Annals of Process Engineering and Management

www.apem.reapress.com

Ann. Proc. Eng. Manag. Vol. 3, No. 1 (2026) 34–50.

Paper Type: Original Article

Integrating Smart Card Data and Environmental Factors in Public Transportation Management: A Machine Learning-Based Framework for Mashhad

Shariat Radfar[1,*], Hamidreza Koosha[2], Ali Gholami[3], Atefeh Amindoust[1]

[1] Department of Industrial Engineering, Na.C, Islamic Azad University, Najafabad, Iran; shariatradfar@sin.iaun.ac.ir; aamindoust@iau.ac.ir.

[2] Department of Industrial Engineering, Faculty of Engineering, Ferdowsi University, Mashhad, Iran; koosha@um.ac.ir.

[3] Department of Civil Engineering, Golestan University, Gorgan, Iran; a.gholami@gu.ac.ir.

Citation:

Received: 01 April 2025 *Revised: 08 June 2025* *Accepted:13 September 2026*	Radfar, Sh., Koosha, H., Gholami, A., & Amindoust, A. (2026). Integrating smart card data and environmental factors in public transportation management: A machine learning-based framework for Mashhad. *Annals of Process Engineering and Management, 3*(1), 34-50.

Abstract

Smart cities have emerged as a solution to urban challenges such as population growth, air pollution, traffic congestion, and resource depletion. A key component of these cities is intelligent transportation systems, which leverage advanced technologies and big data to enhance the efficiency, sustainability, and accessibility of public transport. This study investigates the integration of Smart Card Data (SCD) and environmental factors such as weather, air quality, events, and land use in public transportation management in Mashhad, Iran. By applying data mining and machine learning techniques, we develop a data-driven framework capable of identifying travel patterns, predicting demand, and optimizing operational planning. The findings demonstrate that combining SCD with environmental variables significantly improves prediction accuracy, enabling more data-informed decision-making in transport management and urban planning.

Keywords: Estimation of origin-destination, Trip chain model, Smart card data, Predicting public transport trip demand, Spatial and temporal variables.

1 | Introduction

The accelerating growth of urbanization worldwide, particularly in developing countries, has introduced multifaceted challenges across social, economic, and environmental domains. According to the United Nations Fund for Population Activities (UNFPA) [1], over half of the global population now resides in urban areas, with projections indicating this will reach approximately 5 billion by 2030. This transformation necessitates more sophisticated planning and innovative strategies for managing urban services, including

Corresponding Author: shariatradfar@sin.iaun.ac.ir

Licensee System Analytics. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0).

transportation. Widely recognized as a fundamental component of smart cities, smart transportation plays a crucial role in alleviating urban issues [2]. Developing high-quality public transport systems demands innovative planning and operational strategies supported by recent advancements in information technology that enable sophisticated data collection methods for significantly enhanced public transport management.

As an invaluable source of data, smart card transactions enable accurate estimation of public transport network flows. These datasets provide comprehensive insights into urban dynamics and public transport mobility patterns, while also facilitating travel demand prediction [3–5]. The extracted individual-level travel patterns serve as valuable information for network design and optimization. Implementation of these strategies can reduce dependence on private cars, mitigate overcrowding, and alleviate traffic congestion [6–9]. Travel demand forecasting plays a critical role in determining future trip volumes, identifying potential congestion points, and estimating adoption rates of new transport services [10]. However, reliance on these data alone may prove insufficient. A comprehensive analysis requires understanding the key variables that influence travel intentions and behaviors, which shape travel patterns and enable better predictions regarding mode choice and frequency of use. Such enhanced predictions support more data-driven policy development and program implementation. Multiple factors such as holiday periods, social trends, demographic shifts, and meteorological conditions can significantly affect public transport demand and travel patterns [11]. Furthermore, travel patterns exhibit strong linkages with urban land use characteristics. Strategic planning of city land use and the spatial environment can reduce overall travel demand while encouraging modal shift from private vehicles to public or active transportation options. These measures contribute to improved traffic conditions, enhanced road service levels, and progress toward sustainable urban development [12–15].

Access to high-quality, high-performance, and efficient public transportation services requires a variety of planning and operational innovations and the factors affecting them [16], [17]. Understanding travel behaviors, people's activity patterns, and passenger demand forecasting are an integral part of real-world public transportation operations and are very challenging, as demand is simultaneously affected by complex and nonlinear interactions of spatial and temporal factors [18–22]. Machine learning, as one of the powerful tools in big data analysis, can help identify hidden patterns and predict demand more accurately. Focusing on the city of Mashhad, this study presents a machine learning-based framework for integrating Smart Card Data (SCD) and environmental factors.

Delivering high-quality, high-performance, and efficient public transportation services demands innovative planning approaches, operational solutions, and consideration of multidimensional influencing factors [16], [17]. Understanding travel behaviors, analyzing activity patterns, and forecasting passenger demand constitute essential components of public transportation system management - tasks that remain particularly challenging due to the complex, nonlinear interplay of spatial and temporal variables [18–22]. As one of the most powerful tools in big data analysis, machine learning offers capabilities to uncover hidden patterns and generate more accurate demand predictions. This study develops a machine learning-based framework that integrates SCD with environmental factors, using Mashhad as a case study.

2 | Methodology

Fig. 1 illustrates the overall framework of this study by presenting the input data, outputs, and tools used at each stage.

