International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 9 Issue 5, Sep-Oct 2025 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Prediction of Crowding Levels at Different Time Periods for Beijing Subway Line 6 Based on a Combined Model of Autoregression and Linear Regression with Exogenous Variables

Qishi Feng, Yuxiang Chen, Xiang Li, Sifan Zhang, Zhe Tan

School of Systems Science and Statistics, Beijing WUZI University, Beijing, China

ABSTRACT

As a critical traffic artery connecting the city center with the Tongzhou New Town (subsidiary administrative center of Beijing), Beijing Subway Line 6 bears significant passenger flow pressure. Issues such as carriage overcrowding and low passenger comfort are prominent, especially in the section between “Jintai Road” and “Shilipu” stations. To address this issue, this paper employs a hybrid modeling approach combining an Autoregressive (AR) model with an exogenous variable of daily total passenger flow. By integrating statistical data on the real-time crowding levels of Beijing Subway and the total passenger flow data from the preceding week, this study aims to achieve accurate predictions of carriage crowding levels for different time periods and directions. The research findings are expected to optimize passenger travel experience and provide theoretical references and practical insights for the intelligent operation of urban rail transit.

KEYWORDS: *Autoregressive (AR) model; Passenger flow prediction; Beijing Subway.*

How to cite this paper: Qishi Feng | Yuxiang Chen | Xiang Li | Sifan Zhang | Zhe Tan "Prediction of Crowding Levels at Different Time Periods for Beijing Subway Line 6 Based on a Combined Model of Autoregression and Linear Regression with Exogenous Variables" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-9 | Issue-5, October 2025, pp.1011-1016, URL: www.ijtsrd.com/papers/ijtsrd97646.pdf

IJTSRD97646

1. INTRODUCTION

Beijing Subway Line 6, which commenced operation on December 30, 2012, is the second east-west rail line traversing Beijing. With an operational length of 53 kilometers, its eastern section (from Chaoyangmen Station eastward to Lucheng Station) facilitates transportation for residents along the Chaoyang North Road residential areas and Tongzhou commuting to the city center, as well as for city residents traveling to Beijing's subsidiary administrative center. Characterized by relatively long station intervals and high speeds, it has become the preferred choice for many commuters. Passenger travel is highly concentrated during the morning and evening peaks. The eastern section (Chaoyangmen - Lucheng) exhibits a distinct "sparse east, dense west" interchange layout: there are no interchange stations along approximately 20 kilometers east of Jintai Road, leading to a concentrated influx of passengers from residential areas in Tongzhou; whereas, within the short stretch from Jintai Road to Chaoyangmen, three interchange stations (Chaoyangmen, Hujialou, Jintai Road) are densely clustered. This configuration creates a typical commuter pattern for the eastern section: during the morning peak, passengers board primarily from the subsidiary center and alight in the city proper, reversing during the evening peak. In the morning peak, trains departing from Lucheng experience a rapid increase in passenger load after passing through residential areas like Caofang, reaching saturation levels even before arriving at Jintai Road. Consequently, high passenger volumes during peak hours result in significant crowding at stations and inside carriages, leading to a poor passenger experience. In 2024, according to data released by the Beijing Rail Transit Control Center, the average daily passenger volume on Line 6

