

需提交的资料包括以下三部分，请分别压缩后在“会员中心”统一提交：

- 1、论文正文（pdf 格式），提交时间：2016 年 5 月 15 日 24:00 前；
- 2、论文正文（doc 格式）、源数据（组委会提供的源数据外）、过程数据、程序及模型文件，压缩成“附件资料.zip”，提交时间：2016 年 5 月 16 日 24:00 前；
- 3、参赛承诺书打印后签名，并扫描生成 pdf 文件，提交时间：2016 年 5 月 15 日 24:00 前。
- 4、为确保作品能提交且不遗漏，建议大家在此处提交作品的同时，另外再将论文正文.pdf、参赛承诺书.pdf、附件.zip 通过邮件发送到组委会专用邮箱：13570264389@qq.com，请在邮件中注明报名信息：队号、队长名、学校、手机等。

所选题目： B 题铁路旅客流量预测研究

综合评定成绩：

评委评语：

评委签名：

铁路旅客流量预测研究

摘要：为了保持市场竞争力，实现利润最大化，需要对客流进行充分的了解及预测，从而为制定合理的价格、改善客运站组织方式、优化铁路车辆资源配置、提高客运设备的服务能力提供帮助。

首先采用 VB.net 编程对原始数据进行结构化处理，其次对结构化的数据进行预处理，最后分类分析各因素与客流量之间的关系，提取有用信息建立预测模型并优化调整停靠方案。

通过对铁路公司的 ZD190（站）至 ZD111（站）区段的客运专线进行短期客运量的分析，按车次、时段（小时）、车站、区间、天气以及车型等方面分析客流规律，采用相关系数的方法得出天气、气温等因素对短期客流量影响不大，同时得知关联系数较高的因素——上下车人数的时序特征和变化规律，表明短期客运量时间序列是一种非平稳时间序列。

在此基础上，建立 ARIMA 模型（时间序列预测模型），采用 R 软件编程对未来两周铁路旅客流量进行预测；与此同时，利用贝叶斯正规化的 BP 神经网络模型，采用 MATLAB 软件编程同样对未来两周铁路旅客流量进行预测，结果表明贝叶斯正规化的 BP 神经网络模型具有更好的预测精度，适宜做铁路旅客流量的短期客流预测。

对于 D02-D19 的客座率较低以及管内有些站点并无此类车次停靠的问题，我们从两种情况考虑：一种是对单车次而言，时间不变、停靠站变化；二是对于整体而言，发车时间不变，改变停靠站（站点的到发时间可能改变）。由于题目中没有给出票价信息，而票价与客座率呈正相关，我们采用改变停靠方案和车辆配置充分利用资源尽可能满足周转旅客需求同时提高客座率的方式，建立 0-1 整数规划模型使得每个站点都有车辆停靠。鉴于此，首先考虑单次车的停靠，这样能有效提高客座率，但有些车次不能满足旅客周转需求，也不能保证每个站点都有停靠。接着我们从整体考虑得出的停靠方案能保证每个城市每个站点都有停靠，通过 lingo 软件编程求解得到新的停靠方案：有 6 个车次路线没有改变，其余车次路线稍有变动，根据最后的停靠方案路线以及距离等重新规划发车停车时刻表。

从数据处理的过程和分析的结果来看，各因素的时序特征较为合理，从模型角度来看：预测模型中神经网络模型虽然操作难度较大但精度和效果较好更贴近实际；优化方案中第一种情形求解时采用了简化的方式搜索求解，得出的结果不一定是最优，但也能达到提高客座率的效果，第二种情形求解结果虽然是局部最优解，但整体方案效果较好，能提高客座率，满足更多旅客的需求。建议铁路部门可以根据预测结果采用优化停靠方案以此来提高效益。

关键词：铁路旅客流量；短期客流预测；ARIMA 模型；贝叶斯正规化；神经网络；0-1 整数规划

Research of Prediction of Railway Passenger Flow

Abstract: In order to keep market competitiveness and realize the profit maximization, we need to fully understand and predict the flow of passenger. It will offer help for formulating the reasonable price, improving the passenger organization mode, optimizing the resources distribution of railway vehicle, and improving the service ability of passenger equipment.

Firstly, we structured processing of original data using .net programming, and then preprocess the structured data. Finally, we categorical analyze the relationships between various factors and passenger flow, and extract the useful information to establish a prediction model and adjustment the stop program.

Through analysing the short-term passenger flow of the passenger dedicated line which is located in ZD190 (station) to ZD111 (station), and analysing the rules of passenger flow according to trips, period of time (hours), stations, interval, weather and the vehicle, using correlation coefficient method, we conclude that weather and temperature have little effect on the short-term traffic. And at the same time we get that the short-term passenger volume time series is a kind of nonstationary time series according to the correlation coefficient higher factors -- on and off timing characteristics and change law of the number.

On this basis, we establish ARIMA model (time series forecasting model), and forecast the passenger flow in the next two weeks using R software programming. At the same time, we do it using the Bayesian regularization BP neural network model, and MATLAB software programming. The results show that the later method has a better prediction precision, and it is suitable for predicting the railway passenger flow of short-term.

For the problem of lower rate of load factors in D02-D19 and no such trips dock in some sites, we consider from the following two cases: one is for cycling times, time is the same and docking stations are changed. The second is for the whole, departure time invariant and change the dock station (arrival and departure time may change the site).

Since there is no fare information in the title, and fares and guest rate are positively correlated, we adopt the way of changing the dock schemes and vehicle configuration to make full use of resources as far as possible to meet passenger demand and at the same time, improve the rate of guest, and establish the 0-1 integer programming model so that each site has parked vehicles.

In view of this, we first consider the single car dock, which can effectively improve the visiting rate. But some trains can not meet the needs of passenger turnover, and it can't guarantee that each site has docked. Then the program which we consider from the whole drawn ensure that every site in each city has docked. By lingo software programming, we obtain a new docking program that there is no change in 6 train route, and the rest of the train route has changed slightly, we replan start stop time table according to the last stop route and distance.

From the data process and analysis results, the temporal characteristics of each factor is more reasonable. from the model point of view: the operation of neural network model in prediction model is more difficult, but the precision and effect of it are closer to the actual. In the first case of optimization scheme, it is solved by the simplified way search algorithm, and the result is not necessarily the optimal, but it can improve the effect of guest rate. In the second case, although the solving results is the local optimal solution but overall program effectiveness is better and it can improve the guest rate and meet the demand for more passengers. It is suggested that the railway department can improve the efficiency by optimizing the berth plan according to the forecast results.

Key words: railway passenger flow; short-term passenger flow forecasting; ARIMA model; Bayesian regularization; neural network; 0-1 integer programming

目录

1. 挖掘目标.....5

2. 分析方法与过程.....5

2.1. 总体流程.....5

2.2. 具体步骤.....6

2.2.1. 数据结构化.....6

2.2.2. 数据预处理.....6

2.2.3. 数据分析.....8

2.2.4. 预测模型的建立与求解.....17

2.2.5. 车辆停靠优化设计.....23

2.3. 结果分析与反馈.....32

2.3.1. 结果解析.....32

2.3.2. 模型评价.....33

3. 结论.....34

4. 参考文献.....34

1. 挖掘目标

在经济发展进入新常态的背景下，铁路部门也在不断深化铁路投融资改革。铁路作为大众化运输工具是我国国民经济的大动脉，与社会生产、人民生活质量和福利水平息息相关，所以如何科学合理的预测客流量，以提高铁路运营组织效率，保证铁路运行畅通，具有重要意义。

本文通过运用相关系数对天气因素分析得出其对短期客流量影响不大，所以对铁路公司的 ZD190（站）至 ZD111（站）区段的客运专线短期客运量的分析，分别按车次、车型、时段（小时）、车站等条件分析客流规律，得出其时序特征和变化规律。在此基础上，我们还利用了两种数据挖掘工具（ARIMA 模型和贝叶斯正规化的 BP 神经网络模型）去对比探索找到最好的预测模型，最后建立 0-1 整数规划模型，调整停靠方案，制定发车时间，充分利用管内有效资源满足管内更多旅客需求，减少管外车次的压力。

2. 分析方法与过程

2.1. 总体流程

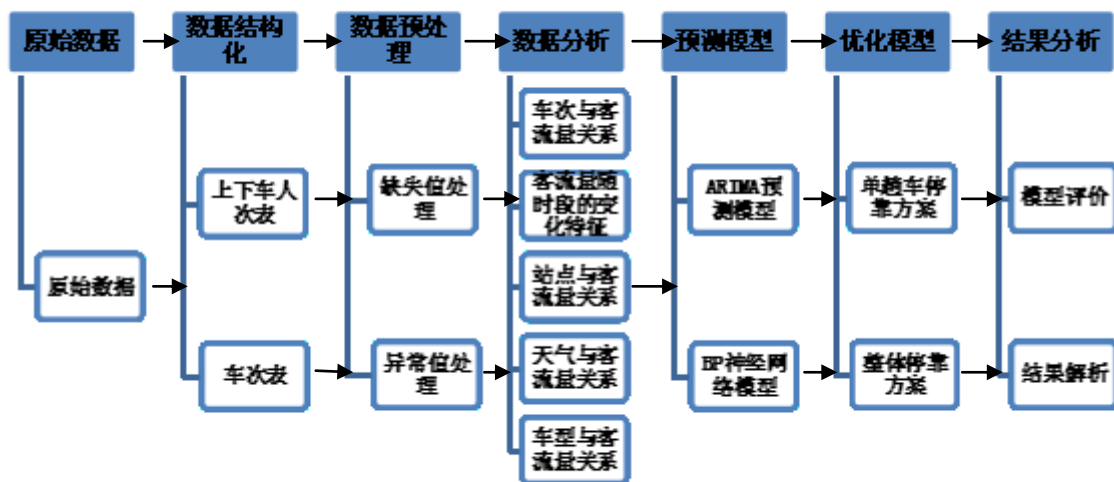


图 1 总体流程图

本论文的分析流程可大致分为以下六步：

- 步骤一：利用 VB.net 编程使原始数据结构化，得到上下车人次表及车次表；
- 步骤二：对获取的数据进行预处理，包括缺失值处理、异常值处理等；
- 步骤三：对数据进行分析，包括对车次、时段、站点、天气、车型与客流量关系分析；
- 步骤四：利用 ARIMA 模型、神经网络模型进行预测；
- 步骤五：针对 D02—D19 设计单趟车停靠优化方案和整体停靠优化方案；
- 步骤六：结果分析与反馈。

2.2. 具体步骤

2.2.1. 数据结构化

由于原始数据是很多张梯形密度表（见附件 1），因此我们首先使用 VB.net 程序将原始数据进行结构化处理(图 2)。



图 2 VB.net 程序运用窗口

得出两张数据表——上下车人次表及车次表（见附件 2）。随后对结构化的数据进行筛选，分出管内（ZD190 至 ZD111 区段）、管外的客运专线数据，并进行分析、验证、检查。

2.2.2. 数据预处理

数据预处理主要包含三个步骤：

第一步是对结构化的数据补录所缺失日期的数据；

第二步是对数据进行异常值处理；

第三步是提取并处理分析及建模需要用到的数据。

- 对结构化的数据补录所缺失日期的数据

由于原始数据缺失某些日期的上下车人次的数据，因此我们采用计算算术平均值的方法填充缺失值，具体实施过程如下：

通过用站点一年的总人数的算术平均值来补录每个缺失日期的上下车人次数据。如图 3：

算术平均值公式如下：

$$\text{缺失日期的上下车总人数} = \frac{\text{15个月上下车总人数}}{\text{15个月总天数}}$$

	A	B	C	D	E	F	G
83	20150323	3826	11633				
84	20150324						
85	20150325	4256	12119				
86	20150326	4227	11945				
87	20150327	5531	13244				
88	20150328	5435	12892				
89	20150329	5272	13237				
90	20150330						
91	20150331						
92	20150401	3643	12101				
93	20150402	5069	15330				
94	20150403	9987	17817				
95	20150404	7563	14736				
96	20150405	6502	13203				

