



OPEN

## Machine learning-aided hybrid technique for dynamics of rail transit stations classification: a case study

Ahad Amini Pishro<sup>1,3</sup>, Shiquan Zhang<sup>2</sup>, Alain L'Hostis<sup>3</sup>, Yuetong Liu<sup>2✉</sup>, Qixiao Hu<sup>2</sup>, Farzad Hejazi<sup>4</sup>, Maryam Shahpasand<sup>5</sup>, Ali Rahman<sup>6</sup>, Abdelbacet Oueslati<sup>7</sup> & Zhengrui Zhang<sup>1</sup>

Accurate classification of rail transit stations is crucial for successful Transit-Oriented Development (TOD) and sustainable urban growth. This paper introduces a novel classification model integrating traditional methodologies with advanced machine learning algorithms. By employing mathematical models, clustering methods, and neural network techniques, the model enhances the precision of station classification, allowing for a refined evaluation of station attributes. A comprehensive case study on the Chengdu rail transit network validates the model's efficacy, highlighting its value in optimizing TOD strategies and guiding decision-making processes for urban planners and policymakers. The study employs several regression models trained on existing data to generate accurate ridership forecasts, and data clustering using mathematical algorithms reveals distinct categories of stations. Evaluation metrics confirm the rationality and accuracy of the results. Additionally, a neural network achieving high accuracy on labeled data enhances the model's predictive capabilities for unlabeled instances. The research demonstrates high accuracy, with the Mean Squared Error (MSE) for regression models (Multiple Linear Regression (MLR), Deep-Learning Neural Network (DNN), and K-Nearest Neighbor (KNN)) remaining below 0.012, while the neural networks used for station classification achieve 100% accuracy across seven time intervals and 98.15% accuracy for the eighth, ensuring reliable ridership forecasts and classification outcomes. Accuracy in rail transit station classification is critical, as it not only strengthens the model's predictive capabilities but also ensures more reliable data-driven decisions for transit planning and development, allowing for more precise ridership forecasts and evidence-based strategies for optimizing TOD. This classification model provides stakeholders with valuable insights into the dynamics and features of rail transit stations, supporting sustainable urban development planning.

**Keywords** Rail Transit Station Classification, Transit Oriented Development, Machine Learning algorithms, Clustering methods, Regression models

### Abbreviations

<i>AGNES</i>	Agglomerative Nesting
<i>ANN</i>	Artificial Neural Network
<i>ARI</i>	Adjusted Rand Index
<i>CH</i>	Calinski-Harabasz Index
<i>DBSCAN</i>	Density-Based Spatial Clustering of Applications with Noise
<i>DDNN</i>	DNN after deleting outliers
<i>DKNN</i>	KNN after deleting outliers

<sup>1</sup>School of Civil Engineering, Sichuan University of Science and Engineering, Zigong 643000, China. <sup>2</sup>School of Mathematics, Sichuan University, Chengdu 610065, China. <sup>3</sup>Univ. Gustave Eiffel, Ecole des Ponts, LVMT, Marne-la-Vallée 77454, France. <sup>4</sup>School of Environment and Technology, University of the West of England, Bristol BS16 1QY, UK. <sup>5</sup>Staffordshire University London, Queen Elizabeth Olympic Park, Here East, London E20 3BS, UK. <sup>6</sup>School of Civil Engineering, Faculty of Engineering and Physical Sciences, University of Leeds, Leeds, UK. <sup>7</sup>Univ. Lille, CNRS, Centrale Lille, UMR 9013 - LaMcube - Laboratoire de mécanique, multiphysique, multiéchelle, Lille 59000, France. ✉email: liuyuetong@stu.scu.edu.cn

<i>DMLR</i>	MLR after deleting outliers
<i>DNN</i>	Deep-Learning Neural Network
<i>GMM</i>	Gaussian Mixture Model
<i>IEW</i>	Information entropy weighting
<i>JC</i>	Jaccard Coefficient
<i>K – Means</i>	K-means clustering
<i>KNN</i>	K-nearest Neighbor
<i>ML</i>	Machine Learning
<i>MLR</i>	Multiple Linear Regression
<i>MSE</i>	Mean of Squared Errors
<i>N</i>	Node value
<i>P</i>	Place value
<i>R</i>	Ridership value
<i>RI</i>	Rand Index
<i>SC</i>	Silhouette Coefficient
<i>SSE</i>	Sum of Squared Errors
<i>T</i>	Time span
<i>TOD</i>	Transit Oriented Development

Transit-Oriented Development (TOD) is a strategic urban planning and design framework that establishes mixed-use, high-density developments close to public transportation infrastructure. The core tenets of TOD entail optimizing land utilization, fostering pedestrian-friendly environments, and facilitating seamless access to efficient public transit systems<sup>1,2</sup>. By concentrating development activities around transit nodes, TOD aims to minimize reliance on private automobiles, curb traffic congestion, and mitigate environmental impacts<sup>3</sup>. In addition to these primary objectives, TOD enhances urban livability by creating vibrant, walkable communities that foster social interaction and a strong sense of place. By reducing the need for long commutes, TOD contributes to better work-life balance and improved public health. Furthermore, the emphasis on mixed-use development supports local economies by attracting diverse businesses and services, providing job opportunities, and stimulating local investment.

TOD's multifaceted significance lies in its capacity to enhance transportation efficiency, bolster sustainability, stimulate economic growth, elevate the quality of life, and promote social equity. By strategically integrating residential, commercial, and recreational spaces within walking distance of transit hubs, TOD optimizes travel distances, cultivates vibrant neighborhoods, reduces energy consumption, mitigates greenhouse gas emissions, attracts business investment, generates fiscal revenues, and fosters inclusive communities that cater to diverse income groups. As a genuine urbanistic concept and hence multiscale by design, TOD conveys a transit line and network dimension that connects to urban and urban region planning perspectives.

Classifying rail transit stations using mathematical methods involves algorithms and models, including machine learning and clustering techniques, to examine and categorize stations based on passenger flow, connectivity, and infrastructure features. K-Means is a renowned method for grouping data into K-distinct sections by reducing the total variance within the clusters<sup>4</sup>. K-Means is suitable for the first attempt at analyzing rail transit station data because its iterative and straightforward nature makes it an easy option. For instance, in a TOD scenario, K-Means can quickly and effectively identify similar stations regarding ridership volume, connectivity, and accessibility. This kind of clustering is of great help for urban planners in listing and counting transport hubs and identifying unfulfilled stations and urban zones that might need some infrastructural improvements<sup>5-7</sup>. It is one of the most effective ways to concentrate resources and pave the path for urban development through smart infrastructural design.

AGNES clustering, a type of hierarchical clustering, creates interior clusters by combining or dividing them based on particular conditions. Unlike K-Means, K-Means does not require knowing the number of clusters earlier; instead, it makes the procedure more flexible and adjustable<sup>8</sup>. In terms of rail transit stations, AGNES is a significant way to identify the network hierarchical structure, finding both macro and micro-clusters. This tool is very efficient in recognizing the direct and indirect connections between stations and in the station's hierarchy, whether it is a main, a branch, or a leaf, and it also gives an overall impact of the network topology on strategic planning for TOD, among other things<sup>9</sup>.

DBSCAN, a density-based clustering algorithm, determines the clusters of various shapes and sizes other than the density of data points. The method of dealing with noise and outliers is one of its most prominent advantages<sup>10</sup>. DBSCAN can reveal sets of spherical stations in transportation stations, which is not the case for natural K-Means. This is the primary TOD application because it can find city blocks that are lastly joined or other problematic places that may result from stations being isolated or performing at lower levels. DBSCAN is an all-around tool that improves transit networking for better passenger service through decisions on the density of station usage and connectivity<sup>11</sup>.

GMM is one of the basic paradigmatic methods of cluster analysis, which is grounded on the a priori assumption of the mixture of several pairs of Gaussian distributions. This action provides an all-around perspective necessary for the group's different shapes, sizes, and orientations, making it highly versatile and robust<sup>12</sup>. By applying GMM technology to rail transit stations, we can also analyze the complex station characteristics and usage patterns and thus gain a deep insight into the transit network. The stochastic feature of the process is also an economic assignment that allows the stations to be put into clusters, with weights exerted by the GMM but part of the weight belonging to the so-called "soft" group that sometimes assigns the stations if they are used for public services or are dwelling in or nearby communities. This ability is vitally important in

the field of TOD, as it is critical for developing multifunctional strategies that consider each platform's different roles and necessities.

Applying innovative clustering methods to rail transit station classification is a new approach in TOD studies. Each contributes differently to the analysis, providing fresh insights and suggestions for city planners and politicians<sup>4-15</sup>. By utilizing the four approaches, namely K-Means, AGNES, DBSCAN, and GMM, the TOD projects will get a much clearer picture of transit networks, better resource allocation, eco-friendly urban development, and data utilization in decision-making. These modes of operations can fully delineate the transit system, whereby the mind can be steered to the primary nodes, connections, and areas that need to be developed. Technical approaches such as modeling the travel time of linear vibrating PETs (public electric transportation) and the angle of environmental noise sources at the same place coordinates can be utilized for safety and shared streets. The concept of “equal access” can be best understood by inspectors of social safety in real-time in mixed and shared traffic and pedestrians through digital twins of ADA/VDAs. Nevertheless, the ultimate goal is to ensure that users and dependents experience transportation benefits without safety incidents. In particular, these ways of examination not only enrich the information necessary for the formation of the most optimum distribution of available transit system facilities but also provide very relevant and effective instruments for urban development that are both sustainable and efficient.

In this context, classifying railway stations is critical to the design of urban and regional planning strategies. Accurate classification enables planners to tailor development projects to each station's specific characteristics and needs, ensuring efficient resource allocation and optimal land use. This precision in planning facilitates the integration of transportation networks, promotes seamless connectivity, and supports the creation of cohesive and sustainable urban environments. Furthermore, such classifications provide valuable insights for policymakers, helping them implement policies that align with the goals of TOD and address the unique challenges different regions face.

## Literature review

The classification of rail transit systems plays a crucial role in Transit-Oriented Development (TOD) as it profoundly impacts the scale, design, and planning considerations of development projects centered around transit stations. This classification informs decisions on development intensity, station design, infrastructure requirements, and transportation network integration<sup>16-18</sup>. Urban planners and developers leverage this classification to optimize land use, design efficient stations, plan for TOD projects, and create comprehensive transportation networks that seamlessly connect different modes of travel<sup>19,20</sup>.

Using clustering methods from mathematics enhances rail transit classification by extracting meaningful patterns and objectively categorizing transit systems. Clustering algorithms enable the recognition of similarities and shared characteristics among systems, employing unsupervised learning to identify inherent structures and relationships within large datasets<sup>21</sup>. By leveraging clustering algorithms, transportation professionals can make data-driven decisions, improving the reliability and scalability of the classification process. Clustering results provide valuable insights for future planning, benchmarking, and resource allocation, facilitating the development of efficient and optimized rail transit systems<sup>22,23</sup>.

The application of Artificial Neural Network (ANN), including some aspects of mathematics such as Multiple Linear Regression (MLR), is a hybrid tool to study the relationship between different parameters in a model<sup>15,24,25</sup>. The utilization of Machine Learning (ML) in rail transit classification is paramount due to its ability to enable data-driven decision-making, handle complex and multidimensional data, uncover hidden patterns, adapt to changing conditions, and support continuous improvement. Machine learning algorithms excel at processing vast volumes of data and extracting meaningful insights that may not be apparent through traditional analysis methods<sup>26-29</sup>. By leveraging machine learning techniques, transportation professionals can make informed decisions based on the intricate relationships and dependencies identified within rail transit data. These algorithms facilitate the identification of clusters or groups of similar transit systems, leading to more accurate and granular classification outcomes.

Jingru Huang et al.<sup>30</sup> investigated the relationship between the physical environment and subway ridership in Beijing. The results indicated that greater employment density and enhanced accessibility to public transportation contribute to higher ridership during morning peak hours. The research also highlighted variations in the built environment's impact on ridership across different areas and confirmed the reliability of the model used in the analysis. Liu Yang et al.<sup>31</sup> investigated different methods for categorizing TOD in cities, with Ningbo as the focus. Their research introduced a novel approach to improve TOD based on metro stations, expanding the traditional node-place model with additional attributes and utilizing machine learning for station selection. The study's findings provided insights for other cities with urban rail transit systems and suggested avenues for future research on analyzing traffic behavior and adjusting land-use strategies in TOD. Enrica Papa et al.<sup>32</sup> classified station areas and promoted public transport use as a solution to mitigate the negative impacts of private car usage in cities. Through cluster analysis, factors influencing these areas were examined, and a method for identifying different station area typologies was proposed. The approach was demonstrated in Naples, showcasing the integration of land-use and transport planning strategies to improve rail stations. Shiliang Su et al.<sup>33</sup> analyzed the specific impact of TOD on metro ridership in Shanghai. The study identified spatial and temporal variations in the influence of TOD factors and neighborhood demographics by combining the node-functionality-place model with interpretable machine learning. The findings underscored the importance of functionality, revealed key interactions, and offered insights for urban planning in high-density cities. Dan Qiang et al.<sup>9</sup> analyzed the metro stations in Shanghai using new urban datasets and identified five clusters based on 15 indicators for TOD. The study found strong correlations between transportation, pedestrian-oriented accessibility, and urban development indicators with ridership, emphasizing the importance of population density in metro passenger traffic.

Xin Yang et al.<sup>34</sup> focused on applying neural networks in TOD and specifically addressed the complex task of short-term prediction of passenger volume in urban rail systems. The study proposed an improved spatiotemporal long short-term memory model (Sp-LSTM) that utilized deep learning techniques and big data, outperforming other prediction methods such as LSTM, ARIMA, and NAR, as demonstrated through a case study on the Beijing Metro Airport Line. In another research<sup>35</sup>, they introduced the Wave-LSTM model, a combination of LSTM and wavelet techniques, which demonstrated superior prediction accuracy compared to other algorithms in an empirical study using practical data from Dongzhimen Station in the Beijing Subway system. The study concluded that the newly proposed model holds great potential as a reliable approach for predicting precise short-term inbound passenger flow in urban rail systems. Jinlei Zhang et al.<sup>36</sup> highlighted the importance of short-term origin-destination (OD) flow prediction in urban rail transit (URT) for real-time operation and management. They introduced a channel-wise attentive split-convolutional neural network (CAS-CNN) that addressed the challenges of data availability, dimensionality, and sparsity. Through testing on real-world datasets from the Beijing Subway, the CAS-CNN model, incorporating innovative components such as channel-wise attention and inflow/outflow-gated mechanisms, outperformed existing benchmarking methods. Chunyan Shuai et al.<sup>37</sup> proposed a pattern match algorithm, TSNE-KNN, for accurate short-term origin-destination (OD) demand prediction in urban rail transit. The TSNE-KNN model outperformed other approaches, including deep neural network models, and identified similarity indicators as universal indicators reflecting the time-space properties of OD flow and the shifting patterns of rail transit stations.

Regarding the limitations in model accuracy, predictive capability, and scalability in the literature, we first acknowledge that previous studies have made significant contributions by proposing methods to enhance accuracy, provide objective evaluations, enable predictive capabilities, and facilitate scalability. However, these limitations and challenges have often been inadequately addressed or only superficially mentioned. Therefore, this study seeks to clarify these aspects and provide concrete details regarding the limitations faced in existing models and methodologies.

Although several studies<sup>9,30,34,36</sup> have demonstrated that machine learning models such as ANN, LSTM, and Sp-LSTM improve predictive accuracy for rail transit systems, the accuracy of these models still faces challenges due to the inherent complexity of urban transit systems. Existing methods, particularly those relying on traditional clustering algorithms (K-Means, DBSCAN), are limited in accurately classifying stations with mixed or transitional characteristics where different clusters overlap. These models may exhibit high accuracy under specific conditions but struggle with highly dynamic and non-linear ridership patterns, especially when faced with new, unseen data. Our approach addresses this limitation by integrating more advanced techniques like deep neural networks (DNNs), which handle complex patterns better, but we also note that model accuracy may still be affected by factors such as data quality and the availability of real-time datasets.

While predictive capabilities in previous research have been advanced by introducing models like Wave-LSTM<sup>34</sup> and CAS-CNN<sup>36</sup>, limitations remain in their ability to predict ridership fluctuations under varying temporal conditions and in stations with irregular passenger volumes. The short-term prediction models, although effective for certain types of data, often fail to generalize across different station types (e.g., those with fluctuating passenger volumes during peak and off-peak hours). Moreover, the scalability of predictive models for short-term origin-destination flow prediction is constrained when moving from small datasets to larger, city-wide applications. Our work seeks to improve this by incorporating predictive models considering a more nuanced range of variables, including socio-economic factors and real-time ridership data, which enhance predictive capabilities over broader datasets. However, predictive accuracy is still limited by the availability of comprehensive data across all stations.