reached 857,800, ranking fourth among all Beijing rail transit lines. Although the line was constructed with forward-lookingly 8-car Type B trainsets to handle large passenger flows, it still struggles to cope with such high demand. Regarding morning peak (7:00-9:00) entry volumes on weekdays, three stations on the eastern section-Wuzi University road, Caofang, and Tongzhou Beiguan-ranked among the top 20 in Beijing for passenger volume. The morning peak passenger flow at these three stations accounted for 48.69%, 49.13%, and 45.31% of their respective daily totals, underscoring the extreme concentration of ridership on Line 6 during peak hours. Furthermore, the typical commuter pattern results in substantial passenger volumes at the three interchange stations in the city proper. According to 2024 statistics on morning peak transfer volumes at Beijing urban rail transit stations on weekdays, Hujialou and Jintai Road on Line 6's eastern section ranked second and ninth, with average daily transfer volumes of 42,300 and 28,500 person time, respectively. Such immense passenger flow exacerbates crowding levels in carriages and on platforms, impairing passenger comfort and increasing the risk of incidents like stampedes. Therefore, short-term passenger flow prediction models are crucial. For operators, these models can forecast periods of high demand in advance, enabling proactive deployment of staff for crowd control measures or even adding extra services to meet demand, thereby effectively mitigating risks associated with large passenger flows. For passengers, these models can facilitate off-peak travel choices, allowing them to avoid crowds and enjoy a more comfortable journey. In recent years, the transportation sector has made significant progress in open data sharing. In 2016, the Ministry of Transport issued the "Implementation Opinions on Promoting Open Sharing of Data Resources in the Transportation Industry." Years of development have gradually established a coordinated and efficient data resource management system, laying a solid foundation for in-depth mining and application of transportation data. Concurrently, new productive forces represented by artificial intelligence are integrating into the transportation industry at an unprecedented pace. Sichuan Province, for instance, has achieved initial success in applying AI within transportation, such as using drones and deep learning models for disaster damage identification on mountain highways and developing "inspection robots" to enhance bridge and tunnel maintenance efficiency. In intelligent transportation systems, advanced technologies including information technology, data communication transmission, electronic sensing, control, and computer technology are effectively integrated. This equips transportation systems with powerful capabilities for perception, interconnection, analysis, prediction, and control, offering novel ideas and methods for addressing traffic challenges. Against this backdrop, our research team employs a combined model integrating Autoregressive (AR) and linear regression with exogenous variables. By analyzing crowding level data for different directions and time periods, the model aims to provide passengers with information to choose off-peat travel times for a more comfortable experience. Simultaneously, it can assist operators in proactively deploying resources for crowd management and scheduling additional services as needed, effectively preventing risks associated with large passenger flows. This work aligns with the trend of large-scale intelligent transportation development and seeks to contribute innovative solutions to urban traffic congestion management.

2. Research on Passenger Flow Data Correlation

Analysis of the daily data and same-period crowding level data for Beijing Subway Line 6 over the past month reveals a strong correlation among these datasets.

2.1. Total Daily Passenger Flow Data for Beijing Subway Line 6 over Four Weeks

Figure 2: Total Daily Passenger Flow of Beijing Subway Line 6 and Corresponding 19:30 Crowding Level in the Jintai Road–Shilipu Section (August 11 – August 17, 2025)

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Total Passenger	92.05	92.23	95.63	94.85	96.56	68.58	56.37
Crowding Level	Yellow (Mildly Crowded)	Yellow (Mildly Crowded)	Yellow (Mildly Crowded)	Yellow (Mildly Crowded)	Red (Crowded)	Green (Comfortable)	Green (Comfortable)

Note: 1.Crowding level data is sourced from the Beijing Subway mini-program.
2. Total Passenger Flow is “in 10,000 persons”

Following the organization of the daily passenger flow data for four consecutive weeks from July 21 to August 17, 2025 (see figure above), preliminary observations indicate that the passenger volume on weekdays (Monday to Friday) is significantly higher than on weekends. Furthermore, the passenger flow on corresponding weekdays demonstrates high stability across different weeks. For instance, the Monday passenger volume consistently remains within the range of 872.6 thousand to 920.5 thousand, while the weekend passenger flow is generally below 700 thousand and exhibits relatively greater volatility. We employed the Pearson correlation coefficient to calculate the correlations among the data from these four weeks.

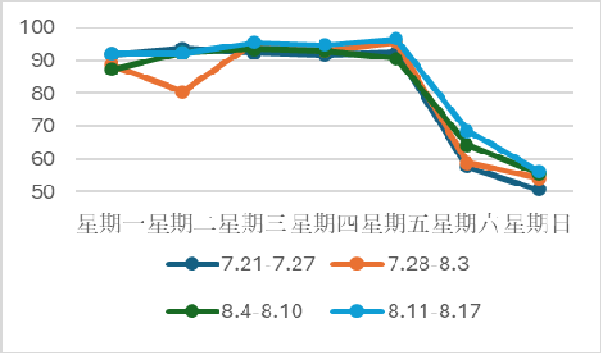

Figure 1: Total Passenger Flow of Beijing Subway Line 6 (July 21 – August 17)

Pearson Correlation Coefficient:

Given time series $X - x_1, x_2, \ldots, x_n$ and $Y - y_1, y_2, \ldots, y_n$, the calculation formula is:

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where $\bar{x}$ and $\bar{y}$ are the means of the respective series. The correlation coefficient $r_{XY} \in [-1,1]$, with positive values indicating a positive correlation and negative values indicating a negative correlation.The analysis results show a significant positive correlation among the passenger flow data (r = 0.905). This result indicates that the variation trends of passenger volume for the same weekday across different weeks (e.g., among all Mondays) are highly consistent, demonstrating a strong periodic pattern. This statistical conclusion aligns well with the descriptive characteristics presented in the figure above, confirming that the daily passenger flow of Beijing Subway Line 6, influenced by rigid travel demands such as commuting, possesses a high degree of predictability.