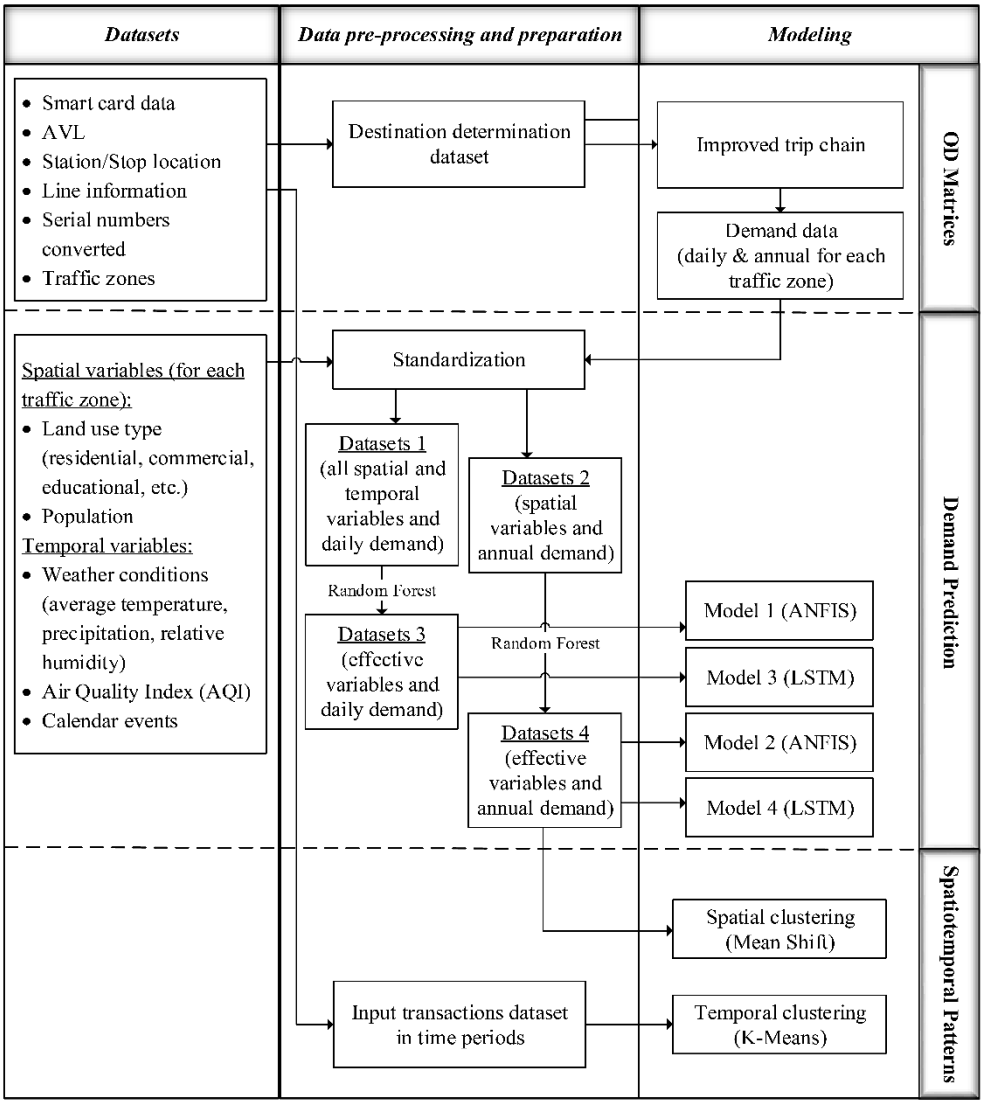


Fig. 1. The general framework diagram of this research.

2.1 | Research Environment and Data

Mashhad, as Iran's second-largest metropolitan area with a population exceeding 5 million, presents a distinctive urban context for public transportation demand analysis. The city's remarkable population density (>10,000 inhabitants/km², substantially higher than global urban averages) creates unique challenges and opportunities for transit system planning. Our analysis focuses on 253 officially designated traffic zones [23], which form the spatial framework for this study.

The research leverages transaction records from "Man-Card" – Mashhad's integrated smart fare payment system. This is an entry-only Automated Fare Collection (AFC) system, requiring passengers to validate their cards when boarding buses or accessing metro stations. As a result, it provides precise origin data, while passenger destinations must be inferred through modeling approaches. The complete anonymized dataset was acquired through official collaboration with Mashhad Municipality. *Fig. 2* illustrates the city's multimodal transit network, depicting bus routes, metro lines, and station locations.

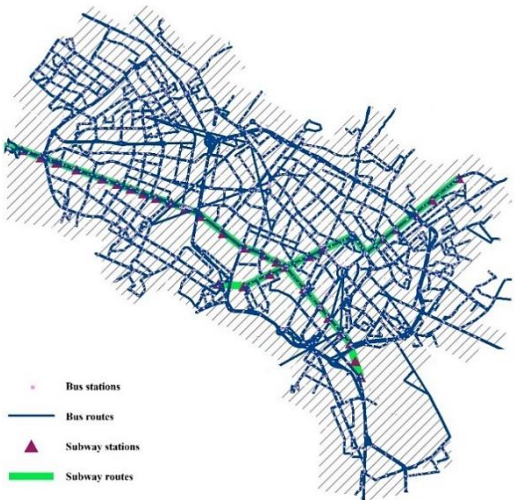


Fig. 2. The bus and subway stations and routes.

In this study, the dataset comprises Smart Cards Data, Automatic Vehicle Location (AVL) data, traffic zones data, land uses data, population data, weather data, air quality data, and calendar data. These diverse datasets, collected from March 2019 to March 2020, were made available to researchers in raw format.

2.2 | Data Preprocessing

The preprocessing steps involved removing incomplete data, consolidating data, and converting the data into a suitable format for analysis. These preprocessing steps were performed at each stage of the modeling process. We created two datasets -a daily dataset and an annual dataset- by merging all spatial and temporal data. Then, using the random forest algorithm [24–26], the best features were identified and selected, resulting in two datasets: Dataset 3 and Dataset 4.

2.3 | Modeling

2.3.1 | Estimating current public transport trip demand

Trip Chain Model (TCM): In this study, we employed the improved TCM [27] to infer public transport trip destinations by traffic zone. The key enhancements to the model included:

I. Utilizing one year of smart card transaction data
II. Including only cards with a single daily transaction
III. Incorporating actual vehicle arrival times rather than scheduled timetables
IV. Excluding transactions unsuitable for destination inference in Origin-Destination (OD) matrix estimation

Validation of destination inference results: Validation was performed by calculating the Root Mean Square Error (RMSE) between bus passenger alighting counts and TCM outputs.

2.3.2 | Predicting future public transport trip demand

Two machine learning approaches - Adaptive Neuro-Fuzzy Inference System (ANFIS) and Long Short-Term Memory (LSTM) networks - were employed to predict public transport demand across different time horizons. The models incorporate available spatial and temporal variables to generate both short-term and long-term forecasts for each traffic zone.

Adaptive Neuro-Fuzzy Inference System Model: The ANFIS model is based on the Takagi-Sugeno type-3 fuzzy inference system, where the output of each rule is a linear combination of input variables and the final output represents the weighted average of all rule outputs [28], [29]. This system combines human knowledge

with input-output datasets to construct a fuzzy inference system whose membership function parameters are automatically tuned through learning algorithms. The model's architecture dynamically adapts to the number of inputs, outputs, and rules employed in the system. For parameter optimization, ANFIS employs a hybrid learning approach that integrates gradient descent with backpropagation for parameter updates and least squares estimation for error minimization [30], [31]. The gradient descent method, widely used in neural network training, benefits from backpropagation's efficient gradient computation. Simultaneously, the least squares estimation method, commonly applied in regression models, determines optimal error coefficients by minimizing the sum of squared residuals [29], [32–34].

