Paper Type: Original Article

Analysis and Prediction of Railway Passenger Flow Patterns Based on the ARIMA Model

Bo Jiao

School of Mathematics, Yunnan Normal University, Kunming 650500, PR China

Abstract

This article uses time series data obtained by the Railway Bureau from January 1, 2015 to March 20, 2016. According to the passenger line passenger flow data from the ZD190 (station) to ZD111 (station) section of the railway company, the degree of passenger number is influenced by seasonal changes, holidays, climate and other factors, and python tools are used for data processing and visual description. This paper summarizes the change of passenger number by analyzing the properties such as train type, station traffic, passenger rate and station time period. The ARIMA season model is established to predict the future passenger flow according to the historical data, and the railway staff can be provided with reference data, so as to facilitate the railway departments to make corresponding structural adjustmentins in time and make full use of railway resources.

Keywords: Time Series; Python; Visualization; ARIMA Season Model; Passenger Flow Forecast.

1. Introduction

China has undergone a rapid transformation from "green-skinned trains" to maglev trains, with the national railway operating mileage reaching 162,000 kilometers by the end of 2024. Railway operations are vital to people's livelihoods, and the rapid expansion of the railway network reflects the success of China's reform and opening-up as a monumental policy.

However, with societal development, railway passenger traffic experienced a significant decline between 2019 and 2023 (as shown in Figure 1.1) and continues to decrease. This indicates that railway development planning now faces substantial challenges. Passenger volume is a key metric that reflects the transportation industry's service to the national economy and people's livelihoods, as well as an essential indicator for studying the scale and pace of transportation development. Therefore, research on railway passenger flow forecasting has become a critical focus for railway passenger services.

Figure 1.1 Statistical Chart of the total passenger and freight Transport volume of the whole society from 2018 to 2023

China, as the world's most populous country, handles an enormous volume of railway passenger traffic. At the same time, railway passenger flow is influenced by various factors. For instance, during holidays such as the Spring Festival travel rush ("Chunyun") and National Day, railway passenger traffic surges rapidly, straining operational capacity and leaving many passengers unable to secure tickets. This also creates significant challenges and pressure for railway staff.

Conversely, during off-peak travel seasons and on weekdays, some routes experience low seat occupancy rates, with certain trains operating at less than 10% capacity, leading to severe wastage of railway resources.

Therefore, analyzing and predicting railway passenger flow patterns can help in:

Setting reasonable ticket prices, improving station organization methods, optimizing the allocation of railway vehicle resources, and enhancing the service capacity of passenger transport facilities.

This research holds significant importance for improving the efficiency of railway passenger transport.

2. literature review

Since the launch of China's reform and opening-up policy, the nation's economy has achieved remarkable growth, with the railway system - as the most widely used transportation means - experiencing particularly rapid development. In recent years, the emergence of maglev train technology has elevated China's railway technology to the global forefront.

However, the development process inevitably encounters challenges. With increasing transportation alternatives and continuous improvement in people's living standards, conventional train services can no longer meet public demands. Our key priorities now lie in

optimizing railway organizational structures and enhancing service quality, while ensuring rational allocation of railway resources to promote balanced economic development across all regions of the country.

2.1 current situation of overseas research

Amidst information and economic globalization, populous countries have placed particular emphasis on railway reform and construction. The fundamental approach to railway industry reform in foreign countries involves separating government functions from enterprise operations, weakening planned regulations, supporting healthy competition, stimulating the business development of railway enterprises, and realizing the marketization and privatization of the railway industry.

Early on, foreign countries had already formulated certain plans for high-speed railway development, establishing a basic framework for overall high-speed train planning. By comprehensively considering operational costs and passenger travel time consumption, mathematical models and mathematical programming methods have been employed to optimize train operation plans. In recent years, optimization theories such as systems engineering, operations research, and intelligent algorithms have been widely applied to the optimization of train operation plans. The optimization typically focuses on operational sections, station service frequency, train frequency (number of train pairs), train stop patterns, train formation numbers, and train schedules. Among these, stop patterns are a key factor in evaluating transportation service quality. In practical optimization processes, train stop patterns are usually based on passenger flow predictions and require continuous adjustments to meet the dynamic demands of passengers.

As the world's second most populous country, the Indian Railway Management Committee released the "Indian Railways Vision 2020," outlining an unprecedented development blueprint for Indian railways. The plan includes constructing 25,000 kilometers of new railway lines by 2020, bringing the total length of Indian railways to over 89,000 kilometers. This includes building at least four high-speed railways with operating speeds of 250–350 km/h, covering all four major regions of the country. Additionally, at least eight high-speed transport corridors will be constructed to connect India's commercial, tourist, and pilgrimage centers, with six of these corridors currently undergoing technical feasibility studies. According to this plan, India will also build 50 world-class railway stations and over 200 fully functional large stations. All these projects are planned to be invested in and constructed through public-private partnerships (PPP).

By 2030, the French government in the West plans to achieve a freight turnover of 100 billion ton-kilometers, shift 2.5 billion passenger-kilometers from road to rail, and shift 2 billion passenger-kilometers from air to rail, thereby reducing greenhouse gas emissions by 2 million tons annually. In terms of passenger transport, French railways have developed diversified and multi-dimensional passenger products based on factors such as train speed class, operational scope, and service quality, perfectly meeting the needs of the French market. In recent years of

economic development, passenger railways, represented by France, have shown favorable growth, with both passenger turnover rates and market share increasing. In recent years, the operational characteristics of SNCF's passenger services have included vigorously developing TGV high-speed passenger transport, actively expanding the international passenger market, striving for branded operations, and continuously innovating service content through technological means.

2.2 status quo of domestic research

In recent years, with the introduction of policies such as China's "13th Five-Year Plan for Railway Development" and "Urban Rail Transit Operation Management Regulations," China's railway construction has entered a new development cycle. Currently, the national rapid railway network, with the "Eight Vertical and Eight Horizontal" high-speed rail lines as its backbone, has been basically completed. The railway network framework in central and western regions is being accelerated, and comprehensive passenger transport hubs are being simultaneously improved. By the end of 2018, China's railway operating mileage reached 131,700 kilometers (as shown in Figure 2.1), including 29,000 kilometers of high-speed railway lines, ranking first in the world.

Figure 2.1 The total operating mileage of China's railways from 2009 to 2018

The railway transportation system plays a crucial role in annual passenger and freight transportation due to its extensive network coverage and relatively low cost. According to the latest data released by the National Railway Administration, China's railways carried 3.375 billion passengers in 2018, representing a year-on-year increase of 9.4%. National railways transported 3.317 billion passengers, up 9.2% year-on-year, including 2.005 billion passengers carried by EMU (Electric Multiple Unit) trains, which saw a 16.8% increase compared to the previous year.

As China has entered the critical period of its "13th Five-Year" railway development plan, 2019 and 2020 are expected to witness peak periods for new railway line openings. According to the

working conference held by China Railway Corporation in January 2019, the plan includes putting 6,800 kilometers of new lines into operation in 2019, including 3,200 kilometers of high-speed rail, both figures showing significant growth compared to 2018. By 2020, China's passenger traffic is projected to reach 4 billion trips, with freight volume reaching 3.7 billion tons.

The railway development plan includes expanding commercial and travel services at stations and on trains to meet passengers' diverse and personalized needs. It will coordinate the planning of station commercial operations, upgrade service standards, enrich service offerings, and provide comprehensive, full-process services for travelers.

3. Theoretical models and data sources

3.1 ARIMA Differential integrated moving average autoregressive model

The ARIMA model is one of the fundamental and important models in the Box-Jenkins modeling approach. It consists of an Autoregressive Model (AR) and a Moving Average Model (MA). By employing corresponding mathematical models, it explains the autocorrelation among a set of random variables that changes over time. The model predicts future values based on past and present values of the time series under study, thereby representing the development process of the predicted object. This model has wide-ranging applications and is suitable for many fields including medicine and economics.

AR：autoregression, p is an autoregressive term;

MA：moving averages, q is the moving average term;

D: he number of differences performed when the time series is stable.

Principle: The model is established by converting a non-stationary time series into a stationary one, and then reviewing the dependent variable only for its lag value (p-order), as well as the present value and lag value of the random error term.

ARIMA(p, d, q)The common mathematical manifestations of the model are:

$$\left(1+\sum_{i-1}^{p}\emptyset_i L^i\right)(1-L)^d X_t=\left(1+\sum_{i-1}^{q}\theta_i L^i\right)\varepsilon_t \qquad 3\text{-}1$$

Among them, L is the lag operator（Lag operator）, $d \in \mathbb{z}, d>0$.

The dimensions corresponding to seasonality also have different differences, at the levels of quarters, months, weeks, etc. Suppose there is ∇ being the difference operator, that is:

$$\nabla^2 y_t=\nabla(y_t-y_{t-1})=y_t-2y_{t-1}+y_{t-2} \qquad 3\text{-}2$$

Suppose there is a delay operator B, that is:

$$y_{t-p}=B^p y_t,\ \forall p\geq 1 \qquad 3\text{-}3$$

The result that can be obtained is:

$$\nabla^k=\left(1-B\right)^k \qquad 3\text{-}4$$

Suppose there is order d、nonstationary time series y_t, $\nabla^t y_t$is a stationary time series, we can get:

$$\lambda(B)(\nabla^t y_t) = \theta(B)\varepsilon_t \quad 3\text{-}5$$

among the rest:

$$\lambda(B) = 1 - \lambda_1 B - \lambda_1 B^2 - \ldots - \lambda_p B^p \quad 3\text{-}6$$

$$\theta(B) = 1 - \theta_1 B - \theta_1 B^2 - \ldots - \theta_q B^q \quad 3\text{-}7$$

To sum up, the common mathematical manifestations of the ARIMA(p,d,q) model can be obtained.When the difference order d is 0, the ARIMA model is equivalent to the ARMA model. That is, based on whether the sequence is statinary or not, the former is a non-stationary time series, while the latter is a stationary time series.

3.2 Research methods and data sources

This study applies the ARIMA model for railway passenger flow forecasting. Through appropriate data processing methods, the collected data is integrated, with the p and q parameters determined according to the AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) principles. Using time-series data obtained from railway authorities covering the period from January 1, 2015 to March 20, 2016, the research examines passenger flow data on the dedicated passenger line section between ZD190 (station) and ZD111 (station), analyzing how seasonal variations, holidays, weather conditions and other factors affect passenger numbers. Python and MySQL tools were employed for data processing and visualization.

Five database tables were constructed:

Regional weather table (meteorology_table)
Passenger flow information table (passenger_flow_table)
Station mileage table (station_info_table)
Train information table (train_info_table)
Daily seat occupancy rate table (train_kezuolv_table)

These tables record various data including dates, train numbers, stations, departure/arrival times, seat occupancy rates, average daily capacity allocations, and weather conditions. The tables are relationally linked through station IDs and train IDs, enabling cross-table queries.

The passenger flow information table was compiled by processing daily data across multiple months. With the dataset spanning over a year (totaling 430 days), it contains 394,100 records in total (partial data shown in the table below). By analyzing entry and exit timestamps, hourly passenger volumes at each station can be calculated, allowing for optimized daily work prioritization and human resource allocation.

Table 3.1 Passenger Flow Information Table (passenger_flow_table section)

date	trainId	station	stationNum	Inbound Time	Outbound Time	Up Num	Down Num

2015/1/1	D02	ZD326	2	07:52:00	07:53:00	277	254
2015/1/1	D03	ZD192	4	07:22:00	07:24:00	78	27
2015/1/1	D03	ZD190-01	1	05:46:00	05:46:00	189	0
2015/1/1	D03	ZD190-02	2	06:02:00	06:03:00	77	1
2015/1/1	D03	ZD111-01	6	08:33:00	17:08:00	0	599
2015/1/1	D02	ZD250	3	08:29:00	08:30:00	511	619
2015/1/1	D03	ZD250	3	06:59:00	07:01:00	182	49
2015/1/1	D02	ZD190-01	5	09:48:00	16:19:00	0	832
2015/1/1	D02	ZD111-01	1	07:10:00	07:10:00	1233	0
2015/1/1	D02	ZD062	4	08:57:00	08:58:00	122	438

The regional meteorology table contains daily weather-related data for ZD111 City, ZD326 City, ZD250 City, and ZD190 City over a one-year period. Each day's weather condition varies and can be broadly categorized into: fog, haze, sunny, mostly sunny, cloudy with light rain, cloudy with showers, light to heavy rain, heavy rain to storm, sleet, etc.