The scalability of rail transit classification methods has been significantly challenged by the complexity of urban transit systems, particularly when expanding from city-level studies to regional or national networks. Studies such as Su et al.<sup>33</sup> and Shuai et al.<sup>37</sup> have focused on specific cities (e.g., Shanghai and Beijing), and the proposed models excel at handling localized datasets but struggle to scale effectively across diverse transit networks with different ridership behaviors and infrastructure designs. Furthermore, the computational demands of machine learning algorithms, such as deep learning models, can become a bottleneck when applied to larger datasets. While our research incorporates scalable algorithms such as GMM and hierarchical clustering methods, we acknowledge that further optimization is necessary to ensure computational efficiency and practicality when applying these models to large-scale TOD planning.

This study builds upon these previous findings by introducing a more systematic and data-driven method to evaluate and classify rail transit stations into clusters based on their characteristics. The research applies clustering algorithms such as K-Means, AGNES, DBSCAN, GMM, and other applicable mathematics and machine learning tools<sup>26-29</sup> to provide a precise classification method for rail transit stations. By enhancing accuracy, providing objective evaluations, enabling predictive capabilities, and facilitating scalability for large-scale analyses, this innovation empowers urban planners and policymakers with valuable insights for sustainable and efficient urban development, optimizing TOD outcomes. This approach leverages existing methodologies and addresses their limitations by incorporating advanced machine-learning techniques to achieve more accurate and practical results for TOD planning.

## Case study and methodology

### Case study

Chengdu, the capital city of Sichuan province in China, has a well-developed rail transit network that provides efficient and convenient transportation for its residents and visitors. The Chengdu rail transit system consists of metro lines and high-speed rail connections. The Chengdu Metro is the backbone of the city's urban transportation system. The city center of Chengdu is located around Tianfu Square, which is considered the heart of the urban area. Key commercial districts, cultural landmarks, and government offices are concentrated

in this central area. The urban region of Chengdu is densely developed with a mix of high-rise buildings, shopping centers, and residential complexes. Beyond the Third Ring Road, the city transitions into suburban areas with lower-density development and a mix of residential and industrial zones. The Fourth Ring Road generally marks the boundary between the suburban and rural areas, where farmland and small villages become more common. The rail transit system in Chengdu has dramatically improved the city's transportation efficiency, reducing congestion and providing a convenient way for people to travel within and outside the city. It has played a significant role in enhancing Chengdu's urban development and supporting its economic growth.

For several compelling reasons, Chengdu's metro network serves as an excellent case study for Transit-Oriented Development (TOD) and the classification of rail transit stations. Firstly, the metro network in Chengdu boasts extensive coverage, encompassing a vast area that includes urban and suburban regions. The network connects vital city areas, including the city center, central commercial districts, educational institutions, and suburban communities. This extensive reach ensures that many people can access efficient public transportation. This expansive reach provides diverse station locations with unique characteristics and development potential. Secondly, many of Chengdu's metro stations are strategically located in areas characterized by mixed land use. These areas combine residential, commercial, and recreational facilities, setting the stage for integrated development around the stations. A diverse range of land uses facilitates the creation of walkable neighborhoods and vibrant urban environments. Moreover, Chengdu's rapid urban growth has presented numerous development opportunities around metro stations. Underutilized or vacant land near stations can be transformed through redevelopment and revitalization initiatives, fostering economic growth and community enhancement.

Chengdu's commitment to TOD is evident in its urban planning and policy framework. The local government has implemented various measures to encourage the integration of transportation, land use, and urban design. Policies promoting high-density development around metro stations, mixed-use zoning, and pedestrian-friendly infrastructure have been critical to the city's TOD success. The government has also invested in public amenities, green spaces, and cultural facilities around transit hubs to enhance the quality of life for residents.

The significant passenger volume of Chengdu's metro system is another compelling aspect. With many daily commuters, the stations attract high levels of foot traffic. This demand creates a favorable market for various businesses and services to thrive around the stations, further stimulating TOD. Additionally, the local government in Chengdu has proactively promoted TOD and urban development around metro stations. Through the implementation of policies and strategies, they encourage the integration of transportation, land use, and urban design. These initiatives aim to create sustainable, livable communities that benefit from well-planned transit-oriented environments.

The success of Chengdu's metro network and TOD initiatives can be attributed to several factors, including strategic planning, strong government support, and community engagement. The city's approach to TOD addresses transportation and land use and focuses on creating vibrant, inclusive, and sustainable urban environments. Lessons from Chengdu's experience can inform TOD projects in other cities, highlighting the importance of comprehensive planning, stakeholder collaboration, and adaptive strategies to local contexts.

Researchers and planners can gain valuable insights into effective station classification and development practices by examining the success of Chengdu's metro network and its associated TOD efforts. The lessons learned from Chengdu can serve as a valuable guide for future TOD projects in other cities, aiding in creating sustainable and vibrant urban areas. Figure 1 presents the Chengdu rail transit network, the CBDs of Chunxi Road, the Third Tianfu Street, and the boundaries of the rural and urban areas.

### Problem statement

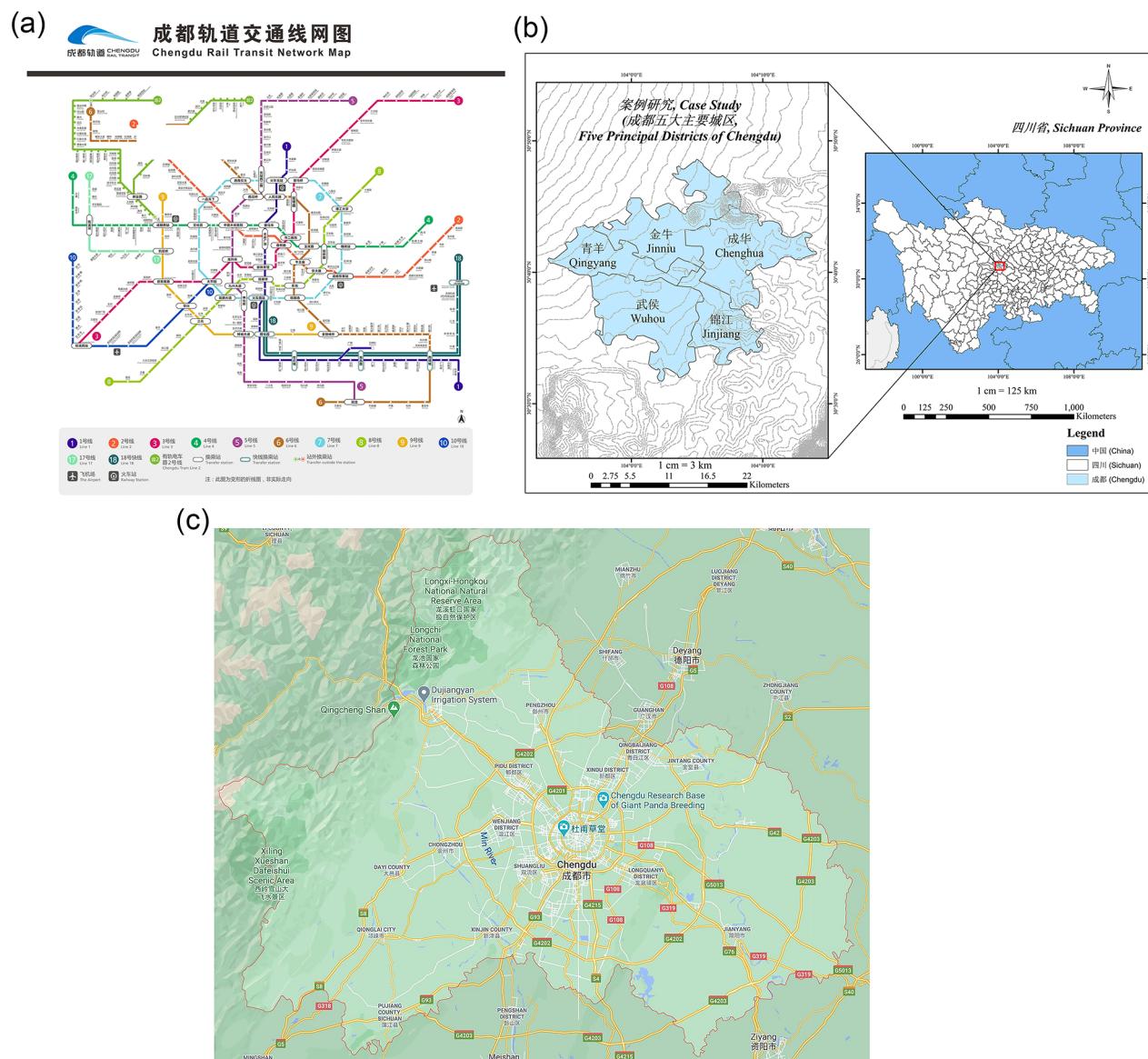
By examining the Chengdu Metro system, the study provides a comprehensive understanding of how various factors influence ridership patterns. This paper's primary focus is divided into ridership fitting and station classification. The first part of the analysis involves fitting ridership data to assess the relationships between ridership levels and variables such as node value, place value, and time factors. This approach helps to elucidate how these variables interact and contribute to the overall ridership figures. In the second part of the study, stations are classified based on a combination of node value, place value, ridership, and time. This classification process reveals patterns and distinctions among stations, offering insights into their operational and functional characteristics. Identifying these commonalities and differences is crucial for tailoring specific strategies to optimize station performance and enhance the overall efficiency of the transit network.

The challenges in solving the problem of predicting subway ridership and categorizing stations stem from the complex, dynamic nature of urban transit systems. Accurately predicting passenger flow at subway stations requires a comprehensive understanding of how factors such as station characteristics, location, and time affect ridership. This study employs different fitting methods to model ridership for Chengdu subway stations, aiming to select the most suitable prediction model. Additionally, subway stations, particularly those intersected by multiple lines, display unique characteristics while sharing common traits. Categorizing these stations through clustering techniques in machine learning helps identify patterns that can inform the design and planning of new stations. However, the unlabeled nature of stations and the varying outcomes from different clustering methods pose challenges, which are addressed by integrating results and using a neural network to classify stations and predict the characteristics of new ones. Figure 2 illustrates the research framework of this study.

### Node, place, ridership, and time indicators

#### *Node indicators*

We evaluate the value of a station's node based on four aspects: the station's facilities, the availability of nearby transportation options, the accessibility to various destinations, and its importance within the network. Table 1 presents eight indicators that fall under these four aspects.



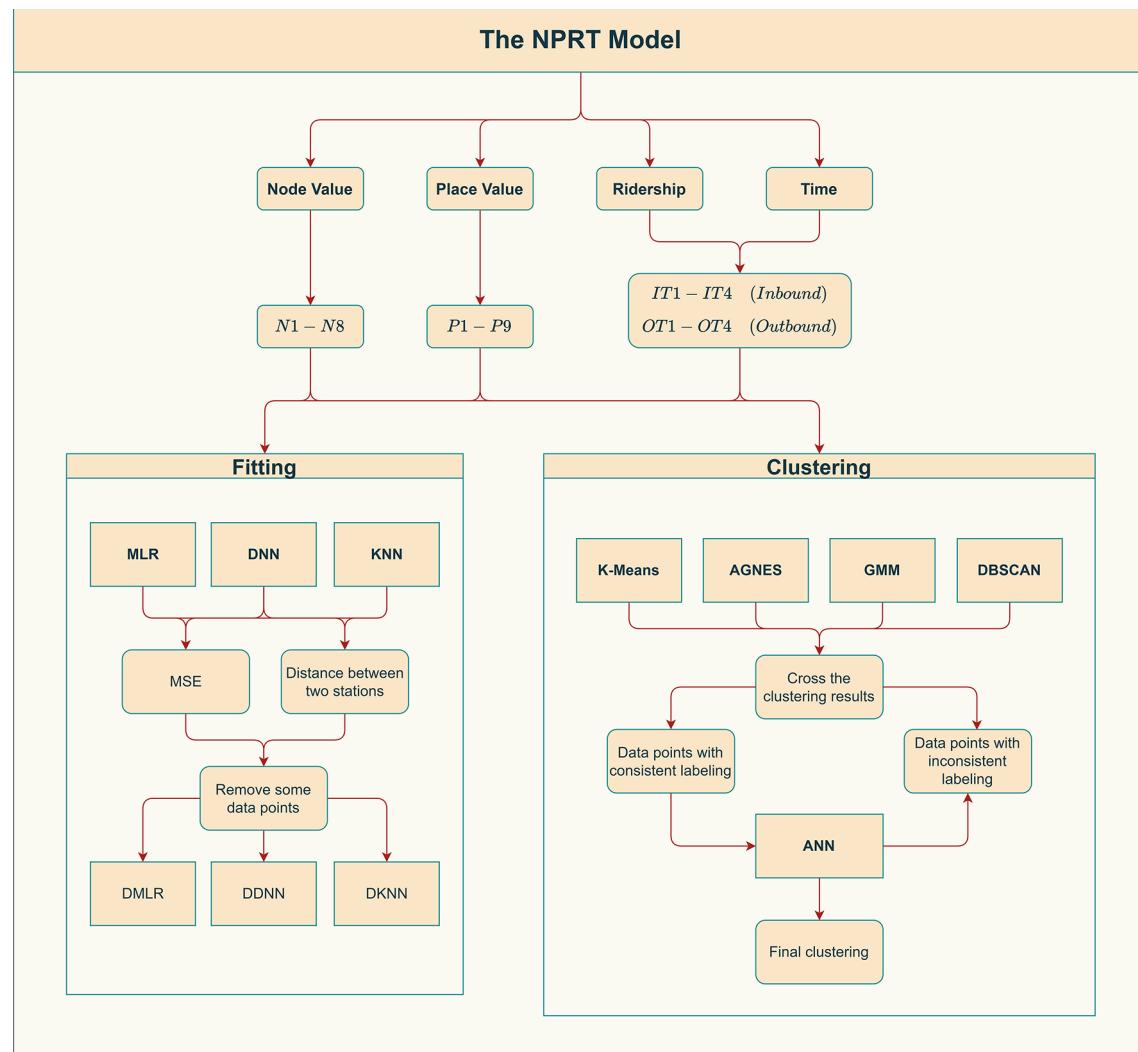
**Fig. 1.** Chengdu: key insights and structure.

The station's facilities are assessed by considering the number of entrances and exits (N1) in each metro station. To determine the accessibility of transits, we consider the number of metro stations (N2) that can be reached within a 20-minute travel time, the number of stations connecting to the central business district (CBD) at Chunxi Road (N3), and the number of stations connecting to the CBD at 3rd Tianfu Street (N4). Since Chengdu has two CBDs, Chunxi Road and 3rd Tianfu Street, we calculated the number of stations and the distance to each CBD. The distance indicates the accessibility of destinations to the CBDs, with Chunxi Road represented by (N5) and 3rd Tianfu Street by (N6). Lastly, the network centrality comprises degree centrality (N7) and closeness centrality (N8).

Using graph modeling, we have employed network centrality to capture the significance of a station within the transit network. To represent the Chengdu rail transit network as a graph  $G = (V, E)$ , we assign vertices in  $V$  to represent the stations, while the set  $E$  consists of edges denoting the connections between stations. The weight of each edge in  $E$  is determined by the transit traveling distance.

To measure the degree centrality (N7) of a transit station  $v \in V$  in the Chengdu network, we consider the number of links connected to station  $v$  in Eq. (1). Here,  $L_{vt}$  represents the linkage between station  $v$  and another station  $t \in V$ , while  $K$  represents the total number of stations in set  $V$ :

$$N7(v) = \sum_{t=1}^K L_{vt} (v \neq t), L_{vt} = \begin{cases} 1 & \text{station } v \text{ is linked to station } t \\ 0 & \text{station } v \text{ is not linked to station } t \end{cases} \quad (1)$$



**Fig. 2.** Structured research framework and methodological workflow.

Closeness centrality indicates how close and accessible a node is within the network component. The closeness centrality ( $N8$ ) measurement for station is determined by the inverse sum of the shortest transit distances from station  $v$  to all other stations in set  $V$ , as shown in Eq. (2). Here,  $d_{vt}$  represents the shortest transit distance between station  $v$  and another station  $t \in V$ :

$$N8(v) = \frac{1}{\sum_{t=1}^K d_{vt}} \quad (v \neq t) \quad (2)$$

#### Place indicators

To account for the low-density nature of certain areas in Chengdu, we establish a transit catchment area using a radius of 500 m and 1000 m. The evaluation of a station's place value involves three factors: design, density, and diversity. Table 1 presents nine place indicators categorized under these three factors.