2.2. Analysis of Intra-week Correlation Between Daily Passenger Flow and Time-Specific Crowding Levels on Beijing Subway Line 6

Following the analysis of crowding levels across different sections and time periods of Beijing Subway Line 6, we observed commonalities in the fluctuation patterns of passenger flow between weekdays and weekends in various sections. However, the passenger flow change in the Jintai Road–Shilipu section at 19:30 was the most pronounced, characterized by a particularly stark contrast between the peak features on weekdays and the trough on weekends. Given its high representativeness of overall passenger flow patterns and the significant variability that provides an ideal sample for in-depth analysis, our research group has selected this specific section and time as a key case study for detailed focus.

Figure 3: Beijing Subway Official Real-time Crowding Level Data Standard

Icon	Crowding Level	Density (persons/m²)
Green	Comfortable	0-2
Light yellow	Relatively Comfortable	2-3
Yellow	Relatively Crowded	3-4
Red	Crowded	>4

Following the organization of daily passenger flow data and the 19:30 crowding data from July 21 to August 11, 2025 (see Figure 2, Figure 3), we quantified the latter according to the standards of Beijing Subway's official real-time crowding data. Guided by the statistical principle of using the midpoint value to represent all data within an interval, we processed the real-time crowding data (e.g., the green zone in the chart, indicating a "Comfortable" crowding level, was quantified as 1.5 persons/m²). The Pearson correlation coefficient was then applied to calculate the correlation between the daily passenger flow and the daily 19:30 crowding level in the Jintai Road–Shilipu section. The analysis results indicate a significant positive

correlation between the two variables (r = 0.925).In time series analysis, while the Pearson correlation coefficient can reveal superficial linear relationships between variables, the comparison of "Total daily passenger flow vs. Daily 19:30 crowding level in Jintai Road–Shilipu section" lacks an intuitive visual diagnosis of the inherent temporal structure within the data, unlike the analysis of "Total passenger flow of Beijing Subway Line 6 from July 21 to August 17." Therefore, to complement the Pearson correlation, our research group decided to employ an autocorrelation model. Visualization tools such as the Autocorrelation Function (ACF) plot can intuitively identify temporal dependency features within the series itself, such as trends and periodicity. This approach helps avoid misinterpreting spurious correlations arising from simultaneous fluctuations as intrinsic relationships, thereby ensuring the reliability of statistical conclusions and the rigor of model construction.

Autocorrelation Coefficient:

The autocorrelation coefficient at lag is given by:

$$r_k = \frac{\sum_{t=k+1}^{n}(x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2}$$

The 95% confidence interval is given by $\pm \frac{1.96}{\sqrt{n}}$. Values falling outside this interval indicate the presence of significant temporal dependency.

Following calculations, the obtained p-value for the autocorrelation coefficient is 0.0016, indicating the absence of autocorrelation. This result demonstrates a significant positive correlation between the daily total passenger flow (the exogenous variable) and the crowding levels at different times in the daily local section. Based on the above research, we can infer that the daily total passenger volume has an influence on the crowding degree at different time periods.

3. Prediction Model Construction Based on the Autoregressive Exogenous (ARX) Framework

The prediction model constructed in this study is based on the Autoregressive Exogenous (ARX) framework, aiming to enhance forecasting accuracy by integrating the autocorrelation characteristics of time series with external influencing factors. The core hypothesis of the model posits that passenger flow at a specific time period is jointly influenced by two factors: firstly, the temporal inertia effect of historical passenger flow from the same period, meaning that the passenger flow at the same time on the previous workday exerts a persistent influence on the current moment; secondly, the synergistic effect of the overall daily operational scale, where the total daily passenger flow, serving as an exogenous variable, reflects the radiating influence of the overall passenger flow trend on specific time periods.

First, it is necessary to construct an expectation for the future total daily passenger flow of Beijing Subway Line 6 based on autoregression. Leveraging the correlation existing in the total daily passenger flow for corresponding days of the week over the past four weeks (e.g., the Fridays across four weeks), an autoregressive model is used to predict the total daily passenger flow for the upcoming week.