Based on the meteorological bureau's classification standards for different weather conditions, the weather is planned to be divided into 14 levels: sunny, cloudy, overcast, showers, light rain, sleet, moderate rain, heavy rain, storm, light snow, moderate snow, heavy snow, fog, and haze.

Table 3.2 Regional Meteorological Table (meteorology_table section)

date	weather	temperature	wind	district
2015/1/1	Sunny/sunny	4℃/-3℃	The south wind is ≤ level 3/The south wind is ≤ level 3	ZD111
2015/1/1	Sunny/sunny	1℃/-3℃	North wind of force 5 to 6/North wind of force 4 to 5	ZD190
2015/1/1	Sunny/sunny	2℃/-5℃	South wind of 3-4 levels/South wind of 3-4 levels	ZD250
2015/1/1	Sunny/sunny	3℃/-7℃	South wind ≤ level 3 / South wind ≤ level 3	ZD326
2015/1/2	Sunny/sunny	6℃/0℃	North wind ≤ level 3 / South wind 3- level 4	ZD111
2015/1/2	Sunny/sunny	4℃/0℃	North wind of force 3-4 / North wind of force 3-4	ZD190
2015/1/2	Sunny/sunny	4℃/-4℃	North wind: 3-4 levels/South wind: 3-4 levels	ZD250
2015/1/2	Sunny/sunny	7℃/-5℃	North wind ≤ level 3 / South wind 3- level 4	ZD326
2015/1/3	Sunny/sunny	12℃/3℃	South wind of 3-4 levels/South wind ≤3 levels	ZD111
2015/1/3	Sunny/cloudy	7℃/2℃	South wind of force 4-5 / South wind of force 3-4	ZD190

The intra-domain station timetable information table is formed through a joint query of the passenger flow information table and daily seat occupancy rate table by comparing time and station names. Through query conditions, only information from intra-domain stations is extracted. For intra-domain stations in each region, combined analysis can be conducted. Most stations outside the domain primarily involve inbound and outbound passenger flows, which exhibit unique characteristics. Therefore, the current analysis focuses solely on the correlation of passenger flows at intra-domain stations.

Table 3.3 Station Schedule Information Table within the Jurisdiction (train_gn Part)

date	train Id	station	Station Num	Inbound Time	Outbound Time	Up Num	Down Num	keZuo Lv
2015/4/3	D01	4	ZD111-01	01:52:00	02:04:00	145	147	97.8
2015/4/3	D01	8	ZD190-02	05:20:00	18:24:00	0	350	97.8
2015/4/3	D01	6	ZD192	03:26:00	03:34:00	0	57	97.8
2015/4/3	D01	7	ZD250	03:59:00	04:06:00	8	168	97.8
2015/4/3	D01	5	ZD326	02:56:00	03:06:00	18	133	97.8
2015/4/6	D01	4	ZD111-01	01:52:00	02:04:00	33	63	98.5
2015/4/6	D01	8	ZD190-02	05:20:00	18:24:00	0	427	98.5
2015/4/6	D01	6	ZD192	03:26:00	03:34:00	3	41	98.5
2015/4/6	D01	7	ZD250	03:59:00	04:06:00	11	120	98.5
2015/4/6	D01	5	ZD326	02:56:00	03:06:00	4	83	98.5

The average daily capacity of train services - Passenger Load factor information table (train_rk) is obtained through the joint query of train service information (train_info_table) and daily passenger load factor table (train_kezuolv_table), and is mainly used to analyze the statistics of the total number of passengers in the stations corresponding to the actual train services. The main focus is on the analysis and comparison of vehicles within three cities that operate for one day.

Table 3.4 Information Table of Average Daily Capacity - Passenger Load Factor for Train Services (train_rk Part)

date	trainId	Start Station	End Station	Open Days	Daily average capacity	passenger load factor
2015/4/3	D01	ZD013	ZD190-02	1	1220	97.8
2015/4/6	D01	ZD013	ZD190-02	1	1220	98.5
2015/4/30	D01	ZD013	ZD190-02	1	1220	106.9
2015/5/3	D01	ZD013	ZD190-02	1	1220	79.5
2015/6/19	D01	ZD013	ZD190-02	1	1220	107.6
2015/6/22	D01	ZD013	ZD190-02	1	1220	94.5
2015/9/25	D01	ZD013	ZD190-02	1	1220	100.6
2015/9/30	D01	ZD013	ZD190-02	1	1220	109.9
2015/10/7	D01	ZD013	ZD190-02	1	1220	53.3
2015/1/1	D02	ZD111-01	ZD190-01	1	1112	95.6

Subsequent analyses, including holiday data analysis, overall passenger turnover analysis for regular dates, and passenger turnover analysis for different stations, all require joint queries across the five foundational tables to obtain the necessary data.

Numerous issues arose during the data processing phase. During data import, challenges emerged due to variations in the number of stations per table, daily train frequency differences, and inconsistent monthly day counts. After researching solutions, these were addressed through multi-layered nested loops with conditional checks. Additional discrepancies were identified: July and August lacked data for the 26th, March and April had formatting issues, March was missing data for the 24th, 30th, and 31st, and February 2016 had no data for the 19th. To optimize code

efficiency, months with identical day counts were grouped into lists, and corresponding if conditions guided the execution of specific for loops.

Each problem was systematically resolved through meticulous cross-checking, re-importing, and data filtering. The accuracy of any analytical outcome hinges 70% on data correctness—a principle deeply reinforced during my university studies and internships.

4. Overall data analysis and comparison

4.1 Comparative analysis of date change data

The overall data package shows that from January 2015 to March 2016, with the change of time, the railway passenger flow will also present corresponding change patterns. Figure 4.1 shows the average daily passenger load factor. Through the observation of the average daily passenger load factor, there are seasonal variations.

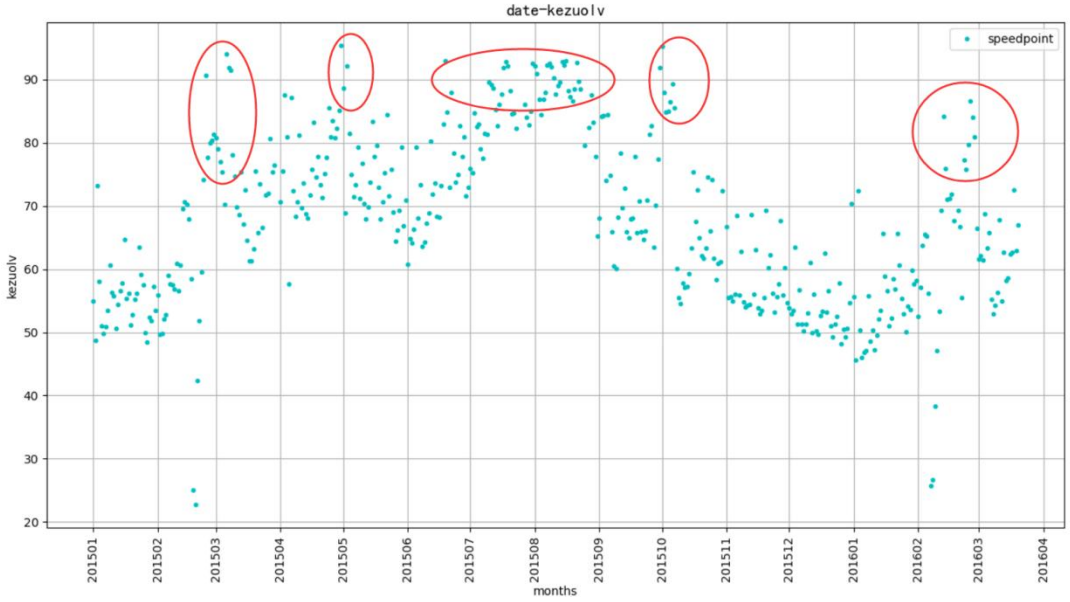

Figure 4.1 Scatter plot of the trend of passenger load factor changes from 201501 to 201603

As indicated by the red circle mark in Figure 4.2, the passenger load factor is mostly as high as 90%. According to the distribution of holidays in China, the longest holidays are the Spring Festival and National Day respectively. It is obvious that the passenger load factor during this period is higher than that in ordinary times. Based on the analysis of passenger attributes, the attributes of passengers vary in each period. During the Spring Festival holiday, the "main force" of passengers is migrant workers, while during the National Day holiday, the "main force" consists of students, blue-collar workers, etc. It can also be found from Figure 4.1 that the passenger load factor is generally higher than 70% in summer and autumn. Besides the influence of holidays, the influence of the tourism industry is also very crucial, thus causing the trend of the passenger load factor waveform. Based on the data time period, there are a total of eight

holidays during this period, including two New Year's Day, two Spring Festival and two Qingming Festival holidays, followed by Labor Day, Dragon Boat Festival, Mid-Autumn Festival, Victory Day of the War of Resistance against Japanese Aggression and National Day. The following is a main analysis and visualization of the data of some holidays.

4.1.1 Passenger load factor changes during ordinary days

By eliminating the data of all holiday dates, the average passenger load factor on regular dates was obtained, as shown in Figure 4.2. After excluding special holidays, the overall trend of passenger load factor changes still shows that summer and autumn occupy a high position, but the overall passenger load factor is approximately at 65%. From an overall perspective, the waste of resources is still very significant. After excluding the passenger load factor during holidays, the overall density has also decreased and is no longer as concentrated as before, reflecting the objective law of changes in passenger flow.