Long Short-Term Memory Model: The LSTM model, as an advanced Recurrent Neural Network (RNN) variant, demonstrates superior capability in time-series data prediction. Its architecture effectively preserves long-term temporal dependencies while establishing optimal mappings between input and output sequences. The key distinction from traditional RNNs lies in the hidden layer composition, which integrates memory cells and gating mechanisms to regulate information flow. Specifically, LSTM incorporates four fundamental components: 1) an input gate, 2) a forget gate, 3) a cell state, and 4) an output gate [35–37].

Performance evaluation of forecasting models: Model performance assessment employs five statistical coefficients: 1) Mean Squared Error (MSE) for prediction accuracy, 2) coefficient of determination (R^2) for variance explanation, 3) Willmott Index (WI) for reliability measurement, 4) BIAS for systematic error evaluation, and 5) Nash-Sutcliffe Efficiency (NSE) for hydrological modeling effectiveness [38–42].

2.3.3 | Identifying temporal and spatial patterns

Temporal patterns: Investigating the status of traffic zones based on the number of transactions made in different time intervals can be effective in providing a general status of the network to transportation decision makers. For this purpose, the K-Means algorithm [43], [44], which is the most well-known distance-based algorithm [45] will be used in common Python libraries, and clusters are determined in three groups: 1) morning, 2) afternoon, and 3) evening.

Spatial patterns: Spatial characteristics of traffic zones, such as land use types, area size, and population density, influence public transportation demand and subsequent planning decisions, including fleet allocation. Understanding these patterns supports optimized land use distribution to minimize unnecessary trips and improve overall public transportation performance. The Mean Shift algorithm, a kernel density estimation-based clustering technique [46], will be used for spatial pattern identification.

Clustering evaluation: Clustering results will be evaluated using the Silhouette Score [47], [48], which measures cluster quality within the range [-1, 1], and the Davies-Bouldin Index (DBI) [49], [50], which assesses cluster separability.

3 | Results

3.1 | Estimating Current Public Transport Trip Demand

The model estimated trip destinations and accounted for previously undetectable transactions to determine trip production and attraction within each traffic zone. Subsequently, it calculated the annual number of trips produced and attracted, along with the annual OD matrix. *Table 1* presents a summary of the descriptive statistical indicators for the annual OD matrix.

Table 1. Summary of the results of the statistical indicators of the annual OD matrix.

Indicators	Trip Production	Trip Attraction
Mean	540,482	540,482
Std. Deviation	670,290	662,863
Kurtosis	14.2	17
Skewness	3.2	3.4

Min	0	0
Max	5,202,395	5,447,070

Fig. 3 presents the amount of trip production and attraction map From March 2019 - 2020 for each of the traffic zones.

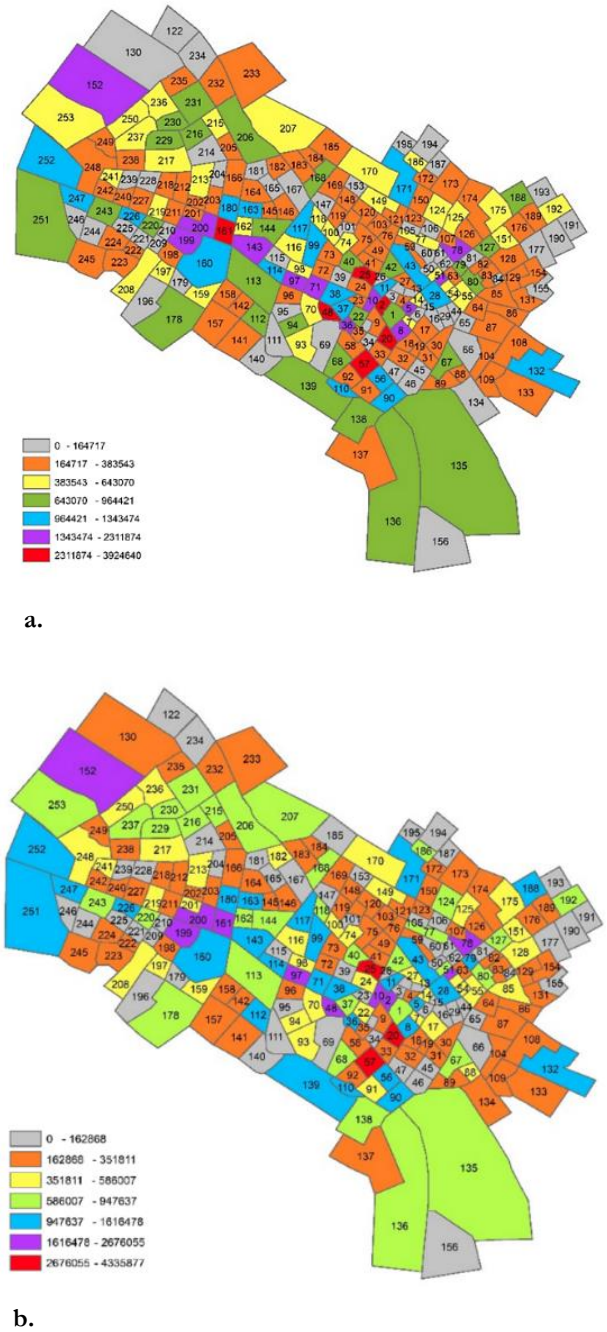


Fig. 3. The number of trip production and attraction map (from March 2019 - 2020); a. trip production, and b. trip attraction.

Large-scale validation of TCMs is difficult, limiting research in this area. To validate proposed model, in this study manually counted passengers boarding and alighting on 64 bus lines and assessed its accuracy against a destination stop inference model using RMSE. Validation results show an error of fewer than 5 passengers per station in most cases, demonstrating the model's promising performance and accuracy. This suggests our

key improvements provide a robust framework for estimating current public transportation demand in zonal travels.

3.2|Predicting Future Public Transport Trip Demand

ANFIS and LSTM models were developed for Datasets 3 and 4 using the complete data. Models were trained and tested on a computer with a 7-core processor and 16 GB RAM, using an 80/20 training/testing split. The prediction models settings are as presented in *Table 2*. Several trials were undertaken to obtain the most-suited combination of values to reach the highest accuracy.

Table 2. Parameters of prediction models.

ANFIS Models Parameters		
Parameters	**Model 1**	**Model 2**
Number of inputs	8	6
Type of membership functions	Gaussian	Gaussian
Fuzzy Inference Method	Sugeno	Sugeno
Number of epochs	50	50
Learning rate	le-4	le-4
Initial parameter values	0.05	0.05
Optimization algorithm	Gradient descent	Gradient descent
LSTM Models Parameters		
Parameters	**Model 3**	**Model 4**
LSTM layers	2	2
Dense layers	1	1
Units in layer1	100	100
Units in layer2	50	50
Learning rate	0.001	0.001
Dropout rate	0.2	0.2
Optimizer	Adam	Adam
Activation function	ReLU	ReLU
Number of epochs	100	100
Batch size	32	32

Table 3 presents the statistical performance results of these models.

