轨道交通智慧客流分析预测

开发团队：该队名已被注册

目录

S2A 项目说明

S2A 1.赛题的价值

1.1 项目背景

在科技革命和产业变革的发展背景下，随着城市间人口流动的增加，中国城市正在兴起轨道交通建设热潮。动售检票系统(AFC)在全线车站启用，改变了以往的经营方式、建设模式和服务方式，标志着城市轨道交通进入了网络信息化时代。2020 年 3 月 12 日发布的《中国城市轨道交通智慧城轨发展纲要》确定城市轨道交通今后的发展方向为智能化发展。

图 1：挤地铁的百态人像

近年来，国内各个城市轨道交通持续的高速发展，轨交乘客数量不断增长。由于轨道交通具有客运量大、设备科技技术含量高、不能随意停靠等特点，轨道交通的客流监测能力不足、缺乏客流精准管控方法及缺乏对突发客流的提前评估和预测等问题日益凸显，对路线各项数据感知落后也使轨道交通管理压力进一步加重。因此，在提高轨道交通经济性和舒适性的同时，增加它的安全性成为当下智能城轨发展的重要内容。

合理分配人力物力资源能让有关部门在遇到紧急情况的时候作出更加有效的策略部署，这些都需要有准确的预测作为前提。轨道交通的客流分析和预测是轨交企业运营管理以及乘客服务的重要一环。因此，基于用户乘车行为大数据的分析，准确预测客流量成为解决该项目的主要问题。

1.2 项目概述

本项目使用服务外包创新创业大赛官网所给的公开数据，包含了脱敏后 2019 年 12 月 26 日到 2020 年 7 月 16 日某城市用户真实的地铁行程数据、8 条线路共 168 个站点的区域信息和上下行信息、2019 年 12 月 1 日到 2020 年 8 月 21 日的城市天气数据、2020 年全年的节假日情况、12 万余的用户个人信息，综合运用数据分析软件 Matlab、SPSS、Excel，对客流进行统计分析，并运用基于神经网络的时间序列模型对未来的客流量进行预测。

1.3 项目意义

动态变化的客流，沿空间和时间的不均衡性分布是城市轨道交通的最基本的特征。对城市轨道交通系统的客流进行分析和预测，并从实际客流规律出发研究运营调度方案的优化问题，可以保障乘客出行体验，提高运营部门服务水平并实现利益最大化。

本项目组借此机会，基于给定的 208 天的 12 万条客流信息，综合出入站点、乘客性别、所在地区等数据，在统计分析的基础上，对接下里 7 天的客流量进行了预测分析，并给出了车次安排等相关方案的建议，对城轨企业日后的改进具有一定的指导意义。

S2A 2. 目标及解决思路

2.1 项目目标

城市轨道交通客流的精确预测对于轨道交通的进一步发展具有极为重要的作用，例如车次的合理安排、各个站点的人员配备。本项目期望参赛者能够基于已知的客流信息数据，结合大数据处理经验和数据本身的特点来建立一套合适的数学模型，并利用此模型进一步建立完善的智能算法，完成基于地铁出行行程大数据的分析建模和算法研究，实现对地铁的线路级别以及站点级别的客流分析和预测。

2.2 任务要求

1) 对题目所给数据进行数据库搭建
2) 基于给定的数据进行客流精确统计，并给出具体的分析
3) 建立具有动态可调节能力的模型，预测未来一段时间的客流量。
4) 开发交互及展示页面，要求系统能够对数据进行实时计算
5) 针对预测结果，对指定时间、指定线路或者站点的客流进行预测和预警，并给出车次安排等相关方案的建议。

2.3 解决思路

2.3.1 思路创意

我们团队研究学习了预测常用的随机森林、神经网络、prophet、时间序列等模型，在大量的文献和学习资料中，我们发现多数的大数据预测方式都是建立单一模型预测，我们也尝试了多种单一模型的预测方式，但结果都不是特别理想。题目中客流量数据带有明显的随时间变化的趋势性，但是，客流量还可能会受到天气、温度等因素影响，用单一时间序列模型预测会丢失这些有价值的信息。如果使用神经网络对输入的指标和输出的客流量进行拟合，则又无法体现出客流量在时间维度上的趋势性。考虑到时间序列由长期趋势、季节变动、循环变动、不规则变动构成，若将由客流量构成的时间序列分解为长期趋势和由剩下三个要素组成的"波动项"，再通过时间序列模型预测长期趋势、通过神经网络和一系列输入指标对剩下的波动项进行拟合，则既可以规避单纯的时间序列模型所造成的指标信息的丢失，也可以消除神经网络的输出项的时间趋势性，获得更佳的拟合效果。

在神经网络模型的选择上，我们采用了遗传神经网络模型，通过遗传算法对神经网络的参数进行寻优，可以解决通过 BP 神经网络容易陷入局部极小值的问题，寻找得到更加合适的网络。

因此，我们团队认为，在时间序列模型的基础上纳入遗传神经网络，是预测未来客流量的较为理想的模型。项目创意如下图 2 所示：

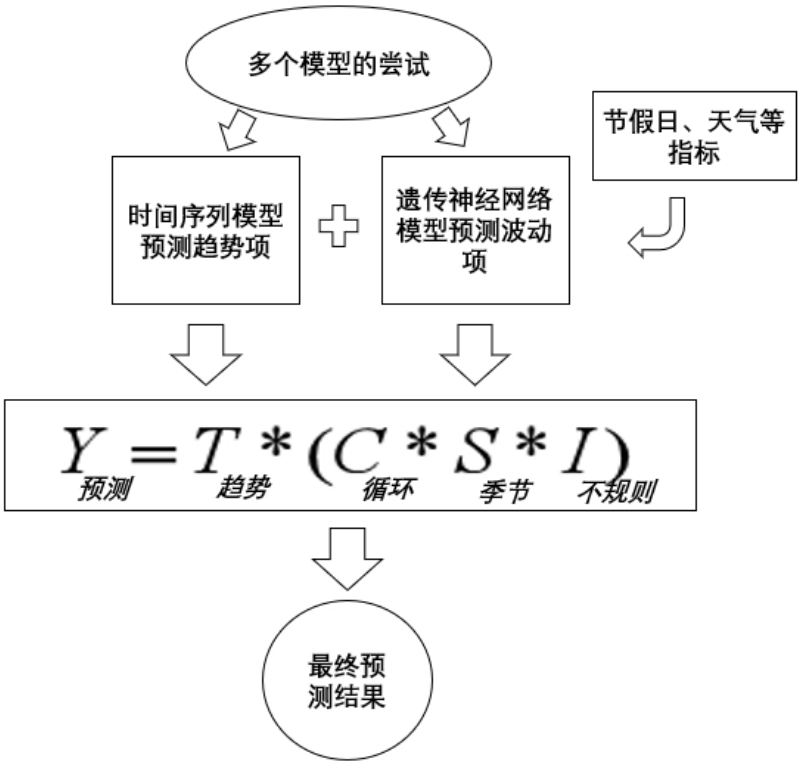


图 2：项目创意图

2.3.2 思路概述

本项目整体可以拆分为两大块任务，第一块任务是对给定数据进行客流量的统计分析，第二块任务是对未来一段时间的客流量进行预测，此外，还需完成的任务包括数据库的搭建与可交互界面的开发与设计。该项目的大致思路入下图 3 所示：

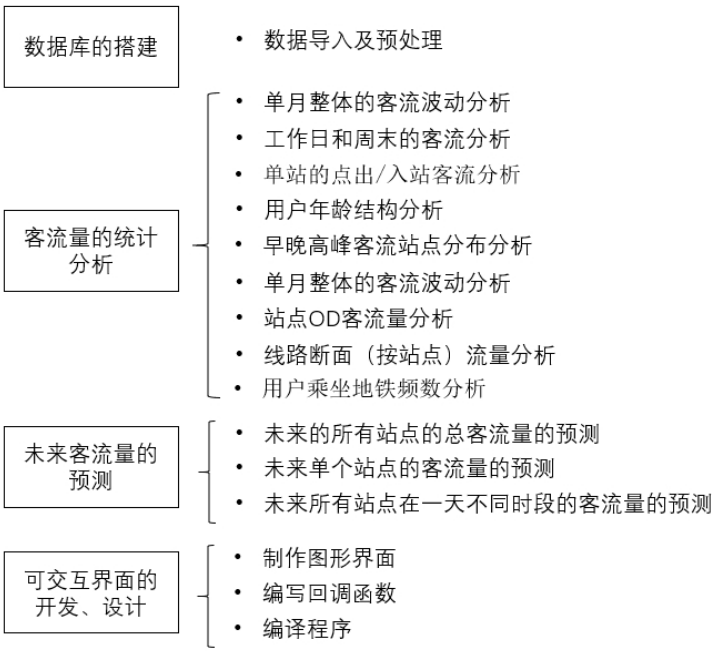


图 3：思路概述图

【数据库的搭建】

数据导入及预处理：在 Matlab 中导入服务外包竞赛组委会提供的 Excel 数据，数据主要包含 trips 数据集中 80 万条行程信息、users 数据集中 12 万条用户信息，station 数据集中的站点信息和 workdays 数据集中 2020 年节假日信息，并对乘车时间异常数据、金额异常数据、同站进出数据进行了删除，对错误的节假日信息进行修改。

【客流量的统计分析】

单月整体的客流波动分析：选取有客流量数据>25 天的月份，计算出这些月份里每一天的客流量大小，画出这些月份的客流量波动的柱状图并进行分析。

工作日和周末的客流分析：分别统计出周一至周日的客流量数据，剔除节假日数据，然后使用箱线图来反映周一至周日客流量样本的均数、离散情况等信息。

单站的点出/入站客流分析：计算出单个站点在一天不同时段的进站客流量和出站客流量，然后在同一张图中用折线分别表示一天中进站客流量和出站客流量随时间的变化情况，并对不同站点进行分类。

用户年龄结构分析：统计出所有用户的年龄，以 20 岁为区间长度，将所有用户划分为 4 个年龄段，用饼状图描绘各年龄段用户人数在总人数中所占比例；再以 5 岁为区间长度，用柱状图反映各年龄段用户人数、用对比柱状图反映各年龄段男性和女性用户的人数。

早晚高峰客流站点分布分析：选取一条线路中所有站点在不同时间段的客流量数据，以站点为 x 轴、时间为 y 轴、客流量为 z 轴画出三维直方图。引入高峰小时比率来反映单个站点客流分布的时间特征，再综合各个站点的客流量，寻找出早晚高峰客流主要分布的站点。

站点 OD 客流量分析：统计出所有行程数据中的出站和入站信息，分别计算出线路与线路间、线路内站点间的 OD 客流量，并进行行分析。

线路断面（按站点）流量分析：理解线路横断面的含义，解决横断面客流分析需要知道线路之间的换乘站。利用最大客流量和乘坐地铁的其他时间这两个指标，建立评价模型，确定线路之间的换乘站。根据换乘站，画出地铁线路的拓扑图。以站点顺序为横坐标，客流量为纵坐标，画出线路横断面的柱状图，与线路站点出/入站总客流量的柱形图进行比较和分析。

用户乘坐地铁频数分析：统计出 12 万余名用户在行程数据集横跨的时间段内的的累计出行频数，并画出乘坐频数的散点图。在散点图的基础上对高频用户、低频用户进行分类。

【未来客流量的预测】

数据集的长度我们选择受疫情影响较小的 2020 年 4 月 1 日-2020 年 7 月 16 日共 107 天的数据，基于我们团队提出的模型，将需要预测的数据集分解出长期趋势项和波动项，长期趋势项用 AR 时间序列模型预测，未来的波动项使用训练好的神经网络预测（对于每一个数据集，均训练出一个较为理想的网络），再通过预测出的长期趋势项和波动项，得到未来的客流量。

未来所有站点的总客流量的预测：统计每天的总客流量作为数据集，并预测。

未来单个站点的客流量的预测：统计单个站点的客流量作为数据集，并预测。

未来所有站点在一天中不同时段客流量的预测：统计一天中不同时段的总客流量作为数据集，并预测。

【可交互界面的开发和设计】

利用 MATLAB 图形用户接口开发环境（GUIDE）建立 GUI 对象，作为用户的交互界面。

制作图形界面：利用一些列 GUI 对象制作出满足功能需求的图形界面。

编写回调函数：利用 CreateFcn、Callback 等回调函数，实现图形界面的交互功能。

编译程序：通过 MATLAB 的 deploytool 工具箱打包，产生可以在 Windows 操作系统中运行的可执行文件。

2.4 方案亮点

1、建立评价模型，从两个角度入手，寻找可能性最大的换乘点。

2、创新性地将遗传神经网络和时间序列结合在一起，扬长避短。

3、采用遗传神经网络，优化权重和阈值。

4、对指标采用独热编码，巧妙地解决了指标的赋值问题。

5、所编写的预测系统具有实时计算功能，能对相同类型的数据进行分析预测，具有适用性。

综上所述，本项目组的方案亮点由下图所示：

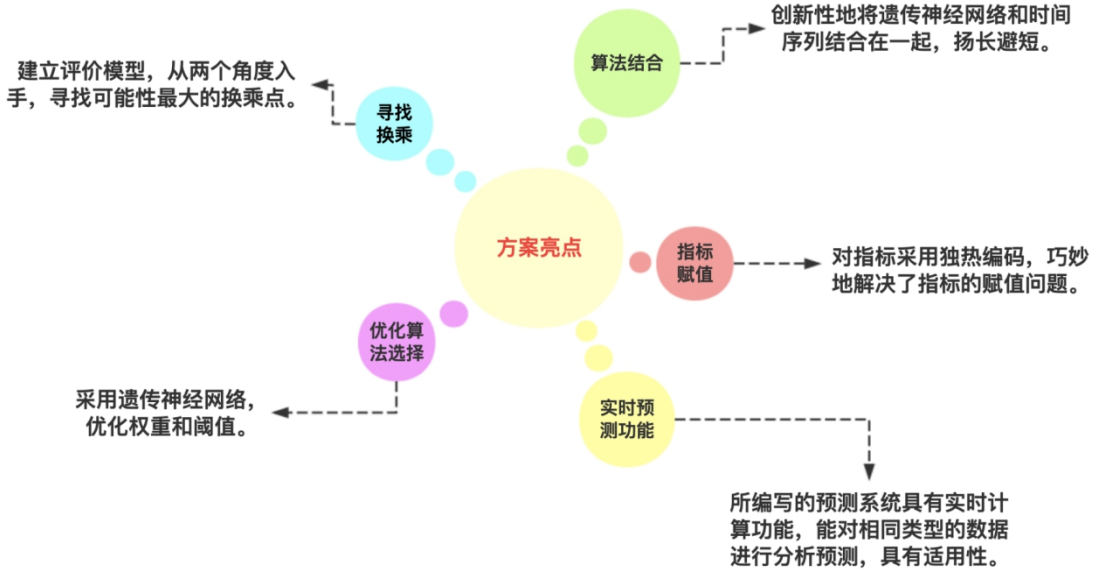


图 4：方案亮点

S2B 技术路线及实现方案

S2B 1.方案总流程图

本项目的方案总流程图如下图所示，主要包括数据导入、数据处理、客流量分析和客流量预测和预警等。

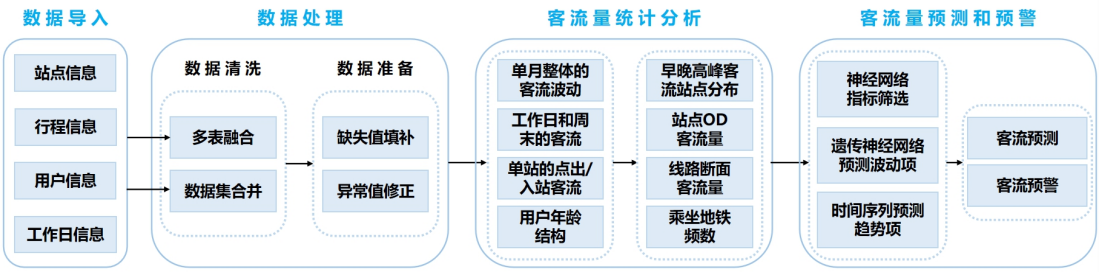


图 5：方案总流程图

S2B 2.数据获取与处理

2.1 数据获取

本文使用的数据来自服务外包竞赛组委会提供的赛题数据。如图 5 所示，数据主要包含 trips 数据集中 80 万条行程信息、users 数据集中 12 万条用户信息，station 数据集中的站点信息和 workdays 数据集中 2020 年节假日信息，涉及用户乘车记录、站点列表、用户基本信息以及日期的工作日和节假日属性等。数据已经过脱敏处理，即保证无法通过数据追溯到用户本人真实信息，但基本保持原数据趋势。具体内容如下图所示。具体内容如下图所示：