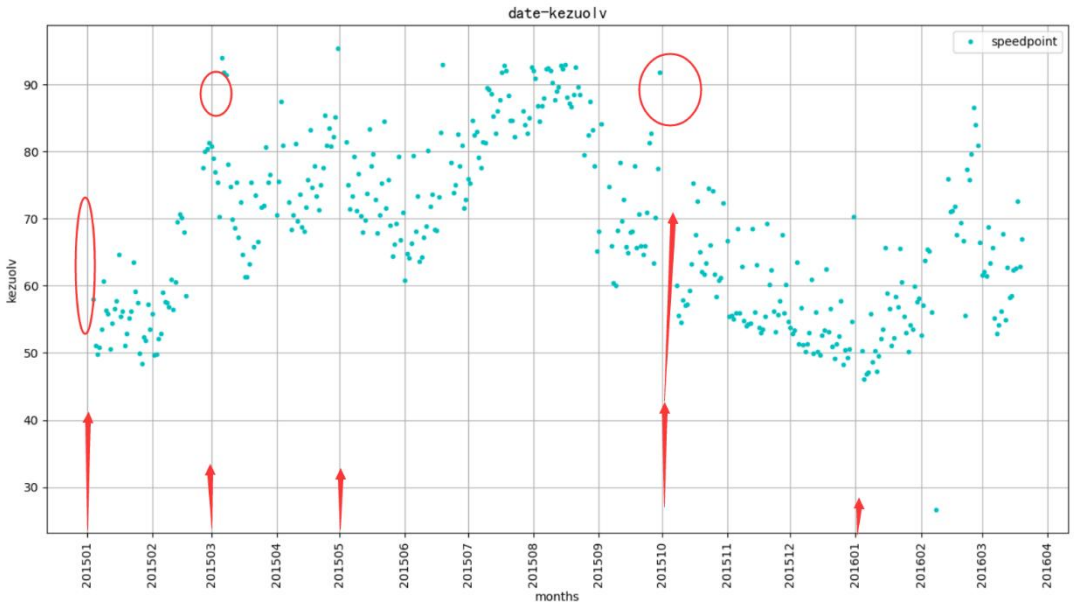

Figure 4.2 Passenger Load Factor - Non-holiday Chart from 201501 to 201603

4.1.2 The change in passenger load factor during the New Year's Day holiday

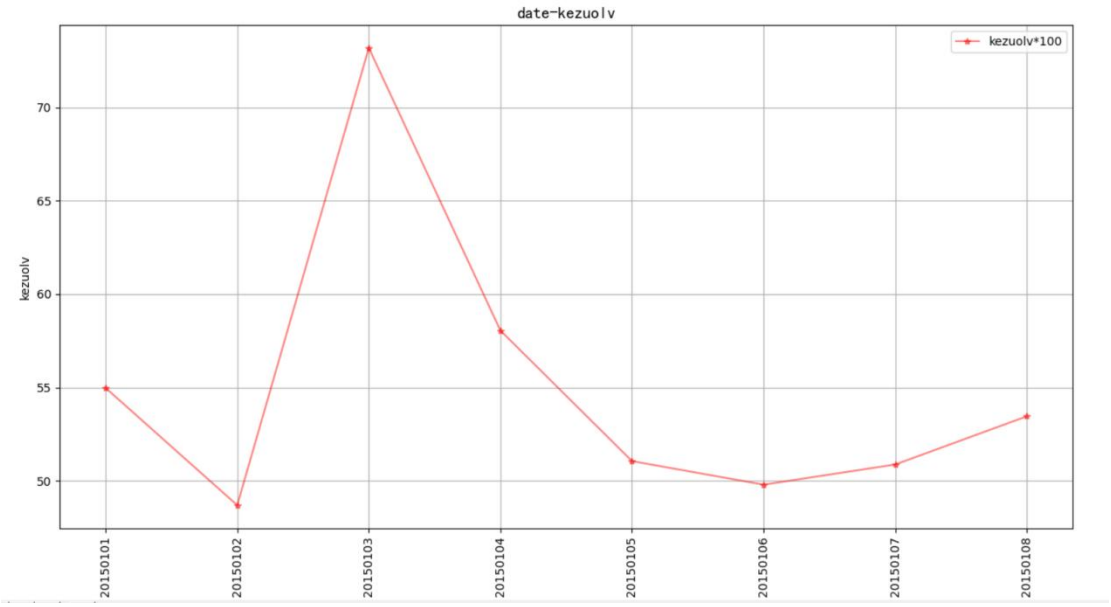

Figure 4.3 New Year's Day 2015 - Passenger Load Factor Change Chart

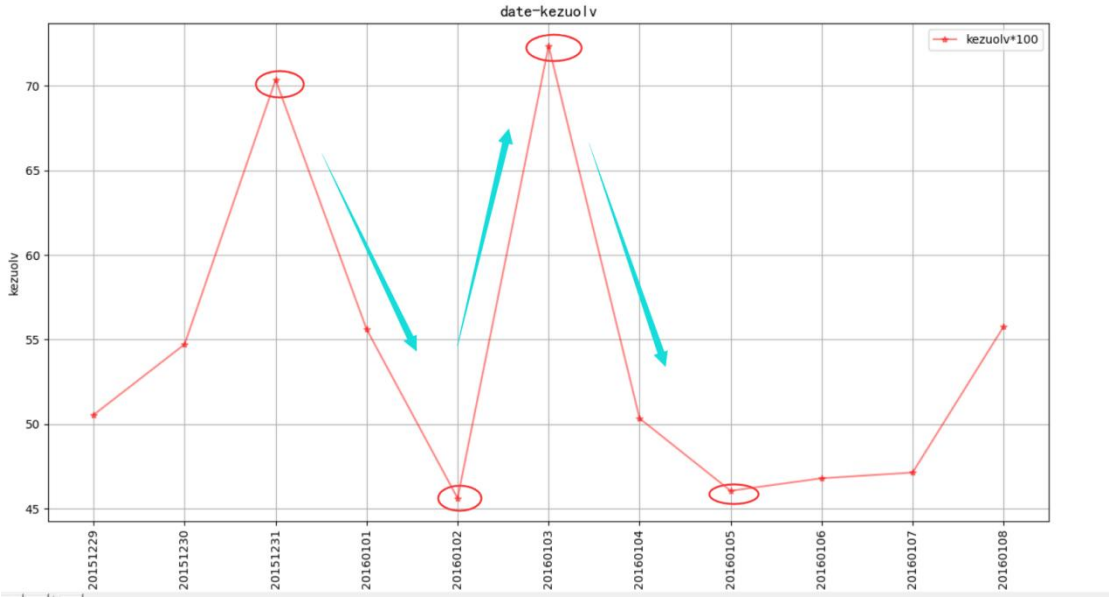

Figure 4.4 New Year's Day 2016 - Passenger Load Factor Change Chart

As shown in Figures 4.3 and 4.4 above, the seat occupancy rate trends during the New Year holiday periods of 2015 and 2016 are almost identical, fully demonstrating the significant role historical data plays in future decision-making. It can also be observed that the seat occupancy rates exhibit abrupt changes both before and after the holiday period. Particularly in the pre-holiday period of the 2016 New Year, the seat occupancy rate reached as high as 70%, but during the holiday itself, it dropped to only 46%, only to rebound rapidly to 73% on the second day after the holiday. In contrast, the post-holiday period of the 2015 New Year saw a slight decline in seat occupancy rate. This reflects how China's rapid economic development has greatly

enhanced national consumption awareness and living standards, leading people to increasingly prioritize higher-quality travel options.

4.1.3 Passenger load factor changes during the Spring Festival and National Day holidays

The Spring Festival is the most grand and solemn festival in China. Considering the actual situation, every time people go home for reunion, they either set off earlier or return later. Therefore, the data of the first six days and the last sixteen days of the holiday were collected, which is in line with the principle of data analysis. The Spring Festival holiday lasts for 7 days, from February 18th to 24th, 2015 and from February 7th to 13th, 2016. According to the comparison between Graph 4.5 and Graph 4.6, it can be concluded that there are obvious inflection points in the round-trip passenger load factor. The lowest point occurs one or two days before the holiday, and the highest point occurs one or two days before the end of the holiday, which conforms to the objective law. The comparison between 2015 and 2016 is almost the same. The trends of the line charts are almost the same.

Figure 4.5 Graph of Passenger Load Factor Changes during the Spring Festival of 2015

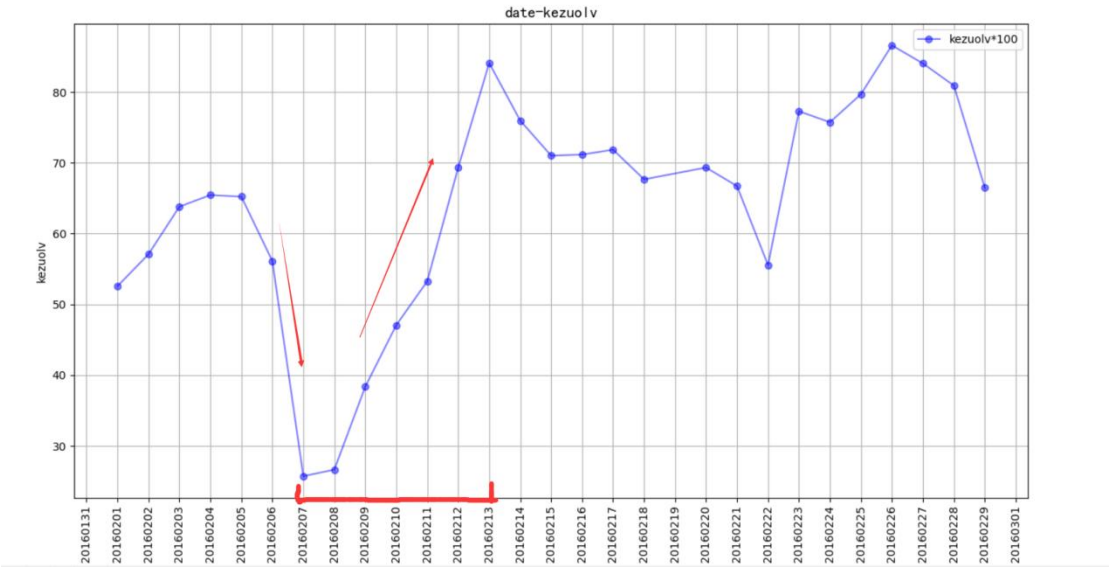

Figure 4.6 Passenger Load Factor Change Chart during the Spring Festival of 2016

The National Day holiday is the longest vacation period besides the Spring Festival. There exists significant difference in travel purposes between these two holidays. As shown in Figure 4.7, the seat occupancy rate gradually increases before the holiday and peaks on National Day itself. Unlike the Spring Festival pattern which shows an immediate sharp decline after the peak, the National Day curve demonstrates only a slight decrease followed by stabilization. This indicates that 80%-90% of passengers travel for tourism purposes, reflecting strong tourism development potential in this urban area - a fact worthy of promotional efforts by both government and railway authorities to further boost local economic growth.

The data reveals two distinctive characteristics compared to other holidays: First, the gap between the lowest and highest occupancy rates is the largest, suggesting Chinese citizens are increasingly prioritizing life enjoyment and showing growing enthusiasm for tourism. Second, the pre-holiday occupancy trend rises gradually while the post-holiday decline is much steeper. This pattern provides accurate reference for determining key work periods and making operational predictions during this season.

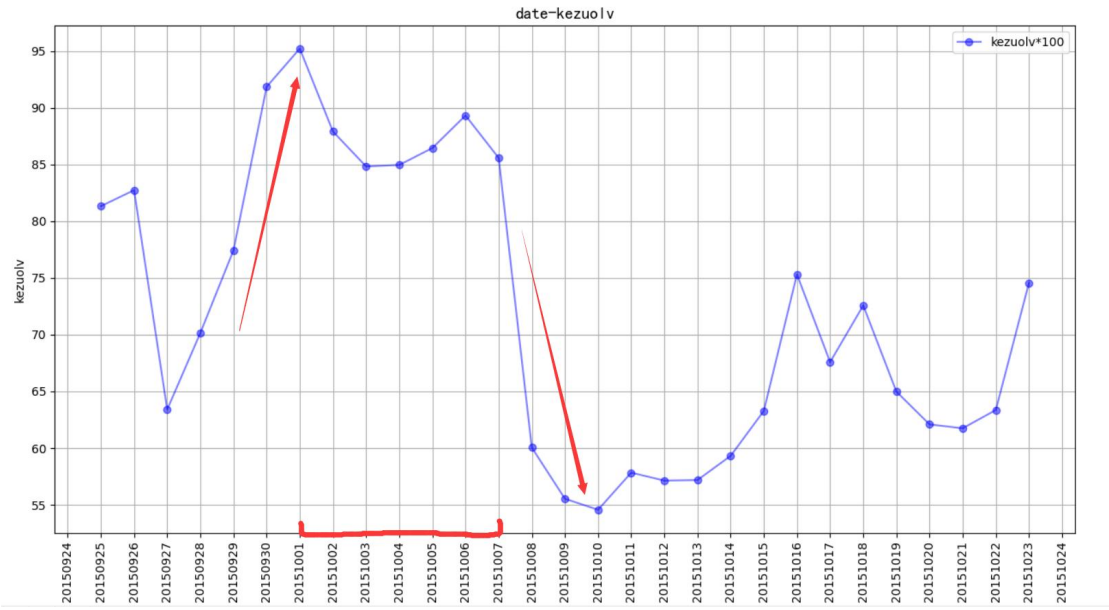

Figure 4.7 Passenger Load Factor Change Chart during the National Day of 2015

4.1.4 Comparison of passenger load factor changes during other holidays

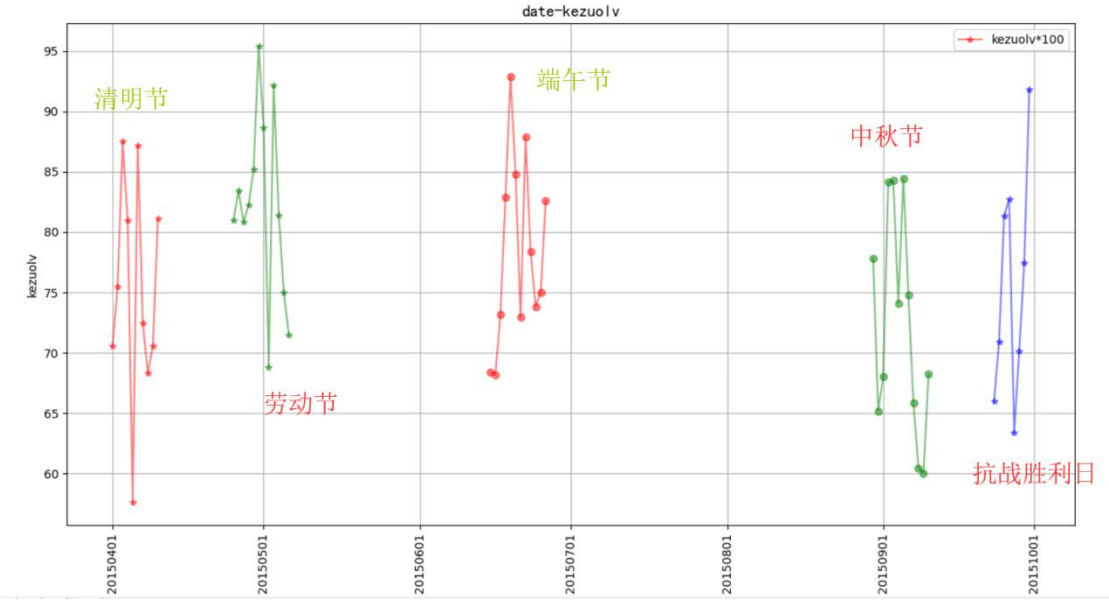

Figure 4.8 Changes in Passenger Load Factor during Other Holidays from 201501 to 201512

By observing seat occupancy rate trends during other holidays (as shown in Figure 4.8), the overall pattern demonstrates abrupt changes - a sharp increase before the holiday followed by a steep decline afterward. However, holidays with family reunion characteristics typically exhibit buffer periods, showing either a delayed pre-holiday increase or post-holiday decrease. For short vacations, which mainly involve short-distance travel and visiting, the trends align more closely with general holiday patterns.