Table 3. Performance of prediction models using statistical coefficients.

ANFIS Model		**MSE**	**BIAS**	**R^2**	**WI**	**NSE**
No. 1	Training data	1.0114	12.4008	0.222	0.1429	0.0222
	Test data	0.9688	158.03382	0.297	0.1643	-0.0297
No. 2	Training data	0.0024	0.0229	0.9415	0.9730	0.8738
	Test data	0.0090	0.0376	0.7729	0.9253	0.7283
LSTM Model		**MSE**	**BIAS**	**R^2**	**WI**	**NSE**
No. 3	Training data	0.0006	0.9308	0.9047	0.7591	0.9047
	Test data	0.0006	0.6402	0.9029	0.7587	0.9029
No. 4	Training data	0.0144	3.4655	0.0341	0.0248	0.0341
	Test data	0.0303	-22.5075	-0.0133	0.0984	-0.0133

Statistical coefficient analysis reveals the following:

I. Model 1 is unsuitable due to poor performance, high systematic error, and poor generalization. Major revisions or alternative approaches are necessary.

II. Model 2 provides robust forecasting for the given dataset and is a reliable predictor of annual demand. However, modifications and overfitting prevention are recommended to improve generalizability.

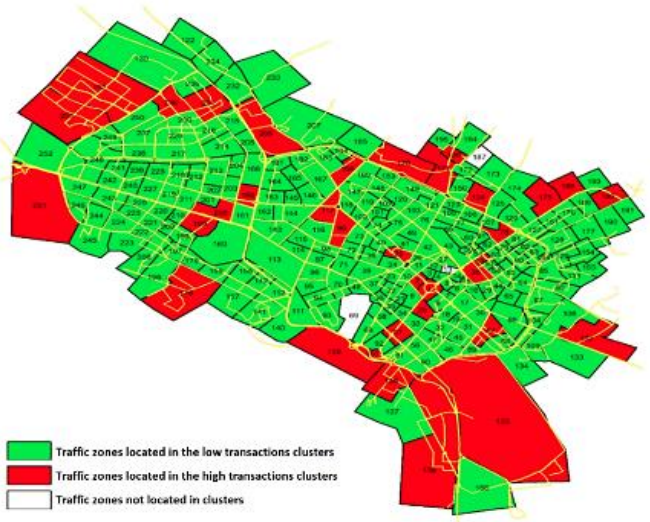
III. Model 3 excels with excellent accuracy, minimal error, and strong generalization, making it the best-performing model for this task.

IV. Model 4 performs inadequately due to high errors and poor generalization, necessitating significant re-evaluation, tuning, or redesign.

Therefore, the ANFIS model is more effective for predicting annual travel demand, while the LSTM model is superior for predicting daily travel demand.

3.3 | Identifying Temporal and Spatial Patterns

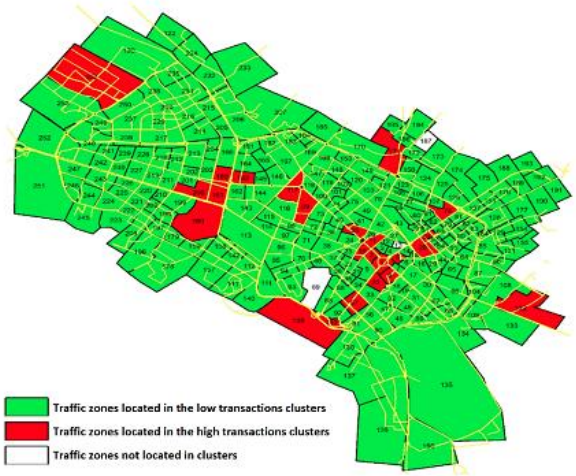
The K-Means algorithm divides the traffic zones for bus transportation into two clusters for each time interval morning, noon and evening: Cluster 0 (Low transaction) and Cluster 1(High transaction). For metro data, the K-Means algorithm divides the traffic zones into seven clusters, representing transaction intervals for morning, noon and evening time intervals.



The identified zonal clusters for the morning, noon, and evening periods are displayed separately for bus and metro in *Fig. 4* and *Fig. 5*, respectively.

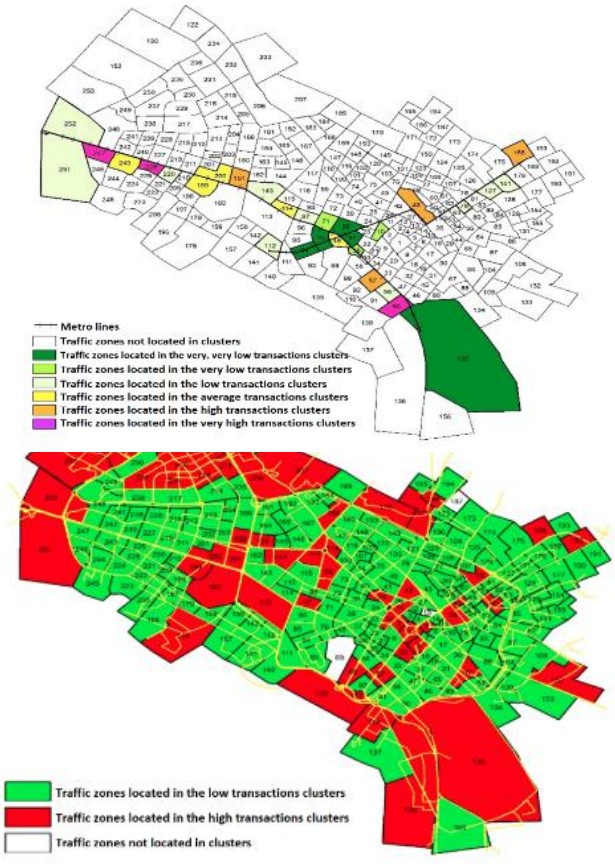
a.

b.



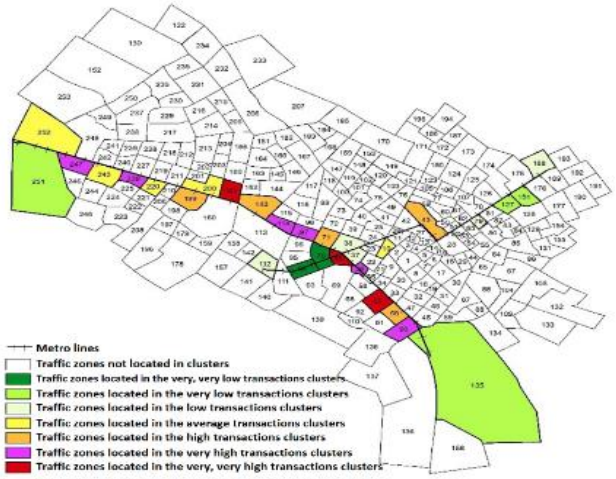
c.