The design aspect is measured using various metrics. These include the average price of office land within the 1000 m-radius catchment area ( $P1$ ), the average price of commercial land within the 1000 m-radius catchment area ( $P3$ ), the average price of residential land within the 1000 m-radius catchment area ( $P5$ ), the number of parking lots within the 500 m-radius catchment area ( $P8$ ), and the number of bus stops within the 500 m-radius catchment area ( $P9$ ). Furthermore, design is also assessed by considering the number of offices within 1000 m ( $P2$ ), the number of shops within 1000 m ( $P4$ ), and the number of residences within 1000 m ( $P6$ ). The diversity factor encompasses public facilities such as parks, cultural facilities, schools, and hospitals within the 1000 m-radius catchment area ( $P7$ ).

Dimension	Branch	Indicator	Max	Mean	Min
<a href="https://www.amap.com">https://www.amap.com</a> , <a href="https://map.qq.com">https://map.qq.com</a> , <a href="https://www.chengdurail.com/index_en.html">https://www.chengdurail.com/index_en.html</a> , <a href="https://chengdu.anjuke.com">https://chengdu.anjuke.com</a> , <a href="https://cd.newhouse.fang.com">https://cd.newhouse.fang.com</a>					
Node Value	Station Facility	N1. Number of entrances and exits in each metro station (unit)	10.0000	4.6535	2.0000
	Accessible Transits	N2. Number of metro stations that one station can reach within 20 min (unit)	88.0000	41.9257	8.0000
		N3. Number of stations to CBD (Chunxi Road) (unit)	23.0000	10.0792	0.0000
		N4. Number of stations to CBD (3rd Tianfu Street) (unit)	33.0000	14.0446	0.0000
	Accessible Destinations	N5. Distance to CBD (Chunxi Road) (km)	45.3230	13.6033	0.0000
		N6. Distance to CBD (3rd Tianfu Street) (km)	43.7770	18.3596	0.0000
	Network Centrality	N7. Degree centrality	6.0000	2.4653	2.0000
		N8. Closeness centrality (1/1000 km)	0.0004	0.0003	0.0001
Place Value	Design	P1. The average price of office land inside the 1000 m-radius catchment area (CNY/m <sup>2</sup> )	74000.0000	11118.2658	5550.0000
	Density	P2. Number of offices within 1000 m (unit)	197.0000	26.7673	0.0000
	Design	P3. The average price of commercial land inside the 1000 m-radius catchment area (CNY/m <sup>2</sup> )	50480.0000	21285.5855	8571.0000
	Density	P4. Number of shops within 1000 m (unit)	397.0000	117.4554	1.0000
	Design	P5. The average price of residential land inside the 1000 m-radius catchment area (CNY/m <sup>2</sup> )	42663.3077	18405.8081	8423.0000
	Density	P6. Number of residences within 1000 m (unit)	552.0000	110.7970	1.0000
	Diversity	P7. Number of public facilities (parks, cultural facilities, schools, hospitals) inside the 1000 m-radius catchment area (unit)	41.0000	10.9208	0.0000
	Design	P8. Number of parking lots inside the 500 m-radius catchment area (unit)	132.0000	21.4851	0.0000
		P9. Number of bus stops inside the 500 m-radius catchment area (unit)	26.0000	7.3515	1.0000

**Table 1.** NP indicators.

Time	Definition	Days	Hours	Max	Mean	Min
IT1	Inbound traffic during working hours	Monday to Friday	6:00–9:00	27654.3478	4451.6051	53.2609
IT2	Inbound traffic during off-hours	Monday to Friday	17:00–20:00	46668.6087	4281.0174	113.6957
IT3	Inbound traffic during the rest of the day	Monday to Friday	9:00–17:00 / 20:00–23:00	54702.0870	5156.2546	140.3043
IT4	Inbound traffic on two days of the weekend	Saturdays & Sunday	6:00–23:00	51955.6250	4607.0829	122.3750
OT1	Passengers leaving the station during working hours	Monday to Friday	6:00–9:00	56982.3478	5456.5258	151.3913
OT2	Passengers leaving the station during off-hours	Monday to Friday	17:00–20:00	26532.4783	4367.8530	69.4348
OT3	Passengers leaving the station during the rest of the day	Monday to Friday	9:00–17:00 / 20:00–23:00	33976.5217	4064.4983	72.0000
OT4	Passengers leaving the station on both days of the weekend	Saturdays & Sunday	6:00–23:00	55496.8750	4607.0829	126.6250

**Table 2.** Time class definition.

#### Ridership and time indicators

To address the limitations of the NPR model, which fails to consider the impact of time and overlooks variations in ridership between departures and arrivals, we have introduced the recording of tapped-in and tapped-out trips to construct an NPRT model that accounts for different conditions.

As previously mentioned, ridership is closely tied to time. Hence, we have classified passenger traffic into inbound traffic (I) and outbound traffic (O). We have further divided the time into peak, off-peak, regular, and weekends, denoted as T1 to T4. This results in eight distinct conditions. For example, IT1 represents inbound traffic during working hours, IT2 represents inbound traffic during off-peak hours, IT3 represents inbound traffic during the remaining hours of the workday, and IT4 represents inbound traffic on two weekend days. Similarly, OT1 denotes the ridership of passengers leaving the station during working hours, OT2 represents ridership during off-peak hours, OT3 represents ridership during the remaining hours of the workday, and OT4 represents ridership on two weekend days.

Table 2 provides a detailed breakdown of each class and the corresponding time intervals from IT1 to OT4.

#### Methodology

##### Information entropy weighting (IEW)

To facilitate the analysis of the data and compose the indicators, Information Entropy Weighting (IEW)<sup>17</sup> was used to integrate  $N_1 - N_8$  into node value ( $N$ ) and  $P_1 - P_9$  into place value ( $P$ ). When the information entropy is lower, the significance of the index in providing information decreases, resulting in a minor role in the comprehensive evaluation, and thus, a lower weight should be assigned to it. Therefore, information entropy can be used to calculate the weight of each index, taking  $N$  as an example. And similarly, we can get  $P$ . If there are  $n$  stations and  $m$  node value indicators, then we get  $X$ , where  $x_{ij}$  represents the value of the indicator  $j$  at station  $i$ .

$$X = \{x_{ij}\}_{n \times m} \quad (3)$$

Station	N1	N2	N3	N4	N5	N6	N7	N8
Weijianian	0.375	0.4375	0.304348	0.545455	0.197427	0.447861	0	0.628906
Shengxian Lake	0.25	0.5625	0.26087	0.515152	0.164023	0.413276	0	0.721591
North Railway Station	0.5	0.8875	0.217391	0.484848	0.128875	0.376887	0.5	0.859327
Renmin Rd.North	0.625	0.825	0.173913	0.454545	0.10284	0.349933	0.5	0.877811
Wenshu Monastery	0.5	0.7875	0.130435	0.424242	0.073031	0.319072	0	0.933192
Luomashi	0.375	1	0.086957	0.393939	0.053527	0.298878	0.5	0.976904
Tianfu Square	1	0.975	0.043478	0.363636	0.031066	0.275624	0.5	1

**Table 3.** Node indicators values normalized by the Min-Max normalization method.

Station	P1	P2	P3	P4	P5	P6	P7	P8	P9
Weijianian	0.056403	0.025381	0.366665	0.131313	0.211279	0.038113	0.146341	0.007576	0.28
Shengxian Lake	0.047561	0.015228	0.307309	0.244949	0.199965	0.116152	0.219512	0.037879	0
North Railway Station	0.073676	0.152284	0.331775	0.578283	0.201768	0.297641	0.219512	0.204545	0.28
Renmin Rd.North	0.074668	0.324873	0.349351	0.616162	0.198684	0.444646	0.585366	0.234848	0.24
Wenshu Monastery	0.079775	0.553299	0.501338	0.558081	0.244897	0.658802	0.365854	1	0.24
Luomashi	0.079775	0.664975	0.333729	0.691919	0.418737	0.885662	0.585366	0.424242	0.24
Tianfu Square	0.103221	0.84264	0.361517	0.69697	0.550529	0.664247	0.463415	0.575758	0.32

**Table 4.** Place indicator values normalized by the Min-Max normalization method.

Algorithm 1 provides the IEW method applied in this study.

#### Algorithm 1 IEW.

**Step 1.** 0-1 normalized the matrix  $X$ .

$$z_{ij} = \frac{x_{ij} - \min_i \{x_{ij}\}}{\max_i \{x_{ij}\} - \min_i \{x_{ij}\}} \quad (4)$$

**Step 2.** Calculate the proportion of each station for indicator  $j$ .

$$p_{ij} = \frac{z_{ij}}{\sum_{i=1}^n z_{ij}} \quad (5)$$

**Step 3.** Calculate the information entropy of each indicator. If  $p_{ij} = 0$ , specify  $\ln(0) = 0$ .

$$e_j = -\frac{1}{\ln n} \sum_{i=1}^n p_{ij} \ln(p_{ij}) \quad (6)$$

**Step 4.** Calculate the imbalance coefficient.

$$d_j = 1 - e_j \quad (7)$$

**Step 5.** Calculate the weight of each indicator.

$$w_j = \frac{d_j}{\sum_{j=1}^m d_j} \quad (8)$$

**Step 6.** The weighted sum of the values for each station is normalized to get the node value.

$$s_i = \sum_{j=1}^m w_j z_{ij} \quad (9)$$

$$N_i = \frac{s_i - \min_i \{s_i\}}{\max_i \{s_i\} - \min_i \{s_i\}} \quad (10)$$

Tables 3 and 4, and 5 present the normalized values of node, place, and integrated ridership-time indicators of some subway stations provided by the Min-Max Normalization method.

Station	IT1	IT2	IT3	IT4	OT1	OT2	OT3	OT4
Weijianian	0.250627	0.023075	0.04088	0.050935	0.049918	0.155809	0.023308	0.040717
Shengxian Lake	0.113181	0.022086	0.037096	0.032339	0.04168	0.081328	0.034306	0.029183
North Railway Station	0.250685	0.129159	0.182756	0.168179	0.204013	0.233525	0.171207	0.167093
Renmin Rd.North	0.189969	0.17323	0.166083	0.148266	0.155033	0.222095	0.224951	0.13873
Wenshu Monastery	0.174754	0.138479	0.150118	0.116576	0.141487	0.179883	0.214244	0.112685
Luomashi	0.158382	0.313183	0.270683	0.153328	0.219143	0.202612	0.565992	0.155882
Tianfu Square	0.100609	0.438796	0.352778	0.26899	0.321523	0.220588	0.622054	0.263616

**Table 5.** Integrated ridership-time values normalized by the Min-Max normalization method.

#### Cluster methods

Unsupervised learning is necessary when dealing with training samples lacking labeling information. The goal is to assign labels to stations based on node value, place value, and ridership at various times, creating a tagged dataset. This labeled data can be further used to train the rail transit stations' classification model with classification techniques to predict new stations.

Clustering methods are employed to partition existing stations and unveil the underlying structure of the data to accomplish this objective. The paper utilizes four clustering methods, which are described below.

The **K-Means** represents a prototyping-based clustering algorithm<sup>4,5</sup>. In this algorithm, we randomly select  $k$  data from the dataset as the initial centers of the clusters and assign each point in the dataset to the cluster to which the point belongs to the nearest center. A new cluster center can be computed from the most recently delineated cluster; thus, the cluster to which the data point belongs can be reassigned. Repetition of such operations until convergence allows the realization of K-Means. Therefore, when training the model for K-Means, we need to consider tuning the hyperparameters for the number of clusters  $k$ .

The **AGNES** is a hierarchical clustering algorithm that operates on a bottom-up aggregation approach<sup>8</sup>. It begins by treating each object as an individual cluster and subsequently merges clusters step by step based on specific criteria. The similarity matrix can be obtained by calculating the distance between any two clusters, and thus, the two closest clusters  $C_i$  and  $C_j$  can be found. We merge them into the same cluster  $C_h$  and also update the distances related to both to get a new similarity matrix. The process is repeated until the algorithm stops when all data points belong to the same cluster. Since the distance metrics available for calculating the distance between clusters are ward connected, wholly connected, average connected, etc., the distance calculation method *linkage* needs to be considered in addition to the number of clusters  $k$ .

The **DBSCAN** stands for Density-Based Spatial Clustering of Applications with Noise and is a density-based clustering algorithm<sup>10,11</sup>. Its fundamental concept revolves around assessing whether data points belong to the same cluster by identifying adjacent points around the data points and recursively exploring the adjacent points around those neighbors.

Two essential parameters are at the core of the DBSCAN algorithm: radius ( $\epsilon$ ) and minimum sample size (MinPts). The parameter  $\epsilon$  defines the distance threshold for a sample's neighborhood, while MinPts specifies the minimum number of samples within that neighborhood (defined by  $\epsilon$  distance) to be considered part of a cluster.

The **GMM** (Gaussian Mixture Model) is a probabilistic model that characterizes the cluster prototype. The cluster partition is determined by the prototype corresponding to the posterior probability<sup>12</sup>. GMM uses

$$p(x|\Theta) = \sum_{j=1}^k \alpha_j P(x|\mu_j, \Sigma_j) \quad (11)$$

where  $P(x|\mu_j, \Sigma_j)$  is a Gaussian distribution with mean  $\mu_j$  variance  $\Sigma_j$  and with a non-negative weight  $\alpha_j$  constituting the overall distribution  $p(x|\Theta)$ , and  $\sum_{j=1}^k \alpha_j = 1$ ,  $\Theta = \{\alpha_j, \mu_j, \Sigma_j | j = 1, 2, \dots, k\}$ .

$k$  represents a parameter to be tuned.

According to the rule, the samples  $x_i$  are assigned to the cluster  $\lambda_i$  corresponding to the Gaussian distribution with the highest probability of belonging to it. Based on the dataset  $D$ , the parameter  $\Theta$  of the GMM can be solved iteratively by the EM algorithm maximizing its log-likelihood function.

#### Performance measurements

Two main types of clustering performance measures exist. The first type involves comparing the clustering results with a reference model, which can be the division results provided by domain experts or inherent data categories. This category of measures is referred to as the external index. The second type of measure evaluates the clustering results directly without relying on any reference model, known as the internal index.

External index For data set  $D = \{x_1, x_2, \dots, x_m\}$ :

It is assumed that the cluster given by clustering is divided into  $C$ , and the cluster given by the reference model is divided into  $C^*$ . Accordingly,  $\lambda$  and  $\lambda^*$  represent the cluster label vectors corresponding to  $C$  and  $C^*$ , respectively. When pairing the samples in pairs, the number of unique pairs, denoted by  $a + b + c + d = \frac{m(m-1)}{2}$ .

$$a = |SS|, SS = \{(x_i, x_j) | \lambda_i = \lambda_j, \lambda_i^* = \lambda_j^*, i < j\} \quad (12)$$

$$b = |SD|, SD = \{(x_i, x_j) | \lambda_i = \lambda_j, \lambda_i^* \neq \lambda_j^*, i < j\} \quad (13)$$

$$c = |DS|, DS = \{(x_i, x_j) | \lambda_i \neq \lambda_j, \lambda_i^* = \lambda_j^*, i < j\} \quad (14)$$

$$d = |DD|, DD = \{(x_i, x_j) | \lambda_i \neq \lambda_j, \lambda_i^* \neq \lambda_j^*, i < j\} \quad (15)$$

Jaccard Coefficient ( *JC* ):

$$JC = \frac{a}{a + b + c} \quad (16)$$

Its result is in the interval [0,1], and the larger the value, the better.

Rand Index ( *RI* ) and Adjusted Rand Index ( *ARI* ):

$$RI = \frac{2(a + d)}{m(m - 1)} \quad (17)$$

$$ARI = \frac{RI - E(RI)}{\max(RI) - E(RI)} \quad (18)$$

*RI* ranges from [0,1]. The larger the value, the better. To realize that the index should be close to 0 when the clustering results are generated randomly, *ARI* is proposed. Its value range is [-1,1], and the larger the value, the more consistent the clustering result is with the actual situation.