Subsequently, the predicted total daily passenger flow is utilized as an exogenous variable, employing its global perspective to calibrate the prediction results. The crowding level data from the same time period over the past week are used as autoregressive coefficients, representing the persistent influence of the crowding level at the same time each day of the previous week on the current moment. Using this ARX framework, the autoregressive prediction model forecasts future time-period crowding level data. The autoregressive coefficients reflect the continuation strength of historical passenger flow, while the exogenous variable coefficient embodies the degree of diffusion of global passenger flow to local time periods. This design not only preserves the ability of time series models to capture periodic patterns but also incorporates real-time operational status information through exogenous variables, ultimately forming a prediction mechanism with dynamic adaptability.

3.1. Autoregressive Prediction Model for Total Daily Passenger Flow (Exogenous Variable)

This model employs linear regression to implement autoregressive time series prediction. The core concept involves using lagged values of the time series as feature variables to construct a linear prediction model. Given time series dataX_1,X_2,X_3,X_4, lagged features are first generated through feature engineering:

$$lag_1 = X_{t-1}, \quad \mathbf{lag_2} = X_{t-2}, \quad \mathbf{lag_3} = X_{t-3}$$

A linear regression model is established as follows:

$$X_t = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} \cdot \mathbf{lag_1} + \boldsymbol{\beta_2} \cdot \mathbf{lag_2} + \boldsymbol{\beta_3} \cdot \mathbf{lag_3} + \boldsymbol{\epsilon}$$

Where β_0 is the intercept term, $\beta_1, \beta_2, \beta_3$ are the regression coefficients, and ϵ is the error term. During the model training phase, parameters are estimated using the least squares method to minimize the sum of squared prediction errors:

$$min_{\beta} \sum_{i=1}^{n} (X_i - \hat{X}_i)^2$$

The prediction phase performs single-step forecasting based on the most recent p observations:

$$hat{X_5} = \widehat{\beta_0} + \widehat{\beta_1} X_4 + \widehat{\beta_2} X_3 + \widehat{\beta_3} X_2$$

This method implements the functionality of a traditional autoregressive model within a linear regression framework. It transforms the temporal dependencies of the time series into linear relationships within the feature space, reducing implementation complexity while maintaining prediction accuracy.

Figure 4: Total Passenger Flow of Beijing Subway Line 6 (July 21 – August 17, Unit: 10000 persons)

	MON	TUE	WED	THU	FRI	SAT	SUN
7.21-7.27	91.87	93.7	92.5	91.68	92.5	57.53	50.8
7.28-8.3	88.76	80.6	95.39	93.06	95.61	58.93	54.19
8.4-8.10	87.26	92.3	93.56	93	90.65	64.24	55.57
8.11-8.17	92.05	92.2	95.63	94.85	96.56	68.58	56.37

Figure 5: Total Passenger Flow of Beijing Subway Line 6 (August 18 – August 24, Unit: 10000 persons) (Predicted)

	MON	TUE	WED	THU	FRI	SAT	SUN
8.18-8.24	89.36	88.41	94.86	93.74	94.27	63.92	55.38

3.2. Autoregressive Prediction of Time-Period Passenger Flow

This model analyzes the provided crowding level data at 19:30 from the past week and the corresponding daily total passenger flow data to train a linear regression model. Concurrently, it incorporates the predicted future total daily passenger flow data from the Autoregressive Prediction Model for Total Daily Passenger Flow (from Section 3.1) as an exogenous variable, thereby constructing an Autoregressive Exogenous (ARX) model.

$$y_t = \beta_0 + \beta_1 \cdot y_{t-1} + \beta_2 \cdot \text{Total}_t + \epsilon$$

where:
y_{t-1}: Passenger flow at 19:30 on the previous day (lagged feature)
Total_t: Total passenger flow on day (exogenous variable)
$\beta_0, \beta_1, \beta_2$: Coefficients learned by the model from the data
ϵ: Error term

After providing the daily total passenger flow and the daily 19:30 crowding level data for Beijing Subway Line 6 from July 21 to August 17, the following results were obtained in **Figure6.**

Figure 6: Predicted Data for Total Passenger Flow and 19:30 Crowding Level in the Jintai Road–Shilipu Section of Beijing Subway Line 6 on August 18

Total Passenger Flow	**89.36(10^4 persons)**
Crowding Level	4.1 (persons/m²) - Red