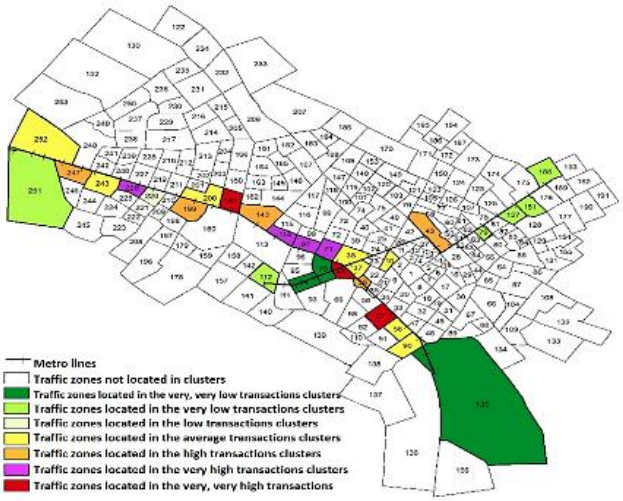
Fig. 4. The spatial distribution of clusters of traffic zones for the bus; a. morning period, b. noon period, and c. evening period.



a.

b.


Metro lines
Traffic zones not located in clusters
Traffic zones located in the very, very low transactions clusters
Traffic zones located in the very low transactions clusters
Traffic zones located in the low transactions clusters
Traffic zones located in the average transactions clusters
Traffic zones located in the high transactions clusters
Traffic zones located in the very high transactions clusters
Traffic zones located in the very, very high transactions clusters



c.

Fig. 5. The spatial distribution of clusters of traffic zones for the metro; a. morning period, b. noon period, and c. evening period.

Mean Shift clustered traffic zones into eight distinct groups, as shown in *Fig. 6*. A Silhouette Index of 0.42 and a DBI of 0.73 indicate satisfactory clustering performance.

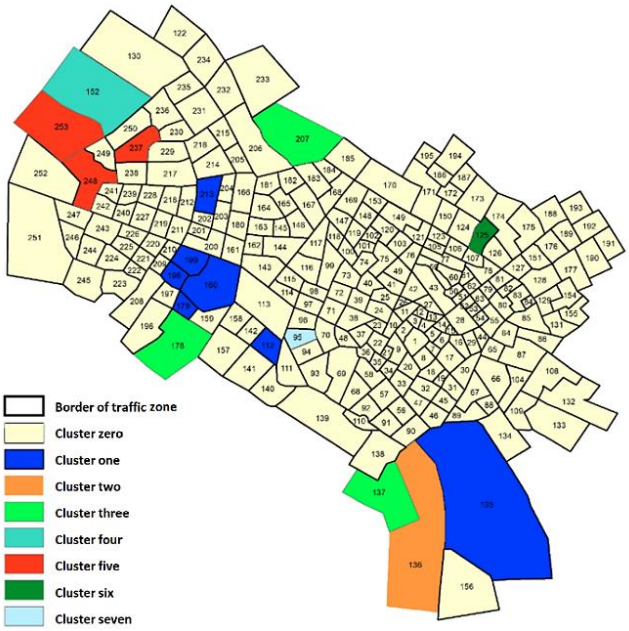


Fig. 6. Spatial location of clustering of traffic zones.

4 | Analysis of Results and Discussion

Of the 261.9 million public transportation transactions in Mashhad from March 2019 to 2020, 29.9 million were excluded from destination station analysis, primarily due to bus drivers' cash card transactions and urban train QR code usage. Consequently, the destination station estimation algorithm processed 232.1 million transactions, successfully estimating the destination for over 98% of them.

Trip chain algorithms achieve 60-85% accuracy in inferring travel destinations [51], but large-scale validation remains challenging, limiting research. Existing validation efforts focus on specific stages of OD matrix estimation, such as destination station or transfer characteristics. The improved TCM presented by Alsger et al. [52] shows 72.4% accuracy, and the destination station estimation of non-link trips by He and Trépanier [53] shows a further 11% improvement. In this study, the improved TCM served as the base algorithm. The destination station of unconnected trips was estimated. Furthermore, undetectable transactions were determined after calculating the proportion of each origin and destination from the total transactions on that route. The volume of data used is the total transactions of smart cards in both bus and metro systems during the year from March 2019 - 2020, which represents a comprehensive survey and coverage of public transportation in all urban areas. Finally, in order to produce the OD matrix, the production and attraction values of each traffic zone are aggregated separately, which can reduce the impact of possible errors in estimating the destination of each transaction.

The matrices derived from smart card transaction data accurately reflect the current network state by providing direct estimates of user trip aggregation within specified time intervals. These current demand values and analyzed variables formed the basis for developing predictive models of future demand. Statistical evaluation results revealed that:

I. Model 2 (ANFIS-based) effectively predicts annual travel demand

II. Model 3 (LSTM-based) demonstrates strong performance in daily demand prediction

Key advantages of the models include:

For Model 2:

I. Flexibility in modeling nonlinear, complex spatial-temporal relationships

II. Dynamic parameter updating capability with new data

For Model 3:

I. Exceptional ability to learn long-term sequential dependencies

II. High accuracy in time-series prediction

This analysis demonstrates how land use patterns, population distribution, Bus Rapid Transit (BRT) network configuration, and major urban attractions collectively shape both spatial and temporal demand variations across traffic zones. The application of random forest algorithms has enabled us to validate and build upon existing research findings regarding key determinants of public transport demand. The random forest algorithm application confirms several established relationships between independent variables and public transport demand, while also revealing some divergent findings. Our analysis validates the population effects demonstrated by Zhang and Xu [54], along with the air quality impacts documented by [55–57], particularly regarding potential vehicle restrictions, reduced active transportation, and facility closures. The special event patterns identified by [58–62] are similarly confirmed within Mashhad's unique socio-religious context. Furthermore, the land use influences reported by [12–15], [63–65] are strongly supported by our results. Notably, the observed positive correlation between population density, land use diversity, and trip generation rates finds parallel in studies examining urban spatial structure, including [8], [66] regarding job-housing imbalances, as well as [67] on accessibility factors. The temporal dimension of these patterns further aligns with [68] regarding residential versus workplace identification and [69] concerning attraction point dynamics.