图3 处理缺失数据

● 对数据进行异常值处理

由于原始数据存在异常值，我们对异常值进行修改，具体实施过程如下：

通过筛选车次，再对异常时间进行查找替换，来处理异常值，如图4：

	A	B	C	D	E	F
92	ZD326	16:13	147			
93	ZD250	16:51	104	19		
94	ZD062	17:20	93	11	37	
95	ZD190-01	18:12	68	13	15	
96	上车人数合计		412	43	52	
97	D13 ZD190-01-ZD111-01 开行 1 天 日均定员: 506 客座率: 44.3 %					
98	上车站		ZD190-01	ZD250	ZD192	ZD326
99	下车站	开到点	13:37	14:51	14:01	15:28
100	ZD190-01	13:37				
101	ZD250	14:50	33			
102	ZD192	13:52	154	56		
103	ZD326	15:27	25	14	28	
104	ZD111-01	16:16	38	41	68	
105	上车人数合计		250	111	116	
106	D14 ZD111-01-ZD190-01 开行 1 天 日均定员: 1172 客座率: 18.0 %					
107	上车站		ZD111-01	ZD326	ZD250	ZD062
108	下车站	开到点	17:12	17:59	18:37	18:28
109	ZD111-01	17:12				
110	ZD326	17:54	100			
111	ZD250	18:35	81	10		
112	ZD062	18:27	24	12	29	
113	ZD190-02	19:34	22	6	16	
114	ZD190-01	19:57	50	13	17	
115	上车人数合计		277	41	62	
116	D15 ZD190-01-ZD111-01 开行 1 天 日均定员: 1172 客座率: 43.9 %					
117	上车站		ZD190-01	ZD190-01	ZD250	ZD326

图4 异常值处理

● 提取并处理分析及建模需要用到的数据（主要包括对每个站点上下车人次处理，天气因素处理）

◇ 对每个站点上下车人次总人次处理

在 Excel 上下车人次数据表中，对管内每个站点每一天上下车人次进行加总，实施过程如图 5。

	A	B	C	D	E	F	G	H	I	J
1	车次	始发日期	到站时间	发车时间	上车人数	下车人数	站点名称			
2	G04	20150101	14:05	14:06	17	32	ZD022			
3	G09	20150101	15:15	15:16	29	87	ZD022			
4	D04	20150101	9:30	9:32	98	268	ZD022			

	A	B	C	D	E	F	G	H	I	J
1	站点	始发日期	上车人数	下车人数						
2	ZD022	20150101	144	387						
3		20150102	95	79						
4		20150103	211	65						

图5 管内站点每日人数加总

得到每个站点每一天上下车总人次表（见附件 3），同样对管外数据每一天上下车总人次进行加总，得到管外数据每一天上下车总人次表（见附件 3）。

✧ 天气因素处理

在 Excel 原始数据车站所属地区气象信息数据中,对天气状况对旅客出行的影响程度作出划分,得出天气状况对旅客出行的影响,如表 1。

表 1 天气状况对旅客出行的影响

天气状况	程度分类	影响程度
晴	1	完全不影响
多云、阴、雾、霾等	0.8	不影响
小雨、小雪等	0.6	轻度影响
中雨、中雪等	0.4	中度影响
(雷)阵雨、暴雨、暴雪等	0.2	严重影响

同样对对风力风向对旅客出行的影响作出划分,得出风力风向对旅客出行的影响,如表 2:

表 2 风力风向对旅客出行的影响

风力风向	程度分类	影响程度
风力 ≤ 3 级	1	不影响
风力 > 3 级	0.5	影响

对原始数据车站所属地区气象信息数据中的气温进行最高气温与最低气温取算术平均值的方法,处理出温度对旅客出行的影响,得出温度对旅客出行的影响如表 3:

表 3 温度对旅客出行的影响

气温	程度分类	影响程度
18 以下	0.5	不适宜
18~24	1	适宜
24 以上	-0.5	不适宜

2.2.3. 数据分析

1. 车次与客流量关系

为了方便旅客更好的出行,每年中国铁路总公司都会根据客流数据统一调整,更新列车运行图。其中主要包括:增加车次、停运车次、修改车次等,所以研究车次与客流量的关系是非常有

必要的。我们根据已有的数据，首先将每一个车次一年的平均客座率算出（见附件 4：分析车次），其中客座率 Z_k 计算公式为：

$$Z_k = \frac{\sum_{i_k j_k} y_{i_k j_k} s_{i_k j_k}}{r_k \cdot \max \{s_{i_k j_k}\}}$$

记 $s_{i_k j_k}$ 为第 i_k 站到 j_k 站的距离， r_k 为日均定员，再取算术平均客座率 $\bar{Z}_k$ 。

作出不同车次对应的平均客座率的散点图，反映其车次与客流量的关系，如图 6。

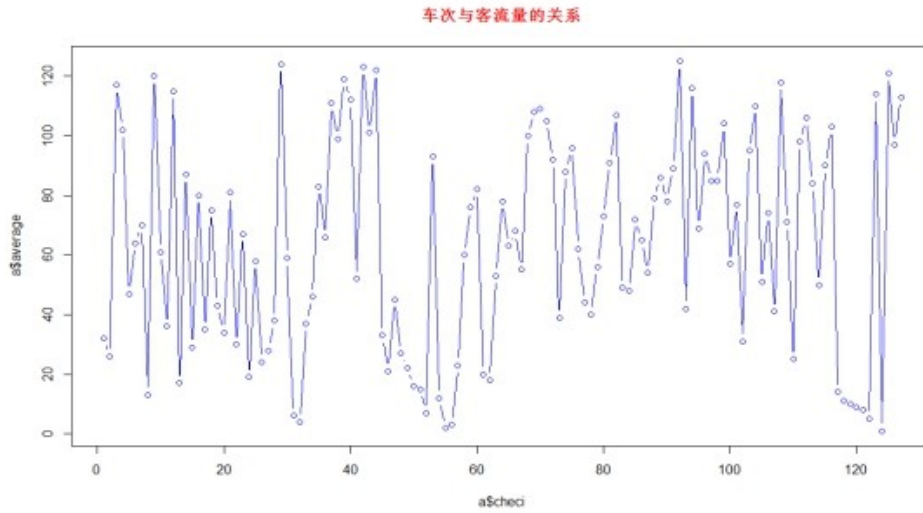


图 6 车次与客流量关系

由图可知，503、D01、D04、D21、D23、D24、D31、D34、D36、D47、D48、G30、G32、K11、Z01 这些车次平均客座率均高达 90%以上，可以增加车次，而 D45，T08 车次平均客座率在 10%以下，可以考虑减少车次或者用其他车次代替，可以进一步研究与其车次有显著相关的其他车次。

2. 客流量随时段的变化特征

对已有客流数据进行统计分析和数据挖掘，把握铁路客流与时段的相关关系，对于研究铁路短期客流量预测具有重要意义。我们将一天分为四个时段，分别为 0：00-6：00，6：00-12:00,12:00-18：00，18：00-24：00（见附件 4：分析时段）。每个时段上（下）车总人数计算公式如下：

每个时段上（下）车总人数=在这个时段区间包含所有上（下）车人数总和，图 7—图 10 反映了其时段与铁路客流量的关系。

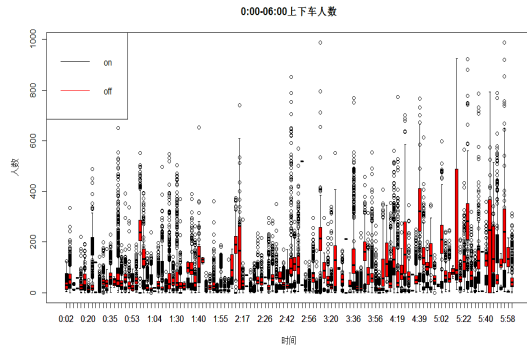


图 7 0:00-6:00

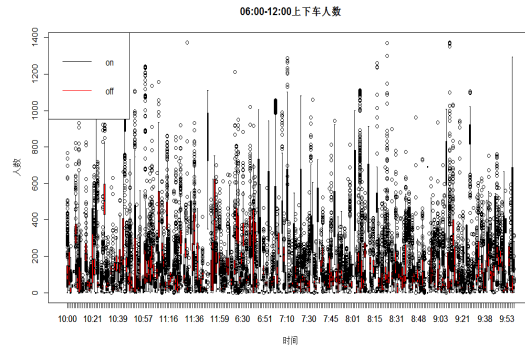


图 8 6:00-12:00

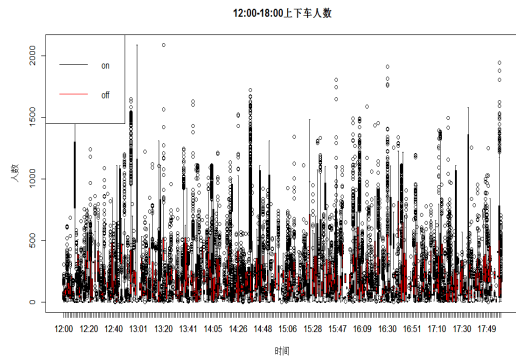


图 9 12:00-18:00

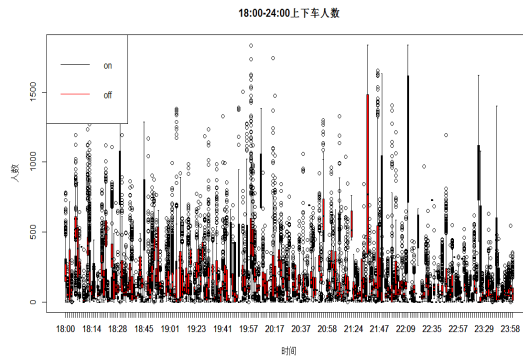


图 10 18:00-24:00

从图中可以看出，各个时段的上下车人数都有一定的差异性，整体来看，一天 24 小时中，早上 6: 00-12: 00 这个时段客流量最密集，其次是 12: 00-18: 00，然后是 18: 00-24: 00，客流量最少的时段是 0: 00--06: 00。就拿 0: 00--06: 00 这个时段来具体分析，它的客流比较小，但是在接近 6: 00 的时候客流量就开始剧增。我们猜想这些变化主要是受到了客流性质的影响，其中客流性质主要受出行目的所决定，包括公务、上班、上学、探亲访友、回家、旅游、经商等，不同性质客流出行时段的要求不同，公务、上班、上学对时段有刚性要求，其他客流对时段要求具有一定弹性，受列车开行方案的影响较大。同时车站客流需求总量也是影响车站高峰小时客流量最直接的因素，也与车站吸引范围的人口，经济发展水平结构，交通设施条件以及在区域经济中的地位有关。

3. 站点与客流量关系

近几年来，在全社会客运量逐年上升的同时，我国铁路旅客发送量和旅客周转量也保持着增长态势，旅客平均运输距离逐渐延长，这些变化不仅受到车型和时段的影响，还受到站点的影响。所以研究站点对于铁路客流量是非常有必要的，其中车站吸引范围的人口，经济发展水平结构和交通设施条件都是站点的影响因素。

我们将各个站点关于每一天的上下车总人数（见附件 4：分析站点）按时间顺序依次建立在同一张图形中。其中，上车总人数我们按照模糊聚类的思想分了四类：很多，多，一般，少。图 11—图 14 分别为不同站点对应的全年每一天的上车总人数。其中将 ZD250,ZD111-01,ZD190-02 为很多的一类，客流量从大到小排列顺序为：ZD190-02、ZD111-01、ZD250。出现这个结果的原因可能在于这几个车站吸引范围的人口基数大，经济发展水平结构比较高和交通设施条件也比较好，其中 ZD190-02 最多，可以考虑在这个站点增加车次。其次 ZD062,ZD192,ZD326 为多的一类，客流量从大到小排列顺序为：ZD326、ZD192、ZD062,其中 ZD192 与 ZD326 随时间走势比较相似。接着，ZD370,ZD143，ZD111-02 分为一般的一类，其中客流量从大到小排列顺序：ZD111-02、ZD143、ZD370。ZD022,ZD120,ZD121 分为少的一类，其中客流量从大到小排列顺序：ZD022、ZD120、ZD121。ZD121 的客流量最小，我们可以考虑在这个站点减少一些车次。