Internal index:

The Sum of Squared Errors ( *SSE* ):

$$SSE = \sum_{i,k} (x_{i,k} - c_k)^2 \quad (19)$$

It represents the sum of the square loss of the distance from the data in the class to the center of the class, which is the optimization goal of K – Means, where  $x_{i,k}$  denotes the  $i$ -th sample point in the  $k$ -th class and  $c_k$  is the center point of the  $k$ -th class.

Silhouette Coefficient ( *SC* ):

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \quad (20)$$

$$SC = \frac{\sum_{i=1}^m s_i}{m} \quad (21)$$

In Eq. (20)  $a_i$  represents the average distance of the sample  $x_i$  from other samples in the cluster and  $b_i$  represents the minimum average distance of  $x_i$  from samples in other clusters. The range of *SC* is [-1,1]. A clustering result with a higher *SC* value, closer to 1, indicates a better clustering effect. This is achieved when the distances between samples of the same class are smaller, and the distances between samples of different classes are larger.

Calinski-Harabasz Index ( *CH* ):

$$B_k = \sum_{q=1}^k |C_q| (c_q - c_*) (c_q - c_*)^T \quad (22)$$

$$W_k = \sum_{q=1}^k \sum_{x \in C_q} (x - c_q) (x - c_q)^T \quad (23)$$

$$CH = \frac{\text{tr}(B_k)(m - k)}{\text{tr}(W_k)(k - 1)} \quad (24)$$

In Eqs. (22) and (23)  $B_k$  and  $W_k$  are covariance matrices for inter-cluster and intra-cluster data.  $c_q$  and  $c_*$  represent the center points of cluster  $q$  and the data set  $D$ , respectively, and  $C_q$  denotes the set of data belonging to cluster  $q$ . The *CH* score represents the ratio between the distance between clusters and the distance within clusters. A higher value indicates a better clustering result because the score has no upper bound. In other words, the larger the *CH* score, the more distinct and well-separated the clusters are, which is considered a better outcome.

*Fitting and classification methods*

**MLR** is a simple linear regression generalization that studies the quantitative dependence between dependent and multiple independent variables<sup>24,25</sup>. The multiple linear regression equation estimated by the sample is:

$$\hat{y} = b_0 + b_1 x_1 + \dots + b_p x_p \quad (25)$$

In multiple linear regression analysis, the regression coefficient is estimated by the least square method, that is, finding the appropriate coefficients  $\{b_0, b_1, \dots, b_p\}$  to minimize the sum of squares of the dependent variable residuals.

**DNN** (Deep Neural Network) is a neural network comprising multiple hidden layers, categorized as input, hidden, and output layers based on their positions<sup>13</sup>. Typically, the first layer serves as the input layer, the last layer as the output layer, and the intervening layers as hidden layers. Each layer is fully connected, meaning every neuron in a layer is connected to every neuron in the subsequent layer (i.e., layer  $i$  is fully connected to layer  $i+1$ ). The input layer solely receives external signals without processing, while the hidden and output layers can perform function processing, and the output layer produces the final result. For a fully connected single hidden layer neural network, with the output layer having only one neuron, the output result can be expressed as:

$$y = g\left(\sum_{k=1}^m v_k f_k\left(\sum_{i=1}^n w_i x_i - \theta_k\right) - \delta\right) \quad (26)$$

where  $g$  represents the activation function of the output layer,  $v_k$  denotes the output weight of the  $i$ -th hidden neuron,  $\delta$  stands for the bias of the output neuron,  $f_k$  is the activation function of the  $k$ -th hidden neuron,  $w_i$  shows the output weight of the  $k$ -th input neuron and  $\theta_k$  is the bias of the  $k$ -th hidden neuron.

When utilizing Deep Neural Networks (DNN), careful consideration must be given to the choice of activation function and the configuration of weights and biases. With the advancements in artificial neural networks, a diverse range of activation functions can now be employed in these networks. Activation functions typically produce values from 0 to 1 or -1 to 1, allowing us to select appropriate ones based on specific requirements. For instance, the Tanh function can be suitable for the hidden layer in a binary classification problem, while the Softmax function proves effective for multi-classification tasks. The ReLU function has gained popularity in deep learning due to its ability to avoid the vanishing gradient problem, and the Softplus function, being smoother than ReLU, also finds utility in specific scenarios.

After deciding on the activation function and configuring the network architecture, the next step is determining the optimal weights and biases to achieve the best performance in problem-solving. A loss function is chosen to evaluate the neural network's performance. Commonly used loss functions include mean square error and cross-entropy error. The neural network is then trained using this loss function, and various training methods can be employed<sup>27,28</sup>.

The primary objective during training is to find the proper parameters that minimize the loss function. This optimization process is essentially an optimization problem. A prevalent approach to address this is using gradient or Stochastic Gradient Descent (SGD) strategies. The weights and biases are updated in the negative gradient direction of the objective (loss function) to minimize the loss and improve the network's performance iteratively. The ultimate goal is to reach a configuration where the neural network performs optimally for the specific task.

**K-Nearest Neighbors (KNN) Regression** operates based on finding the  $k$  nearest neighbors of a given sample and assigning the average value of a specific attribute from these neighbors to the sample<sup>14,38</sup>. In contrast to classification, which produces qualitative outputs, regression yields quantitative results. During the training phase of KNN Regression, three crucial factors are considered: the selection of the  $K$  value, the method for measuring distances, and the decision-making rules. Decision-making rules primarily come in two forms: the average method and the weighted average method. The step-by-step process of the KNN regression algorithm is presented in Algorithm 2.

#### Algorithm 2 KNN Regression.

**Step 1.** Compute the distance between the point to be predicted and the known points.

**Step 2.** Sort the known points in ascending order based on their distances from the point to be predicted.

**Step 3.** Choose the  $k$  points with the shortest distances to the predicted point.

**Step 4.** Return the average or weighted average of the attribute values from the selected  $k$  points as the corresponding attribute value for the predicted point.

## Results and discussion

### Fitting

Actual data has been collected through investigations to track the inbound and outbound passenger flow in Chengdu subway stations at various time periods. The objective is to aid relevant departments in developing improved plans for constructing new subway stations. To achieve this, it becomes essential to forecast the passenger count entering and exiting the subway stations during different time periods. This prediction can be accomplished by utilizing MLR, DNN, and KNN Regression, which will be employed to fit the data based on the known node indicators and place indicators of the subway stations.

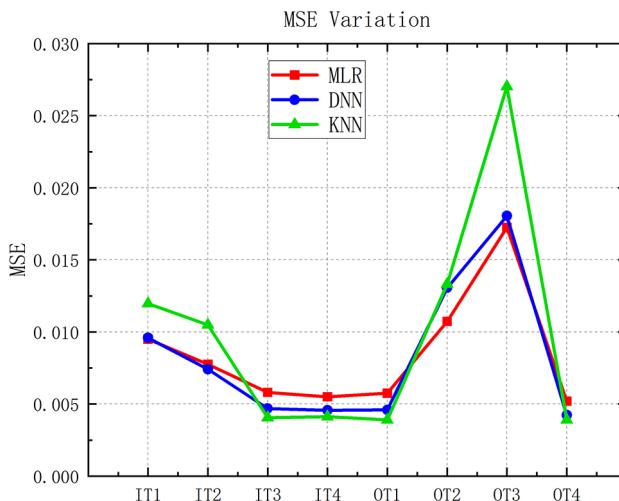
In the context of MLR, we assume the relationship between node indicators, place indicators, and ridership, which is as follows:

$$\text{Ridership} = \alpha + \sum_{i=1}^8 \beta_i N_i + \sum_{j=1}^9 \gamma_j P_j \quad (27)$$

The relationship between node indicators, place indicators, and ridership varies due to the differing number of passengers at the same station during various time periods. As a result, the coefficients associated with this relationship differ accordingly.

Model	K	Number of deleted stations	K <sub>Deleted</sub>
IT1	6	17	6
IT2	10	0	10
IT3	10	1	6
IT4	10	0	10
OT1	10	0	10
OT2	7	14	7
OT3	3	4	9
OT4	10	0	10

**Table 6.** The relevant parameter selection for fitting in each time span.



**Fig. 3.** Line plots of MSE variation on the test set under MLR, DNN, and KNN models for each time period (IT1-OT4).

In the case of DNN utilization, a single hidden layer network is employed to fit the node and place indicators to the ridership data. The input layer comprises 17 neurons, each corresponding to one of the 17 node and place indicators. The hidden layer consists of 10 neurons, and the activation function used for the hidden layer is the ReLU function, as shown in Eq. (28).

$$f(x) = \max(0, x) \quad (28)$$

The output layer of the DNN contains a single neuron, representing the ridership, and it uses the LeakyReLU activation function presented in Eq. (29).

$$f(x) = \begin{cases} x, & x > 0 \\ \alpha x, & x \leq 0 \end{cases} \quad (29)$$

In DNN regression, the loss function employed is the mean square error, and the weights and biases of the DNN are determined through the random gradient descent strategy.

For KNN regression, the Five-Fold Cross-Validation method is utilized. The process involves dividing the training data into five parts, four used as model training sets and the remaining as verification sets. This process is repeated five times to ensure a comprehensive evaluation. The grid search method is then applied to determine the optimal K value, distance measurement method, and decision rules based on the model's performance on the verification set.

The fitting relationship of different data corresponds to different K values, as depicted in the second column of Table 6. The distance measurement method is the Manhattan distance, and the decision rule is the average method.

By employing the three methods mentioned above, we have fitted eight sets of data for IT1-IT4 and OT1-OT4. In each case, 10% of the data is reserved as the test set, while 90% is used for training the models. The effectiveness of the models is assessed by calculating the Mean Squared Error (MSE) on the test set. Figure 3 displays the variation of MSE for different models across the eight test sets.

The outcome is not entirely satisfactory, particularly concerning OT3. To gain insights, we observe the data distribution for this specific case. Subsequently, the distance between the independent variables and the dependent variable of the two stations is calculated using the following methods:

$$\text{Independent}_{\text{Distance}(k,l)} = \max_{i,j} (|N_i^k - N_i^l|, |P_j^k - P_j^l|) \quad (30)$$

$$\text{Dependent(OT3)}_{\text{Distance}(k,l)} = |R_{OT3}^k - R_{OT3}^l| \quad (31)$$

After computing the distance mentioned above between each pair of stations, we can visualize the distribution histograms as shown in Fig. 4.

As shown in Fig. 5, distribution histograms were drawn to depict the distances between data corresponding to every pair of stations within each time span. The data for each time span were then analyzed using these distribution histograms. For OT3, the distance between the variables is predominantly distributed in the range of [0.50, 0.55]. Additionally, the ridership between stations is generally close, except for a few stations where significant differences in ridership exist. It appears that some stations have similar independent variables but substantially different ridership.

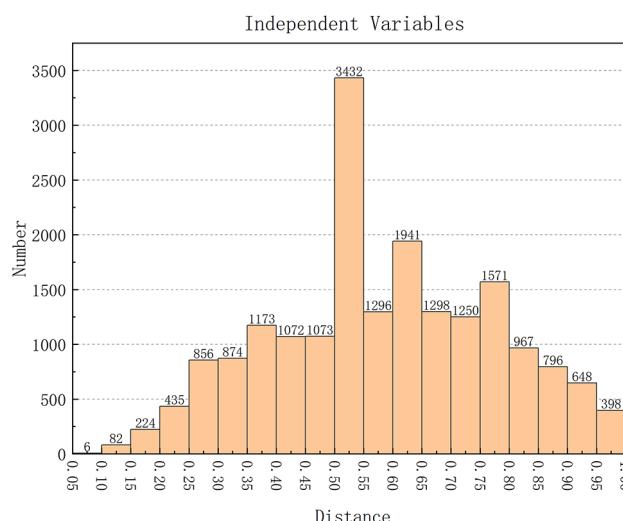
To achieve a more effective fitting outcome, we remove the stations where independent variables are very close, yet ridership differs significantly. This process is repeated for each group, resulting in the number of deleted stations for each group, which is indicated in the third column of Table 6.

After removing some stations, the data was fitted using MLR, DNN, and KNN. The different K values corresponding to the eight data groups are presented in the fourth column of Table 6. Figure 6 shows a comparison diagram of MSE on test sets for different models with distinct data. Specifically, DMLR represents the fitting results with MLR after eliminating contradictory stations, while DDNN and DKNN denote the corresponding results for DNN and KNN, respectively.

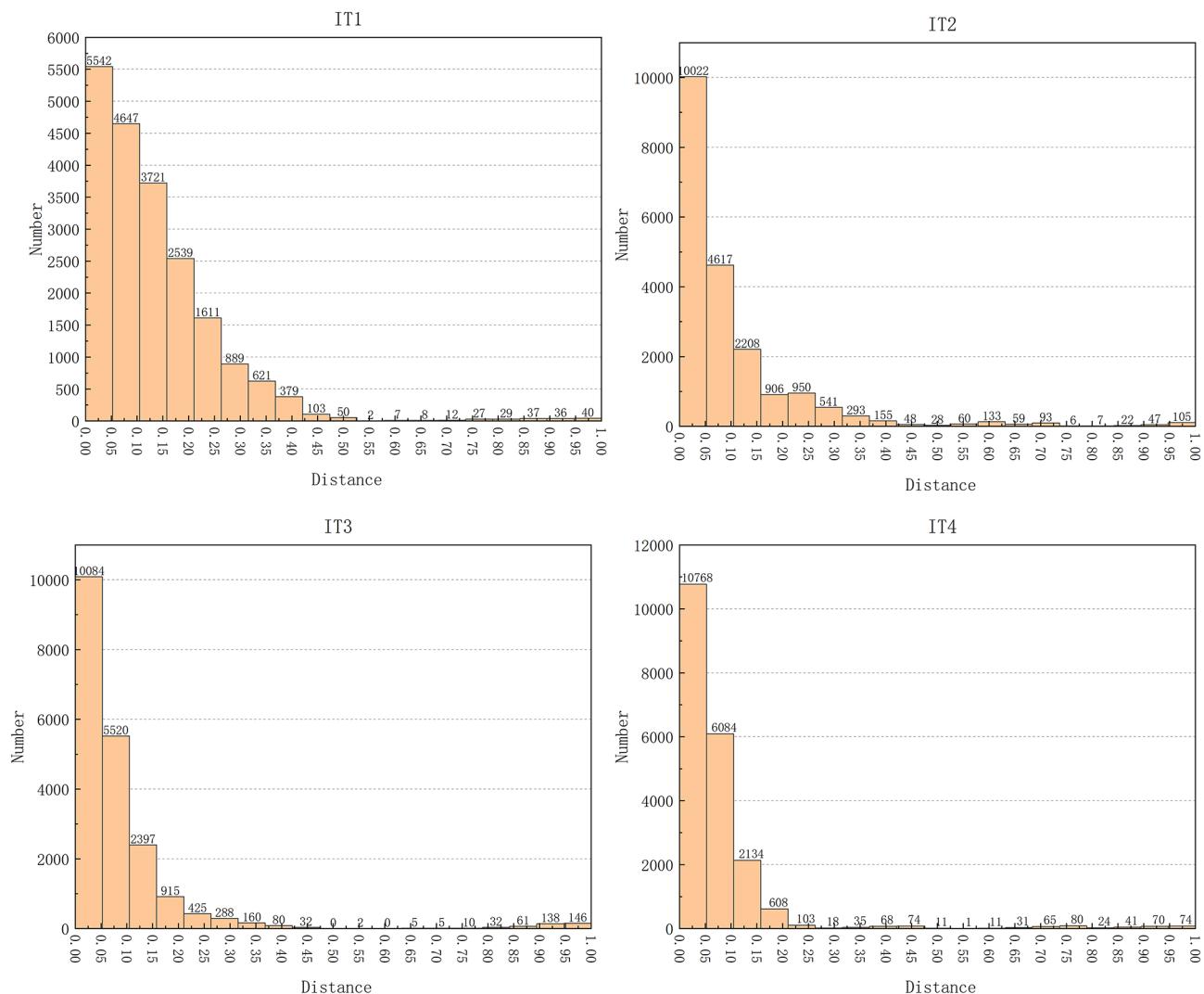
From the observations in Fig. 6, it can be inferred that the overall fitting effects of the MLR, DNN, and KNN models are relatively similar. Notably, the performances of all three models on the IT1, OT2, and OT3 test set improved after removing a certain number of contradictory stations, especially for OT3. However, the MSE values for IT2, OT2, and OT3 are relatively higher when compared to the other five groups of data. Upon analyzing the ridership distribution in Fig. 7 for these three groups, it becomes evident that the ridership still deviates from the distribution range of most stations, indicating relatively poor fitting effects for the models.