However, our findings present an important contrast with [70–72] by showing significantly weaker weather sensitivity in daily demand patterns. This discrepancy appears attributable to Mashhad's substantial population of dedicated public transport users with limited mobility alternatives, combined with the tendency of discretionary riders to switch to private vehicles or taxis rather than persisting with public transport during adverse weather conditions.

5 | Summary and Conclusion

This study presents a comprehensive framework for estimating current and predicting future (short-term and long-term) public transport demand using real-world SCD. Through innovative enhancements to the TCM, we successfully derived trip destinations and generated both daily and annual OD matrices for current demand estimation. Notably, this research represents the first investigation of simultaneous temporal-spatial factor interactions in public transport demand forecasting.

Our modeling approach integrates these multidimensional factors with current demand estimates through two advanced techniques:

I. An ANFIS for handling complex, nonlinear relationships

II. A LSTM neural network for capturing temporal dependencies

The zonal clustering analysis, grounded in smart card transactions and annual OD matrices, revealed distinct temporal and spatial travel patterns. The temporal clustering, implemented via the K-means algorithm, effectively categorized traffic zones into morning, noon, and evening periods based on entry transaction patterns. For spatial clustering, the Mean Shift algorithm incorporated both travel demand (attraction) and key spatial variables -including zonal population characteristics and built-up area distributions across residential, commercial, and educational land uses.

Municipalities in Mashhad and other major cities can practically utilize the findings of this study in urban and transportation planning. The travel demand prediction model presented in this research enables the quantitative assessment of a new land-use permit's impact on travel demand during its operational phase, before the permit is even issued. Planners can forecast the required number of trips for a traffic analysis zone by altering spatial variables (e.g., the area of commercial or residential land use) for long-term horizons or by adjusting temporal alongside spatial variables for short-term horizons. This capability will lead to the design and development of improved land-use and transportation policies through the optimal distribution of land uses across different zones, a reduction in unnecessary trips, efficient allocation of public transportation fleet, and ultimately, increased citizen satisfaction with the transport system.

References

[1] United Nations Fund for Population Activities. (2021). *Urbanization*. https://www.unfpa.org/urbanization

[2] Nassi, C. D., & de Carvalho da Costa, F. C. (2012). Use of the analytic hierarchy process to evaluate transit fare system. *Research in transportation economics*, *36*(1), 50–62. https://doi.org/10.1016/j.retrec.2012.03.009

[3] Wang, Y., de Almeida Correia, G. H., de Romph, E., & Timmermans, H. J. P. (2017). Using metro smart card data to model location choice of after-work activities: An application to Shanghai. *Journal of transport geography*, *63*, 40–47. https://doi.org/10.1016/j.jtrangeo.2017.06.010

[4] Anda, C., Erath, A., & Fourie, P. J. (2017). Transport modelling in the age of big data. *International journal of urban sciences*, *21*(sup1), 19–42. https://doi.org/10.1080/12265934.2017.1281150

[5] Zhao, Z., Koutsopoulos, H. N., & Zhao, J. (2018). Individual mobility prediction using transit smart card data. *Transportation research part c: Emerging technologies*, *89*, 19–34. https://doi.org/10.1016/j.trc.2018.01.022

[6] El Mahrsi, M. K., Come, E., Baro, J., & Oukhellou, L. (2014). Understanding passenger patterns in public transit through smart card and socioeconomic data: A case study in rennes, france. *ACM sigkdd workshop on urban computing*. France: Association for Computing Machinery (ACM). https://hal.science/hal-01053794

[7] Trépanier, M., & Yamamoto, T. (2015). Workshop synthesis: System based passive data streams systems; Smart Cards, Phone Data, GPS. *Transportation research procedia*, *11*, 340–349. https://doi.org/10.1016/j.trpro.2015.12.029

[8] Ma, X., Liu, C., Wen, H., Wang, Y., & Wu, Y. J. (2017). Understanding commuting patterns using transit smart card data. *Journal of transport geography*, *58*, 135–145. https://doi.org/10.1016/j.jtrangeo.2016.12.001

[9] Sportiello, L. (2019). Internet of smart cards: A pocket attacks scenario. *International journal of critical infrastructure protection*, *26*, 100302. https://doi.org/10.1016/j.ijcip.2019.05.005

[10] Owais, M. (2025). How to incorporate machine learning and microsimulation tools in travel demand forecasting in multi-modal networks. *Expert systems with applications*, *262*, 125563. https://doi.org/10.1016/j.eswa.2024.125563

[11] Cools, M., Moons, E., & Wets, G. (2007). Investigating effect of holidays on daily traffic counts: Time series approach. *Transportation research record*, *2019*(1), 22–31. https://doi.org/10.3141/2019-04

[12] Sarkar, P. P., & Mallikarjuna, C. (2013). Effect of land use on travel behaviour: A case study of Agartala City. *Procedia - social and behavioral sciences*, *104*, 533–542. https://doi.org/10.1016/j.sbspro.2013.11.147

[13] Hu, N., Legara, E. F., Lee, K. K., Hung, G. G., & Monterola, C. (2016). Impacts of land use and amenities on public transport use, urban planning and design. *Land use policy*, *57*, 356–367. https://doi.org/10.1016/j.landusepol.2016.06.004

[14] Kim, M. K., Kim, S., & Sohn, H. G. (2018). Relationship between spatio-temporal travel patterns derived from smart-card data and local environmental characteristics of Seoul, Korea. *Sustainability*, *10*(3), 1–18. https://doi.org/10.3390/su10030787

[15] Cai, Z., Li, T., Su, X., Guo, L., & Ding, Z. (2020). Research on analysis method of characteristics generation of urban rail transit. *IEEE transactions on intelligent transportation systems*, *21*(9), 3608–3620. https://doi.org/10.1109/TITS.2019.2929619

[16] Liyanage, S., Dia, H., Abduljabbar, R., & Bagloee, S. A. (2019). Flexible mobility on-demand: An environmental scan. *Sustainability*, *11*(5), 1–39. https://doi.org/10.3390/su11051262

[17] Liyanage, S., & Dia, H. (2020). An agent-based simulation approach for evaluating the performance of on-demand bus services. *Sustainability*, *12*(10), 1–20. https://doi.org/10.3390/su12104117

[18] Tang, J., Wang, X., Zong, F., & Hu, Z. (2020). Uncovering spatio-temporal travel patterns using a tensor-based model from metro smart card data in Shenzhen, China. *Sustainability*, *12*(4), 1–16. https://doi.org/10.3390/su12041475