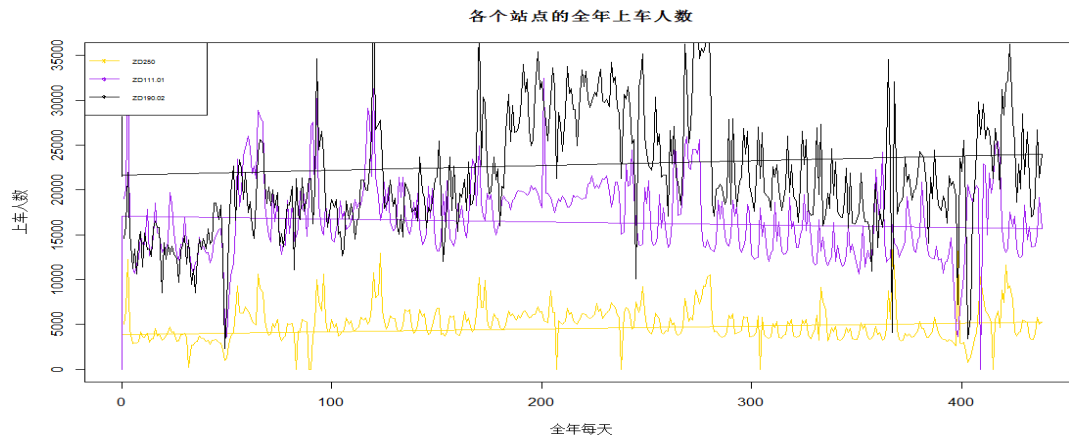


图 11 各站点全年上车人数（类为很多）

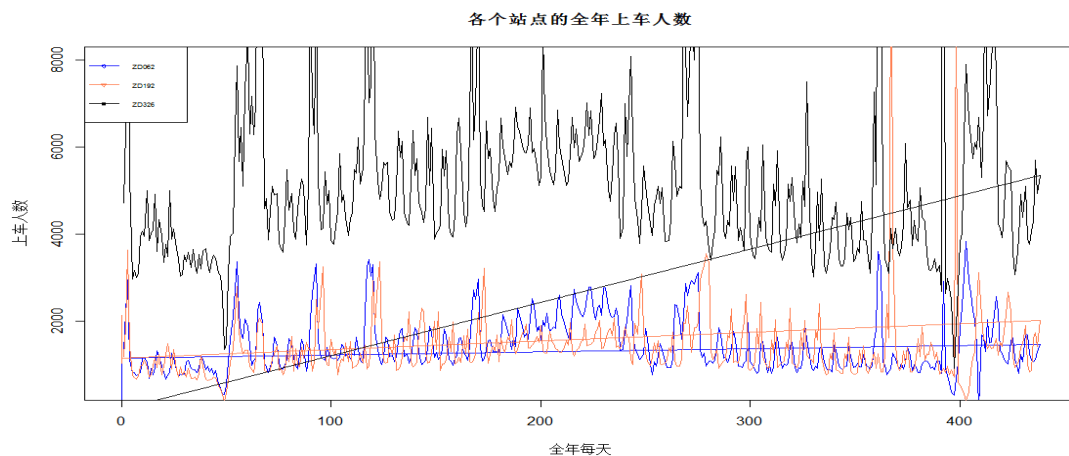


图 12 各站点全年上车人数（类为多）

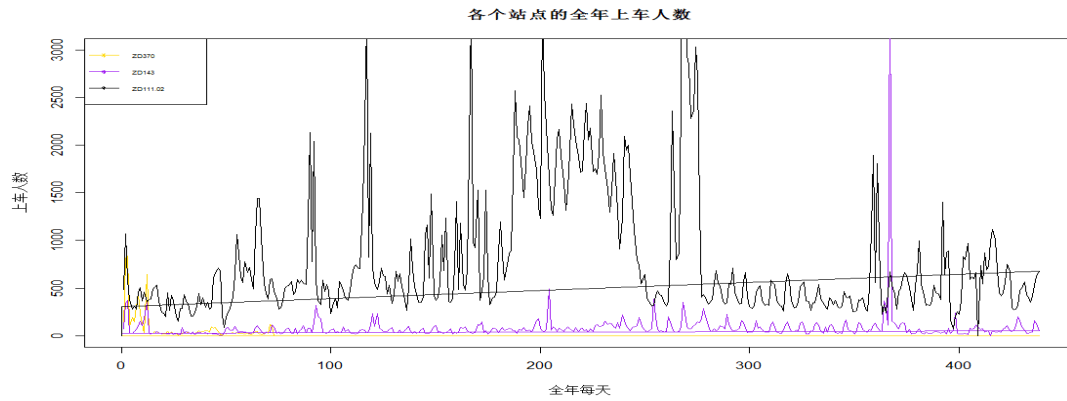


图 13 各站点全年上车人数（类为一般）

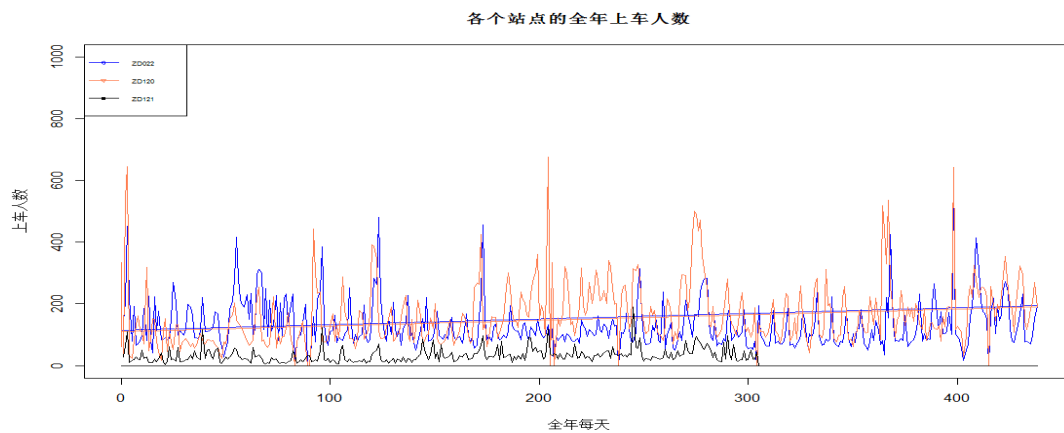


图 14 各站点全年上车人数（类为少）

同样地，下车总人数我们也分了四类：很多，多，一般，少。图 15-图 18 可以看出：ZD111-01，ZD250，ZD326 分为很多的一类，其中 ZD111-01 每天的下车人数最多，当然，也是最不稳定的，波动幅度大，说明受时间的影响比较大。整体来看，这几类站点下车总人数走势相似性也比较高。其中，客流量从大到小排列顺序为：ZD111-01、ZD250、ZD326,由此看来可以考虑在这些站点中增加车次，减少不必要的拥堵。其次 ZD062,ZD192,ZD111-02 分为下车总人数多的一类，也可以考虑增加一些车次。客流量从大到小排列顺序为：ZD192、ZD062、ZD111-02、ZD190-01，下车总人数一般，客流量大小稳定在 500-1500 人次之间。ZD022，ZD120，ZD121，ZD143，ZD370 为下车总人数少的一类。客流量从大到小排列顺序为：ZD120、ZD121、ZD143、ZD370、ZD022。其中 ZD022 最少，总体每天不超过 500 人次，可以考虑减少或者停止此类站点的车次。