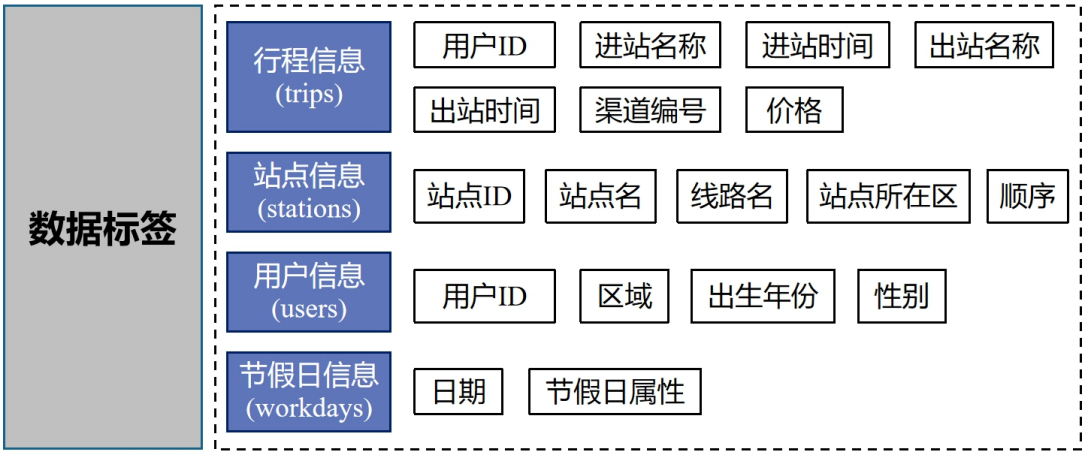


图 6：数据标签集

2.2 数据处理

①数据导入

在 Matlab 软件中运用 readmatrix 函数，将赛题组委会提供的四个 Excel 数据集 station、trips、users 和 workdays2020 导入 Matlab，分别命名为 Data_stations、Data_trips、Data_users 和 Data_workdays2020，如下图所示，此时各个数据集皆转变为字符串矩阵。

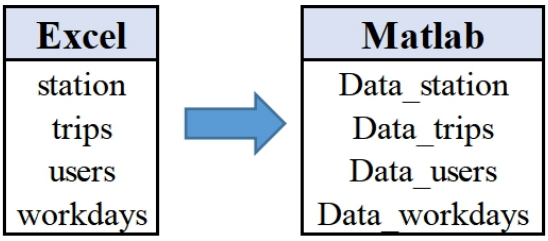


图 7：数据导入名称图

②字符串转换为数字

数据集导入后，皆以字符串的形式存在。Matlab 的计算只有识别数字形式才能运行，所以要将各个数据集中的字符串形式转换为数字形式。转换方法以进站时间为例，具体介绍如下：

Step1：在 Matlab 中将进站时间信息的字符串形式，例如“2019/12/25　23:30:00”，通过 char 函数拆分为一个个的字符。

Step2：利用 str2num 函数把分开的一个个字符转变为数字形式，同时，Matlab 会将这些数字按类别合并，如将表示年份、月份、日号和时间的数字分别合并在一起，最终得到每一条数字形式的进站时间信息，便于后续的统计应用。

③删除乘车时间异常数据

参考组委会答疑和全国各地地铁运营时间的信息，我们团队将数据集 trips 中进站或出站时间在 5:00 以前或 24:00 以后的数据信息予以剔除。基于前文的字符串转数字操作，

利用 find 函数找出 Data_trips 中进站或出站时间在 24:00~5:00 的行程，然后整行删除。

④删除金额异常数据

根据组委会提供的信息显示，正常票价规则最低 2 元 ，最高 10 元。在实行优惠票价的时候 ，最高封顶 7 元。在数据集 trips 中票价以分为单位，所以都应该为 100 的倍数，不存在票价含有十位或个位数字。在 Matlab 中，基于前面字符串转换为数字的操作，同样地，将票价信息的字符串形式也转变为数字形式。然后，我们使用逻辑判别的方法，筛选不符合事实依据的票价信息并且予以剔除。

Step1：利用 mod 函数，计算出票价数据除以 100 所得余数，若余数为 0，在右侧记为“0”；若余数不为 0，则记为“1”。

Step2：利用 find 函数，找出所有低于 200 或高于 1000 的票价，这些数据记为“1”。

Step3：通过上述处理，所有票价数据均在右侧被赋值为“0”或“1”，再次利用 find 函数，找出所有票价数据被赋值为“1”的位置。

Step4：删除票价数据被赋值为“1”的整行数据。

⑤删除同站进出数据

在数据集 trips 中，我们发现存在许多同站进出的用户信息。我们考虑可能是地铁员工每日巡检产生的线上测试数据，也有可能是乘客改变乘车念头进入车站后又从车站离开。考虑到这些数据对统计分析的干扰，我们运用逻辑判别的方式，筛选出同站进出的用户信息并且予以剔除。

Step1：在 Matlab 中运用逻辑判别式“==”，判断 Data_trips 数据集中各用户进出站是否相同。若进出站相同，则记为“1”；进出站不同，记为“0”。

Step2：同样地，利用 find 函数，找出所有进出站记为“1”的用户信息，予以整行删除。

⑥修改节假日属性异常数据

组委会提供的数据集 workdays2020 中，第 115 条记录，20200424 这一天的节假日属性是“lx”。经过查询，发现这天为工作日，属性应为“1”。我们通过 Matlab，将 Data_workdays2020 数据集中的这一错误修改为“1”。

经过核对，我们发现许多日期不能对应其属性，日期的属性是乱序。查阅 2020 年日历和节假日信息，在 Matlab 中对所有日期属性重新进行赋值。

S2B 3.统计分析

2021 年全国综合城市交通调查显示，2020 年地铁乘车占比为 54%，公交占比为 34%，打车和自驾占比为 11%，非机动车占比 1%，这意味着地铁的客流量和机动性比其他系统更大。此外，上述调查还显示，2020 年中国地铁日均客运量为 5662.19 万人次。可以说，地铁已经成为许多通勤者的首选公共交通工具，导致了地铁的高峰期。

城市轨道交通最大的特点即其客流是动态的。从小范围来看，其受到个人的主观意识和外在环境因素影响，会随着时限、天气等因素而变化。这种变化是交通系统本身特

征的反映，也是当下的社会经济活动的体现。而总体表现就是各个不同个体的总和，最明显的特征就是**不均衡性**，而客流的分布不均表现在**时间**和**空间**两个方面。时间不均是指在一段较长时间内，不同时间节点的客流量分布存在明显差异，如因为学习或工作带来的早晚高峰；空间不均则体现在某些位于景观、商业场所附近的站点拥有的远超过其他站点的客流量。

其中影响城轨交通客流量的另外一个重要的特征就是**日期性质**。某一时间段内的客流激增往往发生在国家法定节假日、商场促销、体育比赛或者其他社会性活动举办的时候。

因此将结合以上几点特征，针对用户需求进行统计分析。

3.1 单月整体客流波动分析

客流量是指单位时间进入某个场所的人数，是反应该场所人气和价值的重要指标。在本文中，将客流量定义为一定区域、一定时间内进站人数与出站人数的总和。

由于所提供的数据中部分月份日期信息不完整，存在明显数据缺失，这些月明显不具有广泛预测意义，因此我们进行统计的月份只选取有客流量数据>25 天的月份进行统计。Matlab 操作流程如下:

Step1：用 char 函数将 Data_trips 数据集中的日期字符串拆分为单个可提取的字符。将日期字符串中的年、月、日信息提取出来。

Step2：利用 num2str 函数将日期字符串中的年、月、日信息转变为数字。

Step3：使用 unique 函数，筛选出年、月、日信息中的非重复项，亦即数据集中包含的所有日期；再次使用 unique 函数，筛选出年、月信息中的非重复项，亦即数据集中包含的所有月份。

Step4：创立元胞数组，统计出每一个月份中包含的天数，如果该月份所包含的天数<26，则删除该元胞；如果该月份所包含的天数>25，则使用 histc 函数，计算出该月份中每一天的客流量。

Step5：使用 bar 函数，画出符合条件的月份客流量波动的柱状图。

经计算，12 月份和 7 月份的数据不完整，分别只包含了 6 和 16 天，所以不对此进行统计分析，剩下 6 个月份客流量波动的柱状图如下所示：

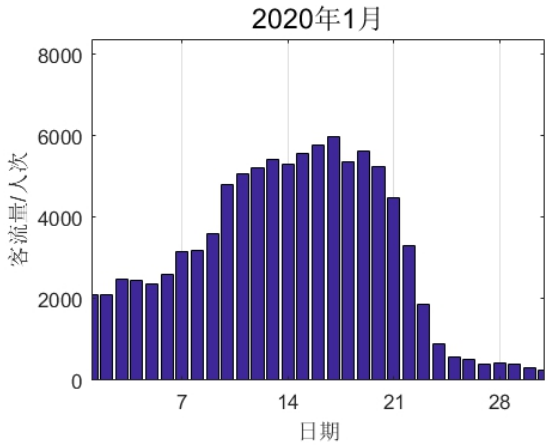


图 8：1 月的客流量

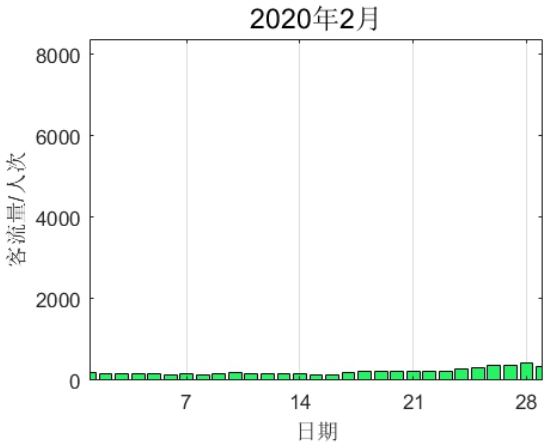


图 9：2 月的客流量

由统计结果观察得该地区 1 月 22 日左右客流量呈线断崖式下降趋势，推测由于该段时间疫情在该地爆发，出行受到影响。同时疫情因素的影响导致二月份的出行人次整月均维持在极低的区间内。该统计情况与实际较为符合。

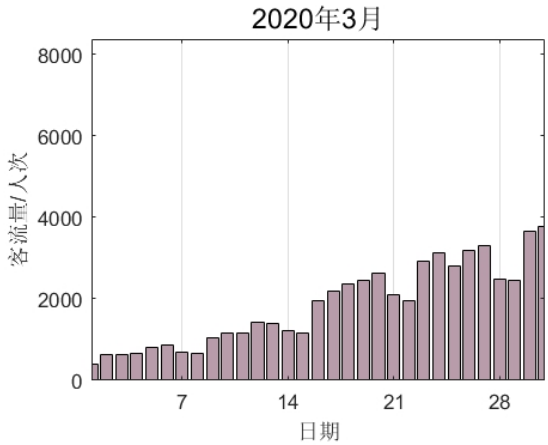


图 10：3 月的客流量

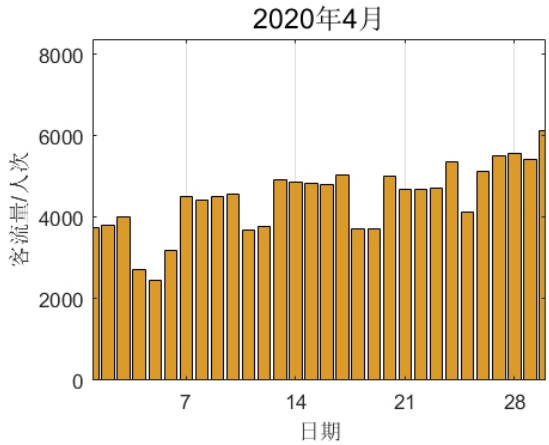


图 11：4 月的客流量

由上图观察得三月客流量随着疫情的快速控制出现了回升的趋势，开始向正常的范围靠近。在疫情防控措施的积极影响下，居民出行积极性逐渐回升。以三月五日为例，实际情况中，工业企业复工率多地提高，如山东达到 99.5%（前日 91.1%）。

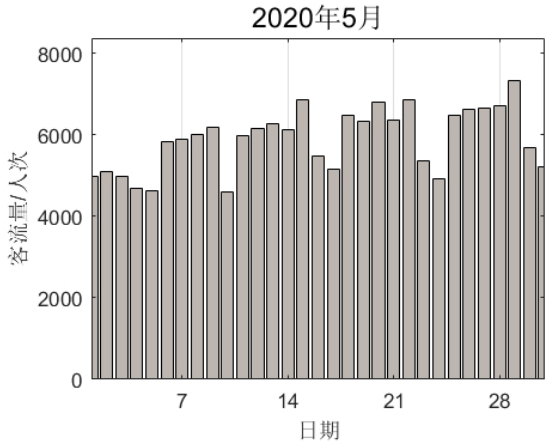


图 12：5 月的客流量

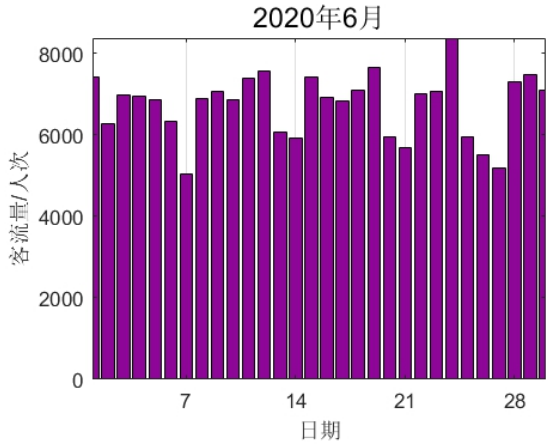


图 13：6 月的客流量

对比图 10，11，12，虽然 4 月的客流量接近平稳但是还是没有完全恢复正常水平，5、6 月客流量已逐渐恢复至疫情前的水准。推测该时间段内，疫情情况有所好转同时复工复产等问题被提上日程，返工人数增加，刺激地铁等出行行业状况回温。6 月疫情得到有效控制，旅游业有所回升，城市人口增加，地铁出行客流量回到一个较为稳定的高值。且可以发现，无论是否考虑疫情因素，客流量始终呈现周期性的波动，每个高峰之间始终有一两天客流量较低但数值相似。

3.2 工作日和周末的客流分析

使用 Matlab 进行操作的流程如下：

Step1：和预测单月整体客流量波动一样，将 Date_workdays 的数据用 char 函数和 stru2num 函数处理，输出为 workdays_table。需要将工作日和周末区分，设立一个因子

w=2 作为基础值。所给第一天 2020.01.01 为星期三，则 $w_{天}$=w+1。而一个星期只有 7 天，因此当 w 天>7 时，需要减去 7 才可以得到正确的星期天数。使用 Matlab 进行操作的流程如下：

Step1：和预测单月整体客流量波动的思路一样，将 data_workdays 的数据用 char 函数和 str2num 函数处理，输出为 workdays_table，包含日期和节假日两列信息。

Step2：利用 char 函数和 str2num 函数，将 data_trips 中的所有行程数据中的年、月、日信息提取出来，再用 unique 函数，找出年、月、日信息中的非重复项，合并成一列，命名为 data_trips_unique。

Step3：使用 histc 函数，计算出 data_trips_unique 中每一天的客流量。

Step4：找出 data_trips_unique 中的所有日期在 workdays_table 中的位置，并将该位置的节假日信息赋值给 data_trips_unique 的对应行。