This observation highlights how railway authorities should rationally adjust staff allocation and service provisions according to different holiday characteristics. Family-oriented holidays require extended service periods before/after the actual dates, while short vacations demand concentrated resources during peak travel days.

4.2 Comparative analysis of train type data

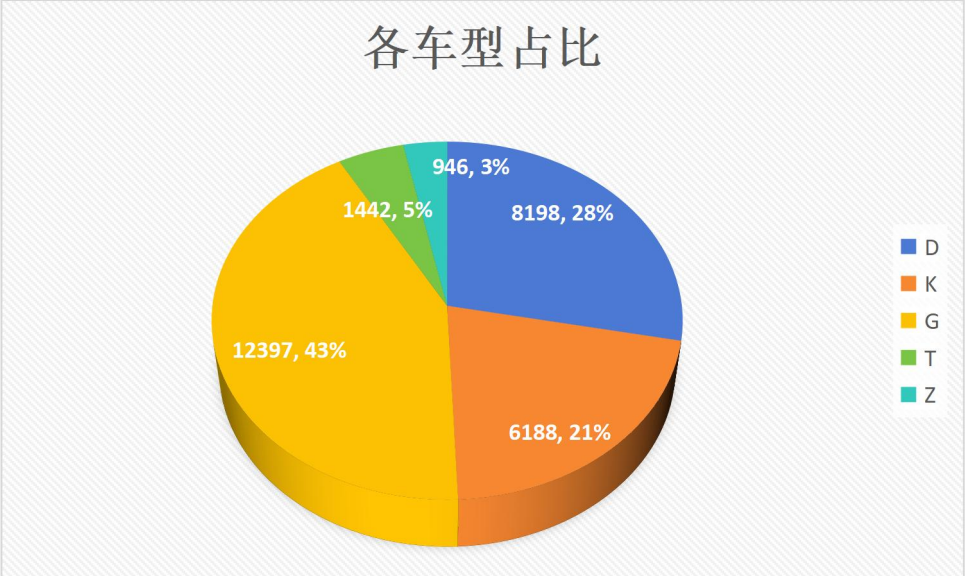

Figure 4.9 Percentage Chart of Train models

(High-speed rail bullet train (G), bullet train (D), direct express train (Z), express train (T), rapid train (K))

During the valid data period, a total of 105 different train services operated within the district. As shown in Figure 4.9, high-speed trains (G), express trains (K), and electric multiple unit trains (D) accounted for 92% of all services, with G trains being the most prevalent at 43%. This distribution indicates robust tourism and economic development in the city, providing significant convenience for travelers while generating substantial revenue, fostering positive urban culture, and stimulating regional economic growth.

Analysis of average seat occupancy rates (Figure 4.10) reveals that 66% of services maintained rates between 40%-85%, while only 9.5% fell below 40%, demonstrating generally favorable occupancy levels. Notably, trains D47, D48, D24, and D23 consistently exceeded 100% occupancy. The high demand for D and G trains reflects passenger preference for superior speed and service quality. In contrast, T trains showed particularly low occupancy rates with extreme fluctuations up to 60% variance, indicating areas requiring operational improvements. The notably underperforming D05 train, averaging below 30% occupancy, necessitates route optimization or potential discontinuation to better utilize resources and meet passenger needs.

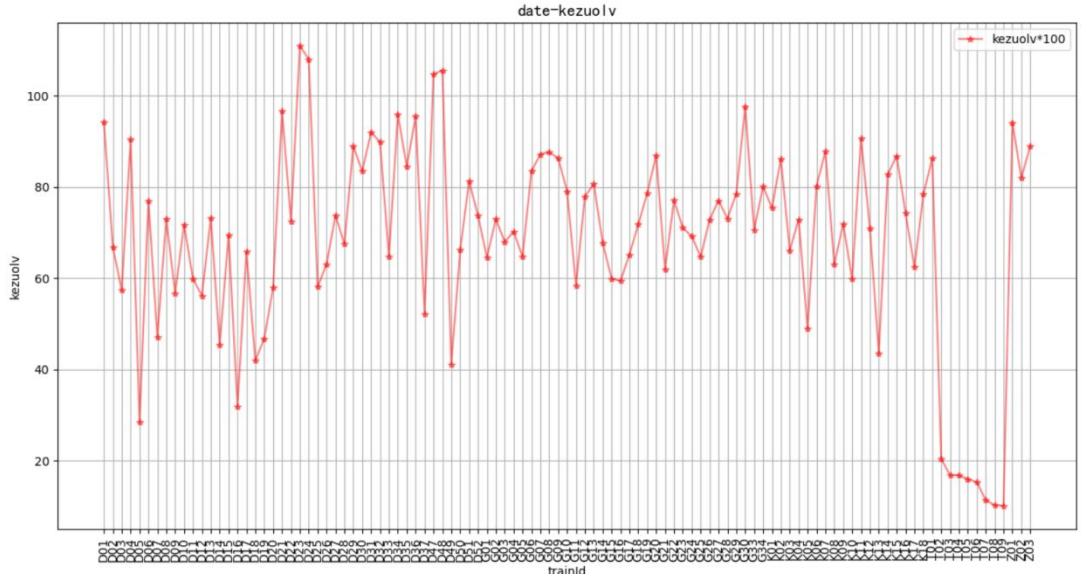

Figure 4.10 Train Number - Passenger Load Factor chart

4.3 Comparison and analysis of data inside and outside the pipe

It can be known from the data that all our data include the stations within and outside the management. Among them, the stations belonging to ZD111 City, ZD326 City, ZD250 City and ZD190 City are within the management, and the rest are outside the management. Through the statistics of the total passenger flow at the two different locations (as shown in Figure 4.11), the overall passenger flow outside the management is generally lower than that inside the management. Through the analysis of objective reasons, one is that local passenger flow accounts for a part. Secondly, the number of tourists visiting this city also keeps increasing with the seasons. It can be seen that the number of people during holidays also increases sharply. The special part, such as the overall passenger flow of some stations outside the management being higher than that inside the management, is caused by the planning of railway routes and the competition from the surrounding tourism industry, as well as the return of some students on vacation.

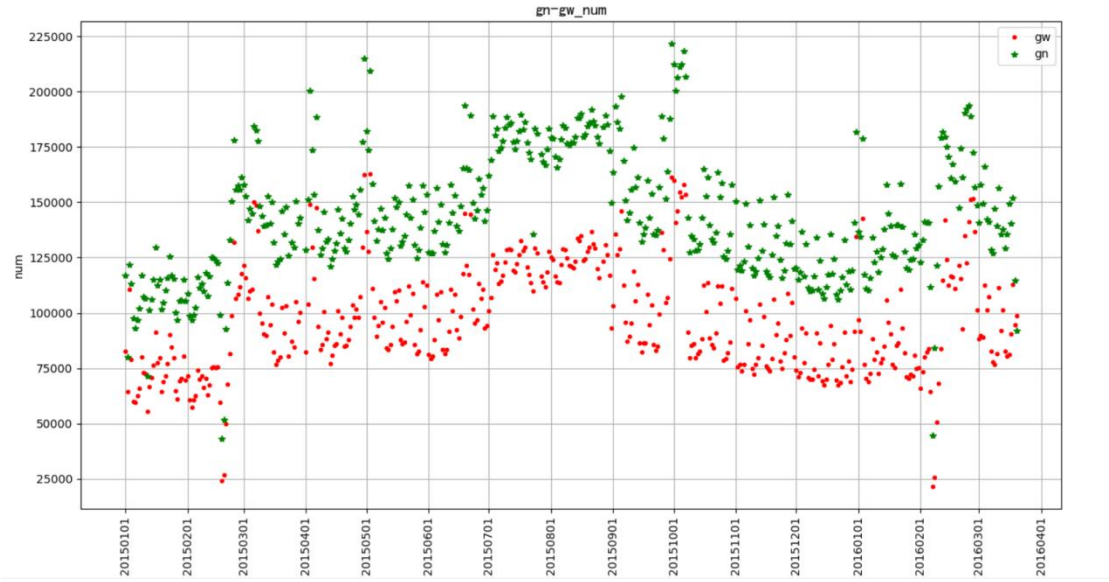

Figure 4.10 Comparison chart of the total passenger flow inside and outside the pipe (gn)

The total passenger flow of different train services was compared and statistically analyzed (as shown in Figure 4.11). Comparatively speaking, Train D has a relatively large number of passengers of one million, while Train D06 has the largest passenger flow, reaching over twelve million. The management and organization of this train service need to allocate more resources and reasonably and appropriately increase the number of train services, etc.

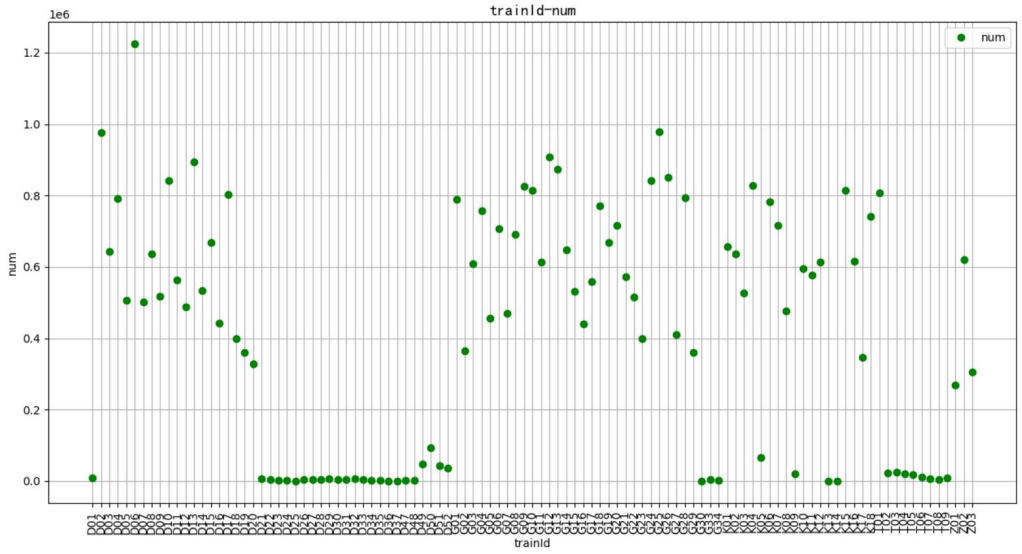

Figure 4.11 Comparison chart of the total passenger flow of each train within the jurisdiction

4.4 Data comparison and analysis of stations within the jurisdiction

From the data, it can be seen that ZD111 City, ZD326 City, ZD250 City and ZD190 City have a total of 14 stations, among which ZD190 City has the most stations. Through comparison, it can be clearly seen which station in each urban area has the largest workload, which is convenient for

decision-makers to make corresponding deployments and also conducive to making reasonable personnel arrangements and train schedules in advance.