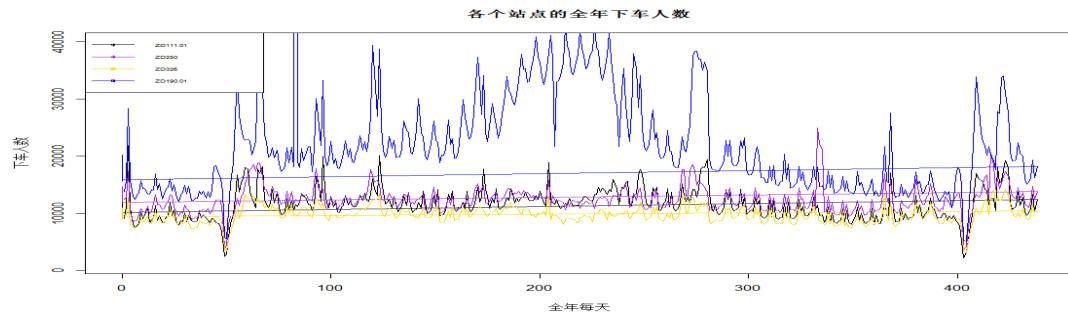


图 15 各站点全年下车人数（类为很多）

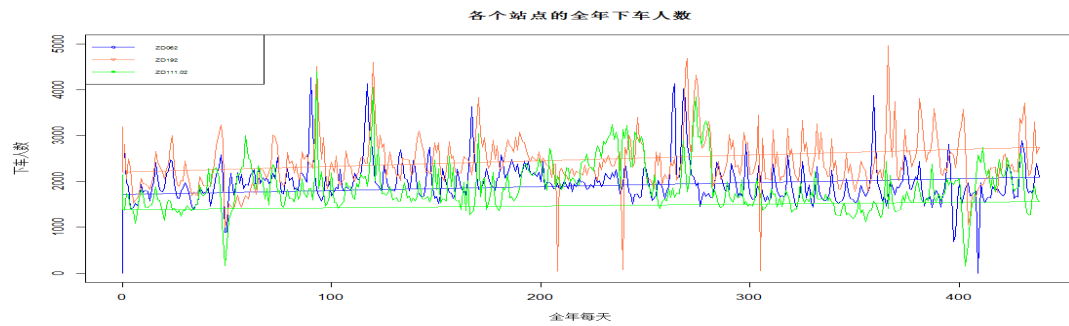


图 16 各站点全年下车人数（类为多）

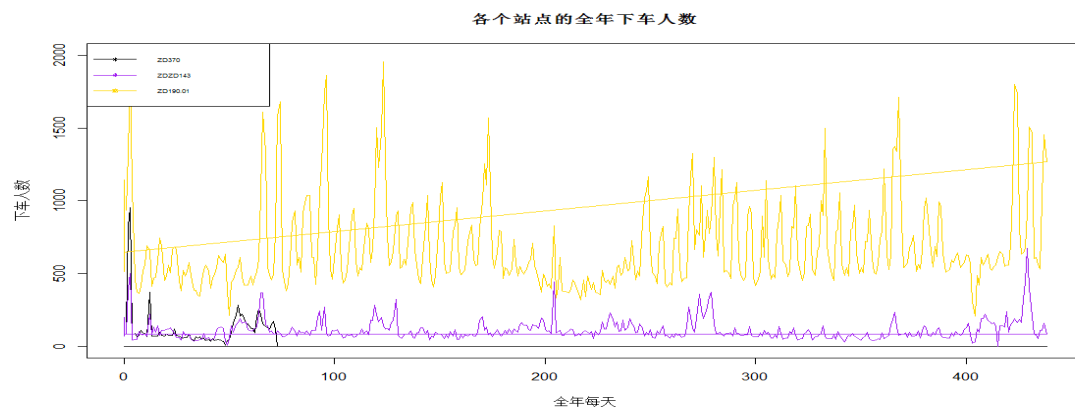


图 17 各站点全年下车人数（类为一般）

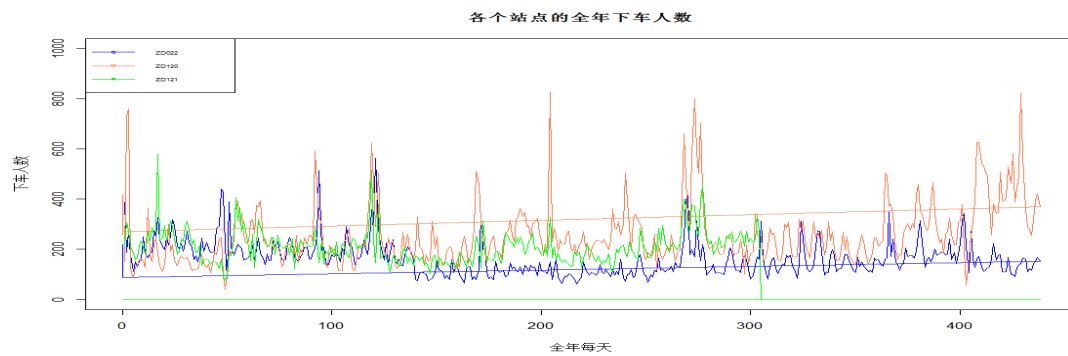


图 18 各站点全年下车人数（类为少）

4.天气与客流量关系

相关系数是用来描述定量变量之间的关系。我们以分析 ZD326 市 15 个月的天气因素为例，其中部分数据如表 4（其余市的天气因素见附件 4：天气因素）。

表 4 D326 市 15 个月的天气因素

station	date	on	off	temperature	weather	wind
ZD326	20150101	4717	9918	0.5	1	1
ZD326	20150102	6396	9728	0.5	1	1
ZD326	20150103	12689	12780	0.5	1	0.5
ZD326	20150104	5363	8050	0.5	0.6	1
ZD326	20150105	2980	7445	0.5	0.8	0.5
ZD326	20150106	3166	7480	0.5	0.8	1
ZD326	20150107	3000	7548	0.5	1	1
ZD326	20150108	3109	7908	0.5	1	1
ZD326	20150109	3935	9095	0.5	1	1
ZD326	20150110	4071	8594	0.5	1	1
ZD326	20150111	3952	8461	0.5	1	1
ZD326	20150112	5005	9393	0.5	1	1
ZD326	20150113	3864	8168	0.5	0.8	1
ZD326	20150114	4024	8279	0.5	0.8	1

我们通过 R 软件对数据集进行 Pearson 相关来衡量其相关程度，通过分析 cor 函数计算出各个因素的 Pearson 积差相关系数，相关系数公式如下：

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

其中

$$\begin{cases} \sigma_{xy} = \sigma^2_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{n} \\ \sigma_x = \sqrt{\frac{\sum (x - \bar{x})^2}{n}} \\ \sigma_y = \sqrt{\frac{\sum (y - \bar{y})^2}{n}} \end{cases}$$

于是

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

通过表 4 计算得到结果如表 5。

表 5 相关系数表

cor	on	off	temperature	weather	wind
on	1.0000000	0.5839301	0.2148079	-0.0459618	0.0012503
off	0.5839301	1.0000000	0.1438550	0.0190191	0.0982793
temperature	0.2148078	0.1438550	1.0000000	-0.2653833	-0.0709618
weather	-0.0459618	0.0190191	-0.2653834	1.0000000	0.0042017
wind	0.0012503	0.0982793	-0.0709618	0.0042017	1.0000000

图 19 为可视化的各个因素的相关性：

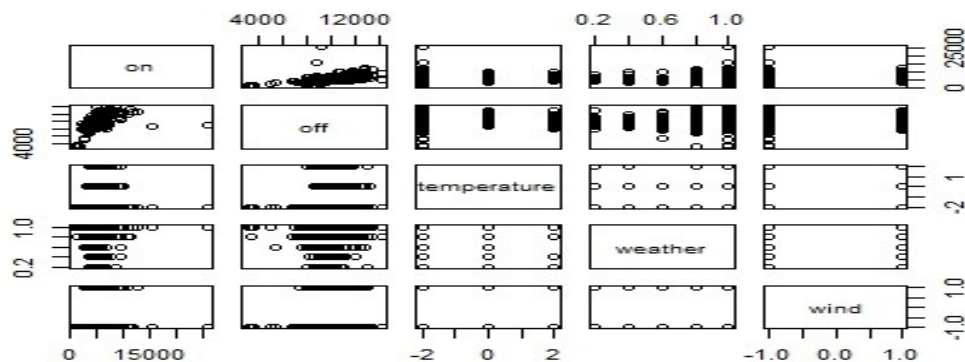


图 19 ZD326 市各因素之间的相关性

通常情况下，我们认为能通过以下取值范围判断变量的相关强度，如表 6。

表 6 相关性判断表

相关系数	相关性
0.8-1.0	极强相关
0.6-0.8	强相关
0.4-0.6	中等程度相关
0.2-0.4	弱相关
0.0-0.2	极弱相关或无相关

我们可以看到上下车人数之间存在一定的相关性，但气象中各个因素和上下车人数之间基本不存在相关性。这是因为在实际中，列车售票会提前 2 个月，而旅客购票时完全不会考虑出行时的天气状况。

5. 车型与上下车人数规律

我们对于车型展开进一步的研究，其方法是按照每个车型分类汇总出各个车型对应的每一个月的上下车总人数（见附件 4：分析车型），具体计算公式如下：

记各个车型对应的每一个月的上（下）车总人数为 E_A ，每月每天上下车总人数为 z_{nk} ：

$$E_A = \sum_{n=1}^N \sum_{k=1}^K z_{nk} \quad (n \text{ 代表月份}, k \text{ 代表日期})$$

再做出每个车型上下车人数与时间的关系图，根据车型种类我们具体分为 G、K、D、T、Z 五类。501—508 车次由于数据太少，我们将这个归因于临时性或者季节性因素，所以直接排除了 501—508 这一类车型。

不同的车型分别刻画了不同的客流规律，首先从整体来看，各个车型存在的共同点是客流量高峰大多聚集在 2015 年 7-8 月，这个可能受到大学生放暑假的影响。而整体客流量在 2016 年 1 月中旬有一次较大的涨幅，这个主要是受到了春运的影响。其次，我们再来分析各个类型车次的客流规律。G 类车型的客流走势比较平缓，说明了月份的变化对于这个车型的影响不大。而 K 类、T 类、D 类车型和 Z 类车型波动比较大，说明月份的变化对于这四个车型的影响占主要地位。接着，通过对比，我们发现客流量从大到小排列顺序为：G 类车型、D 类车型、K 类车型、Z 类车型、T 类车型。由此可以看出 G 类车型比较热门，T 类车型比较冷门。结果如图 20-24 所示：

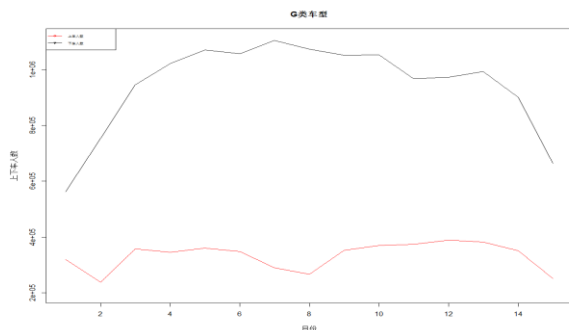


图 20 G 类车次

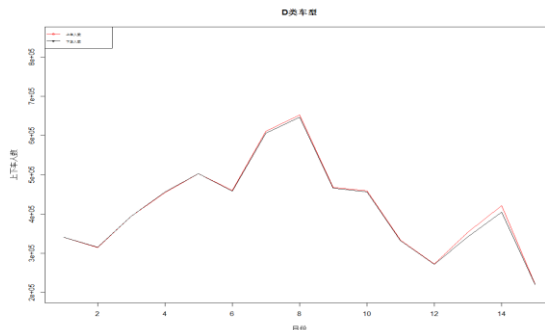


图 21 D 类车次

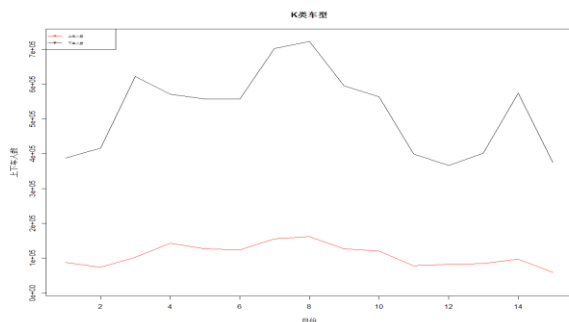


图 22 K 类车次

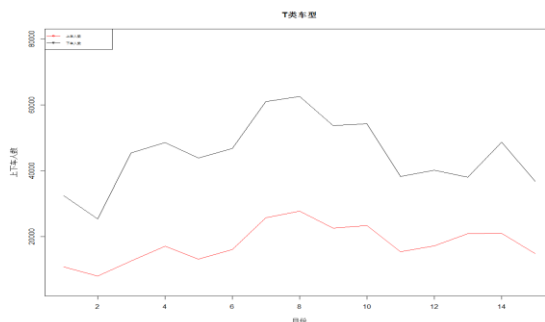


图 23 T 类车次

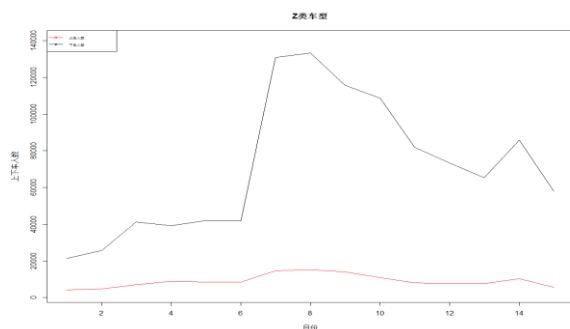


图 24 Z 类车次

2.2.4. 预测模型的建立与求解

1. ARIMA 预测模型^[1]

1) 基本思想

ARIMA 模型全称为差分自回归移动平均模型(Autoregressive Integrated Moving Average Model,简记 ARIMA)，也叫求和自回归移动平均模型，是由博克斯(Box)和詹金斯(Jenkins)于 70 年代初提出的一种著名时间序列预测方法，所以又称为 box-jenkins 模型、博克斯-詹金斯法。其中 ARIMA (p, d, q) 称为差分自回归移动平均模型，AR 是自回归，p 为自回归项；MA 为移动平均，q 为移动平均项数，d 为时间序列成为平稳时所做的差分次数。

所谓 ARIMA 模型，是指将非平稳时间序列转化为平稳时间序列，然后对平稳时间序列建立 ARMA 模型。ARIMA 模型根据原序列是否平稳以及回归中所含部分的不同，可以细分为移动平均过程 (MA)、自回归过程 (AR)、自回归移动平均过程 (ARMA) 以及 ARIMA 过程。通式写作：

$$\phi(B)(1-B)^d X(t) = \theta(B)e(t) \quad (1-1)$$

$d \neq 0$ 时，原序列不是平稳的，因此方差非齐；当我们进行 d 阶差分后，差分后的序列具有方差齐性。而模拟时间序列并不是很常用，一般是用于时间序列的拟合和预测。在 R 中进行时间序列的 ARIMA 建模的步骤是：

- 时间序列的平稳化处理；
- 根据自相关图和偏自相关图选择阶数，建立相应的模型；
- 选好阶数后，进行参数估计；
- 利用建立好的模型做预测分析，并诊断残差序列是否为白噪声。