Step5：编写函数，命名为 week，当输入年、月、日信息时，该函数能返还相应的星期值（例如，输入 20200101，返还 3）。

Step6：创建 7 个变量，命名为 workdays_1、workdays_2、workdays_3、workdays_4、workdays_5、workdays_6、workdays_7，若 data_trips_unique 中日期的节假日类型不为 3，则根据日期所代表的星期几，将该行数据放入对应的 workdays 变量中。

Step7：再将分好的数据用 unique 函数划分为 7 个区间，即为一个星期的七天，输出为 workdays_1、workdays_2、workdays_3、workdays_4、workdays_5、workdays_6、workdays_7,如果选取的一天属于工作日且不包含在 3 内，则就放入对应的集合中。

为了直观明了地识别数据批中客流量的差距及数据批的偏态和尾重，这里使用到箱形图来反映工作日和周末的客流量变动。

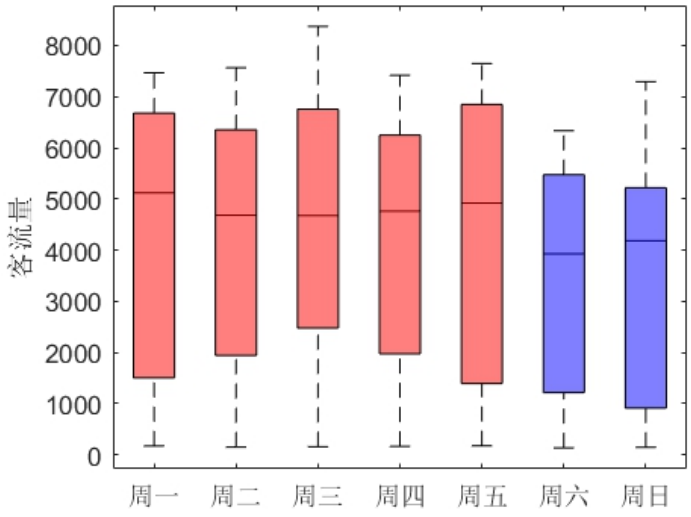


图 14：工作日和周末的客流量分析

如上图所示，最低客流量的差别不是很大，基本持平，工作日的最大客流量除了周三明显高于其他几日之外，周一、二、四、五的最大值差距不是很大，都在 7500 人次左右，但周三的客流量波动也是七天里最大的。周末的最高客流量均低于工作日，其中周六的数据差距特别明显，中位数是 7 天里最低的。

周末的上四分位点相似，都接近工作日客流量的中位数，周日的客流量跨度和工作日几乎没有差别，但总体客流量低于工作日。前五天内，周一和周五工作日客流量普遍大于其他三天，周二至周四的差距并不明显。

考虑周一至周五为工作日，且地铁出行的用户信息年龄分布基本位于 20-40 岁之间，可以推断工作日客流量显著高于周末是因为很大一批上班族和学生都选择了地铁作为出行的交通方式，这与地铁便利的属性有密切的联系。而原本应该作为外出游玩的周末的客流量却在总体上小于前五日，可能是因为周五刚刚结束一周的工作，周六起床大多会比较晚，空出一天的时间来调整自己的状态，而周日出行的频率略微大于周六，但总体上差距并不大，代表只有部分人愿意出门，还有一批人更愿意呆在家里。

3.3 单站的点出/入站客流分析

3.3.1 操作过程

要反映出一天中不同时间段的点出/入站，如果算一天中总的点出/入站数，意义不大。因此我们画出一个具体站点在一天中的不同时段，点出/入站的人数的变化。Matlab 操作步骤概括如下：

Step1：使用 char 函数，将 Data_trips 里的进站和出站信息的字符串转变为独立的字符，再使用 num2str 函数将字符转变为数字，将数字类型的进站和出站信息分别存储在名为 sta_in 和 sta_out 的变量中。

Step2：使用 char 函数，将 Data_trips 里的进站和出站的时间信息的字符串转变为独立的字符，再使用 num2str 函数将包含年、月、日、小时的字符转变为数字，将数字类型的进站和出站的时间信息分别存储在名为 time_in 和 time_out 的变量中。 将 sta_in、time_in 和 sta_out、time_out 合并，命名为 sta_group。

Step3：对于给定的站点，筛选出 sta_group 的所有包含该站点的行，并用 histc 函数，以 1 至 24 为时间间隔，计算出各个时间间隔内的客流量大小。

3.3.2 客流量-时间折线图的类型分析

通过折线图的形状，我们发现单站的点的出/入站的客流分布可以分为三种类型，分别为单峰型、双峰型和多峰型，下面依次举例做简要介绍。

- **单峰型**

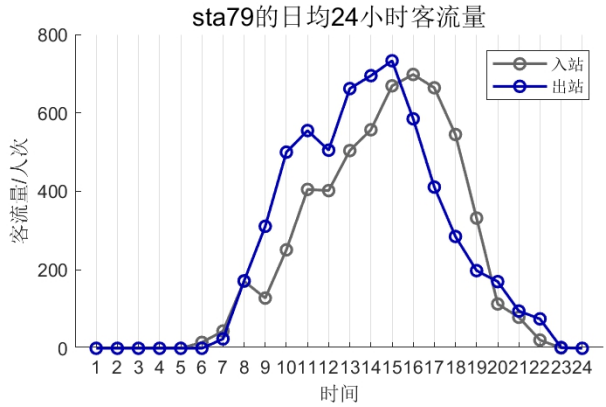


图 15：sta79 的日均 24 小时的客流量

如上图所示，sta79 表现为单峰形，即出/入站均表现为一个客流分布高峰，在所有站点中客流分布为单峰型的站点占比 14.29%。这类型站点的客流量，随时间逐渐增加，一般在 14:00-16:00 达到顶峰后逐渐回落；并且站点的出/入站高峰几乎彼此重叠，客流变化趋势大致相同。该类站点一般位于交通枢纽处，基本上全天的客流量都不低，与时间的变化密切相关。

- **双峰型**

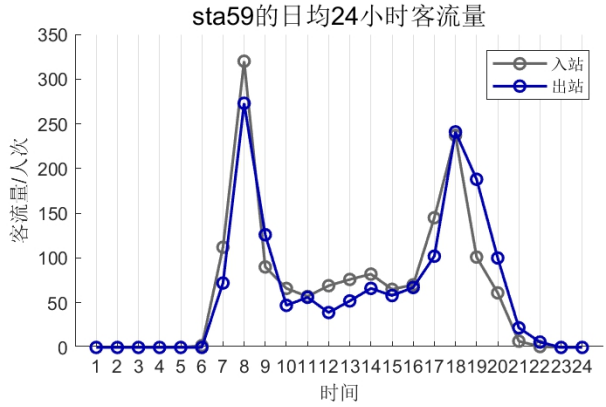


图 16：sta59 的日均 24 小时的客流量

根据所有站点的客流-时间数据，画出的折线图中双峰型站点占到了站点总数的 56.55%，该类站点出/入站均有两个客流高峰。如 sta59 呈现的结果，客流量在上午 8-9 点达到出站和入站客流高峰，17-18 点达到出入站客流的第二高峰，且与“其他时间”段的客流量差距悬殊。双峰型是最为典型的单站的点出/入站客流时间分布类型，该类站点一般位于居民区或者工作区，这些站点的客流量受到早晚高峰的影响很大。

- **多峰型**

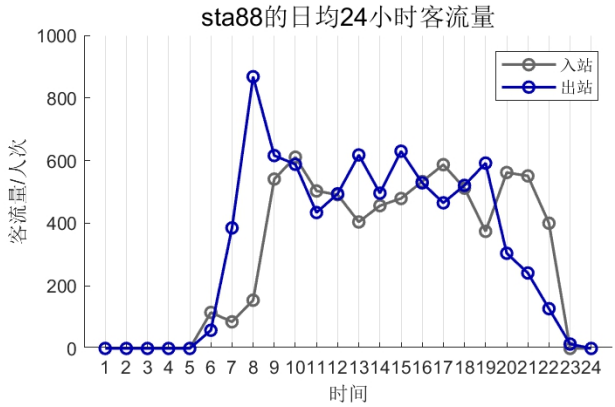


图 17：sta88 的日均 24 小时的客流量

如上图所示，sta88 客流分布呈多峰型，折线图呈锯齿状或波浪状，该类型的站点在 168 个站点的占比有 29.17%。多峰型站点的特点为出/入站的客流趋势大致相同，在达到第一个客流高峰后，客流量呈锯齿状波动，呈现多峰样式，无明显低谷。该类站点一般属于对外枢纽类，客流分布随时间有小幅波动，呈波浪形，但一天的平均客流量处于较高水平。客流高峰时段起止点以及峰值大小均与该类车站所在的枢纽类别（机场、火车站或公路客运枢纽）以及运输组织（到发时刻表及到发量）密切相关。

3.3.3 站点类型分析

根据站点客流-时间折线图的类型，可以将站点的类型进行分类，可分为以下 6 类：

(1) **居住类。**此类站点周围以居民区为主，多数位于近郊区和城市边缘。工作日进站客流时间较为集中，进站时间分布呈单峰形态，早高峰以进站客流为主，晚高峰则反之，以出站客流为主。且早高峰客流量略高于晚高峰。单站的点出/入站客流-时间折线图主要呈现出/入站单峰型。

(2) **工作类。**此类站点周围主要是工作类办公楼，多数位于市区。工作日早高峰以出站客流为主，进站高峰时段主要发生在下班时间，所以单站的点出/入站客流-时间折线图主要呈现出/入站单峰型。

(3) **商业及文体景区类。**此类站点多数位于大型商业中心或体育娱乐中心周边，进、出站高峰出现时段差异较为明显，上午 9:00 之后以出站客流为主，下午 15:00 后以进站客流为主，高峰时段从 16:00—22:00 均有可能。折线图的总趋势为抛物线装，有小幅波动，以单峰型为主。

(4) **混合类。**此类站点周边多为混合用地 (居住+商业、居住+工作、居住+商业+工作等)，客流全日出/进站时间分布均有两个明显的早晚高峰，通常情况下两个高峰峰值存在一定差异；但是，居住与商业混合类出站客流人数分布比较平衡。混合类为大多数站点的类型，客流-时间折线图为单峰型或双峰型。

(5) **交通枢纽类。**此类站点位于城市对外客运交通枢纽地区的车站客流全日时间分布相对均匀，客流分布随时间有小幅波动且无明显低谷。客流高峰时段起止点以及峰值大小均与该类车站所在的枢纽类别 (机场、火车站或公路客运枢纽) 以及运输组织 (到发时刻表及到发量) 密切相关。所以，该类站点的客流-时间折线图为多峰型。

3.3.4 各线路单站的点出/入站客流分析

前文已经就站点客流量-时间折线图进行分类，按照此分类方法将各线路站点的折线图进行归类，并且分析出各站点的类别属性。

➢ **一号线**

就一号线各站点的客流-时间分布折线图，将一号线的 20 个站点分为单峰型、双峰型和多峰型三类。同时，根据具体的出/入站情况考虑该站点的类别属性。

● **单峰型**

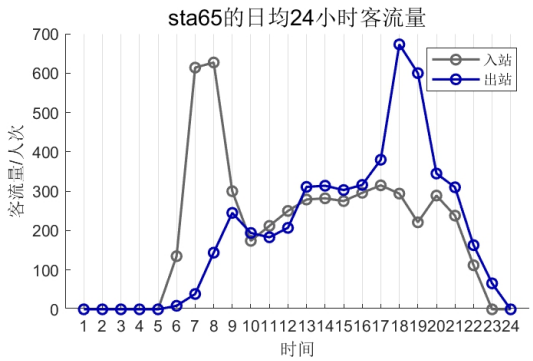


图 18：sta65 的日均 24 小时的客流量

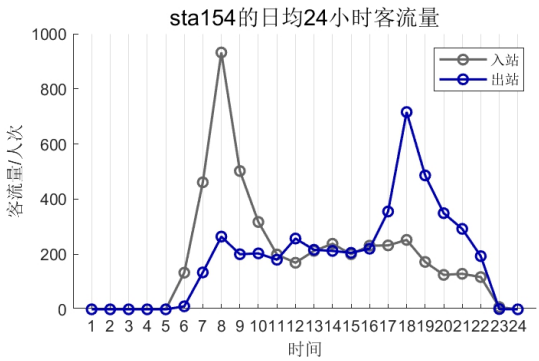


图 19：sta154 的日均 24 小时的客流量

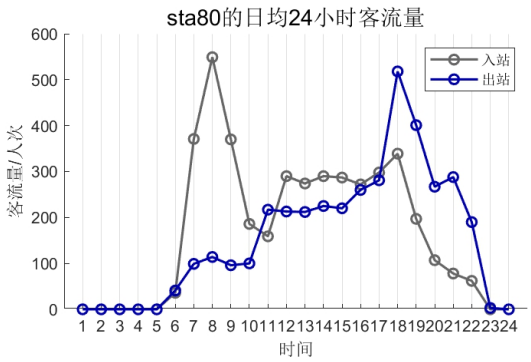


图 20：sta80 的日均 24 小时的客流量

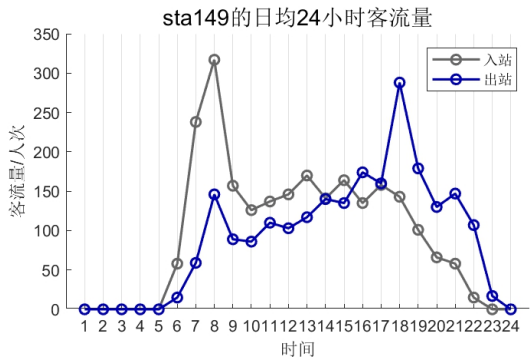


图 21：sta149 的日均 24 小时的客流量

Sta65/154：出/入站均有一个较为明显的高峰，高峰时间段差异显著。上午 7-9 点迎来入站高峰，下午 17-19 点迎来出站高峰，早高峰客流量略高于晚高峰。所以，这两个站点可以看作**居住类**，早晨入站去工作，下午下班出站回家。

Sta80/149：与前面两个站点相似，早晚有明显的高峰时间段，早高峰以入站为主，晚高峰则以出站为主。虽然这两个站点总体上呈现单峰型，但也存在别的峰值较小的时间段，如 sta80 在晚高峰有入站小高峰，sta149 在早高峰有出站小高峰。其中，sta80 客流人数较多可以看作**居住+商业混合类**；sta149 人数较少考虑为**居住类**。

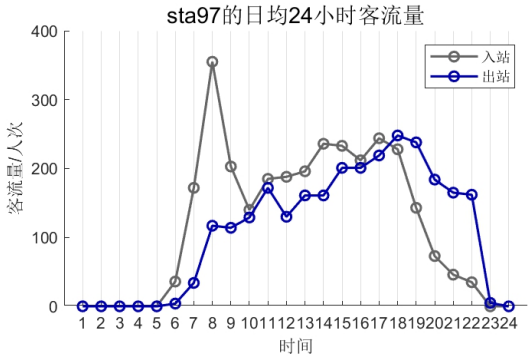


图 22：sta97 的日均 24 小时的客流量

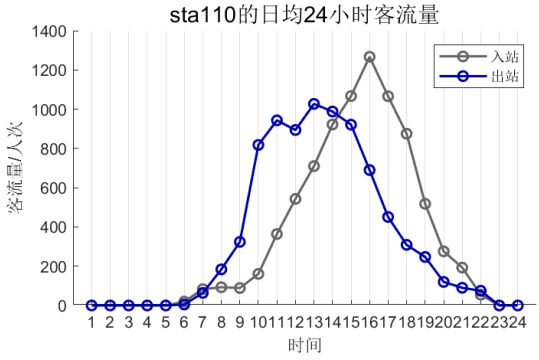


图 23：sta110 的日均 24 小时的客流量

Sta97：入站有一个明显的高峰时间段，在 7-9 点之间；出站没有一个显著的高峰期，随时间客流量呈缓慢上述后比较稳定的维持在一个水平。该站点可以看作**混合类(居住+商业)**，早高峰已入站为主，人们去上班；由于有商业区存在，出站人数比较均衡。

Sta110：出/入站高峰出现时段差异较为明显，上午 9:00 之后以出站客流为主，下午 15:00 后以进站客流为主，客流总量较大高峰时期超过 1000 人次。所以，该站点可以看作**商业及文体景区类**。