[19] Lin, P., Weng, J., Alivanistos, D., Ma, S., & Yin, B. (2020). Identifying and segmenting commuting behavior patterns based on smart card data and travel survey data. *Sustainability*, *12*(12), 1–18. https://doi.org/10.3390/su12125010

[20] Wu, W., Xia, Y., & Jin, W. (2021). Predicting bus passenger flow and prioritizing influential factors using multi-source data: Scaled stacking gradient boosting decision trees. *IEEE transactions on intelligent transportation systems*, *22*(4), 2510–2523. https://doi.org/10.1109/TITS.2020.3035647

[21] Liyanage, S., Abduljabbar, R., Dia, H., & Tsai, P. W. (2022). AI-based neural network models for bus passenger demand forecasting using smart card data. *Journal of urban management*, *11*(3), 365–380. https://doi.org/10.1016/j.jum.2022.05.002

[22] Mitra, A., Jain, A., Kishore, A., & Kumar, P. (2022). A comparative study of demand forecasting models for a multi-channel retail company: A novel hybrid machine learning approach. *Operations research forum*, *3*(4), 58. https://doi.org/10.1007/s43069-022-00166-4

[23] Demographia. (2023). *Demographia world urban areas: 19th annual edition*. http://www.demographia.com/db-worldua.pdf

[24] Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, *20*(8), 832–844. https://doi.org/10.1109/34.709601

[25] Ho, T. K. (1995). Random decision forests. *Proceedings of 3rd international conference on document analysis and recognition* (pp. 278–282). IEEE. https://doi.org/10.1109/ICDAR.1995.598994

[26] Zhou, Y., Thill, J. C., Xu, Y., & Fang, Z. (2021). Variability in individual home-work activity patterns. *Journal of transport geography*, *90*, 102901. https://doi.org/10.1016/j.jtrangeo.2020.102901

[27] Radfar, S., Koosha, H., Gholami, A., & Amindoust, A. (2025). Improved public transport OD matrix estimation using an enhanced trip chain model with smart card data. *International journal of intelligent transportation systems research*, *23*(3), 1341–1356. https://doi.org/10.1007/s13177-025-00513-9

[28] Zhang, X. P. (1997). Adaptive-network-based fuzzy inference system for short term load forecasting. *IFAC proceedings volumes*, *30*(17), 533–539. https://doi.org/10.1016/S1474-6670(17)46460-3

[29] Chopra, S., Dhiman, G., Sharma, A., Shabaz, M., Shukla, P., & Arora, M. (2021). Taxonomy of adaptive neuro-fuzzy inference system in modern engineering sciences. *Computational intelligence and neuroscience*, *2021*(1), 6455592. https://doi.org/10.1155/2021/6455592

[30] Jang, J. S. R. (1993). ANFIS: Adaptive-network-based fuzzy inference system. *IEEE transactions on systems, man, and cybernetics*, *23*(3), 665–685. https://doi.org/10.1109/21.256541

[31] Öztayşi, B., & Bolturk, E. (2014). Fuzzy methods for demand forecasting in supply chain management. In *Supply chain management under fuzziness: Recent developments and techniques* (pp. 243–268). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-53939-8_11

[32] Razavi, S. A., Najafabadi, T. A., & Mahmoodian, A. (2018). Remaining useful life estimation using anfis algorithm: A data-driven approcah for prognostics. *2018 prognostics and system health management conference (PHM-chongqing)* (pp. 522–526). IEEE. https://doi.org/10.1109/PHM-Chongqing.2018.00095

[33] Hassanniakalager, A., Sermpinis, G., Stasinakis, C., & Verousis, T. (2020). A conditional fuzzy inference approach in forecasting. *European journal of operational research*, *283*(1), 196–216. https://doi.org/10.1016/j.ejor.2019.11.006

[34] Chanal, D., Yousfi Steiner, N., Petrone, R., Chamagne, D., & Péra, M. C. (2022). Online diagnosis of PEM fuel cell by fuzzy C-means clustering. In *Encyclopedia of energy storage* (pp. 359–393). Oxford: Elsevier. https://doi.org/10.1016/B978-0-12-819723-3.00099-8

[35] Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2016). LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems*, *28*(10), 2222–2232. https://doi.org/10.1109/TNNLS.2016.2582924

[36] Aggarwal, C. C. (2018). *Neural networks and deep learning*. Springer. https://doi.org/10.1007/978-3-319-94463-0

[37] Abbasimehr, H., Shabani, M., & Yousefi, M. (2020). An optimized model using LSTM network for demand forecasting. *Computers & industrial engineering*, *143*, 106435. https://doi.org/10.1016/j.cie.2020.106435

[38] Willmott, C. J. (1981). On the validation of models. *Physical geography*, *2*(2), 184–194. https://doi.org/10.1080/02723646.1981.10642213

[39] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. Springer. https://doi.org/10.1007/978-1-4614-7138-7

[40] Sandeep, S., Abinash, S., Siddhartha, P., & K., G. D. (2022). Prediction of bed-load sediment using newly developed support-vector machine techniques. *Journal of irrigation and drainage engineering*, *148*(10), 4022034. https://doi.org/10.1061/(ASCE)IR.1943-4774.0001689

[41] Sahoo, A., Parida, S. S., Samantaray, S., & Satapathy, D. P. (2024). Daily flow discharge prediction using integrated methodology based on LSTM models: Case study in Brahmani-Baitarani basin. *HydroResearch*, *7*, 272–284. https://doi.org/10.1016/j.hydres.2024.04.006

[42] Samantaray, S., Sahoo, A., & Baliarsingh, F. (2024). Groundwater level prediction using an improved SVR model integrated with hybrid particle swarm optimization and firefly algorithm. *Cleaner water*, *1*, 100003. https://doi.org/10.1016/j.clwat.2024.100003

[43] McQueen, J. B. (1967). Some methods of classification and analysis of multivariate observations. *Proceedings of the fifth berkeley symposium on mathematical statistics and probability* (pp. 281–297). University of California press. https://cir.nii.ac.jp/crid/1571135649659368064

[44] Cao, F., Liang, J., & Jiang, G. (2009). An initialization method for the K-Means algorithm using neighborhood model. *Computers & mathematics with applications*, *58*(3), 474–483. https://doi.org/10.1016/j.camwa.2009.04.017

[45] Eltibi, M. F., & Ashour, W. M. (2011). Initializing KMeans clustering algorithm using statistical information. *International journal of computer applications*, *29*(7), 51–55. https://doi.org/10.5120/3573-4930

[46] Jin, X., & Han, J. (2017). Mean shift. In *Encyclopedia of machine learning and data mining* (pp. 806–808). Boston, MA: Springer US. https://doi.org/10.1007/978-1-4899-7687-1_532