2) 平稳化处理

如果数据序列是非平稳的，并存在一定的增长或下降趋势，则需要对数据进行差分处理，

如果数据存在异方差，则需对数据进行技术处理，直到处理后的数据的自相关函数（简称 ACF）值和偏相关函数（简称 PACF）值无显著性地异于零。现我们以 ZD326 每天的下车总客流量为例，分析预测经过 ZD326 未来两个星期内的每天下车总客流量，由图 25 可看出 ZD326 每天的下车总客流量是非平稳序列。因此，我们需要对 ZD326 这个数据集进行平稳化处理。

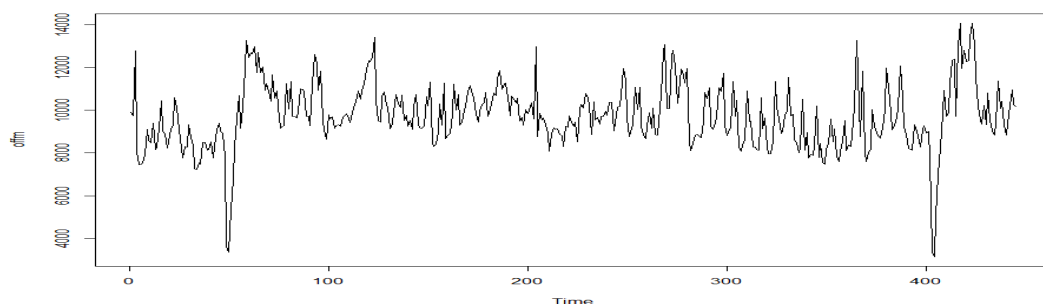


图 25 未经平稳化处理的 ZD326 站点每天的总客流量

首先我们对数据进行一阶差分并绘制差分序列图：

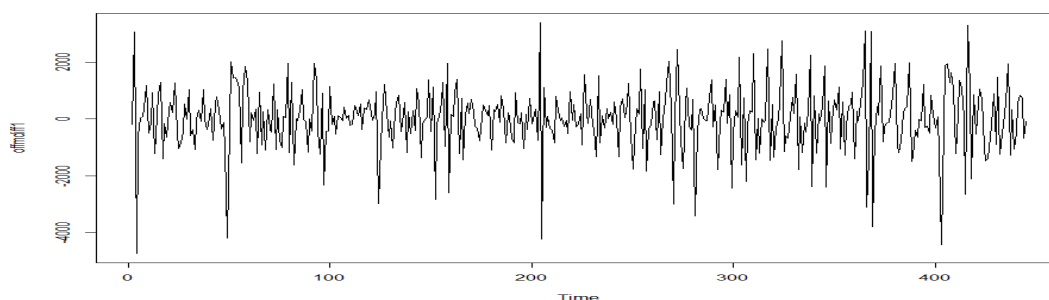


图 26 ZD326 的一阶差分时序图

从图 26 中可以看出，一阶差分时序图趋于平稳，然后我们使用 R 的 `acf` 函数绘制自相关图，来检验差分序列是否平稳，结果如图 27。

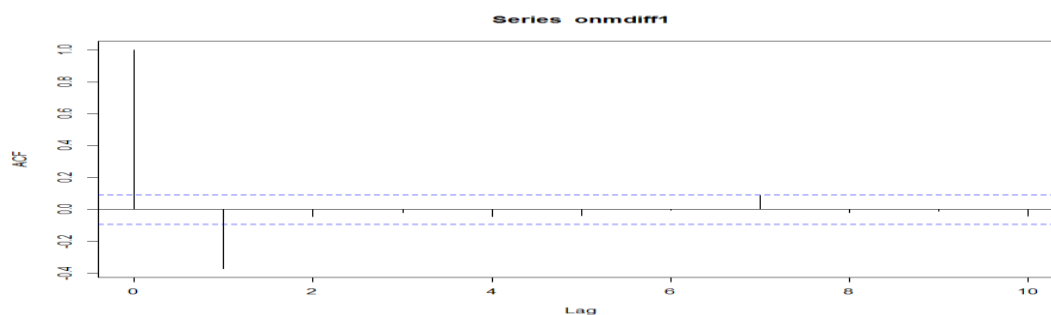


图 27 ZD326 的一阶差分自相关图

从图 27 中可以看出序列的一阶自相关系数略微超过置信界限，但其余各阶自相关均在虚线

以内。

3) 建模并预测

平稳化完成后，要根据时间序列模型的识别规则，对平稳序列建立相应的模型，也就是说，我们要寻找 ARIMA (p,d,q) 中合适的 p 值和 q 值。根据参数估计得到的时间序列模型，模型如下：

$$W_t = -0.1769W_{t-1} + 0.0467W_{t-2} \quad (1-2)$$

利用式 (1-2) 对未来的序列值进行了预测，预测结果如图 28：

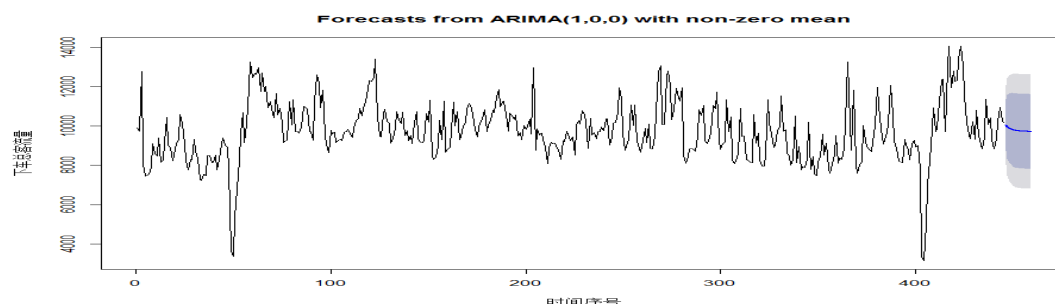


图 28 ZD326 的未来 2 周的预测图

完成 ARIMA 建模的整个过程后，我们对模型进行了假设检验，诊断残差序列是否为白噪声。我们运用自相关图和 Ljung-Box 检验结合，来诊断残差序列的自相关性（图 29）。

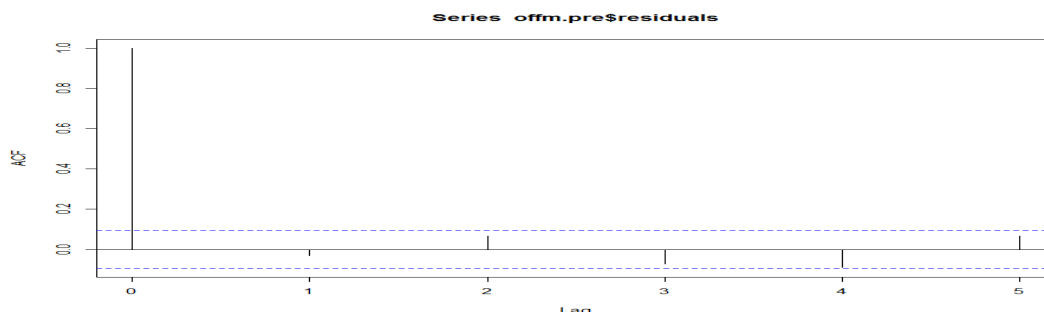


图 29 ARIMA 模型的残差自相关图

再结合自相关图，我们认为此 ARIMA 模型可以通过白噪声检验，适用于对该时间序列的建模和预测。预测结果如表 7。

表 7 时间序列预测结果

第一周	10053	9958	9890	9841	9805	9780	9762
第二周	9749	9740	9733	9728	9725	9722	9721

2. BP 神经网络模型^[5]

1)BP 网络基本思想

BP 网络是一种对非线性可微分函数进行权值训练的多层网络。它的最大特点是仅仅借助样

本数据，无需建立系统的数学模型，就可对系统实现由 m 个输入神经元的模式向量 p 组成的 p_n 空间到 y_n 空间(n 为输出节点数)的高度非线性映射。BP 算法是为了解决多层前向神经网络的权重系数优化而提出来的；在人工神经网络的实际应用中，80%~90%的神经网络模型是采用 BP 网络或它的变化形式。BP 算法是用于前馈多层网络的学习算法。

从数学上对 BP 神经网络模型进行分析，一般情况下可以建立关于 BP 神经网络公式：

$$x_j = \sum_i y_i \omega_{ji} \quad (2-1)$$

对于神经元本身输出的激活函数，一般来说选取 Sigmoid 函数，那么可以得到公式：

$$y_i = \frac{1}{1 + e^{-x_i}} \quad (2-2)$$

通过以上两个公式，分析出来 BP 神经网络中输出结果的计算过程。每个神经元收到刺激 y_i ，然后加权积累（权重 ω_{ji} ）完成后产生 x_j ，再通过激活函数产生刺激 y_j ，向下一层与它相连的神经元传递，依次类推最终输出结果。接下来利用向后反馈机制来修正神经元权重 ω_{ji} ，要修正 ω_{ji} 就需要得到误差量。具体来看，首先用 d_j 来表示真实的正确结果，并且设误差为 e ，那么 $(y_j - d_j)$ 对应的就是 e 对于 y_j 的微分增量，即 y_j 减去 $(y_j - d_j)$ 后就能得到正确值：

$$\frac{\partial e}{\partial y_j} = y_j - d_j \quad (2-3)$$

然后需要知道的是对于权重 ω_{ji} 的误差量是多少也就是 $\frac{\partial e}{\partial \omega_{ji}}$ 的值。而由公式 (1) 中知道 ω_{ji}

与 x_j 相关，那么可以推导出公式：

$$\frac{\partial e}{\partial \omega_{ji}} = \frac{\partial e}{\partial x_j} \frac{\partial x_j}{\partial \omega_{ji}} = \frac{\partial e}{\partial x_j} y_i \quad (2-4)$$

需要求得 ω_{ji} 的误差量，转换为求 $\frac{\partial e}{\partial x_j}$ 的值。而 $\frac{\partial e}{\partial x_j} = \frac{\partial e}{\partial y_j} \frac{\partial y_j}{\partial x_j}$ ；其中 $\frac{\partial y_j}{\partial x_j}$ 可以通过公式 (2-2)

求导得出来：

$$\frac{\partial y_j}{\partial x_j} = y_j(1 - y_j) \quad (2-5)$$

所以最终得到的误差量的值为：

$$\frac{\partial e_j}{\partial \omega_{ji}} = y_j(1-y_j)(y_j-d_j)y_i \quad (2-6)$$

到这里便完成了运用神经网络的输出值 y_j 和正确值 d_j 对最后一层隐藏层 ω_{ji} 的修正，上面的过程由公式（2-3）开始，就可以同理求出其对应其他隐藏层需要修正的权重值误差量。即：

$$\frac{\partial e_j}{\partial y_i} = \frac{\partial e_j}{\partial y_j} \frac{\partial y_j}{\partial x_j} \frac{\partial x_j}{\partial y_i} = y_j(1-y_j)(y_j-d_j)\omega_{ji} \quad (2-7)$$

这样所有的误差量的都可以同理推导完成。最后一步修正 ω_{ji} ，就是加上下面变量了，设置一个（0 到 1 之间）学习率：

$$\Delta W = -l \frac{\partial e}{\partial \omega} \quad (2-8)$$

从而得出完整的 BP 算法数学推导。但是标准的 BP 算法存在着易形成局部极小、收敛速度很慢和过度训练等问题。针对 BP 网络的缺陷，我们使用了 Bayesian 正规化算法改进 BP 神经网络泛化能力，通过拟合含有白噪声的非线性函数做出时间序列图。

2) 贝叶斯正规化 (Bayesian regularization)