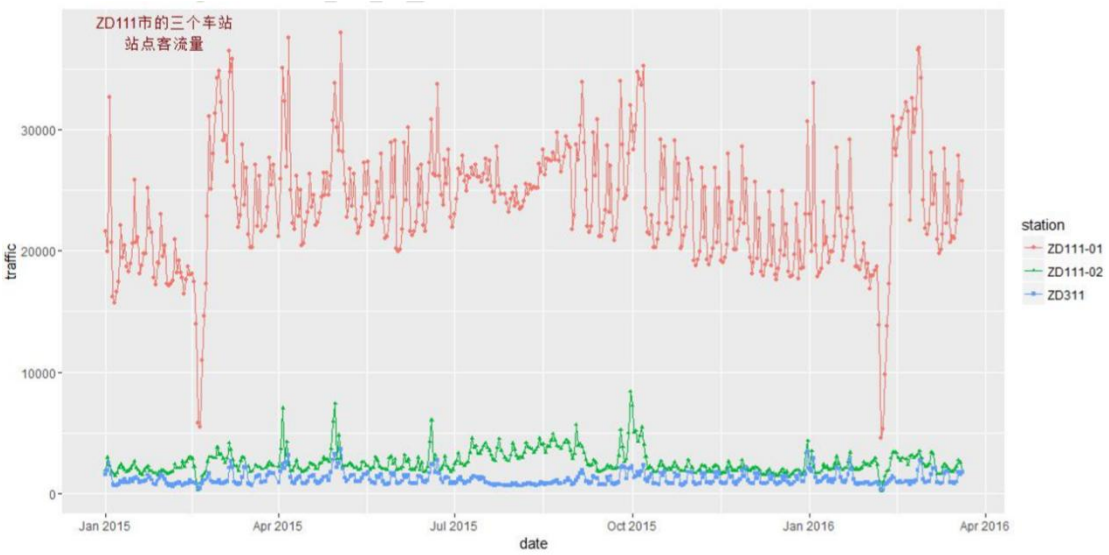

Figure 4.12 Comparison chart of the total passenger flow at all stations in ZD111 City

Figure 4.13 Comparison chart of the total passenger flow at each station in ZD326-ZD250

City

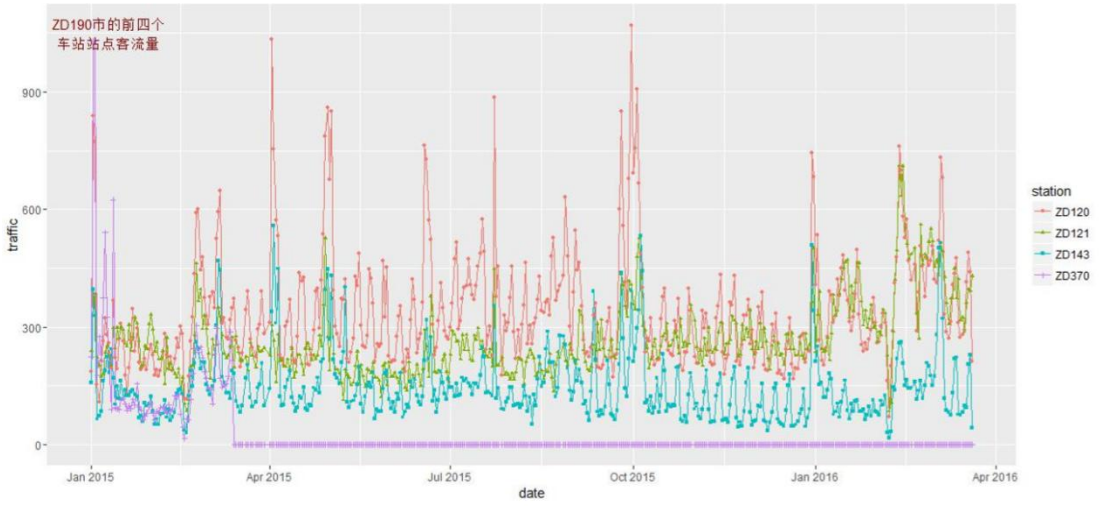

Figure 4.14 Comparison chart of the total passenger flow of the first four stations in ZD190

City

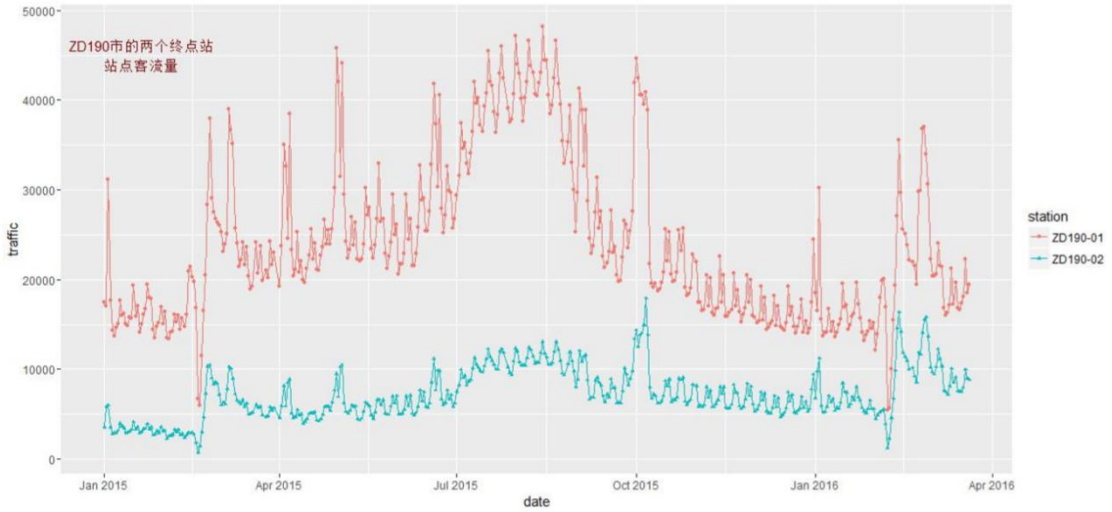

Figure 4.15 Comparison chart of the total passenger flow at Terminal 1 and Terminal 2 of ZD190 City

In the comparison of the stations under the jurisdiction (as shown in Figure 4.16), it can be seen that the total passenger flow of ZD111-01 and ZD190-01 is as high as over 10 million. These two stations are mostly the starting station and the terminal station respectively, which is also one of the reasons for the largest passenger flow. The second reason is that there are more learning and tourist attractions. Among them, ZD250 and ZD326 are also relatively large stations, with a total passenger flow exceeding six million people. For such super-large and large stations, the management and control of personnel in each area are also very dangerous. Especially during the Spring Festival, various vendors and others pass off inferior goods. For tourists from outside, it is necessary to remind them to take good care of their personal finances and pay attention to safety, etc. In addition, a contingency plan for handling emergency situations is also very necessary.

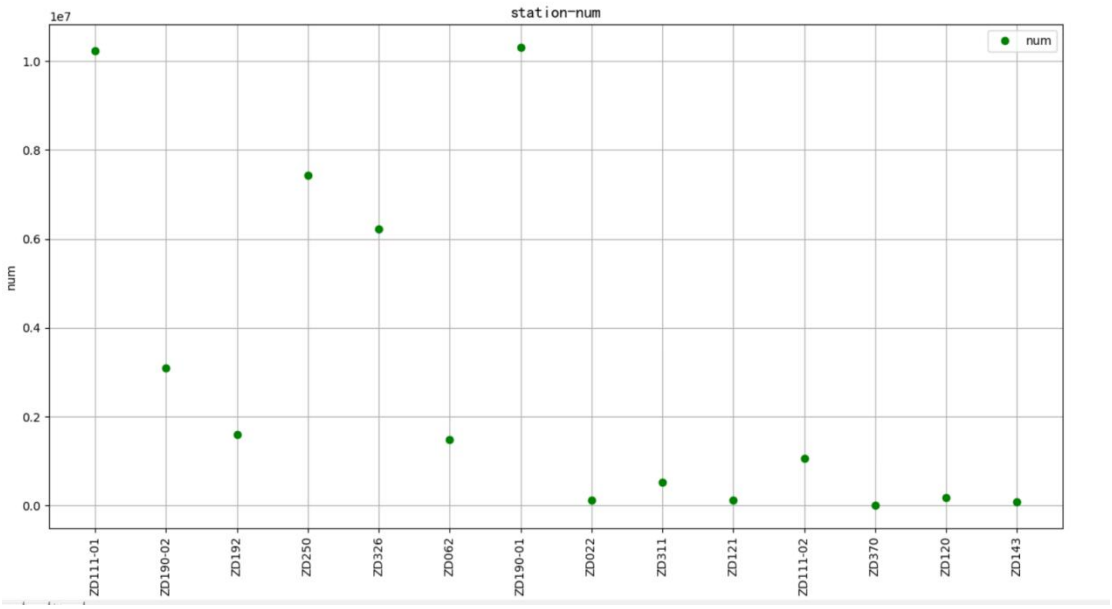

Figure 4.16 Comparison chart of the total passenger flow at each station within the jurisdiction

4.5 Data comparison and analysis of two-way stations

According to the query of the train number information table (train_info_table), Table 4.1 can be obtained. Among them, the trains D02-D19 are all two-way trains (that is, the starting station and the terminal station are in opposite directions). Two-way trains are mostly prepared for the travel of residents between urban areas. It can be analyzed which two urban areas have a greater passenger flow, and the reasons can be analyzed from this. Summarize experiences, etc.

Table 4.1 Two-way Train Information Table (Part)

trainId	Start Station	End Station	Open Days	Daily average capacity
D02	ZD111-01	ZD190-01	1	1112
D03	ZD190-01	ZD111-01	1	613
D04	ZD111-03	ZD190-01	1	586
D05	ZD190-01	ZD111-01	1	1172
D06	ZD111-01	ZD190-01	1	1172
D07	ZD190-01	ZD111-01	1	1226
D08	ZD111-01	ZD190-02	1	586
D09	ZD190-02	ZD111-01	1	586

D10	ZD111-01	ZD190-01	1	1172
D11	ZD190-01	ZD111-01	1	586
D12	ZD111-01	ZD190-01	1	586
D13	ZD190-01	ZD111-01	1	1172
D14	ZD111-01	ZD190-01	1	1226
D15	ZD190-02	ZD111-03	1	586
D16	ZD111-01	ZD190-01	1	613
D17	ZD190-01	ZD111-03	1	1112
D18	ZD111-01	ZD190-01	1	613
D19	ZD190-01	ZD111-03	1	586

By analyzing the line graph of the total initial passenger flow of two-way trains (as shown in Figures 4.17, 4.18, 4.19, 4.20, 4.21, 4.22, and 4.23), it can be concluded that the passenger flow rates of ZD02, ZD04, ZD06, ZD08, and ZD10, which are the starting stations of ZD111 City, are generally high. Moreover, the total passenger flow of the ZD01, ZD03, ZD05, ZD07, ZD09, and ZD11 trains, which depart from ZD190 City, is lower than that of the ZD13, ZD15, ZD17, and ZD19 trains. Therefore, we can draw the conclusion that the passenger flow of two-way trains at different times follows certain patterns, and we can increase or decrease the number of train services and the average daily passenger capacity in a targeted manner.