- **双峰型**

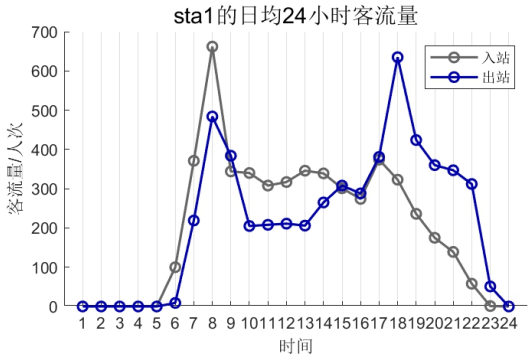


图 24：sta1 的日均 24 小时的客流量

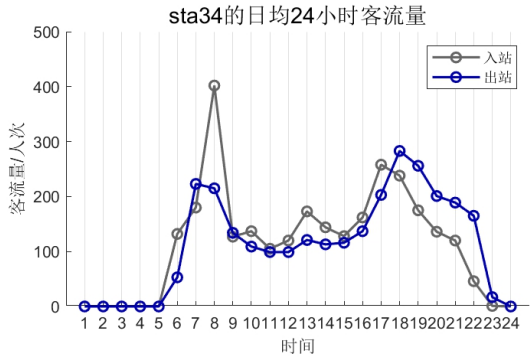


图 25：sta34 的日均 24 小时的客流量

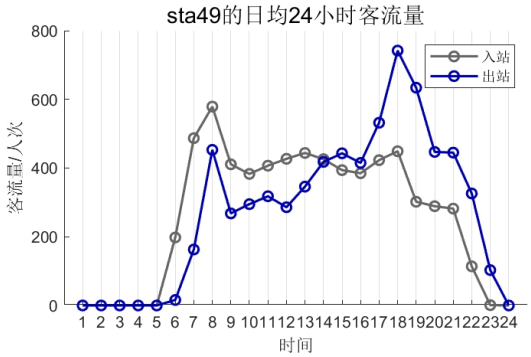


图 26：sta49 的日均 24 小时的客流量

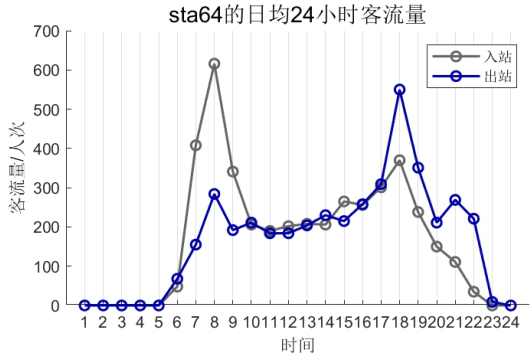


图 27：sta64 的日均 24 小时的客流量

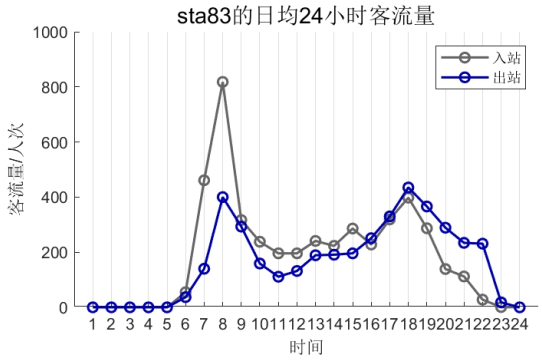


图 28：sta83 的日均 24 小时的客流量

Sta1/34/49/64/83：这些站点的共同特征是出/入站均有两个明显的早晚高峰，且入站早高峰客流量大于出站早高峰，而晚高峰则是出站客流量明显大于入站。同时，这些站点的客流量均较大，高峰时期超过 500 人次，考虑为混合类，且为**居住+工作混合类**，居住区比办公区的覆盖面积要大。

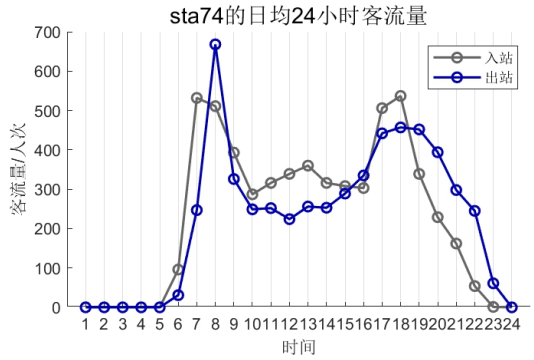


图 29：sta74 的日均 24 小时的客流量

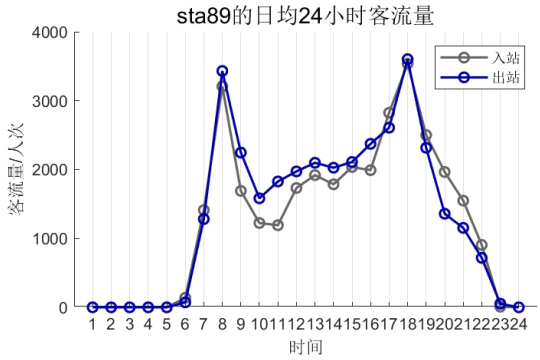


图 30：sta89 的日均 24 小时的客流量

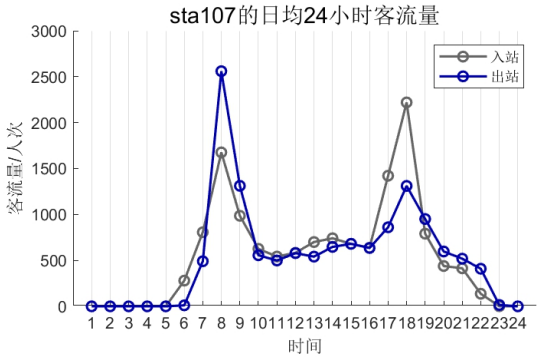


图 31：sta107 的日均 24 小时的客流量

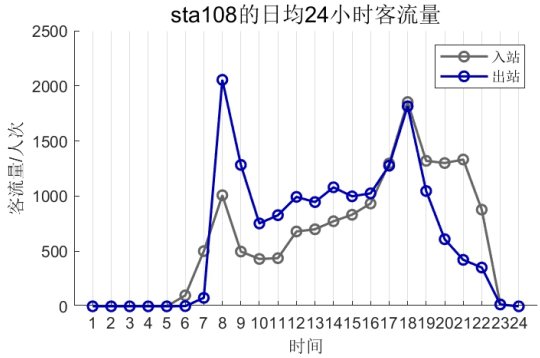


图 32：sta108 的日均 24 小时的客流量

Sta74/89/107/108：这几个站点与前面的相比，共同点是出/入站亦有早晚两个显著的高峰；不同点是出站早高峰客流量大于入站，晚高峰则变为入站客流量大于出站，这与前面的站店恰好相反。值得注意的是，sta89/107/108 客流量总数巨大，超过 1 万人次，并且高峰段内的客流有 2000-3000 人次，是地铁站点中的拥挤站点。可以肯定这些站点位于城市的中心地铁，为混合类型，考虑为**商业+工作+交通枢纽混合类**。而 sta74 人数相对较少，考虑为**工作+居住混合类**。

- **多峰型**

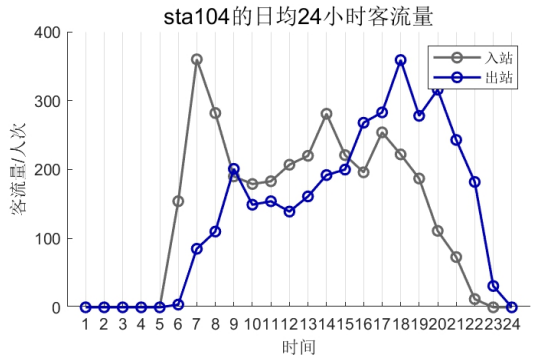


图 33：sta104 的日均 24 小时的客流量

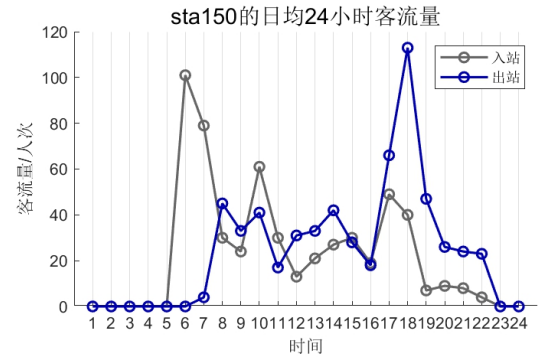


图 34：sta150 的日均 24 小时的客流量

Sta104/150：这两个站点出/入站均呈现锯齿状的多峰型，但出/入站均有一个明显的高峰期，客流人次显著多于其他峰。总体上还是呈现早高峰以入站客流，晚高峰以出站客流为主。所以，这两个站点可以看作是**居住类**。

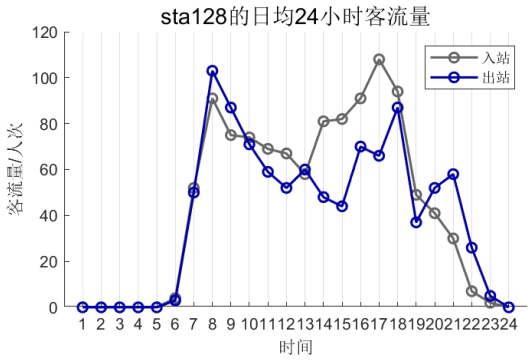


图 35：sta128 的日均 24 小时的客流量

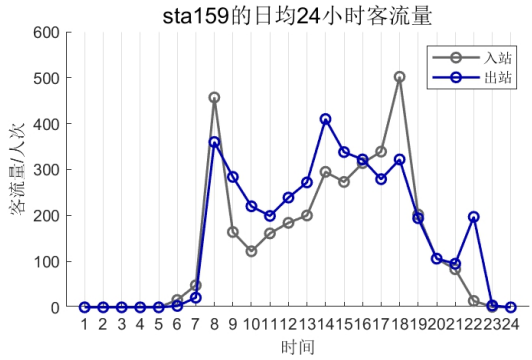


图 36：sta159 的日均 24 小时的客流量

Sta128/159：这两个站点也称多峰型表现，且出/入站均有两个明显的高峰段客流量比其他峰段的多，这两个高峰段主要位于 7-9 点和 17-19 点的早晚高峰时期。由于别的时间段也存在客流人数的波动，但是 sta128 人数较少，波动比例不大，所以考虑这站点为**居住+工作混合类**，而 sta159 人数较多，波动比例大，考虑为**居住+商业混合类**。

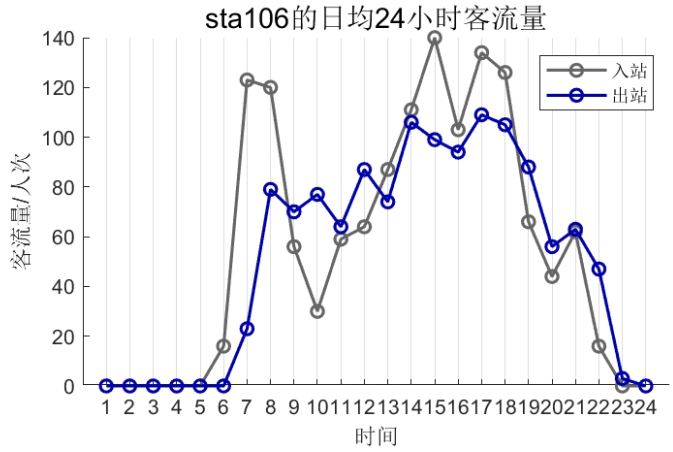


图 37：sta106 的日均 24 小时的客流量

Sta106：该站点入站客流量存在较明显的早晚高峰，而出站客流了则在达到一定水平后维持在这个水平上下波动，没有明显的早晚高峰，类似公交换成站，属于**交通枢纽类**。

3.4 用户年龄结构分析

3.4.1 操作过程

统计分析用户的年龄结构，在 Matlab 对数据的处理步骤如下所示：

Step1：利用 Data_users 数据集中用户的的出生年份信息，用“2020-用户的出生年份”，得到各用户的年龄。

Step2：利用 max 函数找出用户中的最大年龄。

Step3：以每五岁为一个年龄区间，创立元胞数组。并且以用户最大年龄除以 5，向上取整，得到需要创立的元胞数组个数。

Step4:统计每个元胞数组中所包含的年龄值的个数,即这个年龄区间内的用户人数。

Step5：每个年龄区间内的用户人数除以总人数，计算出各个年龄区间用户的占比。

根据上述步骤的数据处理，将得到的数据导入到 Excel 中进行画图操作。

3.4.2 年龄结构分析

根据组委会提供的数据，统计分析数据集 users 中用户的年龄结构，用条形图表示各年龄区间的人数，如下图所示：

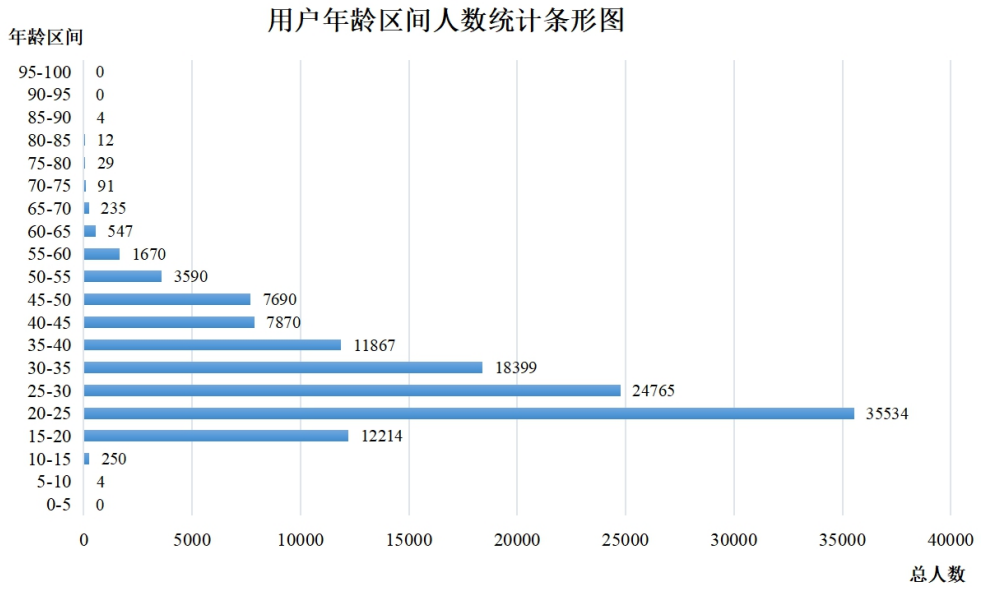


图 38：用户年龄区间人数统计条形图

如上图 38 所示,从整体上看,地铁乘客用户总数达 124771 人,年龄主要集中于 15-60 岁。其中，20-40 岁的青壮年为乘坐地铁的主力军，共有 90565 人。15 岁以下的儿童和 60 岁以上的老人用户乘坐地铁的人数非常少，只有 1168 人。特别是 5 岁以下的婴幼儿和 90 岁以上的老年乘客的人数甚至为 0，该年龄区间的用户几乎不存在独自出行的能力，也不能保证独自乘坐地铁的安全性，符合常理推断。

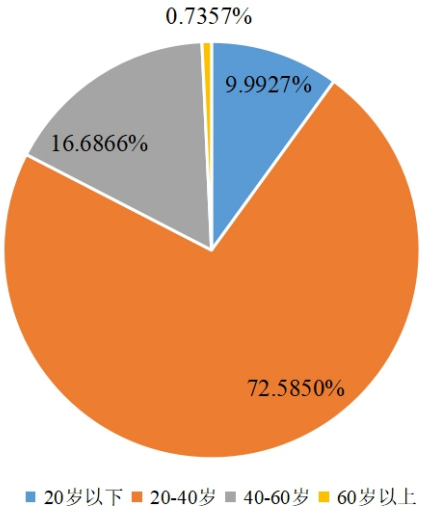


图 39：用户年龄分布饼图

按年龄区间从百分比上分析，由图 39 可知，20 岁以下用户占总人数的 9.9927%；20-40 岁用户占 72.5850%，这一比例远高于其他年龄段的用户占比；40-60 岁用户占比 16.6866%，60 岁以上用户只占 0.7357%。