在训练样本集的大小一定的情况下，网络的推广能力与网络的规模直接相关。正规化方法是通过修正神经网络的训练功能函数来提高其推广功能。一般情况下，神经网络的训练性能函数采用均方误差函数，假设误差函数为 E_D ：

$$E_D = \frac{1}{2} \sum_{k=1}^K \sum_{n=1}^N (y_{nk} - c_{nk})^2 \quad (2-9)$$

上式中 N 为样本数， K 为神经网络输出量个数， y_{nk} 为期望输出， c_{nk} 为网络实际输出。在正规化方法中，网络性能函数改进为：

$$F = \alpha E_\omega + \beta E_D \quad (2-10)$$

其中 E_ω 为网络权重的平方和， E_D 为网络相应和目标值的残差平方和， α 、 β 控制着其他参数（权及阈值）的分布形式，被称为超参数。超参数的大小决定着神经网络的训练目标，若 $\alpha \ll$ ，则训练算法目的在于尽量减小网络的训练误差；若 $\alpha \gg$ ，则训练算法目的在于使用网络产生更为平滑的响应，即尽可能减少有效的网络参数，以弥补较大的网络误差。常规的正规化方法很难确定超参数 α 、 β 的大小，而贝叶斯正规化方法则可以在网络训练过程中自适应地调整超参数的大小，并使其达到最优。

3)建模及预测

我们以 ZD326 为例，通过 MATLAB 建立了经过贝叶斯正规化的 BP 模型（具体代码可详见附件 5），模型流程如下图：

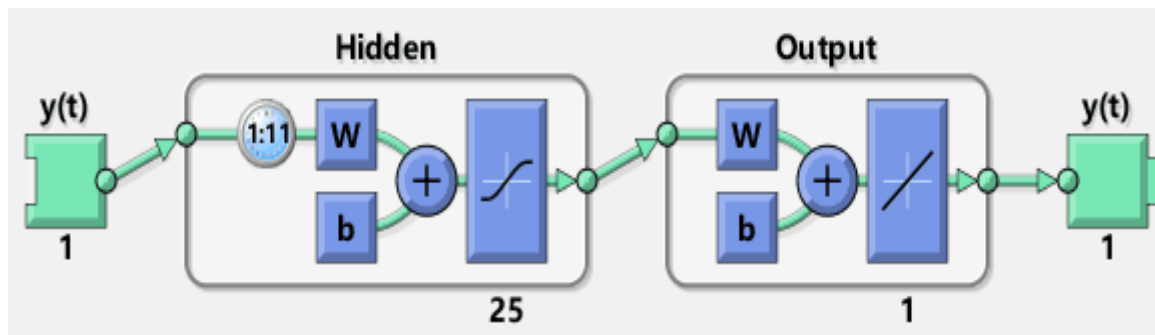


图 30—模型流程图

模型可视化结果如图 31、32：

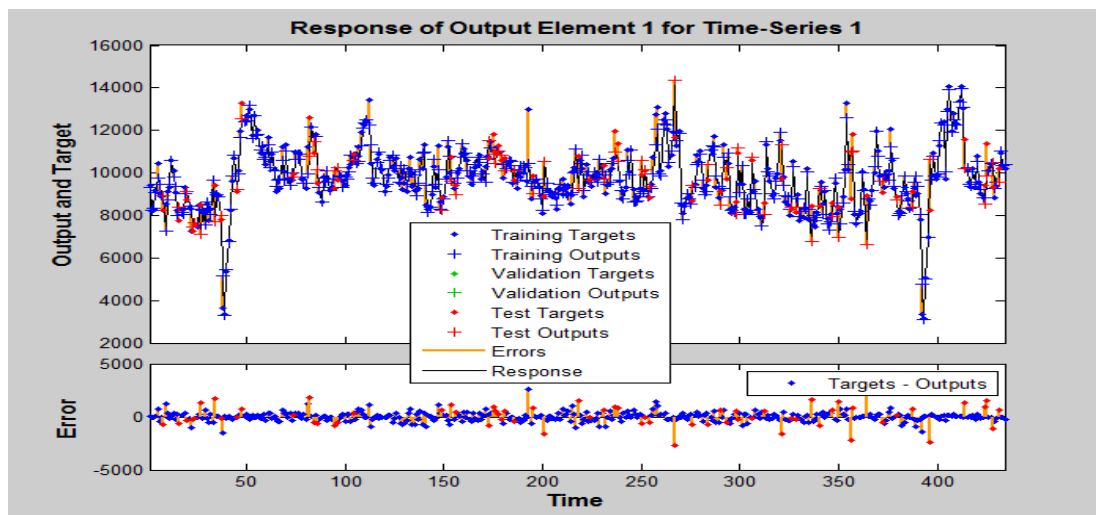


图 31 神经网络中的时间序列图

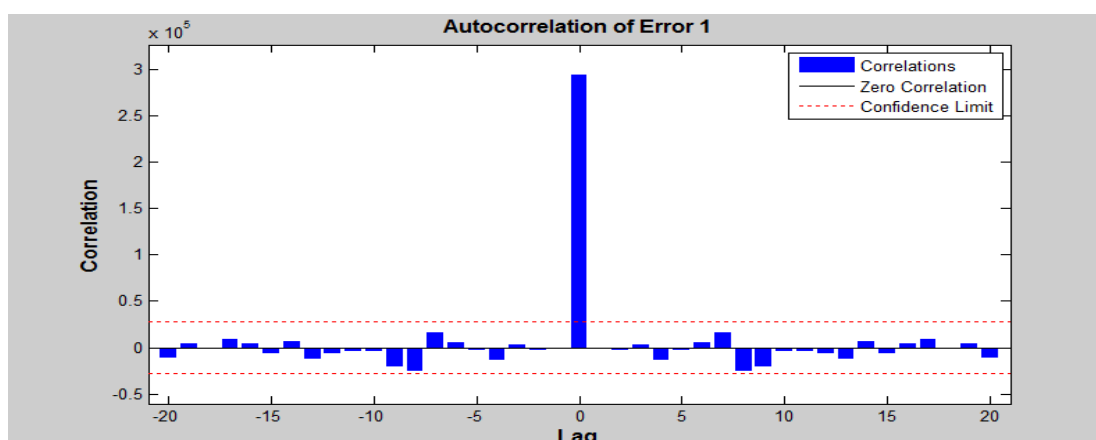


图 32 模型中随机误差项的自相关性

我们可以清晰地看到：经过贝叶斯正规化的时间序列模型的误差值几乎为零，模型中随机误

差项之间的自相关性也完全在可接受的范围内，其拟合效果很不错。训练及预测结果如表 8：

表 8 BP 神经网络预测结果

第一周	9740	9533	10234	10533	11171	10700	10543
第二周	10394	10072	10610	10861	11099	10743	10222

预测的未来两周梯形密度表（见附件 8），其预测结果与旅客尤其是清明节出行规律吻合，表中为 0 的地方是由于区间站点没有车次停靠。其中第 1 天的预测数据如表 9。

表 9 第 1 天的预测数据

ZD190 终到方向														
	管外	ZD111	ZD311	ZD326	ZD192	ZD022	ZD250	ZD062	ZD120	ZD121	ZD143	ZD370	ZD190	下车人数合计
管外														
ZD111	11755													11755
ZD311	237	422												659
ZD326	3499	1155	5086											9740
ZD192	810	267	16	1161										2254
ZD022	60	20	1	16	70									168
ZD250	4268	1409	85	1169	276	4675								11882
ZD062	632	209	13	173	41	3	689							1760
ZD120	93	31	2	26	6	1	31	70						260
ZD121	94	31	2	26	6	1	31	5	67					262
ZD143	38	12	1	10	2	0	12	2	0	27				105
ZD370	0	0	0	0	0	0	0	0	0	0	0			0
ZD190	9121	3011	183	2498	590	50	2989	475	76	57	28	6314		25393
上车人数合计	30607	6567	5389	5079	992	4729	3752	552	144	84	28	6314	0	64238

2.2.5. 车辆停靠优化设计

1. 基本思想

为了优化设计车辆配置及车站停靠方案，通过对 D02-D19 共 18 个车次所经过的站点每天上车人数总和（见附件 9）进行分析，以期了解一年中的热门站点和冷门站点，分析结果如图 33。

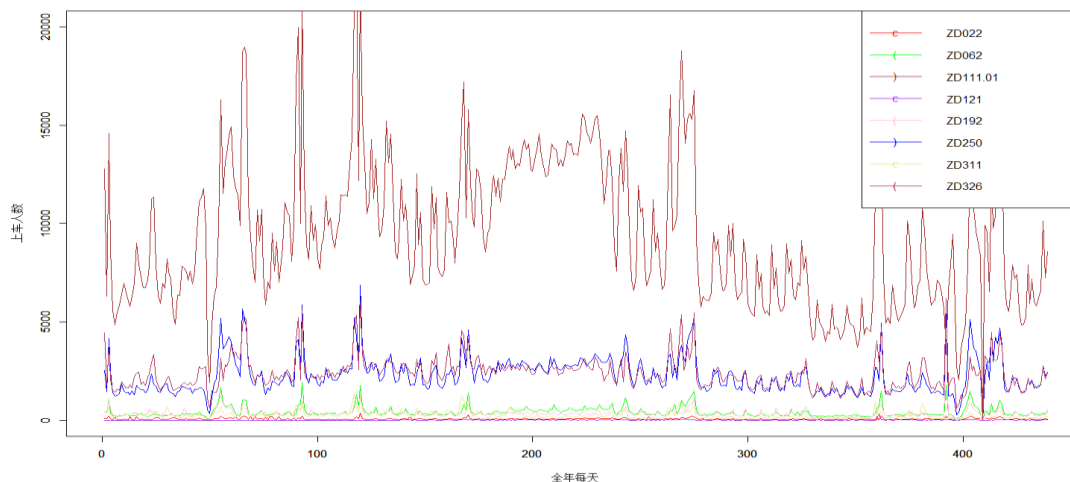


图 33 管内站点每天的人数总和

由图 33 可知，站点热门程度结果排列顺序如下 ZD111-01、ZD250、ZD326、ZD062、ZD192、ZD311、ZD022、ZD121。

通过统计数据得知 D02-D19 这些车次的客座率较低，多数管内站点的旅客周转全靠管外维持。同时通过数据分析得知可以得知在管内的 14 个站点中有 5 个站点（ZD111-02、ZD120、ZD121、ZD131、ZD370）没有此类车次停靠，我们用 0-1 表示是否停靠，具体路线停靠方案见表 10。

表 10 原始路线及停靠方案

	D02	D03	D04	D05	D06	D07	D08	D09	D10	D11	D12	D13	D14	D15	D16	D17	D18	D19
ZD111-01	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ZD111-02	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ZD311	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0
ZD326	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
ZD192	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0
ZD022	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ZD250	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ZD062	1	0	0	1	1	1	0	0	0	0	1	0	1	0	0	0	1	0
ZD120	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ZD121	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ZD143	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ZD370	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ZD190-02	0	1	0	0	0	0	0	0	1	0	0	0	1	1	0	1	1	1
ZD190-01	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

针对这样的问题，我们考虑调整这些车次的停靠方案，以此充分利用管内资源尽量满足管内旅客的需求。我们从两种情况考虑：一种是对单车次而言，时间不变、停靠站变化；二是对于整

体而言，发车时间不变，改变停靠站（站点的到发时间可能改变）。由于题目中没有给出票价信息，而票价与客座率呈正相关，鉴于此我们考虑以客座率最大为目标，车次到某个站点是否停靠为决策，尽量满足管内较多旅客周转的需求建立 0-1 整数规划模型。

2. 单趟车停靠

根据上面分析，我们记 $\Omega = \{1, 2, \dots\}$ 为由近到远的站点集合（其中 1 可以代表 ZD111-01 站点）， $\Omega_k = \{1_k, 2_k, \dots\}$ 为第 k 车次所经过的站点集合（其中 $\Omega_k \subset \Omega$ ）， $s_{i_k j_k}$ 为第 i_k 站到 j_k 站的距离， r_k 为日均定员， s_k^0 为 k 车次的行车路线总距离。