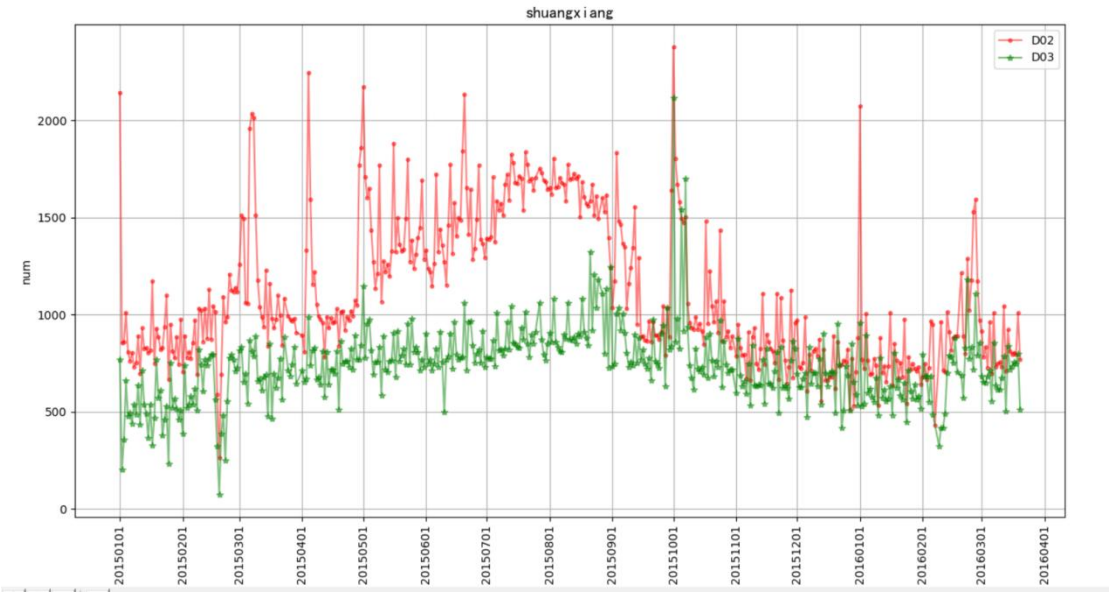

Figure 4.17 Comparison chart of Total Passenger flow from D02 to D03

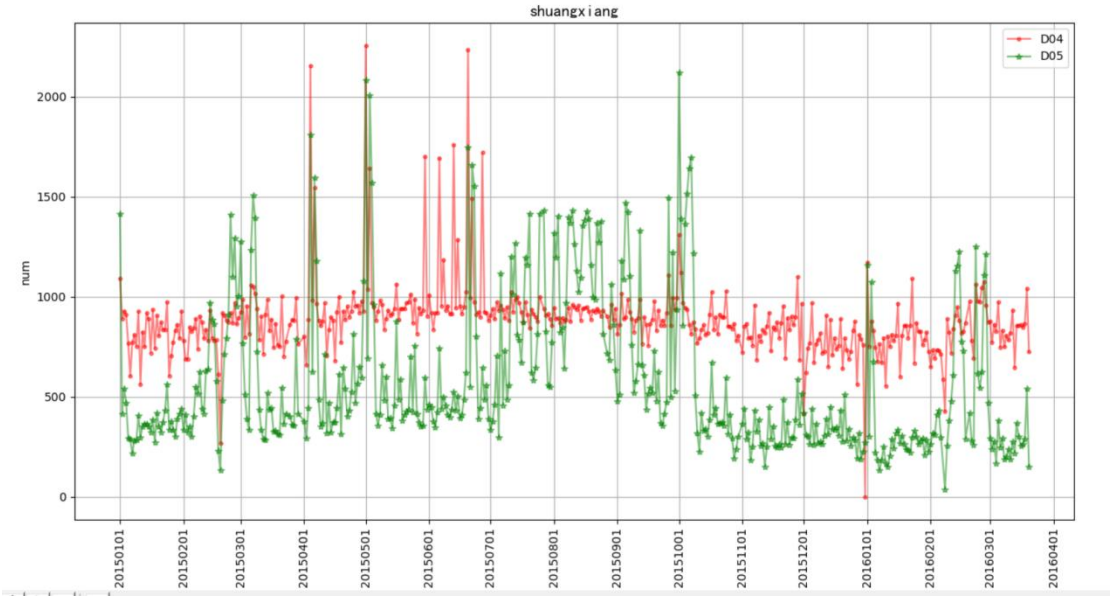

Figure 4.17 Comparison chart of Total Passenger flow from D04 to D05

Figure 4.17 Comparison chart of Total Passenger flow from D06 to D07

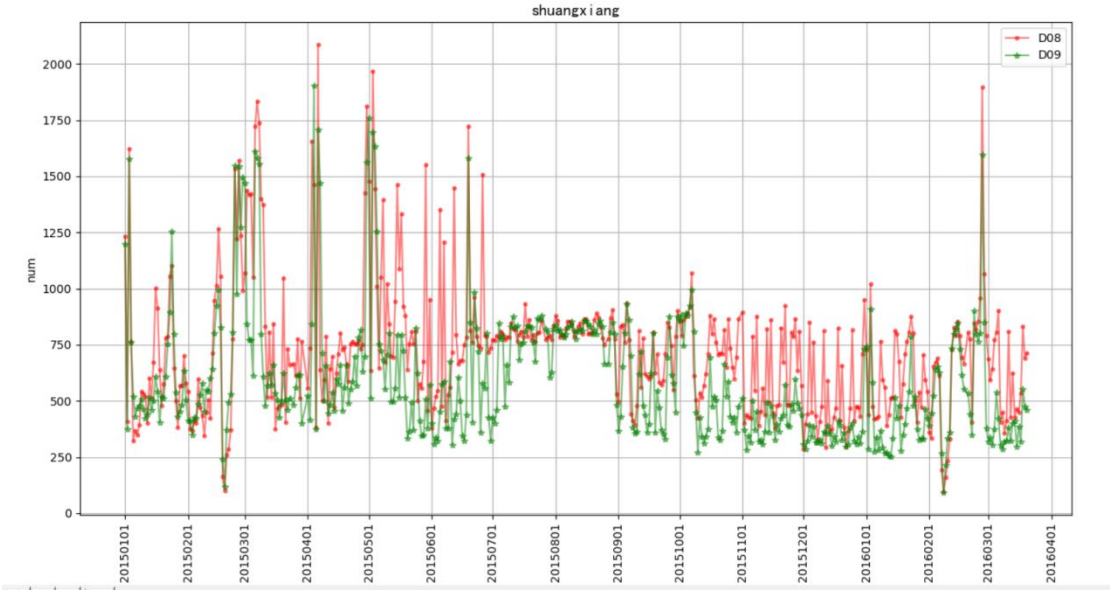

Figure 4.17 Comparison chart of Total Passenger flow from D08 to D09

Figure 4.17 Comparison chart of Total Passenger flow from D10 to D11

Figure 4.17 Comparison chart of Total passenger flow from D12 to D13

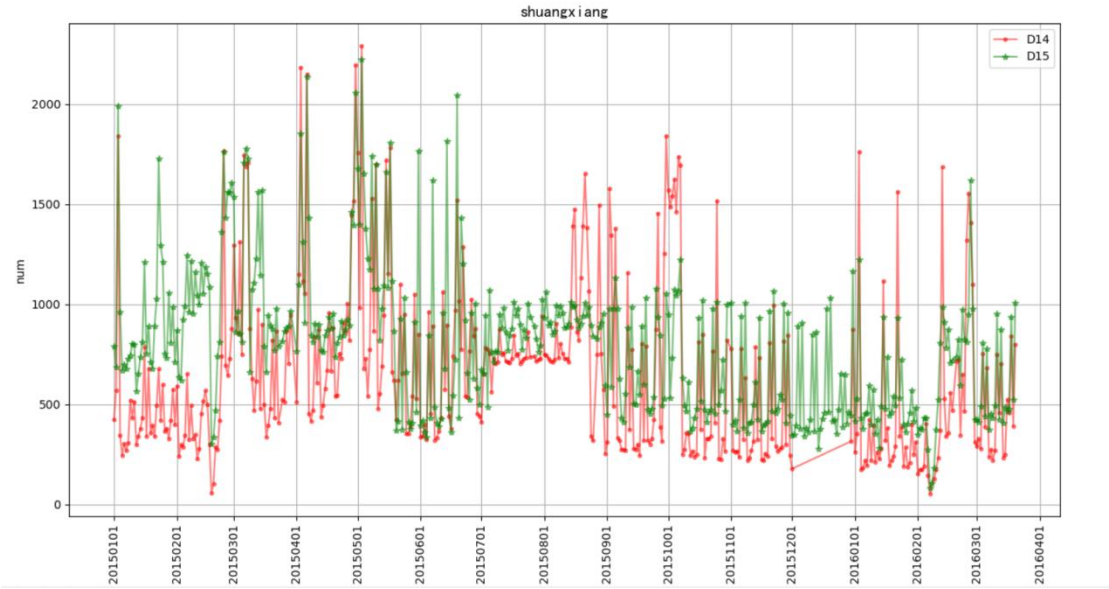

Figure 4.17 Comparison chart of Total Passenger flow from D14 to D15

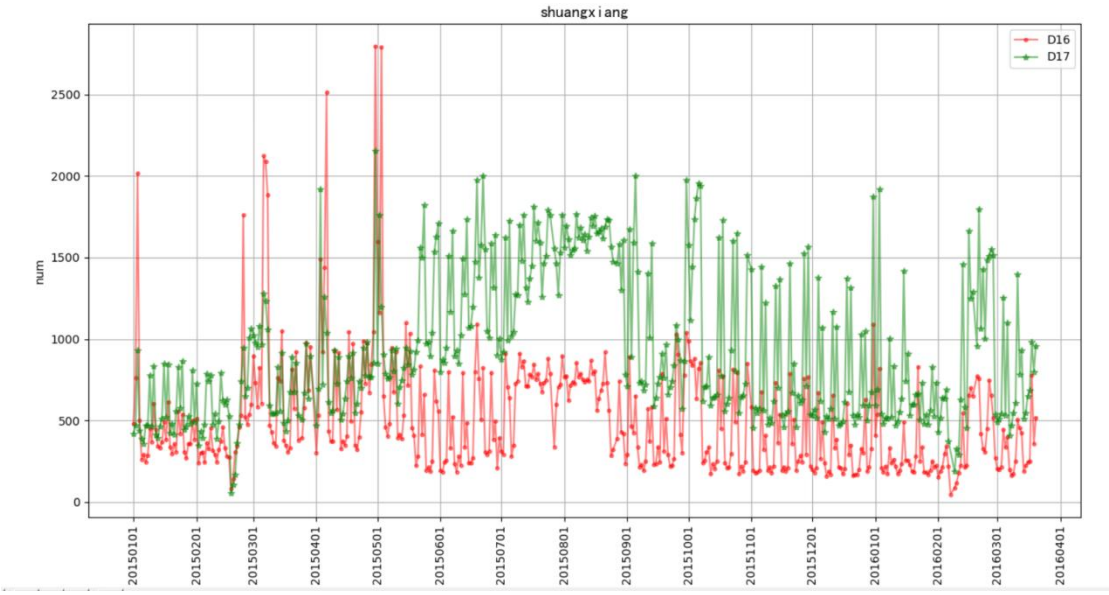

Figure 4.17 Comparison Chart of Total Passenger Flow from D16 to D17

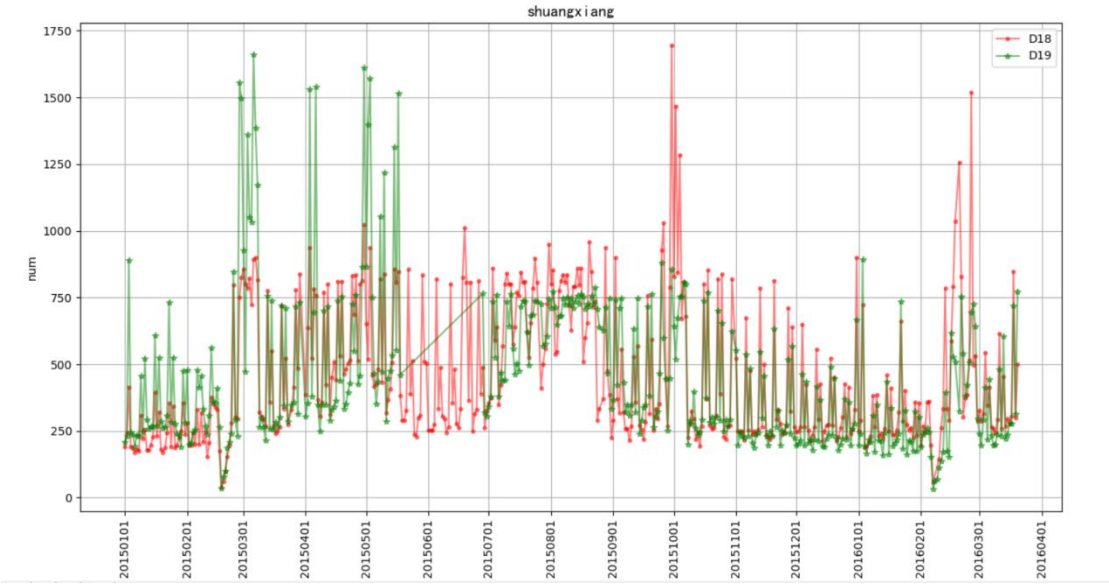

Figure 4.17 Comparison chart of Total passenger flow from D18 to D19

5. Assenger flow prediction and analysis

5.1 Analysis of potential factor influences

In addition to being affected by holidays, passengers' travel may also be influenced by weather conditions, temperature and wind direction, etc. The data provides the weather for a year in ZD111 City, ZD326 City, ZD250 City and ZD190 City. In order to facilitate the later data comparison and calculation, the data is normalized (as shown in Table 5.1). The weather is

classified according to the actual situation and divided into five grades: excellent, good, good, average and poor, with scores of 5, 4, 3, 2 and 1 respectively. The weather conditions are classified and stratified by the meteorological bureau. For example, clear weather is considered excellent and heavy rain is considered poor. Temperature only includes the highest and lowest temperatures. According to the fact that people's activity time is generally during the day, by the proportional averaging method, that is: temperature = (high temperature during the day *70%+ low temperature at night *30) /2, 20℃-30℃ is considered the optimal temperature based on the perceived comfort level. The wind direction varies according to its level, with the optimal level being less than or equal to three levels.