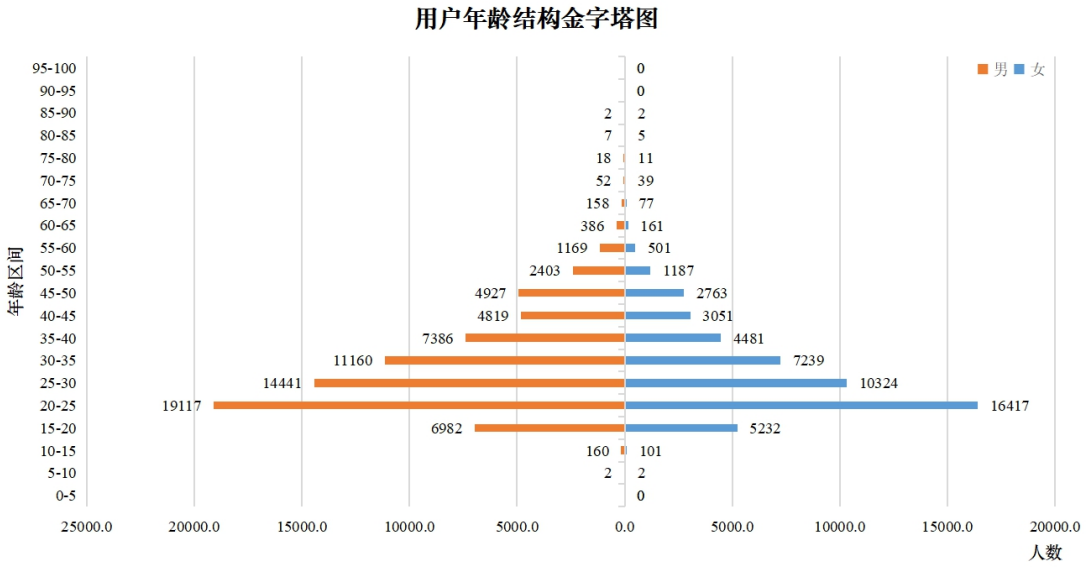


图 40：用户年龄结构金字塔图

图 40 显示了地铁乘客用户的各年龄段性别比例和年龄结构。在乘客用户总人数中，男女用户人数比为 10:7。而各年龄段的男女乘客比例差别不大，各年龄段平均男女乘客人数比为 3:2。可见，男性选择地铁出行的概率略大于女性。

从年龄结构上看，随着地铁逐渐成为许多人外出的首选公共交通工具，地铁乘客用户的年龄变化呈现出二个特征：

其一，**20-40 岁的青壮年是地铁乘客用户的主体。**根据用户的年龄金字塔，可以发现该年龄金字塔塔形下宽上尖，表示年轻人比重很大，为年轻型金字塔。从年龄结构上还能发现，男女比例大致趋同为 3:2，男性用户略多于女性用户，与社会人口的性别比例基本一致。20-40 岁这一年龄段的用户占总用户的 72.5850%，他们是乘坐地铁的主力军。随着城市化的发展，路面交通的压力与日俱增。对于巨大社会竞争压力下的当代年轻人而言，地铁成为便捷出行的不二选择。

其二，**儿童和老年人地铁出行人数较少。**受行动能力、文化水平以及智能设备的影响，15 岁以下的儿童或 60 岁以上以下的老人地铁出行率较青壮年大幅下降。其中，15 岁以下的儿童用户仅有 254 人，占比 0.20357%；60 岁以上用户则有 918 人，且集中于 60-70 岁年龄段。乘坐地铁需要用户有较好的行走能力、良好的地图识别能力等。近年来，智能手机、手表等设备的普及，地铁出行逐渐智能化，进出站都可以使用手机刷码。因此，儿童和老人选择地铁出行的人数较青壮年少的原因不仅是他们出行不便，无法准确选择地铁进站或出站；还因为他们受到智能设备使用的限制，不能便利得乘坐地铁。

3.5 早晚高峰客流站点分布分析

3.5.1 操作过程

Step1：用 char 和 num2str 函数将 Data_station 中的站点-线路信息提取出来，记为 station_line_double。

Step2：用 char 和 num2str 函数将 Data_trips 中的进站、出站站点以及进站、出站时间信息（年、月、日、时）提取出来，命名为 sta_group。

Step3：对于给定的线路，用 find 函数寻找出属于该线路的所有站点，对该线路的所有站点进行遍历，找出在 sta_group 中所有包含该站点的行，再用 histc 函数，以 1：24 为时间间隔，计算出各个时间间隔内的客流量大小。画图并分析。

3.5.2 一号线早晚高峰客流站点分布

选取一号线各站点每天 6:00-22:00 每小时进出车站的总人数，以时间和站点序号为横坐标，客流量为纵坐标，画出一号线各每小时客流站点分布的三维直方图，如下所示：

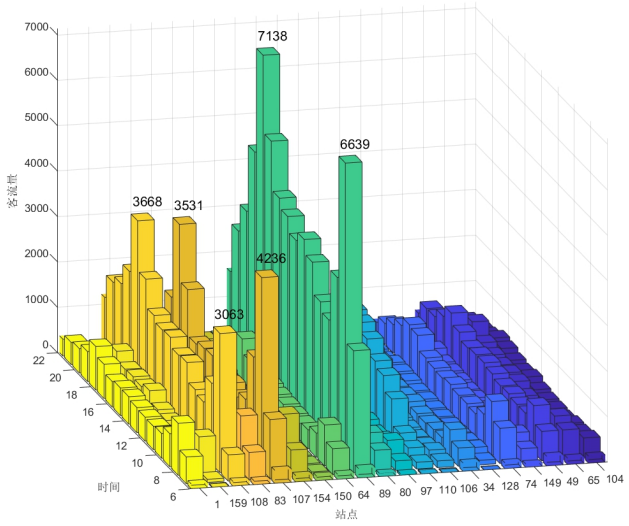


图 41：一号线各站点一天中客流量分布图

由上图所示，一号线大多数站点一天中的客流分布趋势大致呈双峰型，存在早晚高峰；但也有站点只存在早高峰或晚高峰，甚至不存在高峰期。图片显示，站点早高峰时段大多发生在 7:00-9:00，而晚高峰则主要集中于 17:00-19:00。也有一些站点高峰期发生在“其他时间”段。

同时，我们引入高峰小时比率，即站点高峰时间段一小时内的客流量占该站点总客流量的比率，来表示该站点的客流分布的时间特征。

$$\text{高峰小时比率}=\frac{P_{\text{peak}}}{P_{\text{total}}}\times 100\%$$

其中，P_{peak} 表示站点高峰时间段一小时内的客流量；P_{total} 表示该站点总客流量。具体数据如下表所示：

表 1：一号线各站点高峰时间段及客流量数据表

站点序号	104	65	49	149	74	128	34	106	110	97
高峰时间段1	7:15-8:15	7:30-8:30	8:00-9:00	7:45-8:45	8:00-9:00	7:45-8:45	7:45-8:45	7:45-8:45	15:15-16:15	8:00-9:00
高峰时间段客流量1	498	857	1032	482	1179	199	631	223	2038	472
高峰小时比率1	7.447%	8.821%	7.907%	10.635%	10.652%	9.693%	11.437%	8.574%	12.388%	8.474%
高峰时间段2	17:45-18:45	18:45-19:45	18:00-19:00	18:00-19:00	17:45-18:45	17:45-18:45	17:45-18:45	17:45-18:45		17:45-18:45
高峰时间段客流量2	647	939	1191	431	1015	186	572	249		516
高峰小时比率2	9.675%	9.665%	9.126%	9.510%	9.171%	9.060%	10.368%	9.573%		9.264%
总客流量	6687	9715	13051	4532	11068	2053	5517	2601	16452	5570
站点序号	80	89	64	150	154	107	83	108	159	1
高峰时间段1	7:30-8:30	8:00-9:00	7:45-8:45	7:45-8:45	7:45-8:45	8:00-9:00	8:15-9:15	8:00-9:00	8:15-9:15	7:45-8:45
高峰时间段客流量1	677	6639	997	114	1213	4236	1235	3063	839	1157
高峰小时比率1	8.559%	10.301%	12.076%	9.974%	13.296%	15.772%	14.453%	10.057%	11.421%	11.228%
高峰时间段2	18:00-19:00	18:00-19:00	18:00-19:00	17:45-18:45	18:15-19:15	17:45-18:45	18:00-19:00	18:00-19:00	17:45-18:45	18:15-19:15
高峰时间段客流量2	857	7138	920	194	976	3568	833	3668	833	970
高峰小时比率2	10.834%	11.075%	11.143%	16.973%	10.698%	13.285%	9.748%	12.043%	11.340%	9.413%
总客流量	7910	64451	8256	1143	9123	26858	8545	30457	7346	10305

由上表可知，在高峰时间段 1 中，73.69%的客流分布早高峰时间集中在 7:45-9:00。其中，7:45-8:45 占比 42.11%，8:00-9:00 占比 31.58%。同时该高峰时间段的选择符合常规对早高峰的时间定义。而对于晚高峰而言，客流分布时间段有 84.21%处于 17:45-19:00 之间，17:45-18:45 时间段占比 47.37%，处于 18:00-19:00 时间段的有 36.84%。参考国务院规定的 18:00 下班时间，由上表统计分析出的地铁各站点出现晚高峰的时间段也符合客观现实。

然而，部分站点比较特殊，没有出现早晚高峰，站点的客流分布呈单峰型，如一号线的站点 110，高峰时间段出现在 15:15-16:15 之间。推测此类站点位于商业密集区和居住区以外，具有特殊的社会性质。

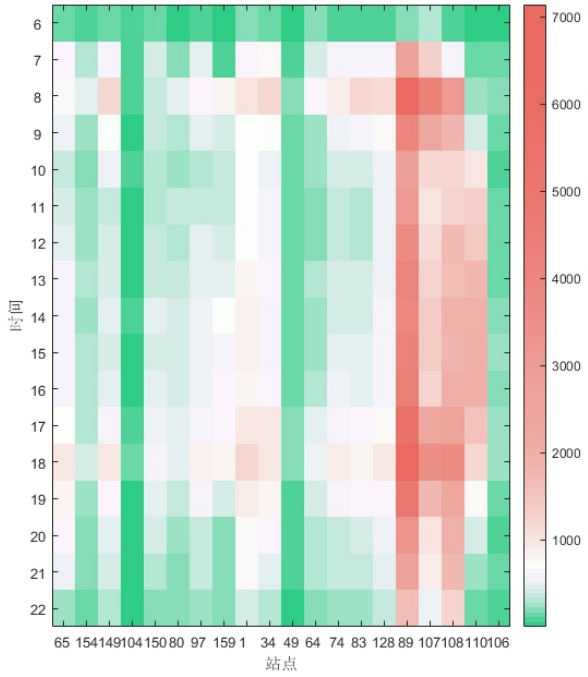


图 42：一号线各站点客流量热力图

由上图 42 可知，从总客流量分析，89、107 和 108 三个站点的客流量明显高于别的站点，分别为 64451、26858 和 30457 人次。所以，在这些站点高峰时间段内的客流量也远高于其余站点。特别是站点 89 在早高峰时间段的客流量为 6639 人次，晚高峰则达到 7138 人次，甚至超过一些站点的总客流量。站点 107 早高峰时间段的客流量为 4236 人次，晚高峰为 3568 人次。而站点 108 早晚高峰时间段内的客流量则分别为 3063 和 3668 人次。

根据上表高峰时间段及客流量统计显示，一号线站点进站客流平均早高峰小时比率为 10.567%，其中，站点 104 早高峰小时比率最低，仅为 7.447%，说明该站点早高峰客

流分布少；107 站点的早高峰小时比率最高，达到 15.772%，站点 108 紧随其后，表明这两个站点早高峰客流分布较多。而对于晚高峰来说，平均晚高峰小时比率为 10.630%，107 和 108 站点依然为晚高峰客流分布比较明显的两个站点。

综上所述，一号线早晚高峰的客流分布主要集中于 89、107 和 108 站点。

3.6 站点 OD 客流量分析

Step1：首先使用 char 函数将 Data_station 中的站点-线路信息中的字符串拆分为独立的字符，再用 str2num 函数将字符转变成数据，得到 8 条线路的各个站点信息，命名为 sta_and_line。

Step2：使用 char 函数将 Data_trips 中的进站站点、出站站点信息的字符串拆分为独立的字符，再用 str2num 函数将字符转变成数据，得到所有行程的进站、出站信息，命名为 sta_OD。

Step3（弦图的绘制）：创立一个 8*8 的零矩阵，命名为 line_to_line_flow，对 sta_OD 进行遍历，计算并在 line_to_line_flow 里存储线路之间的 OD 数据。绘制线路间 OD 客流的弦图。

Step4（弦图的绘制）：手动设定需要画图的线路，创立一个零矩阵，矩阵大小为选定线路的站点数量，命名为 sta_to_sta_flow，对 sta_OD 进行遍历，在 sta_to_sta_flow 里统计出选定线路中站点间的 OD 数据。绘制给定线路站点间 OD 客流的弦图，如下所示：

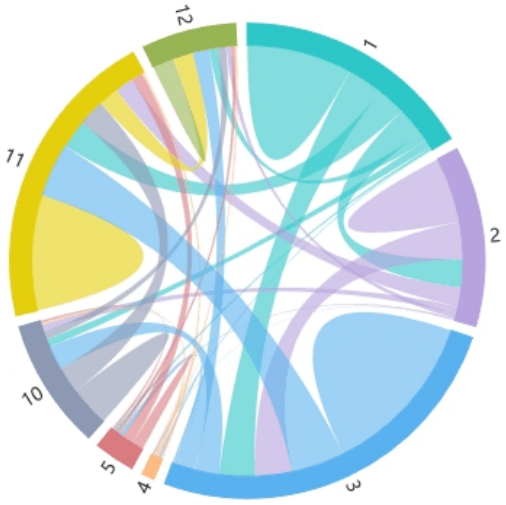


图 43：不同线路之间的 OD 客流量弦图

如上图所示，不同路线中各站点之间总体的 OD 客流量达到了 740741 人次。分析各线路之间的客流，线路 3 和线路 11 的 OD 数量分别为 200285 人次和 159847 人次，均达到了 150000 人次以上，占到了总 OD 数量的 27.038%和 21.579%，两线路基本占据了总客流量的一半左右；线路 1 位居第三，也有 125723 人次，占到总数的 16.973%。相比之下线路 4、线路 5 以及线路 12 的运营压力不是很大，根据数据显示客流量才达到总 OD 的 0.929%、3.073%及 7.100%。

一条线路内，OD 客流量较多的线路，本身线路内的人员流动基本接近总人数的 50%，其中线路 3 依然是流动最多的一条线路，共有 101316 条客流量记录，占到第 3 线路总量的 50.586%。而 OD 客流量少的 4、5、12 则明显表现出自身线路流动数量远远低于 50%的情况，分别占了自身线路的 10.722%、32.163%、24.951%。所有线路中客流

量大的自己线路内部的交流明显更为频繁，而总客流量小的线路与其他线路的来往占据了总量的大多数。

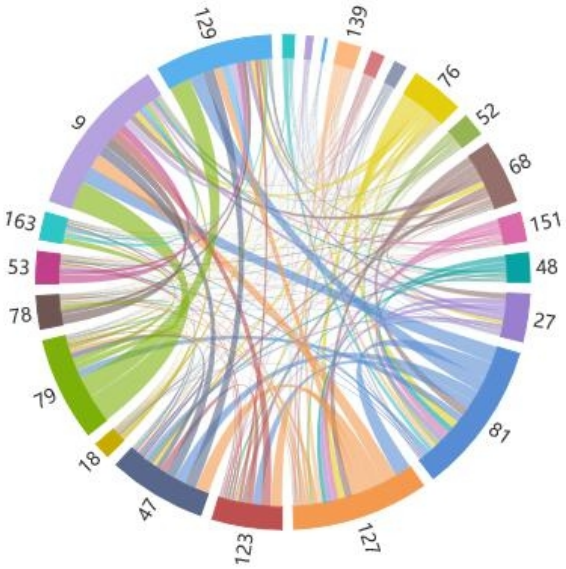


图 44：线路 1 内各站点之间的 OD 客流量数据弦图

单独拿出一条线路的弦图进行分析，我们选取稍微靠近中间但是总客流量又不是太少的线路 1 作为例子。线路 1 总共有 23 个站点，其中 9、79、47、127、81 共 5 个站点的客流量占据了总客流量的 50.247%，而其中客流量数据小于等于 5 的出入站点对共有 145 个，达到了运行线路数量的 27.410%，这里面包括了客流量为 0 的数据，当去除 48 个为 0 的数据之后，满足上述条件的站点对占比 18.336%，依旧有近 1/5 的量。