[47] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

[48] Shahapure, K. R., & Nicholas, C. (2020). Cluster quality analysis using silhouette score. *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)* (pp. 747–748). IEEE. https://doi.org/10.1109/DSAA49011.2020.00096

[49] Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, *1*(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

[50] Ros, F., Riad, R., & Guillaume, S. (2023). PDBI: A partitioning Davies-Bouldin index for clustering evaluation. *Neurocomputing*, *528*, 178–199. https://doi.org/10.1016/j.neucom.2023.01.043

[51] Cheng, Z., Trépanier, M., & Sun, L. (2021). Probabilistic model for destination inference and travel pattern mining from smart card data. *Transportation*, *48*(4), 2035–2053. https://doi.org/10.1007/s11116-020-10120-0

[52] Alsger, A., Assemi, B., Mesbah, M., & Ferreira, L. (2016). Validating and improving public transport origin–destination estimation algorithm using smart card fare data. *Transportation research part c: Emerging technologies*, *68*, 490–506. https://doi.org/10.1016/j.trc.2016.05.004

[53] He, L., & Trépanier, M. (2015). Estimating the destination of unlinked trips in transit smart card fare data. *Transportation research record*, *2535*(1), 97–104. https://doi.org/10.3141/2535-11

[54] Zhang, Y., & Xu, D. (2022). The bus is arriving: Population growth and public transportation ridership in rural America. *Journal of rural studies*, *95*, 467–474. https://doi.org/10.1016/j.jrurstud.2022.09.018

[55] Guang, Z., Yang, J., & Li, J. (2018). Forecast of short-term passenger flow of urban railway stations based on seasonal arima model. *Proceedings of the 3rd international conference on electrical and information technologies for rail transportation (EITRT) 2017* (pp. 759–767). Singapore: Springer Singapore. https://doi.org/10.1007/978-981-10-7989-4_77%0A%0A

[56] Wang, B., & Zheng, S. (2020). Air pollution lowers travel demand in a consumer city. *Transportation research part d: Transport and environment*, *89*, 102616. https://doi.org/10.1016/j.trd.2020.102616

[57] Zhang, J., Chen, F., Cui, Z., Guo, Y., & Zhu, Y. (2021). Deep learning architecture for short-term passenger flow forecasting in urban rail transit. *IEEE transactions on intelligent transportation systems*, *22*(11), 7004–7014. https://doi.org/10.1109/TITS.2020.3000761

[58] Cools, M., Moons, E., & Wets, G. (2010). Assessing the impact of public holidays on travel time expenditure: Differentiation by trip motive. *Transportation research record*, *2157*(1), 29–37. https://doi.org/10.3141/2157-04

[59] Wang, B., Shao, C., & Ji, X. (2017). Influencing mechanism analysis of holiday activity–travel patterns on transportation energy consumption and emissions in China. *Energies*, *10*(7), 1–20. https://doi.org/10.3390/en10070897

[60] Jin, W., Li, P., Wu, W., & Wei, L. (2019). Short-term public transportation passenger flow forecasting method based on multi-source data and shepard interpolating prediction method. *Man-machine-environment system engineering* (pp. 281–294). Singapore: Springer Singapore. https://doi.org/10.1007/978-981-13-2481-9_33%0A%0A

[61] Danfeng, Y., & Jing, W. (2019). Subway passenger flow forecasting with multi-station and external factors. *IEEE access*, *7*, 57415–57423. https://doi.org/10.1109/ACCESS.2019.2914239

[62] Yao, E., Hong, J., Pan, L., Li, B., Yang, Y., & Guo, D. (2021). Forecasting passenger flow distribution on holidays for urban rail transit based on destination choice behavior Analysis. *Journal of advanced transportation*, *2021*(1), 9922660. https://doi.org/10.1155/2021/9922660

[63] Zhou, Y., Qian, C., Xiao, H., Xin, J., Wei, Z., & Feng, Q. (2019). Coupling research on land use and travel behaviors along the tram based on accessibility measurement—Taking nanjing chilin tram line 1 as an example. *Sustainability*, *11*(7), 1–33. https://doi.org/10.3390/su11072034

[64] Liu, X., Wu, J., Huang, J., Zhang, J., Chen, B. Y., & Chen, A. (2021). Spatial-interaction network analysis of built environmental influence on daily public transport demand. *Journal of transport geography*, *92*, 102991. https://doi.org/10.1016/j.jtrangeo.2021.102991

[65] Cheng, J., Liu, G., Gao, S., Raza, A., Li, J., & Juan, W. (2024). Short-term passenger flow prediction in urban rail transit based on points of interest. *IEEE access*, *12*, 95196–95208. https://ieeexplore.ieee.org/document/10591968

[66] Gao, Q. L., Li, Q. Q., Yue, Y., Zhuang, Y., Chen, Z. P., & Kong, H. (2018). Exploring changes in the spatial distribution of the low-to-moderate income group using transit smart card data. *Computers, environment and urban systems*, *72*, 68–77. https://doi.org/10.1016/j.compenvurbsys.2018.02.006

[67] Bautista-Hernández, D. (2020). Urban structure and its influence on trip chaining complexity in the Mexico City Metropolitan Area. *Urban, planning and transport research*, *8*(1), 71–97. https://doi.org/10.1080/21650020.2019.1708784

[68] Briand, A. S., Côme, E., El Mahrsi, M. K., & Oukhellou, L. (2016). A mixture model clustering approach for temporal passenger pattern characterization in public transport. *International journal of data science and analytics*, *1*(1), 37–50. https://doi.org/10.1007/s41060-015-0002-x

[69] Qi, G., Huang, A., Guan, W., & Fan, L. (2019). Analysis and prediction of regional mobility patterns of bus travellers using smart card data and points of interest data. *IEEE transactions on intelligent transportation systems*, *20*(4), 1197–1214. https://doi.org/10.1109/TITS.2018.2840122

[70] Hofmann, M., & O'Mahony, M. (2005). The impact of adverse weather conditions on urban bus performance measures. *Proceedings. 2005 IEEE intelligent transportation systems, 2005.* (pp. 84–89). IEEE. https://doi.org/10.1109/ITSC.2005.1520087

[71] Zhou, M., Wang, D., Li, Q., Yue, Y., Tu, W., & Cao, R. (2017). Impacts of weather on public transport ridership: Results from mining data from different sources. *Transportation research part c: Emerging technologies*, *75*, 17–29. https://doi.org/10.1016/j.trc.2016.12.001

[72] Gössling, S., Neger, C., Steiger, R., & Bell, R. (2023). Weather, climate change, and transport: A review. *Natural hazards*, *118*(2), 1341–1360. https://doi.org/10.1007/s11069-023-06054-2