Figure 8 displays the point plots of MLR, DNN, and KNN-fitted ridership alongside the actual ridership at selected stations for each time period (IT1-OT4). These plots allow for a comparison between the fitted results and the actual ridership. By observing Fig. 8, it is evident that the machine learning model is ineffective in predicting stations with many passengers. However, when the number of passengers at a station is small, the model can accurately predict the passenger count. Therefore, deleting the abnormal stations with similar independent variables but significantly higher passenger numbers can further improve the model's fitting ability. Additionally, the distribution of actual ridership for IT2, OT2, and OT3 remains scattered compared to other time periods, even after removing the abnormal stations. This observation aligns with Fig. 8 and explains why the MSE for these three time periods in Fig. 7 is relatively large.

We obtained promising results on the test set after applying MLR, DNN, and KNN to fit the node indicators, place indicators, and ridership by excluding contradictory stations from the eight data groups. This success allows us to accurately predict station ridership when equipped with the station's node and place indicators at different times.



**Fig. 4.** Histogram of the frequency distribution of the distance between the independent variables (N1-N8 and P1-P9).



**Fig. 5.** Histogram of the frequency distribution of the distance between the dependent variable (i.e., ridership) at any two stations for different time periods (IT1-OT4).

However, two approaches can be adopted to enhance ridership predictions' accuracy. Firstly, we can train the model with more extensive station data. Additional data will help the model capture broader patterns and behaviors, leading to improved predictions. Secondly, investigating other crucial factors that influence ridership is essential. Identifying and incorporating these additional factors into the model can significantly enhance its predictive capabilities and provide more comprehensive insights into the ridership patterns of subway stations.

### Model training and selection

Since stations are unlabeled, the categorization process involves clustering in machine learning, an unsupervised learning technique. Different clustering methods operate on different principles, resulting in varied clustering outcomes. To minimize this effect, the results of multiple clustering methods are integrated. A neural network is trained using consistently labeled results, allowing the trained model to classify stations with inconsistent labeling (thus converting it to a supervised learning task) and predict the class of new stations.

To label the data, we separately employ four clustering methods (K-Means, AGNES, DBSCAN, GMM) and conduct a preliminary evaluation of the clustering effectiveness using internal indexes SC and CH. The ridership data is divided into eight groups based on the time span (T), namely IT1-OT4. The process remains the same for each group, and for illustration purposes, we will use the IT1 data as an example.

The [methodology](#) section provided a brief description of the clustering algorithm. The principles of the algorithm indicate that each one has hyperparameters that influence the clustering effect, with different parameter choices potentially impacting the results positively or negatively. Based on this, the most satisfactory clustering results and their corresponding parameter values are selected by adjusting the parameters during model training and observing changes in the internal index values.

Python is utilized to cluster the given data. The models can be enhanced by adjusting hyperparameters in the algorithms, adhering to the principles of clustering methods. For K-Means and GMM, we need to adjust the

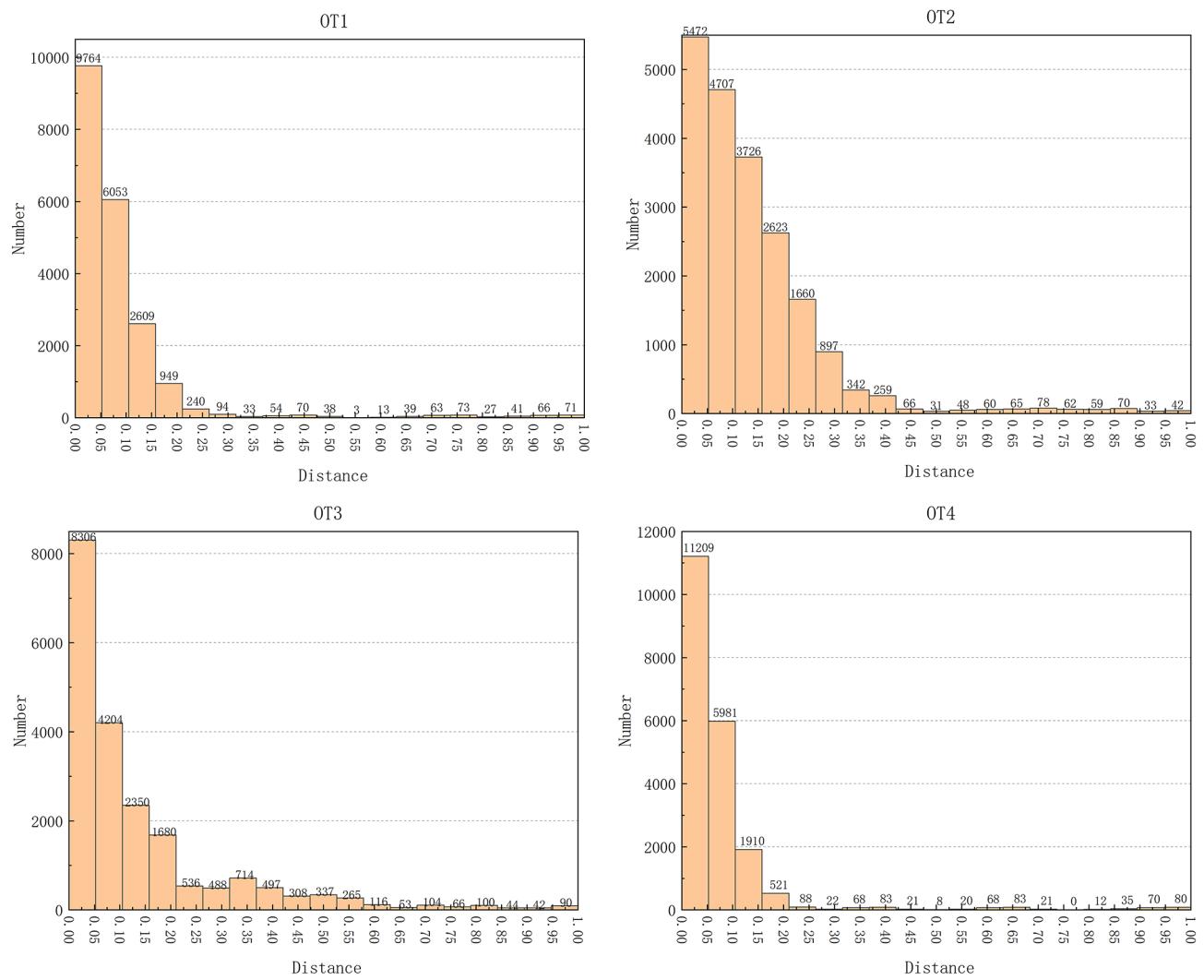


Figure 5. (continued)

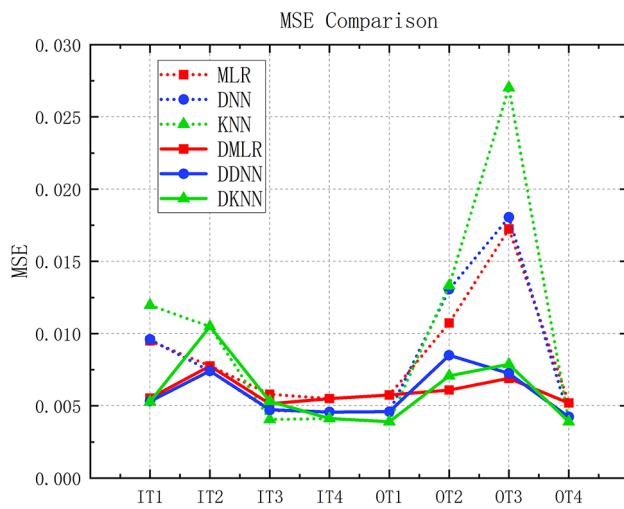
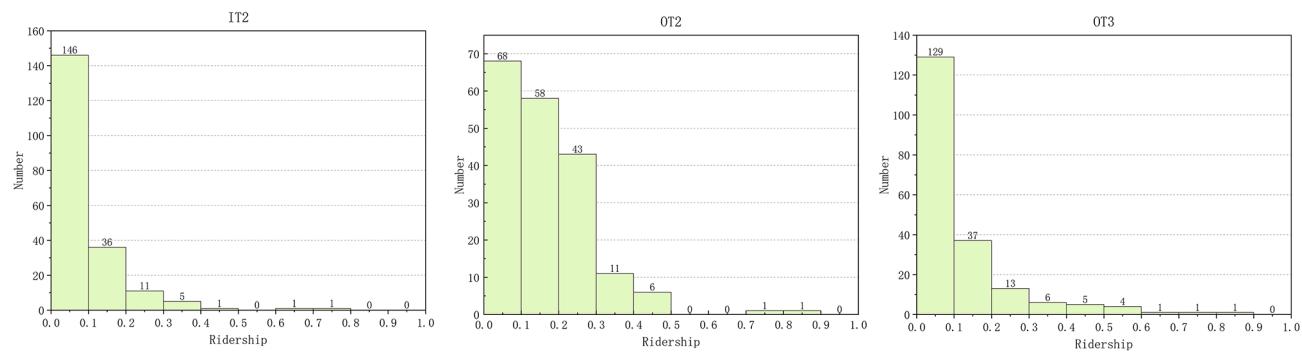
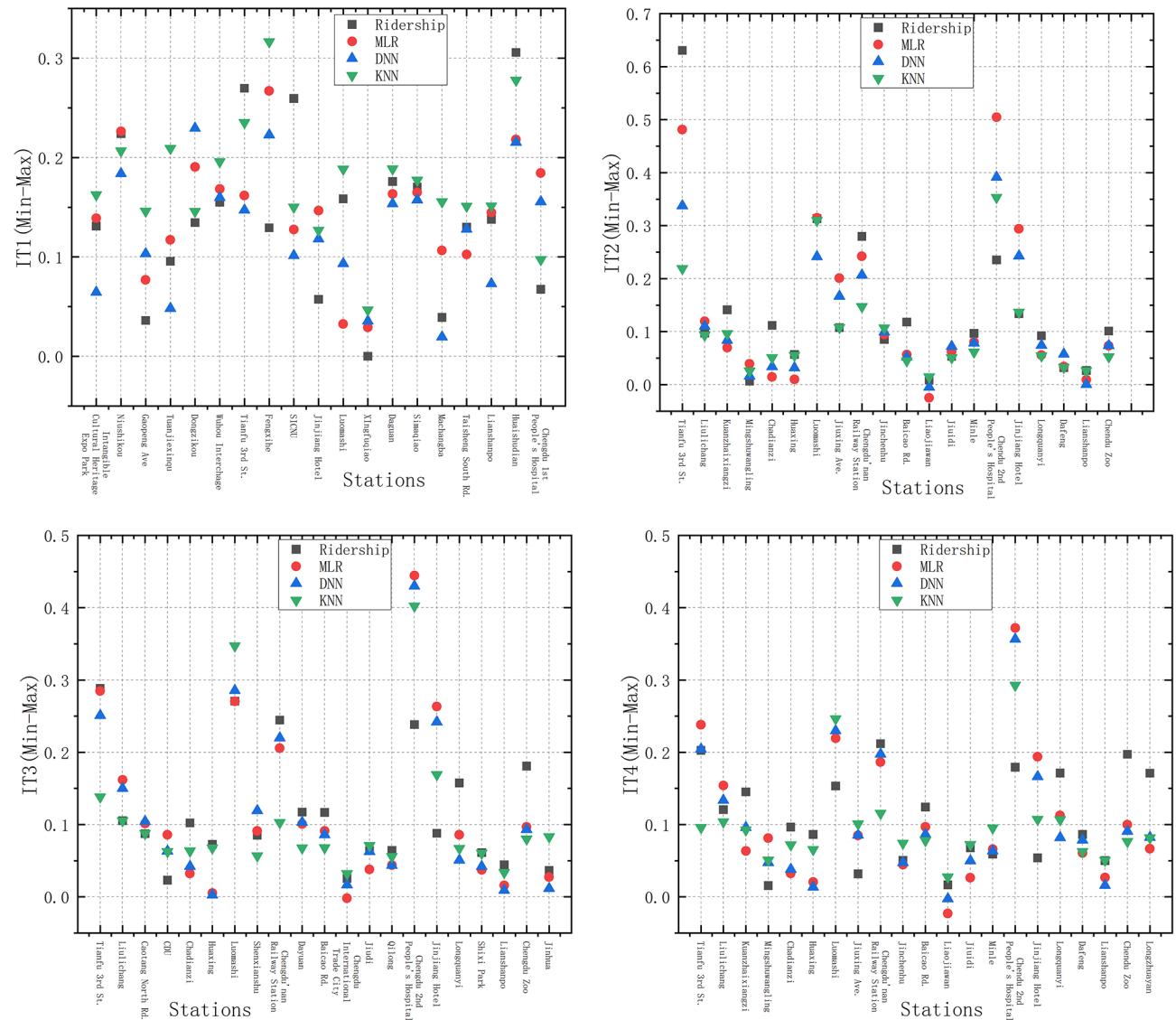


Fig. 6. Line plots of MSE changes in the test set for each time period (IT1-OT4) under MLR, DNN, and KNN, as well as the model with some contradictory stations removed (DMLR, DDNN, DKNN).



**Fig. 7.** Histogram of the frequency distribution of ridership under IT2, OT2, and OT3 time periods for outliers selection.



**Fig. 8.** Comparison and analysis of different fitting results.

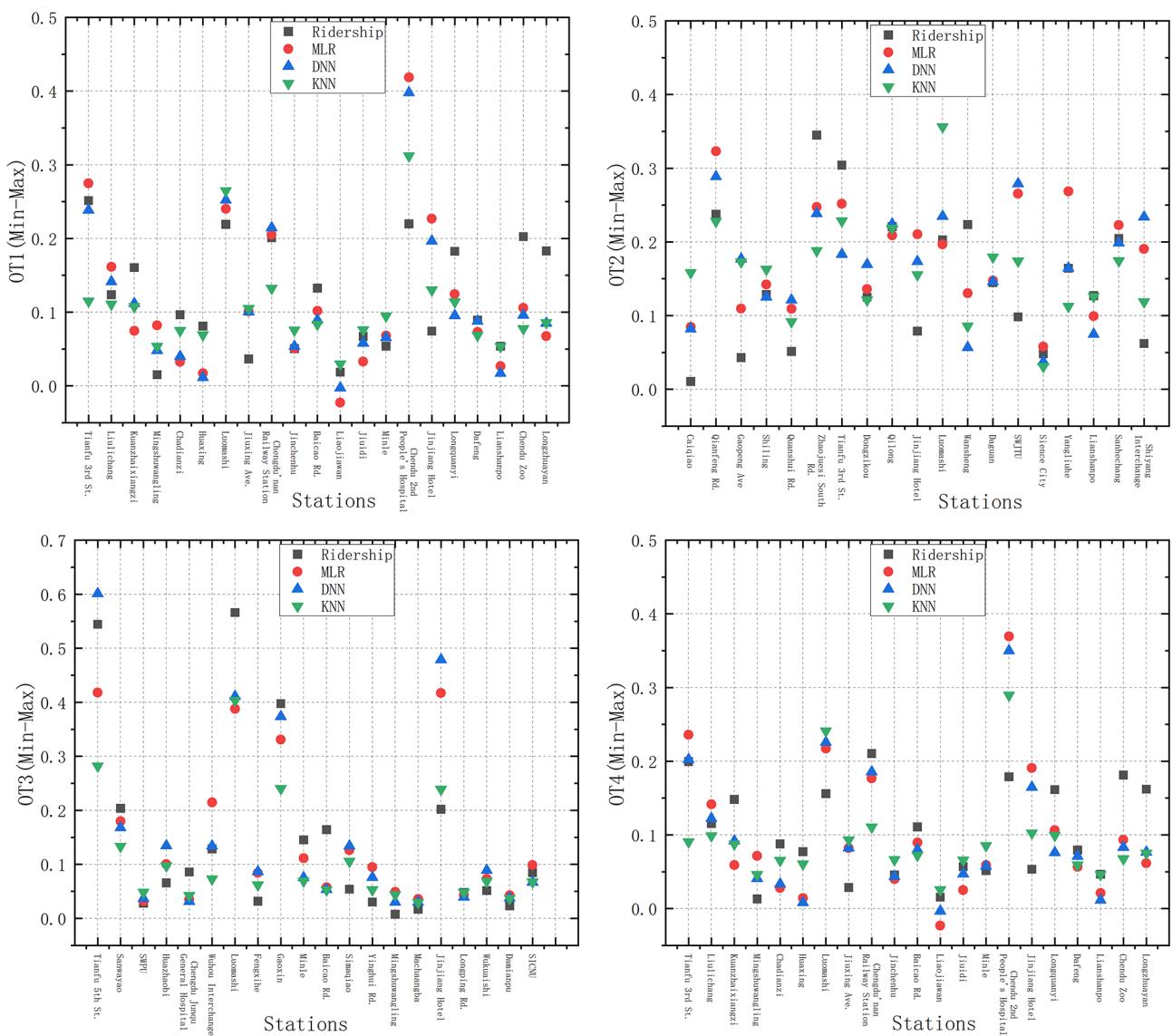


Figure 8. (continued)

number of clusters ( $k$ ) as the critical parameter. In the case of AGNES, both an appropriate  $k$ -value and a suitable distance definition method (linkages) between clusters must be determined. Regarding DBSCAN, the maximum neighborhood radius ( $\text{eps}$ ) and the minimum number of samples within the domain radius ( $\text{min\_samples}$ ) that form the core object must be adjusted for optimum results.