Table 5.1 Normalized Weather Conditions Table (Part)

date	weather	temperature	wind	district
2015-01-01	3	3	5	ZD111
2015-01-01	3	2	2	ZD190
2015-01-01	3	2	4	ZD250
2015-01-01	3	2	5	ZD326
2015-01-02	3	3	5	ZD111
2015-01-02	3	3	4	ZD190
2015-01-02	3	3	4	ZD250
2015-01-02	3	3	5	ZD326
2015-01-03	3	3	4	ZD111
2015-01-03	3	3	3	ZD190

The correlation between weather conditions and passenger flow was tested by rank correlation, and the obtained correlation coefficients were all not high. Secondly, considering that general tourism may focus on the weather, and secondly, the operation of trains is very little affected by the weather. To sum up, the weather is not included in the model.

5.2 Data preparation and processing

5.2.1 Dataset selection and analysis

According to the requirements of the paper, it is necessary to conduct a predictive analysis of the size of passenger flow. Firstly, the data set: from January 1, 2015 to March 20, 2016 is selected.

Considering the particularity of holidays, exclusive analysis is required. Then, the changes in passenger flow during holidays are excluded (as shown in Figures 5.1 and 5.2).

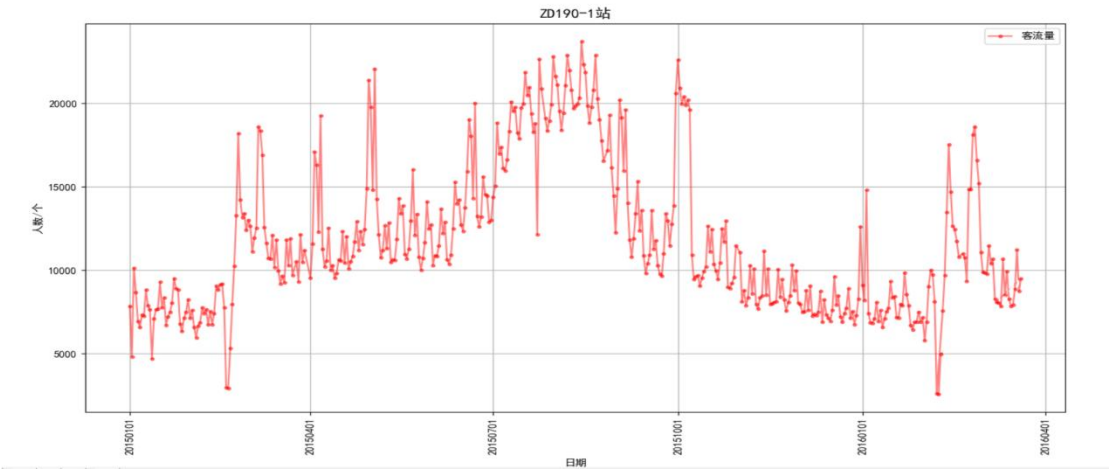

Figure 5.1 Passenger Flow Variation Trend Chart of ZD190-1 Station (including Holidays)

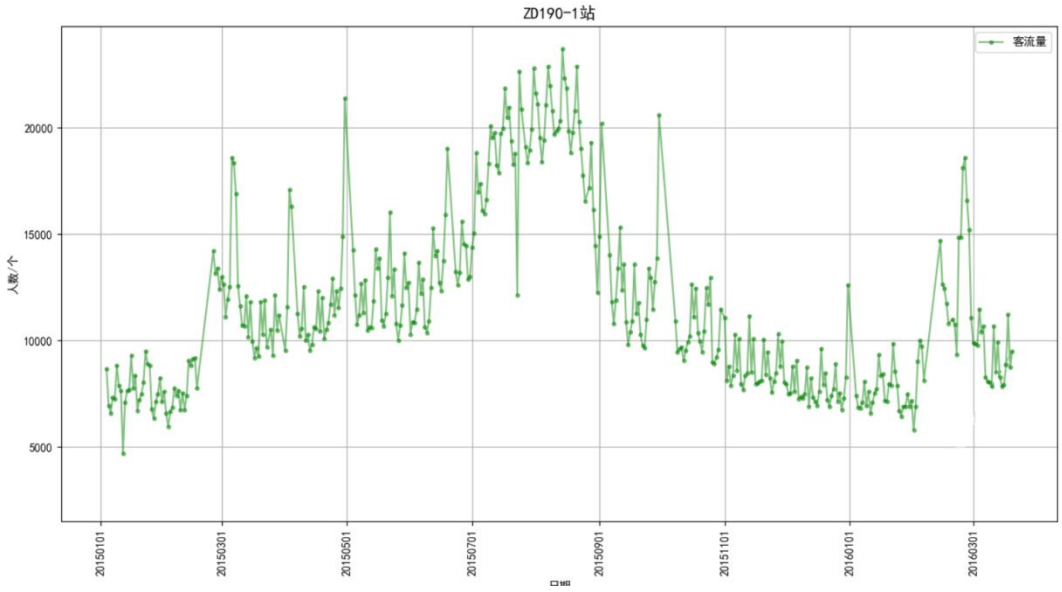

Figure 5.2 Passenger Flow Variation Trend Chart of ZD190-1 Station (excluding Holidays)

As can be seen from the above chart, the overall trend of change is slightly more regular. The daily data predicted by excluding holiday data is more accurate and reasonable. The training samples from 2015/01/01 to 2016/03/20 were selected in the dataset, and the prediction period was from 2016/03/21 to 2016/04/10.

5.2.2 Analysis of sample datasets

Test for the stationarity of training sample data:

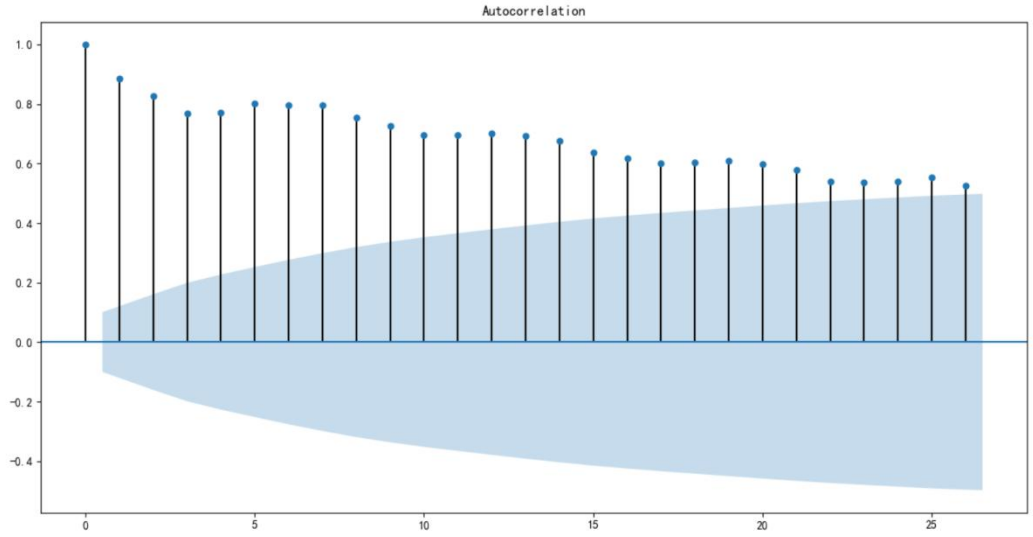

Figure 5.3 Autocorrelation diagram of ZD190-01 Station

from statsmodels.tsa.stattools import adfuller #import package

print(adfuller(num)) #Calculate and print the result

#he return values are in sequence: （adf, pvalue p， usedlag, nobs, critical values critical value , icbest, regresults, resstore）

The test result of the original sequence is: (-1.609501517336602, 0.47881902089316786, 6, 374, {'1%': -3.4479562840494475, '5%': -2.869299109917524, '10%': -2.57090345105665}, 6481.722877053289)

It can be seen from the results that the adf is all greater than the critical values of the three different test levels, and the p value corresponding to the unit detection statistic is significantly greater than 0.05, indicating that the sequence can be determined as a non-stationary sequence.

Perform first-order differences on the sample data (Figure 5.4) :

The ADF test result of the difference sequence is: (-12.708041231974569, 1.0446949084325706e-23, 5, 374, {'1%': -3.4479562840494475, '5%': -2.869299109917524, '10%': -2.57090345105665}, 6465.621461509212)

#The time series graph of the sequence after first-order difference fluctuates relatively smoothly around the mean. The autocorrelation has a strong short-term correlation, and the p-value of the unit root test is less than 0.05. Therefore, it can be said that the sequence after first-order difference is a stationary sequence.

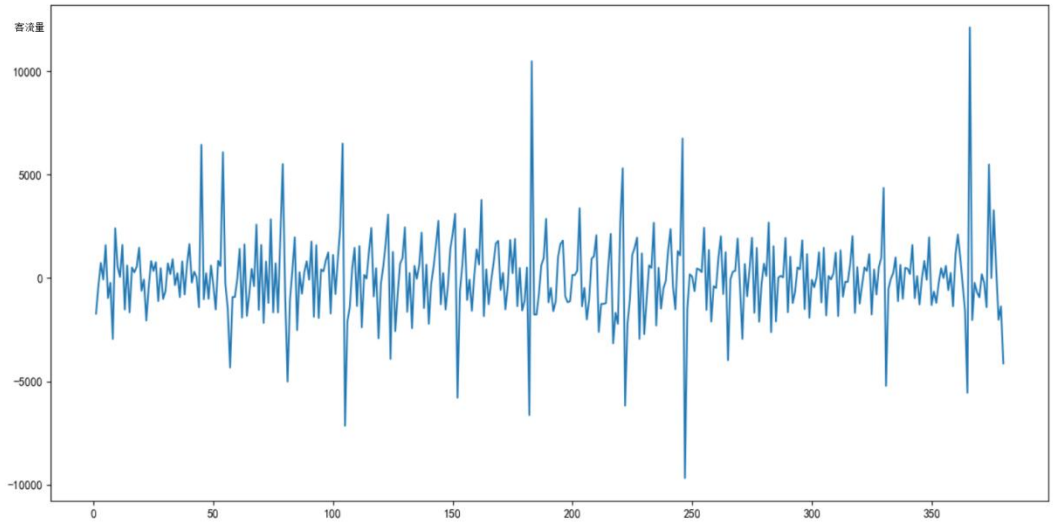

Figure 5.4 The first-order difference timing diagram of Station ZD190-01

Through the partial correlation graph (as shown in Figure 5.5), it can be found that at the positions of lag 1 to 7 orders, the partial autocorrelation coefficient is outside the confidence boundary and, while starting from the lag 7 plane, the value of the partial autocorrelation gradually shrinks to nearly 0.

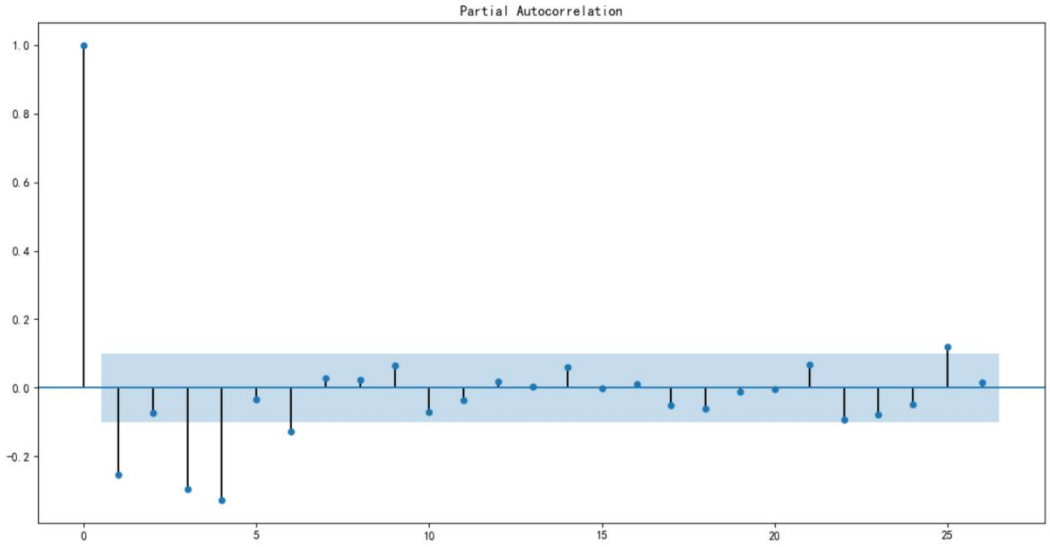

Figure 5.5 First-order differential partial correlation diagram of Station ZD190-01