反观客流量较大的数据，大于 100 的有 122 条，发现与小于等于 5 的数据差别不大，只占到总数的 4.348%，而达到 450 以上的站点对相对前者则更少，只有 15 站，但是每个站点的进出站客流量基本相等，且波动情况不是很大。可以看出客流量数据与数据的稳定性存在一定的反比关系，即客流量越大数据浮动越小，反之客流量总数越小则站点内进出站人数差距越大。

3.7 线路断面（按站点）流量分析

线路断面，定义为在单位时间内，沿同一方向通过轨道交通线路某断面的乘客数量，即通过该断面所在区间的客流量，分为上行断面客流量和下行断面客流量。上行断面客流量为顺着地铁站点顺序乘坐的客流量，下行断面客流量为逆着地铁站点顺序的客流量。

然而，用户通过换乘可以增加换乘后经过站点的客流量，而数据中仅有进出站的信息，并不知道换乘站点，就无法知晓换乘后经过的站点。所以，各线路之间的换乘关系，即**确定换乘站**，是本题统计分析的关键。

换乘站是地铁轨道交通的重要节点，不同线路之间的内部换乘是换乘站的重要功能之一，也是影响客流量分布的重要因素。换乘站的确定存在一定的难度，但我们应用符合实际的原理和评价模型，确定了各线路之间的换乘站，从而可以分析出线路段明的客流量情况。

3.7.1 确定换乘点的原理

①换乘点客流量大

根据常识和生活经验，换乘点一般位于市中心或者每个区域的交界处，人群密集，客流量大。根据前文的统计和分析，可以知道每条线路中各站点的客流量。筛选出客流量较大的站点考虑为换乘点。

②换乘点应使得行程时间符合一定规律

对于用户每次行程的时间可以分为两个部分：一是乘坐地铁的时间，由乘坐的站点数和每站经过的时间确定：

t_出站 - t_进站= t_每站经过的时间*本次经过的站点数 n +其他时间

二是“其他时间”，即站内刷卡、行走、等待和换乘等时间：

其他时间 =t_出站 - t_进站 - t_每站经过的时间 * 本次经过的站点数 n

乘坐地铁的时间每个人大相径庭，无法比较；但是“其他时间”，却存在一定的相似性。换乘用户的“其他时间”一般大于不换乘用户，所以根据不换乘用户在乘坐地铁时花费“其他时间”的频数分布，可以大致确定“其他时间”。然后，设定换成用户的“其他时间”为不换乘用户的两倍来确定换乘用户的“其他时间”。

3.7.2 操作过程

根据上述原理，在无法确定换乘站点的情况下，需要对两条线路的站点进行遍历，设定该站点为换乘站。我们以 1 号线和 2 号线为例，计算这两条线路的换乘站点。方法如下所示：

Step1：在 Matlab 中，根据用户的进出站信息，可以确定该用户是否换乘。若同线路进出则没有换乘，进站和出站的线路不同则存在换乘。

Step2：在用户不换乘的情况下，根据进站点和出站点可以计算出经过的站点数量；同时，根据出站时间和进站时间可以计算出在地铁中的“其他时间”。将所有不换乘用户的“其他时间”作频数分布图，如下所示：

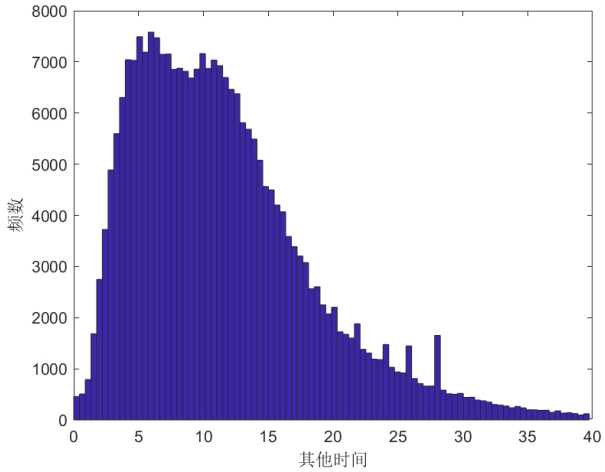


图 45：“其他时间”频数分布直方图

由图 45 可知，80%的“其他时间”集中于 5-15 分钟；其中，7 分钟的“其他时间”频数分布最多。所以，确定不换乘用户的“其他时间”为 7 分钟，而换成用户的“其他时间”设定为不换乘的两倍，为 14 分钟。

Step3：根据 Data_trips 中用户进出站信息，筛选出 1 号线进站、2 号线出站，或者 2 号线进站、1 号线出站的数据，对 1 号线和 2 号线的站点，按照站点顺序依次设定换乘站进行遍历。已知一号线有 20 个站点，2 号线有 23 个站点，所以需要遍历 20×23=460 次。

Step4：由于设定换乘站点，所以每次遍历都已知换乘站，由此可求出换乘用户乘坐的站点数。再根据进出站时间，计算出用户在地铁中的“其他时间”。将此“其他时间”与 Step2 中的“其他时间”，即 14 分钟进行比较，越接近 14 分钟说明该乘客越有可能符合换乘情况。

Step5：将 Step4 中求出的“其他时间”进行处理，取“其他时间”减去 14 分钟的差的绝对值的倒数为评价指标，指标越大，乘客换乘的可能性越大，该站点为换乘站的可能性也越大。同时，对该指标进行归一化处理。

Step6：结合换乘点客流量大的原理，根据前文的统计分析，将各站点的总客流量进行归一化处理，作为另一个评价指标。该指标越大，换乘站的可能性也越大。

Step7：建立评价模型，将上述两个指标的乘积作为评判换乘点的标准，成绩结果最大的作为我们确定的换乘站。结果如下图所示：

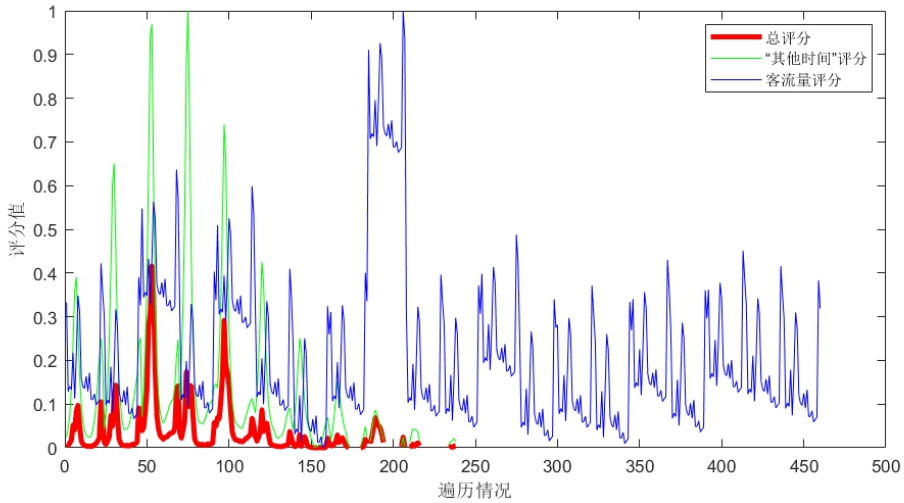


图 46：换乘站遍历情况评分图

其中，横坐标表示换乘点的第 x 种情况，20*23 共有 460 种情况；纵坐标为评分值。绿色曲线表示“其他时间”评分，蓝色曲线表示客流量评分。红色曲线表示两个指标的乘积，即总评分。

由上图所示，第 53 种遍历情况总评分最大。该情况为 1 号线的 **sta108** 和 2 号线的 **sta123** 为两条线的换乘站。

Step8：按照上述方法，对各条线路的换乘站进行计算，得到结果如下表所示：

表 2：各线路换乘站表

线路A	线路B	A中sta	B中sta
1	2	108	123
	3	1	40
	11	159	63
	12	89	136
2	3	47	131
	11	123	63
	12	127	148
3	4	142	62
	5	112	69
	10	142	160
	11	100	63
	12	126	95
4	5	84	96
	10	84	134
	11	38	162
	12	59	95
5	10	43	87
	11	37	20
	12	54	17
10	11	85	23
	12	134	95
11	12	146	31

由上表可知，有些线路之间不存在换乘点，例如线路 1 和 2 与线路 4、5 和 10 之间没有换乘点，即 1 和 2 号线都不能直接到达 4、5 或 10 号线。原因是 1 和 2 号线与 4、5 和 10 号线不在同一个地区内，但线路 4、5 和 10 之间可以互相换乘。

值得注意的是，数据中有 25000 条出行信息是按上述不能乘坐的方式乘坐地铁的，所以无法得知其具体换乘方式，则在断面分析时需要剔除这 25000 条出行数据。

3.7.3 线路断面流量分析

以一号线为例，对线路断面的客流量进行分析，结果如下图所示：

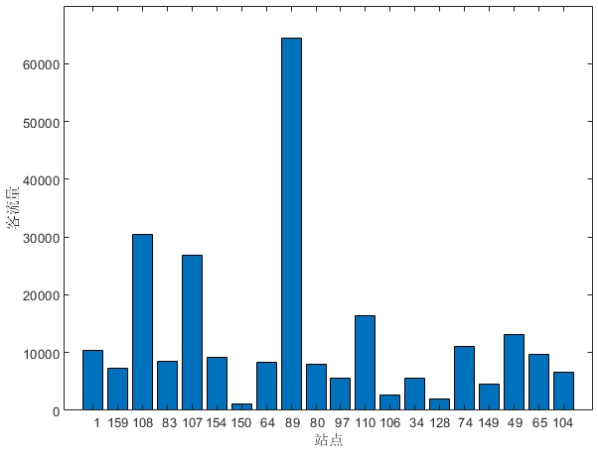


图 47：1 号线各站点进出站总客流量

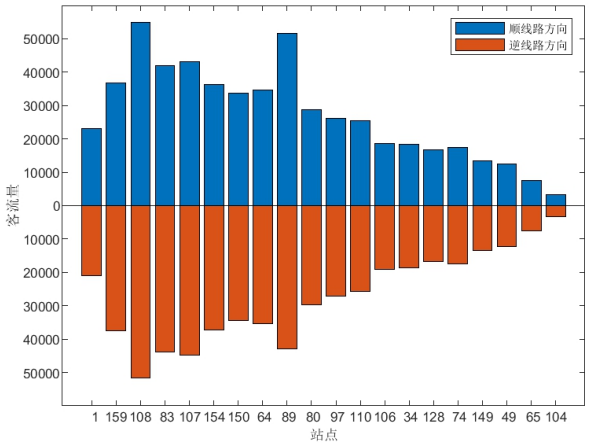


图 48：1 号线各站点断面客流量

由图 47 可知，1 号线中 sta108、107 和 89 进出站总客流量较多，其中 sta89 客流量最多，超过 6 万人次。但这只是统计了在站点中的出/入站人数，并没有统计换乘或不换乘经过站点的客流量。

所以，由上图 48，发现所有站点的断面客流量总数均增加。特别是换乘站 sta108，顺逆线路方向的客流量总数超过 10 万人次，大于 sta89。由于换乘站的存在，sta108 相

邻站点的断面客流量也受到较大影响，与出/入站总客流量相比人数大幅增加，如 sta159、83 等。因为，换乘后用户顺向或者逆向会经过这些站点带去客流统计数量。而与换乘站距离较远的站点，断面客流量无论顺逆方向均较少，且客流总和与出/入站总客流量相差不大，如 sta65、104 等。原因是这些站点处于 1 号线的末站位置，位于偏远的郊区，客流本身较少；而且，许多客流在前面的站点下车较多，不能带去较大的客流统计数量，所以断面客流量与进出站客流量大致相等。

3.8 用户乘坐地铁的频数分析

针对 Data_trips 中重复出现的用户 ID，我们团队考虑对所给的 206 天数据中用户乘坐地铁的次数进行统计分析。具体操作过程如下所示：

Step1：以 Data_trips 的 70 万条用户 ID 和 Data_users 的 12 万条用户 ID 为数据矩阵，利用 ismember 函数，按照 Data_users 中 12 万条用户 ID 的顺序，将 Data_trips 的 70 万条用户 ID 转变为顺序数字。相同的用户 ID 转变为相同的数字，即该用户在 Data_users 中的排列顺序。

Step2：统计相同数字的数量，则可计算出每个用户的地铁出行次数。

Step3：根据统计好的用户出行次数，利用 scatter 函数命令，画出散点图，图下图所示：

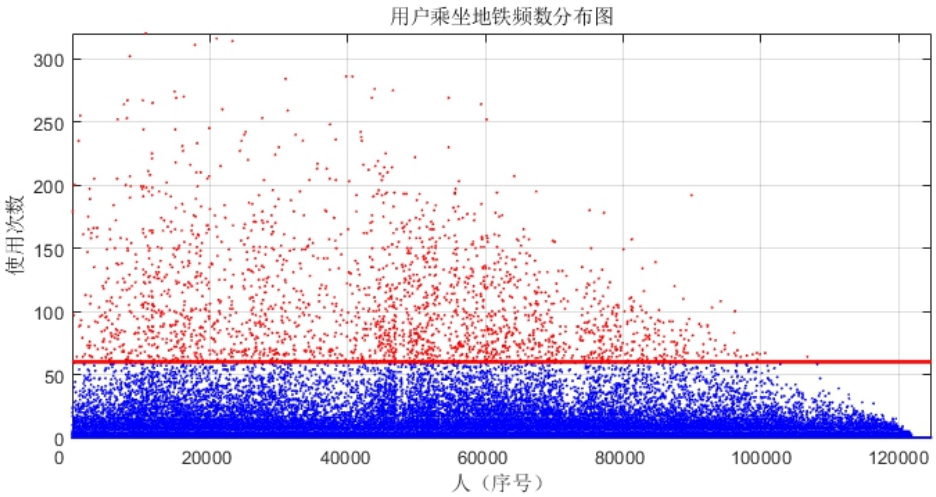


图 49：用户乘坐地铁频数分布图

其中，横坐标表示 Data_users 中 12 万条用户 ID 的序号，纵坐标表示用户出行乘坐地铁的次数，加粗的红色横线表示一周乘坐 2 次地铁的分界线，乘坐频数的均数线。分界线以上，即红色散点部分，表示这些用户乘坐地铁次数超过一周两次，属于高乘坐频率；分界线以下，即蓝色散点部分，表示这些用户乘坐地铁次数低于一周两次，乘坐评率较低。

数据显示，超过分界线的人数为 2241，仅占总人数的 1.796%；而低于分界线的人数为 122541，占比 98.204%，两者差异巨大。由此看出，虽然地铁客流量较大，但绝大多数用户乘坐地铁的次数并不多，集中于一周乘坐两次左右。

现实分析，绝大多数上班族会选择地铁出行。在一周上班 5 天，即一周乘坐五次地铁(包括来回)的情况下，散点应集中于频数 140-150 之间。现实实际与上图结果存在较大的差异，我们认为这是提供的数据存在一定偏差造成的。