To achieve more satisfactory final classification results, better clustering outcomes for each clustering algorithm, which involves parameter selection, are needed. Larger values of the internal indexes SC and CH indicate better model results. The internal index values for models under different parameter settings are calculated separately to select the best model for each algorithm. Taking IT1 as an example, the results of the four basic clustering algorithms are displayed in Tables 7, 8 and 9, and 10. Based on these results, line graphs, presented in Fig. 9, are drawn to intuitively identify the optimal parameter for each algorithm. The diagrams in Fig. 9 show the internal index changes corresponding to the four clustering methods during parameter adjustment. The optimal parameters for each algorithm are then selected, and the model is trained using the clustering results under those parameter settings as the subsequent results for that algorithm. For example, with K-Means, the internal indicators SC and CH reach their maximum when  $k=3$ , indicating that clustering subway stations into three categories is most appropriate. For AGNES, although SC is slightly higher at  $k=2$ , CH is significantly higher at  $k=3$ , suggesting that  $k=3$  should be chosen as the number of clustering categories for AGNES.

After carefully evaluating the results, we selected four models trained by K-Means ( $k=3$ ), AGNES ( $k=3$ , linkages = complete), DBSCAN ( $\text{eps}=0.32$ ,  $\text{min\_samples}=3$ , resulting in two clusters), and GMM ( $k=3$ ). Using three-dimensional scatter plots, we visually represented the clustering effects of the models under their optimal parameters.

k	SC	CH	SSE
2	0.4853	143.7115	12.1938
3	0.5078	170.0935	7.7342
4	0.3951	156.2038	6.2244
5	0.4113	156.6092	5.0135
6	0.4095	153.4773	4.2634
7	0.4262	158.3712	3.5682
8	0.4325	168.1136	2.9657
9	0.4275	167.1136	2.6436

**Table 7.** Values of CH, SC, and SSE with different k values under K-Means (IT1).

k	Linkages	SC	CH
2	Ward	0.5279	150.5990
	Average	0.5779	7.8188
	Complete	0.4981	115.5534
3	Ward	0.4989	162.5765
	Average	0.5264	85.5240
	Complete	0.5304	163.0675
4	Ward	0.3749	145.4261
	Average	0.4974	70.0773
	Complete	0.5335	124.3480
5	Ward	0.3816	148.5483
	Average	0.4827	59.7751
	Complete	0.5305	107.8745
6	Ward	0.4057	145.2241
	Average	0.5157	109.0551
	Complete	0.5111	104.5335
7	Ward	0.4209	151.2981
	Average	0.4987	98.2073
	Complete	0.4934	93.9764
8	Ward	0.4268	164.0516
	Average	0.4755	86.2363
	Complete	0.4184	115.4791
9	Ward	0.3681	163.1856
	Average	0.4769	83.4957
	Complete	0.4245	114.1286

**Table 8.** Values of CH and SC with k value under different linkages in AGNES (IT1).

eps	min_samples	k	outliers	stats	SC	CH
0.32	3	2	1	[160 41]	0.5264	85.5240
0.32	4	2	1	[160 41]	0.5264	85.5240

**Table 9.** Values of SC and CH under the optimal parameter combination in DBSCAN (IT1).

As shown in Table 11; Fig. 10, based on the evaluation index, it is evident that the K-Means model with  $k=3$  provides the ideal clustering result for the data in the IT1 period. Therefore, this clustering result can be chosen as the label for the data. Figure 10 shows that subway stations are divided into three categories by the K-Means, AGNES, and GMM clustering methods and into two categories by the DBSCAN method. The results from each method have similar N, P, and R values, which verify the rationality of the clustering model and the selection of specific clustering numbers.

#### Enhancements to data labeling

During the model selection process, it is evident that choosing the best-performing algorithm's result as the data label is a viable approach. However, since different clustering methods divide the data based on distinct

k	SC	CH
2	0.4739	139.7405
3	0.4427	159.6280
4	0.3823	122.7624
5	0.3178	123.5035
6	0.2909	107.8941
7	0.3048	123.5271
8	0.3097	130.1292
9	0.2974	118.8587

**Table 10.** Values of CH and SC with different k values under GMM (IT1).

principles, they yield varying clustering outcomes. As a result, we can explore methods to combine the results from multiple algorithms. To achieve this, we can first identify samples that consistently belong to the same cluster across all clustering algorithms. These samples can be directly labeled without any modifications. On the other hand, if a data point is assigned to different clusters by different methods, we can temporarily mark it with a distinct symbol and further process it using alternative approaches.

In the provided results, DBSCAN classifies the data into two clusters, while K-Means, AGNES, and GMM classify the data into three clusters. Considering the abovementioned approach, we focus on K-Means, AGNES, and GMM results. Based on this, we label the points that do not consistently fall into the same cluster across these three methods. The figure depicts 25 red points in IT1 with different labels due to the discrepancies in the clustering algorithms. These points are not included in the three clusters formed by the other points. Figure 11 presents the scatter plot of cluster results after crossing. Appendices 1, 2, 3, 4, 5, 6, 7, and 8 illustrate the complete results from IT1 to OT4.

The remaining data points are categorized into three clusters, leaving 25 points without corresponding labels. These points may have suboptimal clustering performance, and their obtained labels may lack strong interpretability. We can employ supervised learning using the accurately labeled data to train a classification model to address this. Subsequently, this model can be used to predict the labels of the unlabeled points, providing them with more accurate labels.

Using the classification algorithm trained on data generated from the results of multiple clustering algorithms, we can group these poorly classified points into three clusters. As a result, all the data points in the dataset will have their respective classes. This approach leads to improved classification results compared to a single algorithm alone. Through this process, we effectively leverage limited data information to enhance the accuracy and interpretability of the labels assigned to these data points.

## Classification

In the previous section, data points with consistent labels were clustered, while those with inconsistent labels were set aside. However, further classification methods are necessary to cluster all stations accurately, including those with inconsistent labels. To address this, a Deep Neural Network (DNN) is employed for station classification. The model was trained using consistently labeled data points, with 70% of the data serving as the training set, and its performance was evaluated based on accuracy on the validation set.

The DNN architecture consists of an input layer with 3 neurons to accommodate the 3-dimensional data features, namely N, P, and R. It includes a hidden layer with 10 neurons that utilize the LeakyReLU activation function. The output layer has 3 neurons, each representing one of the three clusters into which the stations were grouped. This layer employs the Softmax activation function for gradient logarithmic normalization. The Cross-Entropy loss function was used during training to optimize the model's performance.

$$H(p, q) = - \sum_{i=1}^n p(x_i) \log(q(x_i)) \quad (32)$$

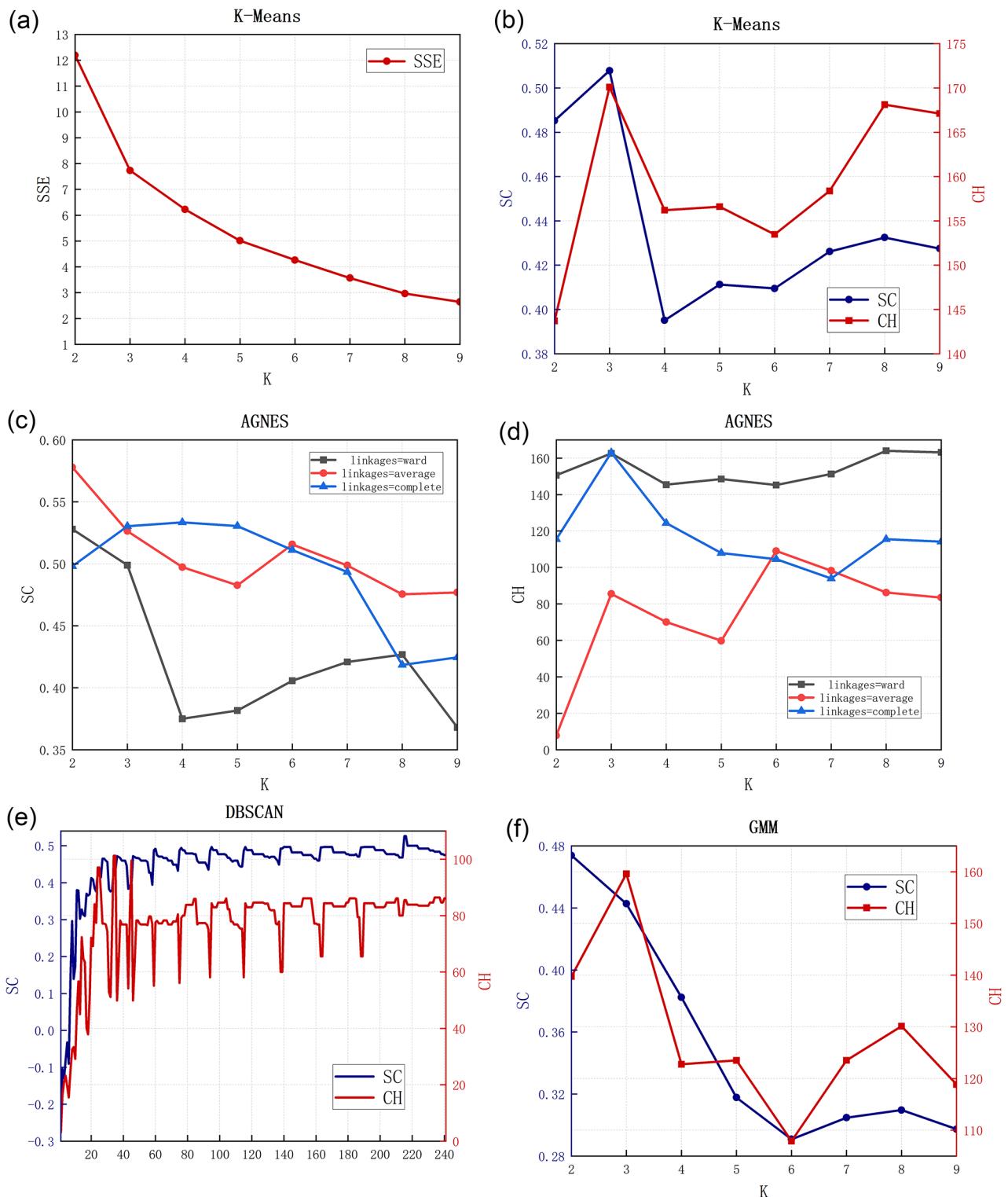
The weights and biases in the DNN are determined using the stochastic gradient descent strategy, which iteratively updates them based on the data. With iterations set to 10,000, the neural network model is trained under each time period (IT1-OT4) based on the above architecture. Figure 12 visualizes the architecture of the neural network used in this paper.

Table 12 summarizes the accuracy results for eight data groups on the validation set.

The models have shown exceptional performance on the test set, with seven out of the eight groups achieving an impressive 100% accuracy, except for OT3. This demonstrates the model's excellent training and reaffirms the effectiveness of the clustering methodology employed in the previous section.

Given this success, the model can accurately classify new stations when provided with relevant information. The trained model can also classify unlabeled sites and compare the results with various clustering methods. This comparative analysis will offer valuable insights into the model's superior performance and advantages over alternative clustering approaches.

Table 13 presents the final cluster results for 30 randomly selected stations for different time periods. These classifiers were trained using the DNN on test sets, ensuring robust evaluation across various time spans. The DNN model demonstrates its ability to make accurate predictions for the selected stations, providing valuable insights into the performance of the classifiers over different periods. This data allows for assessing the model's

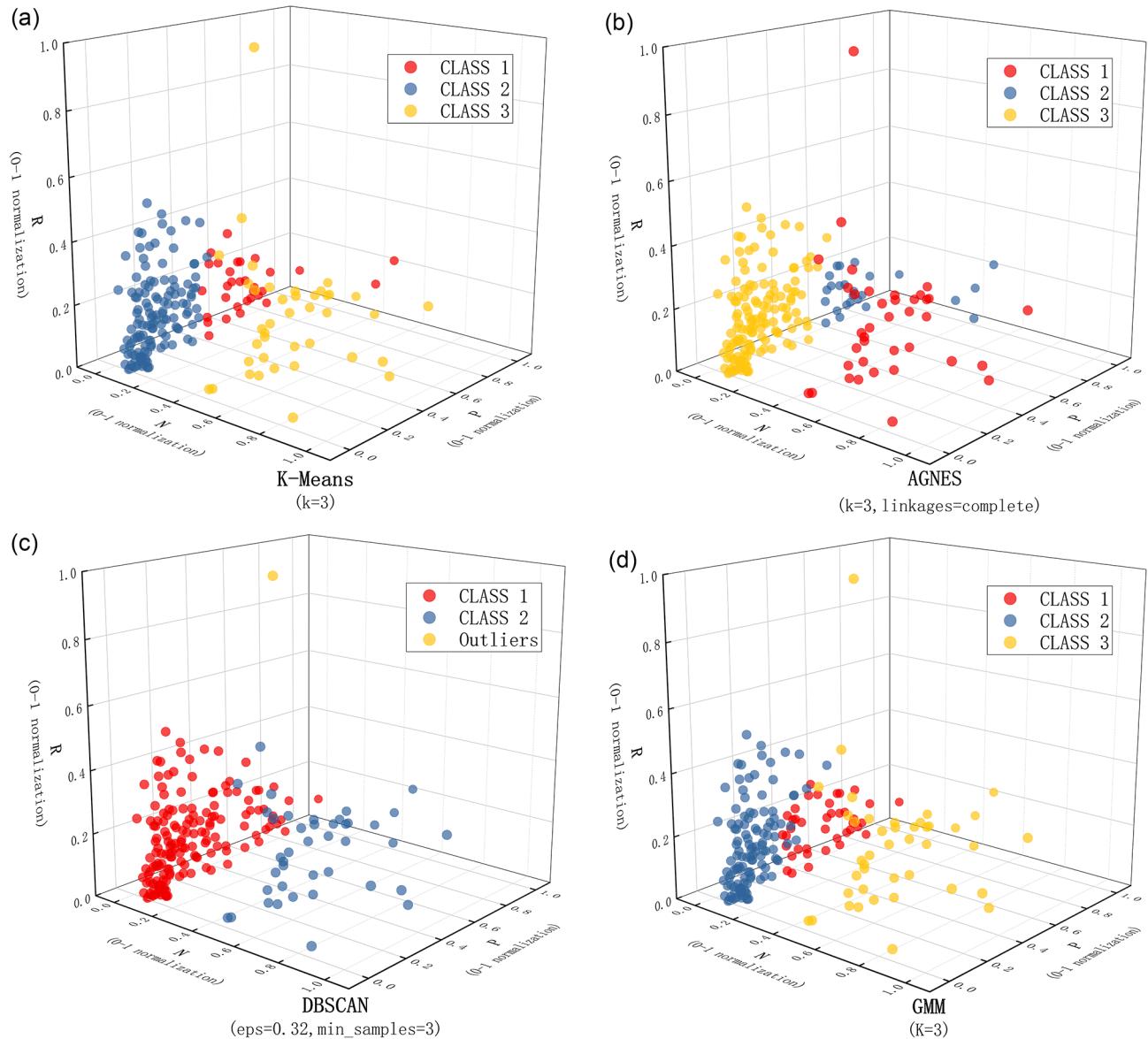


**Fig. 9.** IT1 parameter-based evaluation of clustering methods.

effectiveness in handling temporal variations and offers crucial information for understanding the overall stability and reliability of the predictions.