引入 0-1 变量便是是否在原有站点停靠即：
$$x_{i_k j_k} = \begin{cases} 1, & \text{第 } k \text{ 趟车从 } i_k \text{ 站点出发到 } j_k \text{ 站点停靠} \\ 0, & \text{第 } k \text{ 趟车从 } i_k \text{ 站点出发到 } j_k \text{ 站点不停靠} \end{cases}$$

$y_{i_k j_k}$ 表示第 k 趟车从 i_k 站点出发到 j_k 站点停靠时流动人数。于是有：
$$y_{i_k j_k} = \begin{cases} 0, & x_{i_k j_k} = 0 \\ \text{其他}, & x_{i_k j_k} = 1 \end{cases}$$

上面式子可转化为：
$$(1 - x_{i_k j_k}) y_{i_k j_k} = 0 \quad (3-1-1)$$

M_k 为第 k 车次的周转客流量（根据统计得到），于是有
$$\sum_{i_k j_k} y_{i_k j_k} = M_k \quad (3-1-2)$$

第一个站点必须停靠，则 $x_{1_k} y_{1_k} = 1 \quad (3-1-3)$

最后 1 个站点也必须停靠：

$$\sum_{j_k} x_{i_k j_k} = 1 \quad (3-1-4)$$

鉴于中途其他站点停靠的问题，则从出发站角度考虑至多停两次，则有

$$\sum_{i_k} x_{i_k j_k} \leq 2 \quad (3-1-5)$$

从停靠站的角度考虑，每个站点至多只能停靠一次，即：

$$\sum_{j_k} x_{i_k j_k} \leq 1 \quad (3-1-6)$$

根据客座率 Z_k 计算公式，可知：

$$Z_k = \frac{\sum_{i_k j_k} y_{i_k j_k} s_{i_k j_k}}{r_k \cdot s_k^0} \quad (3-1-7)$$

由 3-1-1~3-1-7 得到单趟车停靠方案的优化模型（I）：

$$\begin{aligned} \max \quad & \frac{\sum_{i_k j_k} y_{i_k j_k} s_{i_k j_k}}{r_k \cdot s_k^0} \\ \text{s.t.} \quad & \begin{cases} \sum_{i_k j_k} y_{i_k j_k} = M_k \\ x_{1_k 1_k} = 1 \\ \sum_{j_k} x_{i_k \max j_k} = 1 \\ \sum_{i_k} x_{i_k j_k} \leq 2 \\ x_{i_k j_k} = 0, j_k > i_k \\ \sum_{j_k} x_{i_k j_k} \leq 1 \\ x_{i_k j_k} \in \{0, 1\}, i_k, j_k \in \Omega_k \\ y_{i_k j_k} \in N^* \end{cases} \end{aligned} \quad (I)$$

由于模型（I）直接求解困难，在求解的时候采用简化思想，寻找可行停靠方案，并比较客座率的方式，搜索最优方案。

1) 录入该车次的日均定员 r_k ，各站之间的距离 $s_{i_k j_k}$ 。每个站点流动人数 $y_{i_k j_k}$ 。

2) 计算每个可行方案的客座率。如：

该车次原始需要经过 4 个站点：

每个站点都停靠： $x_{11} = 1, x_{21} = 1, x_{32} = 1, x_{43} = 1$

缺失站点停靠： $x_{11} = 1, x_{21} = 1, x_{42} = 1; x_{11} = 1, x_{31} = 1, x_{43} = 1$

计算客座率时将缺失站点流动人数，加在前一站或后一站流动，并选择客座率大的一个作为该可行方案的客座率。

3) 最后比较所有可行的停靠方案的客座率，选择最大的客座率所对应的停靠方案作为最终停靠方案。

通过 MATLAB 编程（具体程序见附件 7），求解出单趟车停靠方案如表 11。

表 11 单趟车优化结果

车次	新路线
D02	ZD111-01->ZD326->ZD250->ZD190-01
D03	ZD190-01->ZD250->ZD192->ZD111-01
D04	ZD111-03->ZD111-01->ZD326->ZD250
D05	ZD190-01->ZD250->ZD326->ZD111-01

D06	ZD111-01->ZD311->ZD326->ZD250->ZD190-01
D07	ZD190-01->ZD062->ZD250->ZD111-01
D08	ZD111-01->ZD326->ZD190-01
D09	ZD190-01->ZD250->ZD326->ZD111-01
D10	ZD111-01->ZD326->ZD190-02->ZD190-01
D11	ZD190-01->ZD250->ZD326->ZD111-01
D12	ZD111-01->ZD326->ZD250->ZD190-01
D13	ZD190-01->ZD250->ZD192->ZD111-01
D14	ZD111-01->ZD250->ZD062->ZD190-01
D15	ZD190-01->ZD190-02->ZD250->ZD111-01
D16	ZD111-01->ZD311->ZD326->ZD111-01
D17	ZD190-01->ZD250->ZD192->ZD111-01
D18	ZD111-01->ZD250->ZD190-02->ZD190-01
D19	ZD190-01->ZD190-02->ZD250->ZD111-01

从表 10 可以看出在公司以客座率为最大为目标的情况下，有些车次的停靠站点将被取消，会导致有些旅客在就近的站点下车，不能满足部分旅客的需求。

3. 整体停靠

根据分析并考虑到单车次优化的缺点，我们从整体考虑并将管内其他没有这类车次停靠的站点也加入停靠方案中，保证每个城市每个站点都有车次停靠，以总客流率最大为目标，以是否在该站停靠为决策，尽可能满足旅客周转需求（图 34）。

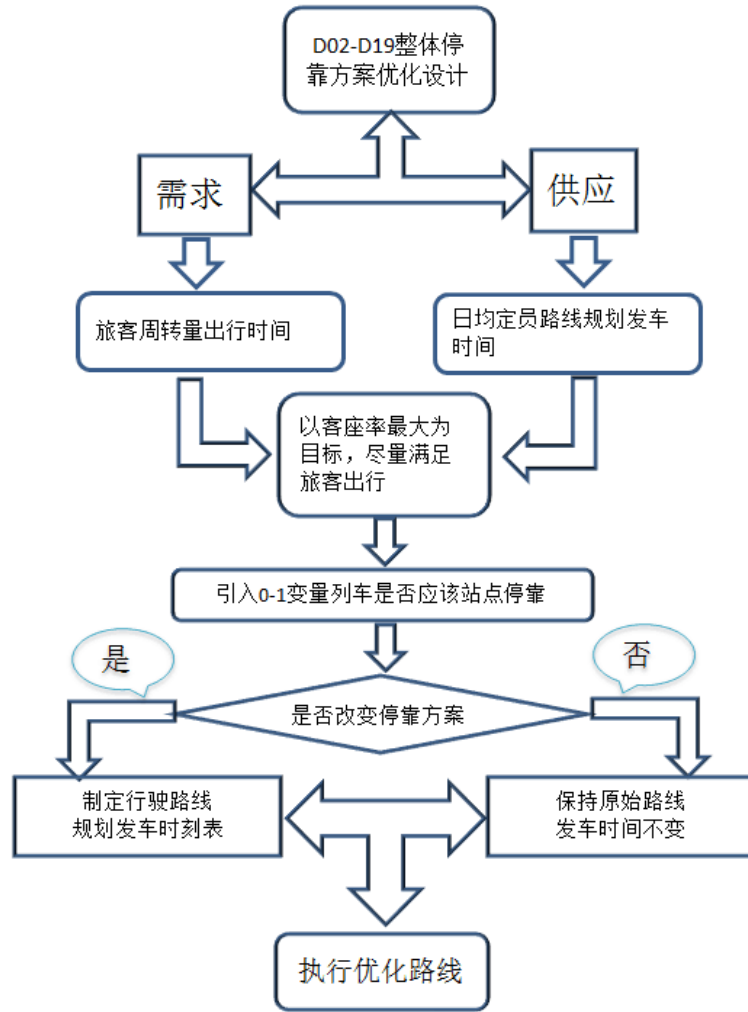


图 34 整体停靠方案流程图

我们记 k_j 为第 j 车次的客座率； a_{ij} 为原始停靠方案； r_j 为第 j 车次的日均定员； s_i 为第 i 站距离起始站的距离； c_i 为站点 i 的总周转数量； x_{ij} 为第 j 车次是否在站点 i 停靠，当停靠时取 1，不停靠取 0； y_{ij} 为第 j 车次在站点 i 停靠时的周转客流量，即有：

$$y_{ij} = \begin{cases} 0, & x_{ij} = 0 \\ \text{其他}, & x_{ij} = 1 \end{cases}$$

上面式子可转化为

$$(1 - x_{ij})y_{ij} = 0 \quad (3-2-1)$$

每趟车周转客流量不能超过日均定员，即

$$\sum_{i=1}^{14} x_{ij}y_{ij} \leq r_j (j=1,2,\dots,18) \quad (3-2-2)$$

每个站点周转客流量不能超过总需求周转量（部分可以采用管外车次周转）。即，

$$\sum_{j=1}^{18} x_{ij} y_{ij} \leq c_i (i=1,2,\dots,14) \quad (3-2-3)$$

起始站点和终点站必须全部停靠：

$$\sum_{j=1}^{18} x_{1j} = \sum_{j=1}^{18} x_{14j} = 18 \quad (3-2-4)$$

每个站点至少有一个车次停靠：

$$\sum_{j=1}^{18} x_{ij} \geq 1 (i=1,2,\dots,14) \quad (3-2-5)$$

每个站点在原有的方案停靠车次数量上可以增加或减少趟车次：

$$\sum_{j=1}^{18} a_{ij} - m \leq \sum_{j=1}^{18} x_{ij} \leq \sum_{j=1}^{18} a_{ij} + m (i=1,2,\dots,14) \quad (3-2-6)$$

每趟车停靠站点数目可以在原有方案上增加或减少个站点：

$$\sum_{i=1}^{14} a_{ij} - n \leq \sum_{i=1}^{14} x_{ij} \leq \sum_{i=1}^{14} a_{ij} + n (j=1,2,\dots,18) \quad (3-2-7)$$

根据客座率的计算公式可知第 k 车次的客座率（其中每个车次经过路线距离都是 393）：

$$k_j = \frac{\sum_{i=1}^{14} x_{ij} y_{ij} s_i}{r_j \cdot 393} \quad (3-2-8)$$

根据 3-2-1~3-2-8 建立优化模型（II）：

$$\begin{aligned}
 & \max \sum_{j=1}^{18} k_j \\
 & \left\{ \begin{aligned}
 & k_j = \frac{\sum_{i=1}^{14} x_{ij} y_{ij} s_i}{r_j \cdot 393} \quad j=1,2,\dots,18 \\
 & \sum_{i=1}^{14} x_{ij} y_{ij} \leq r_j \quad j=1,2,\dots,18 \\
 & \sum_{j=1}^{18} x_{ij} y_{ij} \leq c_i \quad i=1,2,\dots,14 \\
 & \sum_{j=1}^{18} x_{1j} = \sum_{j=1}^{18} x_{14j} = 18 \\
 & s.t. \left\{ \begin{aligned}
 & \sum_{j=1}^{18} x_{ij} \geq 1 \quad i=1,2,\dots,14 \\
 & x_{ij} \in \{0,1\} \quad i=1,2,\dots,14, j=1,2,\dots,18 \\
 & y_{ij} \in N \quad i=1,2,\dots,14, j=1,2,\dots,18 \\
 & \sum_{j=1}^{18} a_{ij} - m \leq \sum_{j=1}^{18} x_{ij} \leq \sum_{j=1}^{18} a_{ij} + m \quad i=1,2,\dots,14 \\
 & \sum_{i=1}^{14} a_{ij} - n \leq \sum_{i=1}^{14} x_{ij} \leq \sum_{i=1}^{14} a_{ij} + n \quad j=1,2,\dots,18
 \end{aligned} \right.
 \end{aligned} \right. \quad (II)
 \end{aligned}$$

利用 Lingo 软件编程（程序见附件 7）得到调整后的路线如表 12:

表 12 原路线调整后的路线

原路线																		
	D02	D03	D04	D05	D06	D07	D08	D09	D10	D11	D12	D13	D14	D15	D16	D17	D18	D19
ZD111-01	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ZD111-02	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ZD311	1	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0
ZD326	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ZD192	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0
ZD022	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ZD250	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ZD062	1	0	0	1	1	1	0	0	0	0	1	0	1	0	0	0	1	0
ZD120	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ZD121	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ZD143	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ZD370	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ZD190-02	0	1	0	0	0	0	0	0	1	0	0	0	1	1	0	1	1	1
ZD190-01	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
调整后的路线																		
	D02	D03	D04	D05	D06	D07	D08	D09	D10	D11	D12	D13	D14	D15	D16	D17	D18	D19

ZD111-01	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ZD111-02	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ZD311	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	1	0
ZD326	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1
ZD192	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1
ZD022	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ZD250	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ZD062	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0
ZD120	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ZD121	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
ZD143	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ZD370	0	1	0	1	0	0	0	0	0	0	0	1	1	0	0	0	1
ZD190-02	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1
ZD190-01	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

从上表可知有 6 个车次没有发生变化，其余只是停靠站点发生变化没有减少站点，仅有 D15 车次增加了一个停靠站点，根据距离重新规划发车时刻表具体计算如下：

$$V = \frac{S}{T} \text{ (} V \text{代表该车次运行速度, } S \text{代表起点站到终点站的距离, } T \text{代表运行时间)}$$

$$t = \frac{s}{V} \text{ (} t \text{代表到达站点所需时间, } s \text{代表到达该站点的距离)}$$

部分结果如表 13（详见附件 7）：

表 13 D02-D19 重新规划发车时刻表（部分）

车次	原始路线	发车时间	到站时间	新路线	发车时间	到站时间
D02	ZD111-01	7:10	7:10	ZD111-01	7:10	7:10
	ZD326	7:53	7:52	ZD111-02	7:13	7:12
	ZD250	8:30	8:29	ZD326	7:55	7:54
	ZD062	8:58	8:57	ZD250	8:32	8:31
	ZD190-01		9:48	ZD190-01		9:48
D03	ZD190-01	5:46	5:46	ZD190-01	5:46	5:46
	ZD190-02	6:03	6:02	ZD370	6:02	6:01
	ZD250	7:01	6:59	ZD250	7:05	7:03
	ZD192	7:24	7:22	ZD022	7:17	7:16
	ZD326	7:45	7:43	ZD326	7:47	7:45
	ZD111-01		8:33	ZD111-01		8:33
D04	ZD111-03	7:58	7:58	ZD111-03	7:58	7:58
	ZD111-01	8:17	8:14	ZD111-01	8:17	8:14
	ZD326	9:02	9:00	ZD326	9:02	9:00

	ZD022	9:32	9:30	ZD022	9:32	9:30
	ZD250	9:46	9:43	ZD250	9:46	9:43
	ZD190-01		11:00	ZD190-01		11:00
D05	ZD190-01	6:43	6:43	ZD190-01	6:43	6:43
	ZD062	7:30	7:28	ZD370	6:58	6:57
	ZD250	7:59	7:57	ZD120	7:17	7:16
	ZD326	8:39	8:35	ZD250	8:03	8:01
	ZD311	9:06	9:04	ZD326	8:45	8:43
	ZD111-01		9:27	ZD111-01		9:27
D06	ZD111-01	9:57	9:57	ZD111-01	9:57	9:57
	ZD311	10:22	10:21	ZD311	10:22	10:21
	ZD326	10:46	10:44	ZD326	10:46	10:44
	ZD192	11:06	11:05	ZD192	11:06	11:05
	ZD250	11:24	11:22	ZD250	11:24	11:22
	ZD062	11:52	11:51	ZD062	11:52	11:51
	ZD190-01		12:48	ZD190-01		12:48
D07	ZD190-01	9:10	9:10	ZD190-01	9:10	9:10
	ZD062	9:56	9:55	ZD062	9:56	9:55
	ZD250	10:24	10:23	ZD250	10:24	10:23
	ZD326	11:02	11:00	ZD111-01		11:50
	ZD111-01		11:50			

备注：颜色突出的部分是没有改变停靠方案的车次。

2.3. 结果分析与反馈

2.3.1. 结果解析

1、数据分析

通过对铁路公司的 ZD190（站）至 ZD111（站）区段的客运专线进行短期客运量的分析，我们研究了车次、时段、车站以及车型与铁路客流量的关系。分析可知，503、D01、D04、D21、D23、D24、D31、D34、D36、D47、D48、G30、G32、K11、Z01 这些车次平均客座率比较高，可以适当增加某些车次，但是有些车次可能是受到季节或者临时性的影响要另作处理；而 D45、T08 车次平均客座率在 10%以下，可以考虑减少车次或者用其他车次代替。从各个时段的上下车人数分析可以看出，时段对于客流量的影响有一定的差异性，客流量从大到小顺序排列：6: 00-12: 00、12: 00-18: 00、18: 00-24: 00、0: 00-06: 00。其中 6: 00-12: 00 时段的上下车人数最密集，0: 00-06: 00 的上下车人数最离散。我们综合各个站点关于每一天的上下车总人数来看，

我们得出 ZD250,ZD111-01,ZD190-02 三个站点客流量比较多,可以适当在这些站点增加一些车次, D120、ZD121、ZD022 站点客流量比较少,可以适当减少一些车次。从对车型研究可以看出 G 类车型比较热门, T 类车型比较冷门。从宏观方面考虑主要出现这些结果的主要原因是受到客流性质、车站客流需求、车站吸引范围的人口、经济发展水平结构以及交通设施等条件的影响。

2、预测模型对比

通过对建立的 ARIMA 模型所预测的结果与经过贝叶斯正规化的 BP 神经网络模型的预测结果进行对比,可视化结果如图 35。

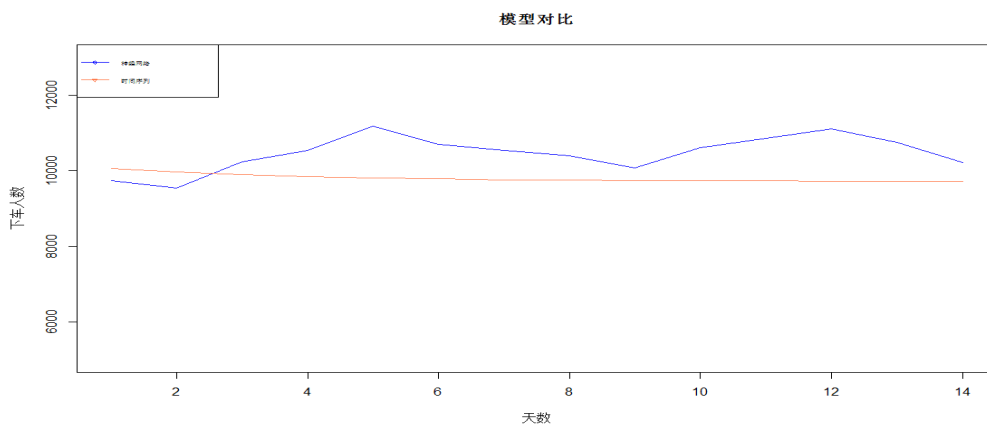


图 35 预测结果对比

我们可以发现 ARIMA 模型所预测的结果过于平稳,而经过贝叶斯正规化的 BP 神经网络模型预测的结果更趋于实际。我们可以认为正规化的神经网络模型预测效果更好,但鉴于神经网络模型是一种非机理模型,无法根据主要因素相互作用的机理,对他们之间的平衡关系进行数学描述。对此,铁路公司可斟酌接受其预测效果。

3、优化结果解释

通过对 D02-D19 车次分析,我们可以发现单趟车优化方案,优点是可以使客座率达到最大,缺点是由于原路线中间某些站点不再停靠,所以不能满足旅客本来出行的规律。整体停靠优化方案中,由于 D02-D19 车次的客座率较低,而且有些站点并未停靠,只能靠管外车次满足,因此我们以总客流率最大为目标,以是否在该站停靠为决策,尽可能满足周转人数。有 6 个车次没有发生变化,其余只是停靠站点发生变化没有减少站点,仅有 D15 车次增加了一个停靠站点,根据距离重新规划发车时刻表。整体方案能满足每个城市每个站点,能满足更多管内旅客的需求,效果较好。

2.3.2. 模型评价

1) 从数据处理的过程和分析的结果来看, 各因素的时序特征较为合理。

2) 从模型角度来看: 预测模型中神经网络模型虽然操作难度较大, 但精度和效果较好, 更贴近实际; 优化方案中第一种情形求解时采用了简化的方式搜索求解, 得出的结果不一定是最优, 但也能达到提高客座率的效果, 第二种情形求解结果虽然是局部最优解, 但整体方案效果较好, 能提高客座率, 满足更多旅客的需求。

3. 结论

通过对铁路公司的 ZD190 (站) 至 ZD111 (站) 区段的客运专线进行短期客运量的分析, 我们得出的结果是: 可以适当增加某些车次, 但是有些车次可能是受到季节或者临时性的影响要另作处理; 可以考虑减少车次或者用其他车次代替。对车型分析的结果是: G 类车型比较热门, T 类车型比较冷门。从宏观方面考虑主要出现这些结果的主要原因是受到客流性质、车站客流需求、车站吸引范围的人口、经济发展水平结构以及交通设施等条件的影响。

通过对建立的 ARIMA 模型所预测的结果与经过贝叶斯正规化的 BP 神经网络模型的预测结果进行对比, 根据具体的预测结果, 我们得出的结论是: 正规化的神经网络模型预测效果更好, 并且预测结果与清明节假日旅客出行规律吻合, 但鉴于神经网络模型是一种非机理模型, 无法根据主要因素相互作用的机理, 对他们之间的平衡关系进行数学描述。对此, 铁路公司可斟酌接受其预测效果。

通过对 D02-D19 车次分析, 我们得出的结论是: 单趟车优化方案可以使该趟车次客座率达到最大, 但从原路线中间某些站点不再停靠, 不能满足旅客本来出行规律方面来看, 还应斟酌是否使用该优化方案。整体停靠优化方案中, 能够覆盖所有城市所有站点停靠, 客座率明显提高, 有 6 个车次没有发生变化, 其余只是停靠站点发生变化没有减少站点, 仅有 D15 车次增加了一个停靠站点, 根据距离重新规划的发车时刻表能满足更多管内旅客的需求, 可以减少管外车次的压力。

建议铁路部门可以根据预测结果采用优化停靠方案以此来提高效益。

4. 参考文献

- [1]李诗羽,张飞,王正林.数据分析:R 语言实战[M].北京: 电子工业出版社,2014.8.
- [2]薛毅,陈立萍.R 语言实用教程[M].北京: 清华大学出版社,2014.

- [3] (美) 拉罗斯 (Larose,D.T) 著; 刘燕权等译.数据挖掘方法与模型[M].北京:高等教育出版社,2011.3.
- [4] (美) 卡巴科弗 (Kabacoff,R.I.) 著;高涛,肖楠,陈钢译.R 语言实战[M].北京:人民邮电出版社,2013.1.
- [5]姜万录.神经网络学习算法存在的问题及对策. 机床与液压,2003,5:29~32.
- [6]魏东,张明廉,等.基于贝叶斯方法的神经网络非线性模型辨识.计算机工程与应用,2005,11:5~11.
- [7]美 RobertI.Kabacoff 著,高涛,肖楠,陈钢译 R 语言实战[M],人民邮电出版社出版.
- [8]美 Jonathan D.Cryer,Kung-Sik Chan 著,潘红宇等译 时间序列分析及应用[M],机械工业出版社.
- [9]JiaweiHan,Micheline Kamber 著,范明,孟小峰译,数据挖掘概念与技术[M],机械工业出版社.