S2B 4.客流量预测模型的构建及实现

构建客流量预测模型的整体思路如下所示：

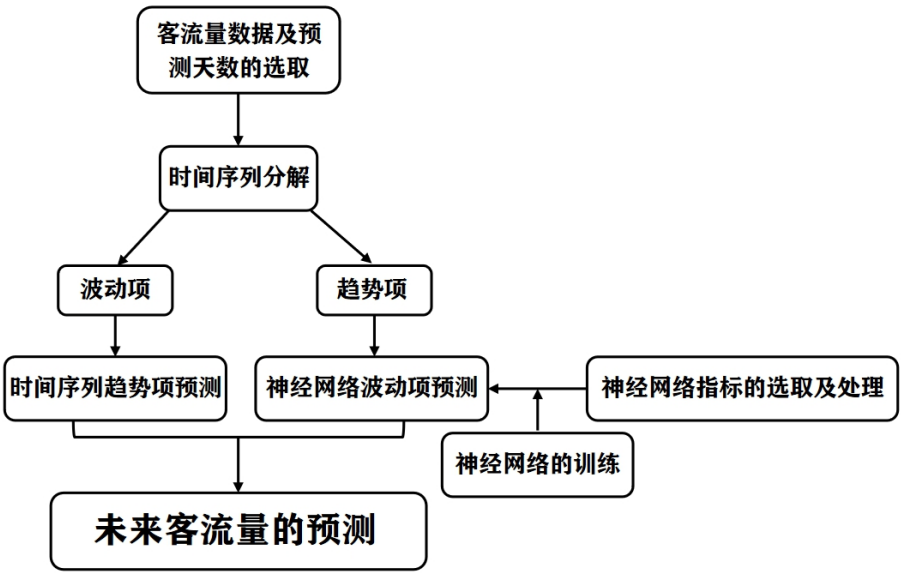


图 50：客流量预测模型构建思路图

4.1 客流量预测模型的构建

4.1.1 时间序列模型介绍

时间序列(Time Series)是按照一系列的时间顺序、时间间隔取得的一系列观测值是由数值和时间的二元对组合的有序集合。它客观地反映了某个事件随着时间变化的状态改变。这些观测数据通常以相同时间间隔来获取，以时间的顺序排列。时间序列数据本质上反映的是某个或者某些随机变量随时间不断变化的趋势。

时间序列分析(Time series analysis)是一种动态数据处理的统计方法。该方法基于随机过程理论和数理统计学方法，研究随机数据序列所遵从的统计规律，以用于解决实际问题。它包括一般统计分析(如自相关分析，谱分析等)，统计模型的建立与推断，以及关于时间序列的最优预测、控制与滤波等内容。经典的统计分析都假定数据序列具有独立性，而时间序列分析则侧重研究数据序列的互相依赖关系。后者实际上是对离散指标的随机过程的统计分析，所以又可看作是随机过程统计的一个组成部分。例如，记录了某地区第一个月、第二个月……第 N 个月的降雨量，利用时间序列分析方法，可以对未来各月的雨量进行预报。地铁客流量随时间变化所形成的序列，即为一个时间序列，在研究客流变化特征时，通过特征分解对时间序列进行降维分析，可以获取时间序列的重复性特征，从而利用时间序列中的规律性和异常性特征进行相关分析。

时间序列分析的应用有以下几个方面的作用：①系统描述。根据对系统进行观测得到的时间序列数据,用曲线拟合方法对系统进行客观的描述。②系统分析。当观测值取自两个以上变量时，可用一个时间序列中的变化去说明另一个时间序列中的变化，从而深入了解给定时间序列产生的机理。③预测未来。一般用 ARMA 模型拟合时间序列，预

测该时间序列未来值。④决策和控制。根据时间序列模型可调整输入变量使系统发展过程保持在目标值上,即预测到过程要偏离目标时便可进行必要的控制。

4.1.2 遗传神经网络模型介绍

神经网络作为是一种模仿生物神经网络的结构和功能的数学模型，能准确且快速地找出蕴含在数据内部的各种规律，已被广泛运用于模式识别、信号处理、判释决策、组合优化、知识工程等领域中，其中也包含了预测功能。神经网络中最基本的成分是神经元模型，在生物神经网络中，每个神经元与其他神经元相连，当它兴奋时，就会向相连的神经元发送化学物质，从而改变神经元内的电位，如果某种神经元的电位超过阈值，那么这一神经元就会被激活，即“兴奋”起来，向其他神经元发送化学物质。在数学模型中，神经元接收到来自 n 个其他神经元传递的输入信号，通过带有权重的连接进行传递，神经元接受到的总输入值将与神经元的阈值进行对比，最后通过激活函数处理以产生神经元的输出。

遗传算法是类比达尔文的生物进化论在计算机领域的扩展应用，以自然选择和生物的遗传为主要的核心思想，为我们提供了一种区域范围内的数据优化模型。该算法是以个体反应种群，通过种群的适用度来筛选适合的个体，最后让这些筛选下来的个体进行交叉和变异，产生下一代，循环迭代后生成最满足生存条件的个体。

本文选择**遗传神经网络**，通过遗传算法对神经网络隐层、输出层阈值以及输入层、隐层权重进行优化，从而寻得最优解。该方法结合了神经网络和遗传算法与进化策略，它通过模仿自然界“适者生存”的原则来赋予神经网络在代际循环中优化的力量，能有效克服传统神经网络在训练过程中的缺点。

使用遗传神经网络的基本算法流程图如下：

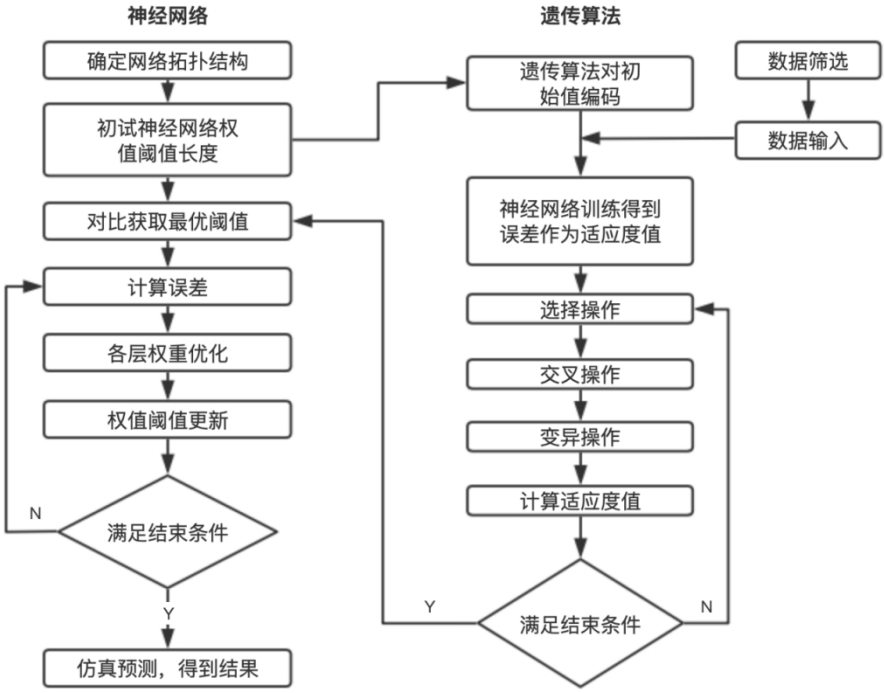


图 51：遗传神经网络的基本算法流程图

4.1.3 基于遗传神经网络的时间序列模型的构建

根据题意，要求我们通过现行数据对站点未来的客流情况进行分析，其中体现了时

间序列预测未来情况的思想。因此本次我们借用时间序列模型对于站点的客流情况进行分析。

由于数据波动会随数据本身的大小而改变，即数据越大，波动越大。因此我们采用乘法模型进行预测。其构成要素有：长期趋势（T）、季节变动（S）、循环变动（C）、不规则变动（I），由于所给的数据时间较短，在此将长期趋势与循环变动耦合，后文直接以 TC 表示，该典型模型表达为：

$$Y = T * C * S * I$$

其中，循环变动 C 包括了天气引起的波动项、节假日等指标产生的波动项。由于未来的节假日信息，与未来短期天气信息易得，所以相比于通过 arima 时间序列模型挖掘其中的波动信息，以神经网络的方式进行波动信息的挖掘更为准确合理。因此各项波动项可重新合并为以下模型：

$$Y = T * (C * S * I) = T * F$$

其中，Y 可以通过时间序列预测得到，F 为可由神经网络模型计算得到的其他波动项，两者的乘积即为预测得到的客流量。

4.2 未来客流量的预测

4.2.1 客流量数据及预测天数的选取

整体客流数据如下图所示：

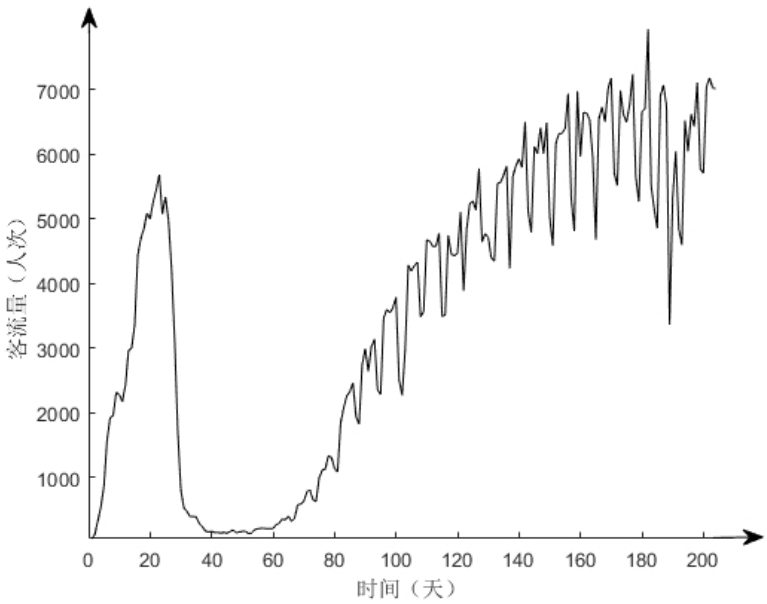


图 52：整体客流数据图

在上图中，第 37 天到第 65 天表示二月份的客流量变动情况，第 66 天到第 86 天表示三月份的客流量变动情况。由于二月受疫情影响过大，整体客流量极少；三月处于逐步回温阶段，客流量整体上升；而四到七月份的客流量数据，整体趋势相对平稳，后期预测效果将较为稳定，因此选择四到七月的出行数据作为用来预测未来客流量的基础数

据（以下简称“基础数据”）。具体为 2020.04.01～2020.07.16，共 107 天数据。

在预测天数的选取上，结合实际情况，我们选择对未来 7 天的客流量进行预测，这是因为在日常生活中，不存在对于未来的地铁出行需要提前预约的情况，因此预测更长一段时间的客流量在实际生活中的应用十分有限，此外，对于影响客流量的天气、温度等不可控指标，预测的时间跨度越短，则越能够预测得到更为精确的结果。

4.2.2 时间序列的分解

在 Matlab 中，利用 smoothdata 函数对基础数据进行平滑去噪处理，可得到去噪后的客流量数据，如下图所示：

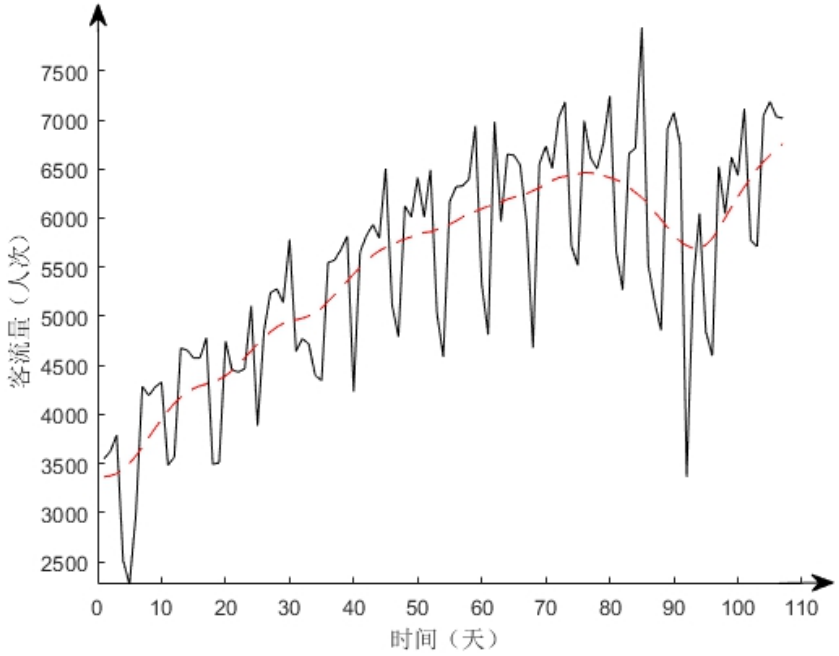


图 53：去噪后的客流量数据图

上图中红色虚线即表示去噪后的客流量数据，可以代表客流量的长期趋势，由于本题中我们采纳了时间序列的乘法模型，故时间序列的波动项可由客流量数据除以趋势项得到，即：

波动项(F) = 原数据(Y) /趋势项(T)

两者相除后得到的波动项，在 1 上下波动，如下图所示：

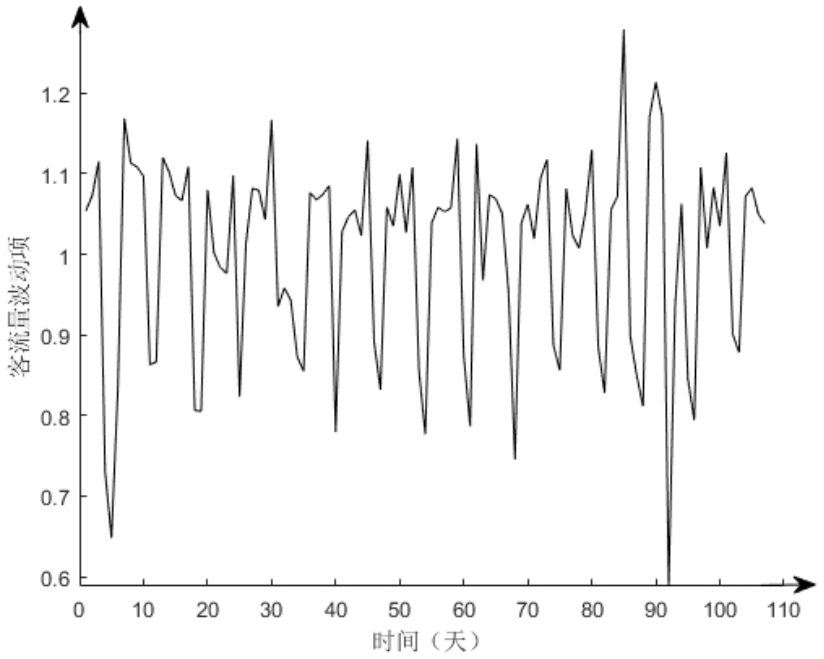


图 54：客流量波动项

4.2.3 时间序列趋势项的预测

由于平滑去噪后得到的时间序列趋势项已经是平稳的，因此我们可以采用自回归模型（AR 模型）来预测。通过网格搜索法，我们发现，当自回归模型的阶数 p=2 时，相比于 p=1，预测效果更加接近于实际值，当阶数 p>3 时，虽然预测值与实际值的误差进一步缩小，但变化十分有限，因此，我们选定 p=2 作为自回归模型的阶数。