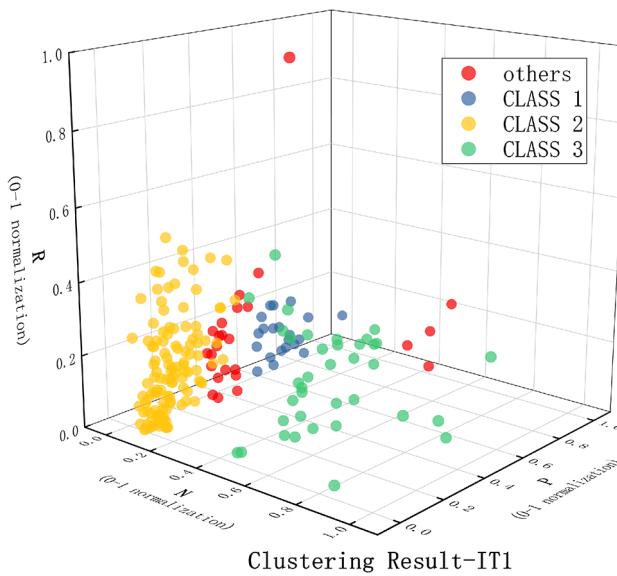
The stations were clustered using each of the three clustering methods. A neural network was then used to re-classify the data points labeled inconsistently under the three algorithms, ensuring that all subway stations had a cluster to which they belonged. This final clustering result is believed to combine some of the clustering results of the three algorithms, especially for the inconsistently labeled data.

Method	SC	CH
K-Means	0.5078	170.0935
AGNES	0.5304	163.0675
DBSCAN	0.5264	85.5240
GMM	0.4427	159.6280

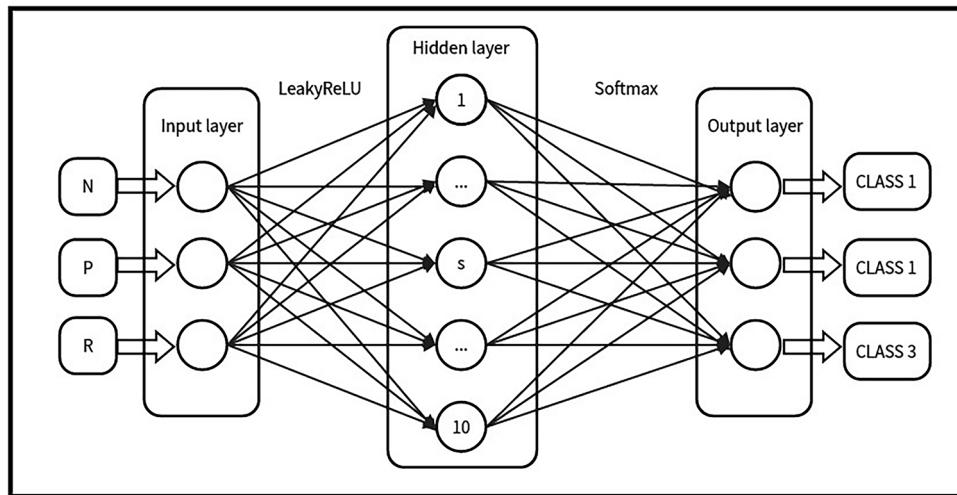
**Table 11.** The internal index results corresponding to the model selected by clustering methods.**Fig. 10.** Scatterplots of the IT1 clustering under the optimal models selected by K-Means, AGNES, DBSCAN, GMM.

Different clustering algorithms produce different results based on their principles. However, it is unclear which set of clustering results is superior. The intersection of the clustering results of the three algorithms was taken to leverage all the clustering results. Data points in the same cluster under all three algorithms were clustered together. Data points that did not fit into the same cluster were considered to be placed into different clusters by algorithms based on different principles.

To address this, a neural network was trained with consistently labeled points. This involved training the model with more accurately clustered points to predict the clusters of those currently unlabeled points. This approach maximized the accuracy of station prediction by using the clustering results of the three algorithms.



**Fig. 11.** Scatter plot representation of cluster results: post-crossing.



**Fig. 12.** The basic architecture of a neural network for training consistently labeled stations and predicting inconsistently labeled stations.

Model	IT1	IT2	IT3	IT4	OT1	OT2	OT3	OT4
Accuracy (%)	100	100	100	100	100	100	98.15	100

**Table 12.** Models accuracy under each time period (IT1-OT4).

### Verification of clustering results

As the initial dataset lacks labels, we are limited to evaluating the cluster performance using internal rather than external indexes. However, by applying the neural network, we could label each point in the dataset, reasonably assuming that these labels are relatively accurate. Leveraging this labeled dataset as a reference model, we can now employ external indexes such as JC (Jaccard Coefficient) and ARI (Adjusted Rand Index) to assess the performance of each clustering algorithm in reverse.

Using the results obtained from the DNN classifier, we calculate the external indexes for the three clustering methods, as presented in Table 14. This approach allows us to verify and compare the performance of each algorithm based on the reference labels generated by the neural network.

All of the external index values for the clustering results are relatively large, indicating a difference, but not a significant one, between the final clustering results and those of a single algorithm. The Jaccard coefficient (JC)

Number	Station	IT1	IT2	IT3	IT4	OT1	OT2	OT3	OT4
1	Weijianian	2	2	2	2	2	2	2	2
2	Shengxian Lake	2	2	2	2	2	2	2	2
3	North Railway Station	3	3	3	3	3	3	3	3
4	Renmin Rd.North	3	3	3	3	3	3	3	3
5	Wenshu Monastery	1	1	1	1	1	1	1	1
6	Luomashi	3	3	3	3	3	3	1	3
7	Tianfu Square	3	3	3	3	3	3	1	3
8	Jinjiang Hotel	1	1	1	1	1	1	1	1
9	Huaxiba	1	1	1	1	1	1	1	1
10	Sichuan Gymnasium	3	3	3	3	3	3	3	3
11	Nijiaqiao	3	3	3	3	3	3	3	3
12	Tongzilin	1	1	1	1	1	1	1	2
13	South Railway Station	3	3	3	3	3	3	3	3
14	Hi-Tech Zone	2	2	2	2	2	1	1	2
15	Financial City	1	2	2	2	2	1	1	2
16	Incubation Park	3	3	3	3	3	3	3	3
17	Jincheng Plaza	2	2	2	2	2	2	2	2
18	Century City	3	3	3	3	3	3	3	3
19	3rd Tianfu Street	2	1	1	2	2	1	1	2
20	5th Tianfu Street	2	2	2	2	2	2	1	2
21	Huafu Avenue	2	2	2	2	2	2	2	2
22	Sihe	2	2	2	2	2	2	2	2
23	Huayang	2	2	2	2	2	2	2	2
24	Haichang Road	3	3	3	3	3	3	3	3
25	Guangfu	2	2	2	2	2	2	2	2
26	Hongshi Park	2	2	2	2	2	2	2	2
27	Luhu Lake	2	2	2	2	2	2	2	2
28	Wuhan Road	2	2	2	2	2	2	2	2
29	Tianfu Park	2	2	2	2	2	2	2	2
30	Western China Int'l Expo City	3	3	3	3	3	3	3	3

**Table 13.** The final clustering results of some random stations.

External Indexes	K-Means	AGNES	GMM
JC	0.9423	0.9238	0.8618
ARI	0.9249	0.9153	0.7859

**Table 14.** The external indexes for the optimal models under K-Means, AGNES, GMM.

and Adjusted Rand Index (ARI) show that K-Means' results are superior, suggesting that K-Means substantially influences the final clustering results.

Table 14 shows that all three clustering algorithms significantly influence the final clustering results. Most data is consistently labeled under all three algorithms, and the DNN prediction utilizes this information. Consequently, the predicted results are partially similar to the clustering results of the individual algorithms. This further confirms the validity and accuracy of our station classification approach.

A portion of the consistently labeled data was selected as a validation set. Since this data has corresponding classes, its results can be used to test the effectiveness of the DNN. For the inconsistently labeled stations, they are used as a test set to predict the classes using the trained neural network. However, since clustering results from three different algorithms are available, the predicted results can be compared against them. The results in Table 14 show that the difference between the predicted and actual clustering results is minimal, verifying the feasibility of the predictions to some extent.

Additionally, the DNN model trained earlier can be used to predict classes for new unclassified stations, ensuring the model's usability. The DNN model can also be updated as station data is continuously updated to achieve better results. Scaling up the training data by constantly adding new inputs improves the neural network's predictive power, enabling it to better predict the classes of new stations and adapt to new changes and trends.

The results of this study have a great deal of meaning for urban planners and policymakers, especially in the case of TOD. The novel classification model explained not only enhances the precision of station classification but also gives implementable ideas that may support more efficient and environmentally friendly urban planning.

With the help of the reborn version of the classification model, town planners, and policyholders can now make informed decisions about new station planning and its improvement. By looking at the unique combinations of property values and the on-time performance for the various station types, the most effective recommendations can be made to be more efficient in land use and the area's transit. The station pair result directs highlights of a set of stations that possess analogous features. Thus, roads and sidewalks, bus stops, and cycle lanes can be constructed whenever necessary. Train stations must be invested only when high in demand rather than just building them all over the city. This step implies the creation of the future city, where technology and GI (geographic information) are used to improve the quality of life of residents, businesses, and the natural environment with less car dependency.

The authorities must apply adaptive TOD strategies that consider the transition of urban growth and transit usage patterns. The estimates of the model can be used to forecast future ridership trends that will enable the change of TOD plans to accommodate the volatile downtown urban structure. If the classification model with the expanded sustainability schemes is combined, then the Sustainable Development Goals (SDGs) will be supported. By concentrating on stations that help use public transport and, on the other hand, do not force people to possess personal vehicles, TOD strategies could lessen urban congestion and carbon emissions.

Another necessary part of community engagement is that local authorities, along with urban designers, can get the community to participate in their projects with the aid of the sorting model. Urban planners can analyze information from the sorting model to interact with local societies efficiently. Planners can get the public's opinion by providing precise and data-supported information about the advantages and future constructions around traffic stations. In addition, they can ensure that the TOD projects align with peoples' needs and preferences.

The main concern is setting up a system for continuously monitoring and examining rail facilities using the classification model. Routine appraisals may help to unearth the new patterns and potential mishaps earlier, which could provide better interventions to keep the efficiency and effectiveness of TOD strategies at the desired level. By encouraging other cities and areas to adopt the classification model, the model will be improved and made accessible to other cities and areas. With the help of the best practices and outcomes from the Chengdu case study that can serve as a model for the same urban environment, promoting a uniform mode of transit station classification and TOD planning would be encouraged.

Urban policies and rules are set to blend and synchronize various urban developments on different scales by transcoding the data from the robot into the already existing urban planning paperwork. The decomplication of these documents can facilitate clarity when trying to find a way to arrange various instruments to achieve sustainable development. Toward this end, urban practitioners working in conjunction with local government leaders and land use stakeholders need to realize that before entering its practical application, the application of the input code is put through deformation, noise addition, erasing some parts, and other special operations. Followed by the deployment of these core measures, by the planners' and politicians' willingness, the classification model is being led to its full potential, which leads to the pulling up of the already inefficient process and the creation of a new and streamlined urban growth utilizing applied Transit-Oriented Development strategy. Global warming and environmental degradation are major issues that must be addressed immediately to prevent the problem from worsening. The necessary proteins are injected into the programming code to produce options for the local government to develop.

According to Table 13, the classification results obtained from the DNN model provide a comprehensive understanding of the ridership patterns across rail transit stations during different time periods. These classifications have significant implications for TOD policymaking and planning, offering valuable insights into how temporal variations in passenger traffic can influence various stages of urban transit planning. The model helps assess the effectiveness of transit infrastructure and services, providing planners with data-driven approaches to improve the efficiency and sustainability of urban transit systems.

One of the key applications of this classification is in demand management and capacity planning. For example, stations like *North Railway Station* and *Renmin Rd. North* exhibit consistently high traffic (classification 3) across all time periods (IT1 to OT4). This suggests that these stations require additional capacity during peak and off-peak hours. The classification data can guide transportation planners in making decisions such as adjusting the frequency of trains, redesigning station layouts to accommodate higher volumes of passengers, and adding amenities to meet increased demand. The insights gained from these classifications allow for better resource allocation, helping to prevent over-crowding at crucial transit hubs and ensuring that stations are equipped to handle peak demand efficiently.

Additionally, the classification results are instrumental in peak-hour optimization. Stations such as *Tianfu Square* and *Sichuan Gymnasium* show high ridership during working hours but a decline in traffic during off-peak or weekend times (OT3, IT4). This pattern indicates the need for peak-hour optimization strategies, such as increasing service frequency during high-demand periods or improving station accessibility to better handle the flow of passengers during peak times. During off-peak and weekend hours, TOD strategies might consider repurposing station spaces for other uses, such as hosting community events or establishing retail pop-ups, which would keep the area vibrant even when ridership decreases.

The classification data also offers insights into infrastructure investments and upgrades. For instance, stations such as *Wenshu Monastery* and *Jinjiang Hotel* consistently show low traffic (classification 1) across all time periods, indicating that these stations experience low passenger volumes even during peak hours. This information is valuable for urban planners when considering prioritizing infrastructure investments. Instead of focusing significant resources on low-traffic stations, planners might decide to allocate more funds to high-demand stations, while low-traffic stations could be maintained with lighter service routes or alternative modes

of transportation such as biking or walking. Furthermore, these low-traffic stations could be targeted with TOD policies to increase ridership through development incentives, such as constructing residential or commercial properties nearby to draw more passengers.

The classification system also provides a framework for service flexibility and scheduling. For example, the *Hi-Tech Zone* shows varying ridership patterns, with low outbound traffic during off-peak hours (OT2) but higher outbound ridership during working hours (OT1). This variation allows transportation planners to adjust transit schedules based on real-time data, reducing operational costs during off-peak hours while maintaining rider satisfaction during periods of high demand. Flexible scheduling, such as reduced service during off-peak times and increased frequency during peak hours, can improve the efficiency of transit systems and better match passenger needs without overburdening the network.

Another critical application of these classifications is in weekend transit planning. Stations like *Tongzilin* and *Financial City* experience notable differences between weekday and weekend ridership patterns, with lower inbound traffic on weekends (IT4=1) compared to higher weekday traffic. Understanding these differences helps planners optimize weekend service routes by reducing service frequency or developing special routes catering to weekend commuters and leisure travelers. This flexibility ensures that resources are allocated efficiently, with transit services tailored to match the specific demands of each station during different time periods.

The classification data is also essential for supporting mixed-use development in high-traffic areas. Stations like *South Railway Station* show consistently high ridership across all time periods (classification 3 for IT1 to OT4). Such stations serve as key commuter hubs and prime locations for mixed-use development, combining residential, commercial, and recreational spaces. TOD strategies could promote the development of high-traffic areas into vibrant, multi-use communities, fostering economic growth while supporting commuters' daily needs. TOD policies can encourage private investment in real estate, retail, and entertainment in these areas, ensuring that the station is a focal point for residents and visitors.

The classification system also highlights opportunities for public-private partnerships. Stations like *Century City* exhibit consistent high traffic across all time periods (classification 3), indicating their importance as key transit hubs. These high-traffic stations are ideal candidates for public-private partnerships (PPP) that could promote commercial ventures or real estate projects, leveraging the high footfall to attract private investment. By working with private stakeholders, TOD strategies can integrate retail spaces, office buildings, and residential complexes into the transit network, maximizing these high-demand stations' economic and social benefits.

Finally, these classifications contribute to the development of sustainability and environmental policies. Stations with moderate but consistent traffic, such as *Weijianian* and *Shengxian Lake* (classification 2 across all conditions), can be focal points for eco-friendly initiatives. By understanding which stations experience moderate traffic, planners can design policies to reduce the carbon footprint of transit systems. These stations may be integrated with pedestrian-friendly zones, bicycle lanes, or green spaces to encourage environmentally sustainable transportation options. Furthermore, TOD policies can promote using renewable energy sources and energy-efficient technologies in transit stations, contributing to broader sustainability goals while ensuring that these stations continue to serve the community's needs.

According to the mentioned strategies and examples, classifying passenger traffic into distinct categories based on time periods and traffic types provides valuable insights into various aspects of TOD planning. These classifications help address challenges in demand management, peak-hour optimization, infrastructure investments, service flexibility, and sustainability. By leveraging this data, urban planners can create more efficient, resilient, and sustainable transit systems that cater to the evolving needs of commuters and align with long-term urban development goals.