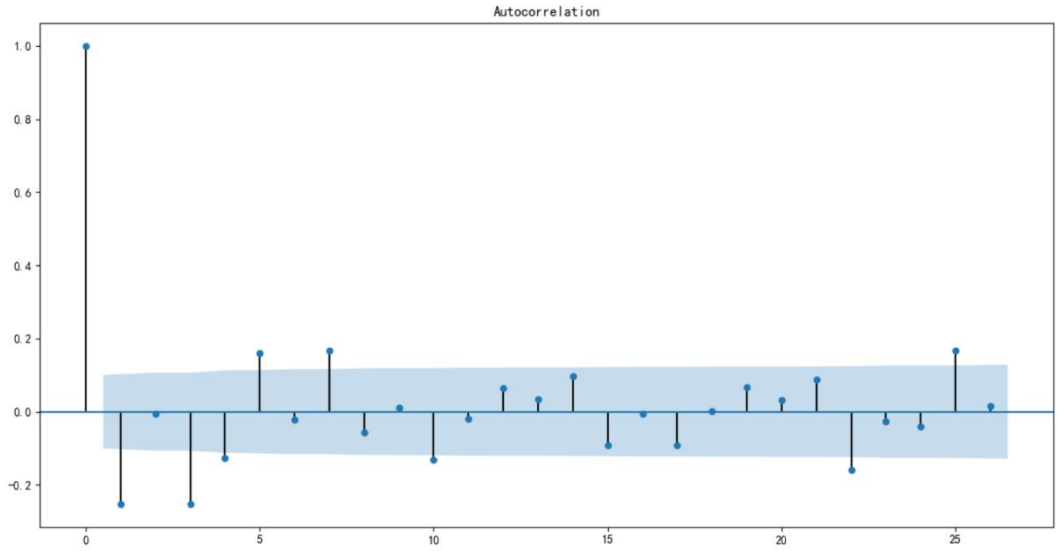

Figure 5.6 First-order difference autocorrelation diagram of Station ZD190-01

```
#对一阶差分后的序列做白噪声检验
from statsmodels.stats.diagnostic import acorr_ljungbox
print(u'差分序列的白噪声检验结果：',acorr_ljungbox(D_data, lags= 1)) #返回统计量和 p 值
```

Figure 5.7 White noise test of first-order difference sequences

The test results of white noise for the differential sequence of sample data at ZD190-01 station:

(array([24.38502465]),array([7.88794854e-07]))

inspection result: (array([11.30402222]), array([0.00077339])) The second item is the p value, which is much less than 0.05.

5.3 ARIMA Model Design and Verification

Based on the data after the first-order difference and the graphical analysis of autocorrelation and partial correlation, the appropriate values of p, q and d are selected, and the order of the model is determined according to the AIC rule of the ARMA model. AIC supports excellent data fitting while avoiding the problem of Overfitting. First of all, the model should have the smallest AIC value. Meanwhile, the Bayesian information quantity (BIC) and the HQ value should also be considered to be the smallest. As shown in the statement in Figure 5.8, Table 5.2 can be obtained. Through comparison, it can be concluded that the ARIMA (7,1,0) model is more appropriate.

```
arma_mod20 = sm.tsa.ARMA(D_data,(7,0)).fit()
print(arma_mod20.aic,arma_mod20.bic,arma_mod20.hqic)
arma_mod30 = sm.tsa.ARMA(D_data,(0,1)).fit()
print(arma_mod30.aic,arma_mod30.bic,arma_mod30.hqic)
arma_mod40 = sm.tsa.ARMA(D_data,(7,1)).fit()
print(arma_mod40.aic,arma_mod40.bic,arma_mod40.hqic)
arma_mod50 = sm.tsa.ARMA(D_data,(8,0)).fit()
print(arma_mod50.aic,arma_mod50.bic,arma_mod50.hqic)
```

Figure 5.8 Model order comparison chart

Table 5.2 Comparison Table of AIC, BIC and HQ Values

AIC	BIC	HQ
6777.830756725519	6813.292298000003	6791.902040066418
6840.804899353432	6852.625413111594	6845.495327133732
6778.696014039314	6818.097726566518	6794.33077330698
6779.314208706943	6818.715921234147	6794.948967974608

model test:

According to the DW test method (as shown in Figure 5.9), it is known that when the DW value is significantly close to 0 or 4, there is autocorrelation, while when it is close to 2, there is no (first-order) autocorrelation. It can be seen from the figure that the calculated DW value is 1.992, close to 2, and the model test is reliable.

```
#对模型进行定阶
arma_mod20 = sm.tsa.ARMA(D_data,(7,0)).fit()
#print(arma_mod20.aic,arma_mod20.bic,arma_mod20.hqic)
print(sm.stats.durbin_watson(arma_mod20.resid.values))
```

```
time_xu ×
CONVERGENCE: REL_REDUCTION_OF_F_<=_FACTR*EPSMCH
1.9920784490358372
```

Figure 5.9 Model DW Test

5.4 data forecasting

Based on the original data date, predict the total passenger flow at Station ZD190-01 in the next 20 days, that is, the specific time is from March 21, 2016 to April 10, 2016. Through further review and comparison of the model report, generate the model using the ARIMA function in Python. The passenger flow after 20 days, as well as the standard error and confidence interval,

were calculated based on the forecast() function (as shown in Figure 5.10). Through comparison, the results were in line with the expected values, and the results are presented in Table 5.3.

```
model = ARIMA(D_data, (p,1,q)).fit()
print(model.summary2())      #生成一份模型报告
print("************")
result = model.forecast(20)  #为未来20天进行预测，返回预测结果，标准误差，和置信区间
print(result)
```

```
time_xu
------------------------------------------------------------------
              Coef.   Std.Err.     t      P>|t|    [0.025    0.975]
------------------------------------------------------------------
const          0.3920   26.6164    0.0147  0.9882  -51.7751  52.5592
ar.L1.D.num   -1.1671    0.0507  -23.0271  0.0000   -1.2664  -1.0678
ar.L2.D.num   -1.1108    0.0736  -15.0875  0.0000   -1.2551  -0.9665
ar.L3.D.num   -1.2239    0.0847  -14.4427  0.0000   -1.3900  -1.0578
ar.L4.D.num   -1.1518    0.0878  -13.1231  0.0000   -1.3239  -0.9798
ar.L5.D.num   -0.8002    0.0847   -9.4432  0.0000   -0.9663  -0.6342
ar.L6.D.num   -0.5947    0.0740   -8.0400  0.0000   -0.7397  -0.4497
ar.L7.D.num   -0.2190    0.0511   -4.2832  0.0000   -0.3193  -0.1188
------------------------------------------------------------------------
               Real          Imaginary          Modulus          Frequency
------------------------------------------------------------------------
AR.1           0.6100           -0.9161           1.1006           -0.1565
AR.2           0.6100            0.9161           1.1006            0.1565
AR.3          -0.1098           -1.2539           1.2587           -0.2639
AR.4          -0.1098            1.2539           1.2587            0.2639
AR.5          -1.0254           -0.6145           1.1954           -0.4141
AR.6          -1.0254            0.6145           1.1954            0.4141
AR.7          -1.6649           -0.0000           1.6649           -0.5000
================================================================
```

Figure 5.10 Model Report

Table 5.3 Prediction Results of ZD190-01 Station from March 21, 2016 to April 10, 2016

date	num	date	num
2016/3/21	20569	2016/4/1	19049
2016/3/22	23776	2016/4/2	23145
2016/3/23	19420	2016/4/3	34181
2016/3/25	21025	2016/4/4	32592
2016/3/26	18608	2016/4/5	24577
2016/3/27	24300	2016/4/6	38511
2016/3/28	20964	2016/4/7	22567
2016/3/29	22388	2016/4/8	20415

2016/3/30	21596	2016/4/9	21140
2016/3/31	36513	2016/4/10	25078

6. Conclusions and Suggestions

6.1 Conclusion Summarizes The Whole

By processing, importing and analyzing the obtained raw data, the ARIMA model was established and verified. The model has a good effect and can predict the passenger flow of the station in the future period of time, achieving the expected goal. During the completion process of the entire thesis, there were significant problems in the processing of the initial data. Some date data were missing and there were issues with the data table format. For the one-time program import of data, it is accomplished through multiple nested loops. There are problems such as program design and data acquisition in the process of data analysis and modeling. With the patient guidance of the teacher and my own extensive research, all the problems were solved.

6.2 Relevant suggestion

Based on the results of data analysis, it can be found that the passenger flow in normal times follows certain patterns. According to the model, the passenger flow in the next week or two weeks can be reasonably predicted, facilitating the railway bureau to make corresponding adjustments. For large stations with a population of over one million, emergency situation plans should be well prepared, etc. For holidays, make reasonable plans by referring to historical data; By analyzing the changes in passenger flow at different stages, summarize the problems existing in services and planning, make timely adjustments, and optimize both railway planning and resource utilization.

References

[1] Li Jie, Peng Qiyuan, Yang Yuxiang. Passenger Flow Prediction of Guangzhou-Zhuhai Intercity Railway Based on SARIMA Model [J/OL]. Journal of Southwest Jiaotong University :1-11[2020-02-14].

[2] Duan Ran, Pang Jianhua, Zhang Liangjun. Research on Passenger Flow Prediction of Railway Stations Based on SARIMA Model [J]. Practice and Understanding of Mathematics,2019,49(09):1-10.

[3] Shuai Minwei. Research on Optimization Scheme of High-Speed Railway Trains Based on Dynamic Passenger Flow Distribution [D] Tianjin Polytechnic University, 2019.

[4] Zhang Lin. Research on Metro Passenger Flow Prediction System Based on Deep Neural Network [D]. Beijing Jiaotong University,2019.

[5] Cheng Qiang. Research on the Ticket Allocation Model of Intercity High-Speed Railway Based on Passenger Flow Prediction [D]. Chongqing Jiaotong University,2018.

[6] Yuan Wenjun. Research on Hybrid Prediction Model of Railway Passenger Flow Based on Data Characteristics [D]. Guangxi Normal University, 2018.

[7] Pan Shan. Railway Passenger Flow Prediction Based on Time Series [D]. South China University of Technology,2017.

[8] Ma Peixian. Research on the Impact of Service Level Improvement in High-Speed Railway Train Operation Plans on Passenger Flow [D] Beijing Jiaotong University,2018.

[9] Xian Min. Research on Railway Passenger Volume Prediction Method Based on Grey Neural Network [D]. Southwest Jiaotong University, 2016.

[10] Wang Weiwei. Research on Railway Passenger Flow Prediction under the Influence of High-Speed Railway [J]. Railway Transport and Economy,2016,38(04):42-64+51.

[11] Cold, Warm, Warm. Passenger Flow Prediction and Analysis of High-Speed Railway Based on Analogy Method [D]. Beijing Jiaotong University,2015.

[12] Fan Dongxue. Research on Railway Passenger Flow Prediction Method Based on Optimized Grey Markov Chain Model [D] Chongqing Jiaotong University,2015.

[13] Cao Cheng. Short-term Passenger Flow Prediction of High-Speed Railway Based on EMD-BPN Method [D] Lanzhou Jiaotong University,2015.

[14] Lu Xiaojuan, Ma Baofeng, Zhang Wujuan. Prediction Analysis and Research on Railway Passenger Flow [J]. Journal of Lanzhou Jiaotong University,2013,32(06):28-31.

[15] Xia Qing. Analysis of the Fluctuation Pattern of Railway Passenger Flow during Holidays and Its Application in Passenger Flow Prediction [D] Beijing Jiaotong University,2011.

[16] Chang Guozhen, Zhang Qiandeng. Railway Passenger Flow Prediction in China Based on Time Series [J]. Statistics and Consulting,2008(04):20-21.

[17] Ruben van Loon,Piet Rietveld,Martijn Brons. Travel-time reliability impacts on railway passenger demand: a revealed preference analysis[J]. Journal of Transport Geography,2010,19(4).

[18] Nataša Glišović,Miloš Milenković,Nebojša Bojović,Libor Švadlenka,Zoran Avramović. A hybrid model for forecasting the volume of passenger flows on Serbian railways[J].

Operational Research,2016,16(2).

[19] Sukmun Oh,Seongho Kim,Jaisung Hong. Comprehensive analysis of the influence of door width on the passenger flow time on Korean urban railways[J]. Proceedings of the Institution of Mechanical Engineers,2016,230(6).

[20] Genetic Algorithms; Recent Research from University of Pardubice Highlight Findings in Genetic Algorithms (A hybrid model for forecasting the volume of passenger flows on Serbian railways)[J]. Computers, Networks & Communications,2016.

[21] Feng Shi,Zhao Zhou,Jia Yao,Helai Huang. Incorporating transfer reliability into equilibrium analysis of railway passenger flow[J]. European Journal of Operational Research,2012,220(2).