2018.06.018

Wang, H., 2022. Transportation-enabled urban services: A brief discussion. Multimodal Transportation 1, 100007. https://doi.org/10.1016/j.multra.2022.100007

Wang, Y., Tang, T., 2023. A simulation-based model for evacuation demand estimation under metro unconventional emergencies. J Transp Eng A Syst 149, 1–14. https://doi.org/10.1061/JTEPBS/TEENG-7682

Wang, Y., Zhang, W., Tang, T., Wang, D., Liu, Z., 2022. Bus OD matrix reconstruction based on clustering Wi-Fi probe data. Transportmetrica B: Transport Dynamics 10, 864–879. https://doi.org/10.1080/21680566.2021.1956388

Wei, M., Liu, Y., Sigler, T., Liu, X., Corcoran, J., 2019. The influence of weather conditions on adult transit ridership in the sub-tropics. Transp Res Part A Policy Pract 125, 106–118. https://doi.org/10.1016/j.tra.2019.05.003

Welch, T.F., Widita, A., 2019. Big data in public transportation: a review of sources and methods. Transp Rev 39, 795–818. https://doi.org/10.1080/01441647.2019.1616849

Wu, W., Liu, R., Jin, W., 2017. Modelling bus bunching and holding control with vehicle overtaking and distributed passenger boarding behaviour. Transportation Research Part B: Methodological 104, 175–197. https://doi.org/10.1016/j.trb.2017.06.019

Wu, W., Xia, Y., Jin, W., 2021. Predicting Bus Passenger Flow and Prioritizing Influential Factors Using Multi-Source Data: Scaled Stacking Gradient Boosting Decision Trees. IEEE Transactions on Intelligent Transportation Systems 22, 2510–2523. https://doi.org/10.1109/TITS.2020.3035647

Xue, M., 2021. Annual report on Shanghai general transportation performance in 2021. Shanghai.

Xue, R., Sun, D.J., Chen, S., 2015. Short-term bus passenger demand prediction based on time series model and interactive multiple model approach. Discrete Dyn Nat Soc 2015. https://doi.org/10.1155/2015/682390

Xylia, M., Leduc, S., Laurent, A.B., Patrizio, P., van der Meer, Y., Kraxner, F., Silveira, S., 2019. Impact of bus electrification on carbon emissions: The case of Stockholm. J Clean Prod 209, 74–87. https://doi.org/10.1016/J.JCLEPRO.2018.10.085

Yang, J., Cao, J., Zhou, Y., 2021. Elaborating non-linear associations and synergies of subway access and land uses with urban vitality in Shenzhen. Transp Res Part A Policy Pract 144, 74–88. https://doi.org/10.1016/J.TRA.2020.11.014

Yousefzadeh Barri, E., Farber, S., Jahanshahi, H., Beyazit, E., 2022. Understanding transit ridership in an equity context through a comparison of statistical and machine learning algorithms. J Transp Geogr 105, 103482. https://doi.org/10.1016/J.JTRANGEO.2022.103482

Zhang, J., Wu, W., Cheng, Q., Tong, W., Khadka, A., Fu, X., Gu, Z., 2022. Extracting the Complete Travel Trajectory of Subway Passengers Based on Mobile Phone Data. J Adv Transp 2022, 1–10. https://doi.org/10.1155/2022/8151520

Zhang, M., Liu, D., Ji, Y., Liu, Y., Wang, W., Chen, Y., He, Z., Jiang, X., 2022. Understanding metro-to-bus transfers in Nanjing, China using smart card data, in: The 11th International Conference on Green Intelligent Transportation Systems and Safety. Springer Singapore, Singapore, pp. 51–68.

Zhang, Y., Cheng, T., 2020. A deep learning approach to infer employment status of passengers by using smart card data. IEEE Transactions on Intelligent Transportation Systems 21, 617–629. https://doi.org/10.1109/TITS.2019.2896460

Zhong, S., Sun, D. (Jian), 2022. Taxi Hailing Choice Behavior and Economic Benefit Analysis of Emission Reduction Based on Multi-mode Travel Big Data, in: Logic-Driven Traffic Big Data Analytics. Springer, Singapore, pp. 227–254. https://doi.org/10.1007/978-981-16-8016-8_11

Zhou, C., Dai, P., Li, R., 2013. The passenger demand prediction model on bus networks, in: Proceedings - IEEE 13th International Conference on Data Mining Workshops, ICDMW 2013. https://doi.org/10.1109/ICDMW.2013.20

Zhou, M., Wang, D., Li, Q., Yue, Y., Tu, W., Cao, R., 2017. Impacts of weather on public transport ridership: Results from mining data from different sources. Transp Res Part C Emerg Technol 75, 17–29. https://doi.org/10.1016/J.TRC.2016.12.001