4. Prediction Results Validation

We combined the prediction results for August 18th with the statistical data from August 11th to August 17th, consolidating them into the dataset presented in **Figure 7.**

Figure 7: Total Passenger Flow of Beijing Subway Line 6 and Daily 19:30 Crowding Level in the Jintai Road–Shilipu Section (August 11 – August 17 and August 18 (Predicted), 2025)

	MON	TUE	WED	THU	FRI	SAT	SUN	8.18
Total Passenger Flow (10^4 persons)	92.05	92.23	95.63	94.85	96.56	68.58	56.37	89.36
Persons per Square Meter (persons/m²)	3.5	3.5	3.5	3.5	5	1	1	4.1

The dataset consolidated in **Figure 7** was analyzed for correlation using the aforementioned model that combines the Pearson correlation coefficient and the autocorrelation coefficient. The results showed a Pearson correlation coefficient of r = 0.912 and an autocorrelation p-value = 0.0016, indicating a statistically significant correlation. This significant correlation, derived from a comparative analysis of the predicted data and the actual observed data, validates the consistency between them. It provides strong evidence for the accuracy of the prediction model, demonstrating that the model can effectively capture the dynamic characteristics of the actual data. Subsequently, we obtained the actual total passenger flow for August 18 and the actual 19:30 carriage crowding level in the Jintai Road–Shilipu section from Beijing Subway records, consolidating them into Figure 8. A comparison reveals that the actual data closely aligns with the predicted data.

Figure 8: Predicted vs. Actual Values of Total Passenger Flow and 19:30 Crowding Level in the Jintai Road–Shilipu Section for Beijing Subway Line 6 on August 18

	Predicted Value	Actual Value
Total Passenger Flow	89.36(10^4 persons)	89.8(10^4 persons)
Crowding Level	4.1 **(persons/m²) - Red**	3.5 **(persons/m²) - yellow**

5. Conclusion

Correlation analysis between the predicted and actual observed data indicates that the prediction model constructed in this study demonstrates a certain degree of effectiveness under limited data conditions. The Pearson correlation test reveals a statistically significant relationship between the predicted and actual sequences ($r = 0.912$, $p < 0.0016$), suggesting that the model can, to some extent, capture the fundamental patterns of passenger flow variation, providing a preliminary theoretical basis for short-term forecasting.

However, it is essential to objectively acknowledge several apparent limitations of the current model. Firstly, the relatively small sample size used in the study somewhat compromises the robustness of the statistical conclusions and the model's extrapolation capability. A small dataset may cause the model to be overly sensitive to random fluctuations, thereby affecting the stability of the predictions. Secondly, the analysis process revealed that the model fails to adequately account for the moderating effects of external environmental variables, particularly the significant impact of specific weather conditions on passenger flow. Meteorological factors such as rain, snow, and extreme temperatures often induce anomalous changes in passenger flow patterns, yet the current model lacks a corresponding meteorological adjustment mechanism.

Furthermore, prediction biases at specific time points expose another critical shortcoming: the model's insufficient adaptability to holidays and special events. Passenger flow patterns during long holidays like the Spring Festival and National Day differ fundamentally from those on regular weekdays. The model currently does not incorporate dummy variables to capture the unique patterns of these special periods. Additionally, the results indicate that the model's characterization of seasonal variations is inadequate, failing to effectively distinguish the periodic fluctuations in passenger flow attributable to different seasons (spring, summer, autumn, winter).In summary, while the current model shows statistical significance in basic correlation tests, proving its possesses a certain degree of predictive capability, enhancing its practical value and prediction accuracy requires future improvements in several areas: expanding the data sample size to increase statistical power; introducing weather factors as moderating variables; establishing a holiday dummy variable mechanism; and refining the seasonal adjustment component. Only through multi-dimensional and multi-factor collaborative modeling can a more accurate and robust passenger flow prediction system be constructed.

6. Fund project

This study was funded by the project "Undergraduate Scientific Research and Entrepreneurship Action Plan (2025)" (Project No. 202501040K021).

References

[1] G. Eason, B. Noble, and I.N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955. (*references*)

[2] Y. Dai, "Big data-based metro passenger flow prediction and scheduling optimization strategies," *People's Public Transportation*, no. 174, pp. 34-36, 2024.

[3] X. Pan, Z. Yao, and J. Wang, "Research on passenger flow prediction of urban rail transit based on time series," *Traffic and Safety*, no. 187, pp. 187-189, 2025.