## Conclusion

Classifying rail transit stations is crucial for shaping Transit-Oriented Development (TOD) and fostering sustainable urban growth. This paper presented a pioneering classification model that seamlessly blends traditional methodologies with state-of-the-art machine learning (ML) algorithms to analyze comprehensively and cluster rail transit stations. By incorporating mathematical models and supervised learning techniques, the model achieves a refined evaluation of station attributes, thereby enhancing the precision of classification techniques in TOD. This holistic approach enables rigorous station performance analysis and facilitates effective planning that aligns with the principles of TOD.

Several results stemming from this study include:

1. MLR, DNN, and KNN regression models were trained using the existing data, and the resulting fitting outcomes can be observed in Fig. 3. Some data clusters display slight generality, prompting the construction of a distribution map illustrating variable distances. Consequently, contradictory points in the data—where independent variables exhibit close distances, yet corresponding dependent variables are significantly distant—were identified and removed. Following the removal process, the three methods mentioned above were applied to fit the data, and the corresponding results are displayed in Fig. 6. Notably, the Mean Squared Error (MSE) for each time interval's data fitting remains below 0.012. This result is satisfactory in comparison to the situation before anomalous point deletion. As a result, accurate ridership forecasts can be generated independently using these well-trained models.
2. K-Means, AGNES, DBSCAN, and GMM algorithms were employed for data clustering. Except for DBSCAN, all algorithms grouped the stations into three distinct categories. Evaluation using internal indices is presented in Table 11. The outcomes indicate that, in terms of the Silhouette Coefficient, the ranking is AGNES > DBSCAN > K-Means > GMM, while for the CH score, the order is K-Means > AGNES > GMM > DBSCAN.

3. Distinct algorithms yield varying clustering outcomes. Consequently, we cross-reference the K-Means, AGNES, and GMM results to categorize points exhibiting consistent clustering patterns. Points with incongruent clustering outcomes are temporarily unlabeled. Refer to Table 12 for the outcomes of the neural network trained on labeled data. Among the neural networks corresponding to the eight temporal spans, only OT3 attains an accuracy of 98.15%. The remaining networks achieve 100% accuracy, indicating satisfactory performance on the test sets. Thus, we employ this neural network to predict the classification of unlabeled data and assign labels to all instances.
4. After labeling the data, the clustering effectiveness is assessed by computing the external indices ARI and JC. Refer to Table 14 for the outcomes. Notably, for both external indices, K-Means > AGNES > GMM. This alignment between the site classification predictions from the neural network and the clustering algorithms highlights the method's rationality and accuracy from an alternative perspective.

### Possible directions for future studies

Several intriguing avenues for future research can be explored to enhance further the accuracy and depth of rail transit station classification within the realm of Transit-Oriented Development (TOD).

One promising direction involves the integration of Partial Differential Equations (PDE) or other advanced mathematical models to capture the intricate spatial and temporal dynamics of station performance, enabling a more comprehensive understanding of the underlying factors influencing TOD outcomes. Additionally, a fruitful path for future investigation lies in harnessing the power of deep learning techniques, such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs), to extract nuanced patterns and correlations from complex station data, potentially uncovering hidden insights that contribute to more refined classification and planning strategies. Furthermore, an interdisciplinary approach that fuses transportation engineering, urban planning, and environmental science could yield a holistic framework for station classification, considering factors like environmental impact, accessibility, and social equity in conjunction with performance metrics. Lastly, exploring the potential integration of real-time data streams, such as crowd flow patterns and energy consumption, could provide a dynamic dimension to rail transit station classification, enabling adaptive and responsive TOD strategies that evolve with changing urban dynamics.

Exploring new research areas can enhance our knowledge of the diverse characteristics of rail transit stations and support the ongoing development of successful and sustainable TOD approaches.

### Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Received: 24 May 2024; Accepted: 7 October 2024

Published online: 13 October 2024

### References

1. Bojing Liao, Pauline, E. W., van den Berg, P. J. V., van Wesemael, T. A. & Arentze Empirical analysis of walkability using data from the Netherlands. *Transp. Res. Part D*. **85**, 102390. <https://doi.org/10.1016/j.trd.2020.102390> (2020).
2. Shiliang Su, H., Zhang, M., Wang, M., Weng, M. & Kang Transit-oriented development (TOD) typologies around metro station areas in urban China: a comparative analysis of five typical megacities for planning implications. *J. Transp. Geogr.* **90**, 102939. <https://doi.org/10.1016/j.jtrangeo.2020.102939> (2021).
3. Hongzan, J., Huang, S. & Zhou, Y. Understanding the land use function of station areas based on spatiotemporal similarity in rail transit ridership: a case study in Shanghai, China. *J. Transp. Geogr. Volume*. **109**, 103568. <https://doi.org/10.1016/j.jtrangeo.2023.103568> (May 2023).
4. Mohiuddin Ahmed, R. & Seraj Syed Mohammed Shamsul Islam. The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. *Electronics* 2020, 9, 1295; <https://doi.org/10.3390/electronics9081295>
5. Ningning Dong, T. et al. A method for short-term passenger flow prediction in urban rail transit based on deep learning. *Multimedia Tools Appl.* **83**, 61621–61643. <https://doi.org/10.1007/s11042-023-14388-z> (2024).
6. Chaojun Dong, Y., Su & Xiankun Liu. Research on TOD Based on Isomap and K-means Clustering Algorithm. 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery. <https://doi.org/10.1109/FSKD.2009.788>
7. Niu, S. et al. Measuring the built environment of green transit-oriented development: a factor-cluster analysis of rail station areas in Singapore. *Front. Architectural Res.*, <https://doi.org/10.1016/j.foar.2021.03.005>
8. Fionn Murtagh, P. & Contreras Algorithms for hierarchical clustering: an overview. *WIREs Data Min. Knowl. Discov.* <https://doi.org/10.1002/widm.53> (2011).
9. Qiang, D., Zhang, L. & Huang, X. Quantitative evaluation of TOD Performance based on Multi-source Data: a case study of Shanghai. *Front. Public. Health.* **10**, 820694. <https://doi.org/10.3389/fpubh.2022.820694> (2022).
10. Schubert, E., Sander, J., Ester, M., Kriegel, H. P. & Xu, X. DBSCAN Revisited, Revisited: why and how you should (still) use DBSCAN. *ACM Trans. Database Syst.* **42** <https://doi.org/10.1145/3068335> (2017). Article 19 (July 2017).
11. Yibin Ren, G., Chen, Y., Han, H. & Zheng Extracting potential bus lines of Customized City Bus Service based on public transport big data. *IOP Conf. Series: Earth Environ. Sci.* **46**, 012017. <https://doi.org/10.1088/1755-1315/46/1/012017> (2016).
12. Zhenyu Mei, J., Gong, C., Feng, L., Kong, Z. & Zhu Assessment of carbon emissions from TOD subway first/last mile trips based on level classification. *Travel Behav. Soc.* **36**, 100792. <https://doi.org/10.1016/j.tbs.2024.100792> (2024).
13. Abiodun, O. I., Jantan, A. & Omolara, A. E. Kemi Victoria Dada, Nachaat AbdElatif Mohamed, Humaira Arshad. State-of-the-art in artificial neural network applications: a survey. *Helijon*. **4**, e00938. <https://doi.org/10.1016/j.helijon.2018.e00938> (2018).
14. AlKhoreibi, A. H., Wakjira, T. G., Kucukvar, M. & Onat, N. C. Predictive Machine Learning Algorithms for Metro Ridership based on Urban Land Use policies in support of transit-oriented development. *Sustainability*. **15**, 1718. <https://doi.org/10.3390/su15021718> (2023).
15. Shiliang Su, C., Zhao, H., Zhou, B., Li, M. & Kang Unraveling the relative contribution of TOD structural factors to Metro ridership: a novel localized modeling approach with implications on spatial planning. *J. Transp. Geogr.* **100**, 103308. <https://doi.org/10.1016/j.jtrangeo.2022.103308> (2022).
16. Pishro, A. A. et al. Lili Zhang. The Integrated ANN-NPRT-HUB Algorithm for Rail-Transit Networks of Smart Cities: A TOD Case Study in Chengdu. *Buildings* **13**, 1944. (2023). <https://doi.org/10.3390/buildings13081944>

17. Pishro, A. A. et al. *WeiYu Li; Node, Place, Ridership, and Time Model for Rail-Transit Stations: A Case Study*1216120 (Scientific Report, 2022). <https://doi.org/10.1038/s41598-022-20209-4>
18. Zhejing Cao, Y. & Asakura, Z. Coordination between node, place, and ridership: comparing three transit operators in Tokyo. *Transp. Res. Part D* **87**, 102518. <https://doi.org/10.1016/j.trd.2020.102518> (2020).
19. Zemp, S., Stauffacher, M., Lang, D. J. & Scholz, R. W. Classifying railway stations for strategic transport and land use planning: Context matters! *Journal of Transport Geography* **19** 670–679. <https://doi.org/10.1016/j.jtrangeo.2010.08.008> (2011).
20. Chorus, P. & Bertolini, L. An application of the node place model to explore the spatial development dynamics of station areas in Tokyo. *J. Transp. Land. Use.* **4** (1), 45–58. <https://doi.org/10.5198/jtlu.v4i1.145> (2011).
21. Shaofei Niu, A., Shen, H. Z., Huang, Y. & Mou, Y. Measuring the built environment of green transit-oriented development: a factor-cluster analysis of rail station areas in Singapore. *Front. Architectural Res.* **10**, 652e668. <https://doi.org/10.1016/j.foar.2021.03.005> (2021).
22. Weiye, X. Yehua Dennis Wei. Assess the non-linear relationship between built environment and active travel around light-rail transit stations. *Appl. Geogr. Volume.* **151**, 102862. <https://doi.org/10.1016/j.apgeog.2022.102862> (2023).
23. Yu, Z., Zhu, X. & Liu, X. Characterizing metro stations via urban function: thematic evidence from transit-oriented development (TOD) in Hong Kong. *J. Transp. Geogr. Volume.* **99**, 103299 (2022).
24. Pishro, A. A. et al. & Qihong Yang; Application of Artificial Neural Networks and Multiple Linear Regression on Local Bond Stress Equation of UHPC and Reinforcing Steel Bars; *Scientific Reports*; 11:15061 ; : 2045–2322; (2021). <https://doi.org/10.1038/s41598-021-94480-2>
25. Pishro, A. A. et al. Comprehensive Equation of Local Bond Stress Between UHPC and Reinforcing Steel Bars; *Construction and Building Materials*; 262 119942; ISSN: 0950–0618 ; (2020). <https://doi.org/10.1016/j.conbuildmat.2020.119942>
26. Pishro, A. A. et al. Yuandi Zhao; advancing ultimate bond stress–slip model of UHPC structures through a novel hybrid machine learning approach. *Structures.* **62**, 106162. <https://doi.org/10.1016/j.istruc.2024.106162> (2024).
27. Pishro, A. A. et al. Siti Juhara Matlan; UHPC-PINN-Parallel Micro element system for the local bond stress–slip model subjected to monotonic loading. *Structures.* **46**, 570–597. <https://doi.org/10.1016/j.istruc.2022.10.053> (2022).
28. Pishro, A. A. et al. Qihong Yang; Structural Performance of EB-FRP-Strengthened RC T-Beams Subjected to Combined Torsion and Shear 15(14), 4852 ; (2022). <https://doi.org/10.3390/ma15144852>
29. Pishro, A. A. et al. Victor Postel; structural behavior of FRP-Retrofitted RC beams under combined torsion and bending. *Materials.* **15**, 3213. <https://doi.org/10.3390/ma15093213> (2022).
30. Jingru Huang, S., Chen, Q., Xu, Y., Chen, J. & Hu Relationship between built environment characteristics of TOD and subway ridership: a causal inference and regression analysis of the Beijing subway. *J. Rail Transp. Plann. Manage.* **24**, 100341. <https://doi.org/10.1016/j.jrtpm.2022.100341> (2022).
31. Liu Yang, X. & Song TOD Typology based on Urban Renewal: a classification of Metro Stations for Ningbo City. *Urban Rail Transit.* **7** (3), 240–255. <https://doi.org/10.1007/s40864-021-00153-8> (2021).
32. Papa, E., Carpenteri, G. & Angiello, G. A TOD classification of Metro stations: an application in Naples. In: (eds Papa, R., Fistola, R. & Gargiulo, C.) Smart Planning: Sustainability and Mobility in the Age of Change. Green Energy and Technology. Springer, Cham. [https://doi.org/10.1007/978-3-319-77682-8\\_17](https://doi.org/10.1007/978-3-319-77682-8_17) (2018).
33. Shiliang Su, Z., Wang, B., Li, M. & Kang Deciphering the influence of TOD on Metro ridership: an integrated approach of extended node-place model and interpretable machine learning with planning implications. *J. Transp. Geogr. Volume.* **104**, 103455. <https://doi.org/10.1016/j.jtrangeo.2022.103455> (2022).
34. Xin Yang, Q., Xue, M., Ding, J., Wu, Z. & Gao Short-term prediction of passenger volume for urban rail systems: a deep learning approach based on smart-card data. *Int. J. Prod. Econ. Volume.* **231**, 107920. <https://doi.org/10.1016/j.ijpe.2020.107920> (January 2021).
35. Xin Yang, Q. et al. A novel prediction model for the inbound passenger flow of urban rail transit. *Inf. Sci.* **566**, 347–363. <https://doi.org/10.1016/j.ins.2021.02.036> (2021).
36. Zhang, J., Che, H., Chen, F., Ma, W. & He, Z. Short-term origin-destination demand prediction in urban rail transit systems: a channel-wise attentive split-convolutional neural network method. *Transp. Res. Part C.* **124**, 102928. <https://doi.org/10.1016/j.trc.2020.102928> (2021).
37. Chunyan Shuai, J. et al. Relationship analysis of short-term origin–destination prediction performance and spatiotemporal characteristics in urban rail transit. *Transp. Res. Part A: Policy Pract. Volume.* **164**, 206–223. <https://doi.org/10.1016/j.tra.2022.08.006> (October 2022).
38. Zhang, M. L. & Zhi-Hua Zhou ML-KNN:Alazy learning approach to multi-label learning. *Pattern Recogn.* **40**, 2038–2048. <https://doi.org/10.1016/j.patcog.2006.12.019> (2007).

## Acknowledgements

Acknowledgments: The authors acknowledge the Sichuan University of Science and Engineering (SUSE), Sichuan University, Université Gustave Eiffel, University of the West of England (UWE Bristol), Staffordshire University London, University of Leeds, and Université de Lille.

## Author contributions

Author Contributions: Conceptualization, Ahad Amini Pishro, Shiquan Zhang, Alain L'Hostis; Data curation, Ahad Amini Pishro, Qixiao Hu, Yuetong Liu; Formal analysis, Ahad Amini Pishro, Shiquan Zhang, Qixiao Hu, Yuetong Liu, Farzad Hejazi, Maryam Shahpasand; Funding acquisition, Ahad Amini Pishro; Investigation, Ahad Amini Pishro, Shiquan Zhang; Methodology, Ahad Amini Pishro, Shiquan Zhang, Alain L'Hostis, Maryam Shahpasand; Resources, Ahad Amini Pishro; Software, Qixiao Hu, Yuetong Liu, Farzad Hejazi; Supervision, Ahad Amini Pishro, Shiquan Zhang, Alain L'Hostis, Farzad Hejazi, Maryam Shahpasand; Validation, Ahad Amini Pishro, Shiquan Zhang, Alain L'Hostis, Farzad Hejazi, Maryam Shahpasand, Ali Rahman, Abdelbacet Oueslati, Zhengrui Zhang; Visualization, Qixiao Hu, Yuetong Liu; Writing – original draft, Ahad Amini Pishro, Qixiao Hu, Yuetong Liu; Writing – review & editing, Ahad Amini Pishro, Shiquan Zhang, Alain L'Hostis, Farzad Hejazi, Maryam Shahpasand, Ali Rahman, Abdelbacet Oueslati, Zhengrui Zhang.

## Funding

This research was supported by the following: (1) Science and Technology Department of Sichuan Province (2022YFWZ0010). (2) Bridge Non-destructive Testing (NDT) and Engineering Computation Sichuan Provincial University Key Laboratory (2022QYY01). (3) The Key Laboratory of Colleges and Universities in Sichuan Province (2020QYJ01).

## Declarations

### Competing interests

The authors declare no competing interests.

### Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-75541-8>.

**Correspondence** and requests for materials should be addressed to Y.L.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

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

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024