在 Matlab 中使用 ar 函数对分离出的趋势项进行预测，预测长度为 7，得到的结果如下所示：

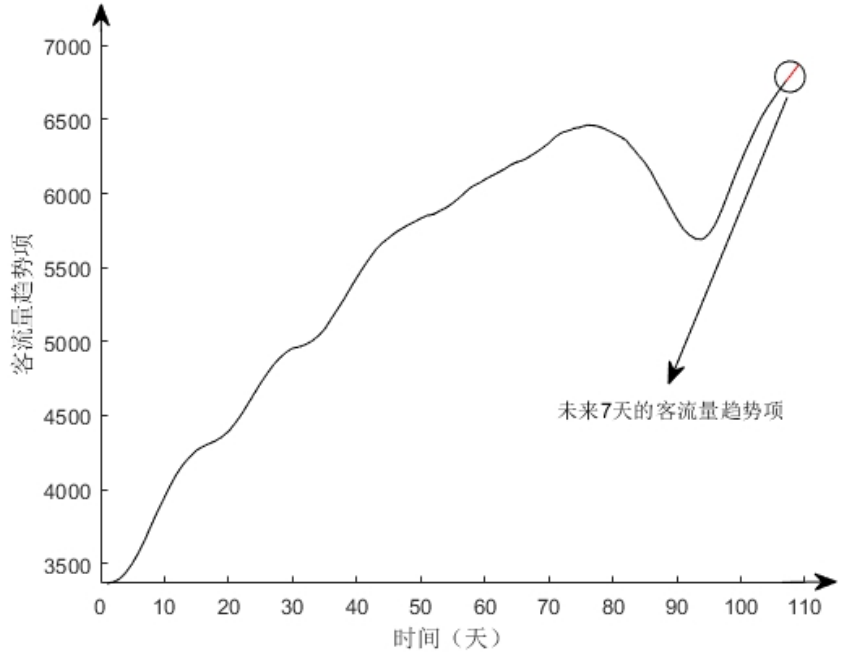


图 55：未来 7 天的客流量趋势项预测

未来 7 天客流量的趋势项的具体数据如下表所示：

表 3：未来 7 天客流量的趋势项预测结果表

天数	第一天	第二天	第三天	第四天	第五天	第六天	第七天
客流量趋势项	6813.3	6868	6921.5	6973.9	7025.1	7075.3	7124.4

4.2.4 神经网络指标的选取及处理

我们考虑了多种可能影响客流量波动项的指标，如天气、温度、节假日、票价总费用、男女性别比等，但我们发现，只有天气、温度、节假日这三种指标是能够获取未来的准确或较为准确的信息的，对于票价总费用、男女性别比等指标，无法准确获取未来的信息，此外，从因果关系上看，天气、温度、节假日这些指标是造成客流量波动的原因，而票价总费用、男女性别比则是受客流量波动的影响，随客流量变化而产生的附带结果，因此，最终考虑舍弃票价总费用、男女性别比等指标，留下天气、温度、节假日这三种指标并作进一步的挖掘与处理。

4.2.4.1 天气指标的选取及处理

从组委会提供的数据来看，天气一共分为 8 种：晴天、多云、阴、小雨、中雨、大雨、暴雨和雷阵雨。对于这 8 种天气类型，若全部归属于天气这一指标，则在赋值的过程中，不同的天气类型会出现数值上的大小差异，这必然会对神经网络的训练造成影响，事实上，这 8 种天气类型是互相独立的，即不同类型的天气之间不能排序，因此，考虑使用独热编码的方式来对诸如天气类型的定性指标进行数值化处理。通过独热编码，将 8 种天气类型分别作为独立的指标，每个指标进行 0-1 赋值处理，“1”表示该天气类型在当天的存在情况为“真”，“0”表示该天气类型在当天的存在情况为“假”。

例如，2020 年 4 月 13 日的天气为“小雨/多云”，构建天气指标集 [“晴天”，“多云”，“阴”，“小雨”，“中雨”，“大雨”，“暴雨”，“雷阵雨”]，则该日的天气经独热编码后的结果为：

$$[0,1,0,1,0,0,0,0]$$

下面对 8 种天气类型的指标进行筛选，利用 SPSS 软件，分别对独热编码后的 8 个指标进行独立样本 t 检验，从而比较存在该天气的情况下代表的样本和不存在的情况下代表的样本的平均数有无显著性差异。分组变量为独热编码后的指标，检验变量为客流量的波动项，得到各个指标独立样本 t 检验后的显著性结果如下：

表 4：8 种天气独立样本 t 检验结果表

天气	晴天	多云	阴	小雨	中雨	暴雨	雷阵雨
p值	0.275	0.233	0.393	0.59	0.846	0.01	0.003

从结果看，多数天气情况的 p 值都比较大，这说明这些天气的存在与否对于结果的影响并不是很大，若以 p=0.05 为判断两样本是否显著的标准，则满足条件的指标只剩下暴雨和雷阵雨两个，且在 107 个样本中，存在暴雨或雷阵雨天气的样本只有 4 个，因此考虑适当放宽标准，将 p=0.3 作为分界线，这样，阴、小雨、中雨三种天气指标可以剔除，晴天、多云、大雨、暴雨、雷阵雨 5 种指标被留下。

接下来考虑对天气指标进行进一步的挖掘，为了更好得显示出各个天气指标与客流量的变化关系，这里使用多元线性回归来分析得到各个天气指标与客流量的波动项的权重系数，如下表所示（标准化系数即为权重系数）：

表 5：八种天气的标准化系数及显著性

天气情况	晴天	多云	阴天	小雨	中雨	大雨	暴雨	雷阵雨
标准化系数	0.134	0.165	0.168	0.022	0.006	0.003	-0.281	0.086
显著性	0.263	0.199	0.167	0.867	0.962	0.967	0.007	0.396

在经过多元线性回归分析后，可以得到每种天气与客流量的波动项的权重系数，由于每一天都包含了两种天气，为了寻找出与客流量波动项相关性更明显的权重系数的数据，在这里进行三种尝试：①.把每一天的第一种天气所对应的权重系数归为第一列；②.将每一天的第二种天气所对应的权重系数做成第二列；③.将每一天的第一种天气所对应的权重系数和每一天的第二种天气所对应的权重系数之和归为第三列。虽然各天气的权重系数的值并不为整数，但每一个权重系数可以代表每一种天气的相对特性，即代表一种离散的等级，因此使用 Spearman 分析来对三列天气数据进行分析，得到结果如下：

表 6：客流量的波动项和三列天气数据的斯皮尔曼分析表

相关系数	客流量的波动项	第一列天气	第二列天气	两列之和
客流量的波动项	1.000	0.216	0.057	0.160
第一列天气	0.216	1.000	0.087	0.617
第二列天气	0.057	0.087	1.000	0.781
两列之和	0.160	0.617	0.781	1.000

当相关性系数的绝对值介于 0.1~0.3 之间时，一般认为变量间存在弱相关；当相关系数的绝对值介于 0.3~0.5 之间时，一般认为变量间存在中度相关；当相关性系数的绝对值大于 0.5 时，一般认为变量间存在强相关。从 Spearman 分析所得相关系数矩阵可以看出，第一列天气和两列天气之和与客流的波动项呈弱相关，第二列天气与客流量的波动项甚至没有相关性，因此最终考虑不将这三列数据纳入神经网络的输入指标层。

4.2.4.2 温度指标的筛选

同样地，为了得到更为明显的指标与结果之间的关系，我们将对温度数据作更多的变换，根据组委会提供的数据，温度数据包含了两列，分别是每天的最高温和最低温，在这里我们将两列数据取平均值，作为处理后的第三列数据。我们将以这三列数据为基础，进行数学上变换，通过相关系数分析寻找出相关系数较高的数据。由于温度数据为连续性而非等级性的数据，因此采用 Pearson 相关系数进行分析。

一、最高温及其变换的相关系数分析。将已经整理好的最高温一列的数据依次进行以下处理：取平方、开根号、取对数、倒数，这之后将各种运算所得到的结果连同最高温数据本身，与客流量波动项进行上述的 Pearson 相关系数的分析，可以得到结果如下：

表 7：客流量波动项与最高温处理数据的 Pearson 相关性分析表

皮尔逊相关性	原始数据	平方	开根号	取对数	倒数	客流量波动项
原始数据	1	0.993	0.998	0.992	-0.968	0.106
平方	0.993	1	0.984	0.970	-0.932	0.096
开根号	0.998	0.984	1	0.998	-0.981	0.111
取对数	0.992	0.970	0.998	1	-0.992	0.116
倒数	-0.968	-0.932	-0.981	-0.992	1	-0.126
客流量波动项	0.106	0.096	0.111	0.116	-0.126	1

从表中最后一行可以看出，将最高温数据取倒数后与客流量的波动项的相关程度最高，为-0.126，将最高温数据平方后与客流量的波动项的相关程度最低，为 0.096。总体而言，最高温数据及其变换与客流量的波动项的呈弱相关，因此考虑舍去这些指标。

二、最低温及其变换的相关系数分析。与上一步处理最高温的过程相似，将最低温的数据进行平方、开根号、取对数、取倒数处理后，连同最高温数据本身，与客流量波动项进行 Pearson 相关系数的分析，可以得到结果如下

表 8：客流量波动项与最低温处理数据的 Pearson 相关性分析表

皮尔逊相关性	原始数据	平方	开根号	取对数	倒数	客流量波动项
原始数据	1	0.995	0.999	0.994	-0.977	0.032
平方	0.995	1	0.988	0.978	-0.950	0.027
开根号	0.999	0.988	1	0.999	-0.987	0.035
取对数	0.994	0.978	0.999	1	-0.994	0.039
倒数	-0.977	-0.950	-0.987	-0.994	1	-0.048
客流量波动项	0.032	0.027	0.035	0.039	-0.048	1

从表中最后一行可以看出，将最低温数据取倒数后与客流量的波动项的相关程度最高，为-0.048，将最低温数据平方后与客流量的波动项的相关程度最低，为 0.027。总体而言，最高温数据及其变换与客流量的波动项无明显的相关性，因此也考虑舍去这些指标。

三、平均温及其变换的相关系数分析。将最平均温的数据进行平方、开根号、取对数、倒数处理后，连同平均温数据本身，与客流量波动项进行 Pearson 相关系数的分析，结果如下：

表 9：客流量波动项与平均温处理数据的 Pearson 相关性分析表

皮尔逊相关性	原始数据	平方	开根号	取对数	倒数	客流量波动项
原始数据	1	0.995	0.999	0.994	-0.976	0.076
平方	0.995	1	0.987	0.977	-0.948	0.067
开根号	0.999	0.987	1	0.998	-0.986	0.081
取对数	0.994	0.977	0.998	1	-0.994	0.087
倒数	-0.976	-0.948	-0.986	-0.994	1	-0.099
客流量波动项	0.076	0.067	0.081	0.087	-0.099	1

与上述两种数据相似，从结果来看，平均温与预测值的相关性处于最高温与最低温的折中，总体而言无明显的相关性，考虑舍去。

4.2.4.3 节假日指标的筛选

已知，日期属性分为工作日、周末和节假日三类。每个类别分别以 1、2 和 3 进行赋值。经过进一步思考，认为这三种日期属性无法充分地体现周一到周日具体的每一天的日期差别。所以，引入“周指标”，即周一到周日七天这七个指标。采取独热编码的

形式对“周指标”进行赋值处理。如果某一天可以对应到具体的星期几，则在此具体的星期几赋值为 1，而在其星期几赋值为 0。

赋值后，采用独立样本 t 检验的方式，检验一个星期内不同日子对客流量波动项的影响。得到结果如下表所示：

表 10：一周不同日子对客流量波动项影响的独立样本 t 检验结果表

星期几	周一	周二	周三	周四	周五	周六	周日
P值	0.032	0.029	0.145	0.022	0.001	0	0

将一周七天这 7 个指标对客流量波动项的影响进行独立样本 t 检验后，由显著性 P 值可以发现，其中绝大多数都存在显著的差异性(P<0.05)，特别是周六和周日这两个指标的 P 值为 0，周五非常趋近于 0，说明这三个指标与其他指标相比对客流量波动项的影响较大。只有周三的 P 值大于 0.05，与其他指标相比对客流量波动项的影响没有显著性差异。

由于节假日对客流量的波动项存在相当的影响，从而导致周一至周五（即一周中的 5 个工作日）、周六和周日两天（即一周中的双休日）的样本与客流量的波动项存在显著的差异。因此，考虑先剔除节假日，比较五个工作日指标对客流量波动项的影响。

表 11：工作日对客流量波动项影响的独立样本 t 检验结果表

星期几	周一	周二	周三	周四	周五
P值	0.454	0.331	0.68	0.064	0.556

由表 11 可知，工作日的 p 值较表 11 明显增加，且都大于 0.05，说明工作日对客流量波动项影响的差异性不明显。但是将五个工作日的 p 值进行比较之后发现，周四的 p 值仍旧处于最低位，仅为 0.064。因此，考虑增加“工作日是否为周四”这一个指标。

与上述讨论工作日的过程相同，对双休日这两个指标进行独立样本 t 的检验，发现 p 值仅为 0.649，未达到显著影响的标准，因此舍去。

受到节假日的影响，人的行为轨迹，行为趋势等等，在节假日前后都会产生总体范围内的显著变化。根据此推测，在基础数据 107 天中，提取出节假日前一天与后一天对应的客流量，将提取出的节假日前一天或后一天的客流量波动项和非这一天的客流量波动项进行比较。

经过提取，发现 107 天中一共有三天属于节假日，则有三个节假日前一天。比较这三个节假日前一天与非这三天的平均值，结果显示三个节假日前一天的客流量波动项的平均值为 1.1967；而非这几天的客流量波动项的平均值为 1.0651。通过独立样本 t 检验，发现二者对客流量波动项的 p 值为 0.04，具有显著性差异。

相同地，三个节假日后一天与非这三天的平均值，结果显示三个节假日后一天的客流量波动项的平均值为 1.1408；非这几天的客流量波动项的平均值为 1.1408。独立样本 t 检验，发现 p 值为 0.05，具有显著性差异。

综上所述，节假日前一天与后一天均对客流量波动项的影响显著，有必要引入这两个指标。因此，节假日指标筛选结果为：工作日、双休日、节假日、是否为工作日的周四以及是否为节假日的前一天。

经过天气、温度、节假日三方面的筛选，将作为神经网络输入层的指标有：

1、是否晴天

2、是否多云

3、是否大雨

4、是否暴雨

5、是否雷阵雨

6、是否为工作日

7、是否为双休日

8、是否为节假日

9、是否为工作日的周四

10、是否为节假日的前一天

11、是否为节假日的后一天

4.2.5 神经网络的训练及波动项的预测

在筛选完指标后，我们将指标数据进行搜集，就拥有了神经网络的输入层，为一个 11 行 107 列的矩阵，11 表示筛选出的 11 个指标，107 表示神经网络训练的样本数量，如下所示：

11x107 double

	1	2	3	4	5	6	7	8	9	10	11	12
1	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	1	1	1	0	1	1
3	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0
6	1	1	1	0	0	0	1	1	1	1	0	0
7	0	0	0	0	0	0	0	0	0	0	1	1
8	0	0	0	1	1	1	0	0	0	0	0	0
9	0	1	0	0	0	0	0	0	1	0	0	0
10	0	0	1	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	1	0	0	0	0	0

同时，我们的输出层，即客流量的波动项，也已计算得到：

1x107 double

	1	2	3	4	5	6	7	8	9	10	11	12
1	1.0540	1.0732	1.1156	0.7303	0.6482	0.8343	1.1687	1.1135	1.1083	1.0968	0.8630	0.8666

使用神经网络模型，对输入层、输出层进行训练，设置神经网络中间隐层神经元的个数为 5，训练集长度为 97，输出集长度为 10，将训练集与输出集划分好后。

我们的神经网络模型一共分为四部分，输入层，隐层神经元，输出层神经元，输出层，如下图所示：

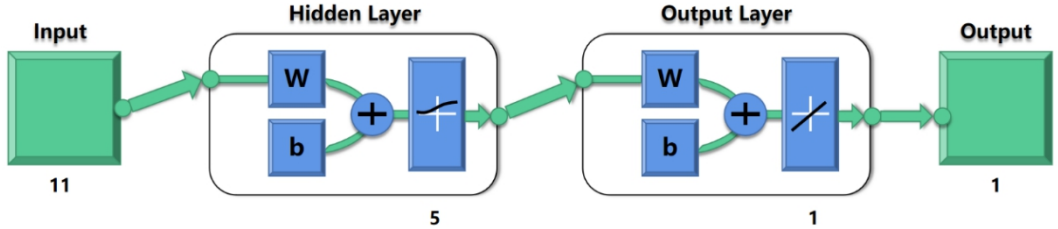


图 56：神经网络组成部分

然后，通过遗传算法优化求解神经网络模型中各个指标在不同隐层中的权重、隐层

神经元的阈值以及隐层神经元在输出层神经元的权重、输出层神经元的阈值。设置初始种群数量为 100，最大迭代次数为 10 次。最终，通过训练好的网络，对训练集和测试集的预测结果和实际值进行画图对比，如下图所示：

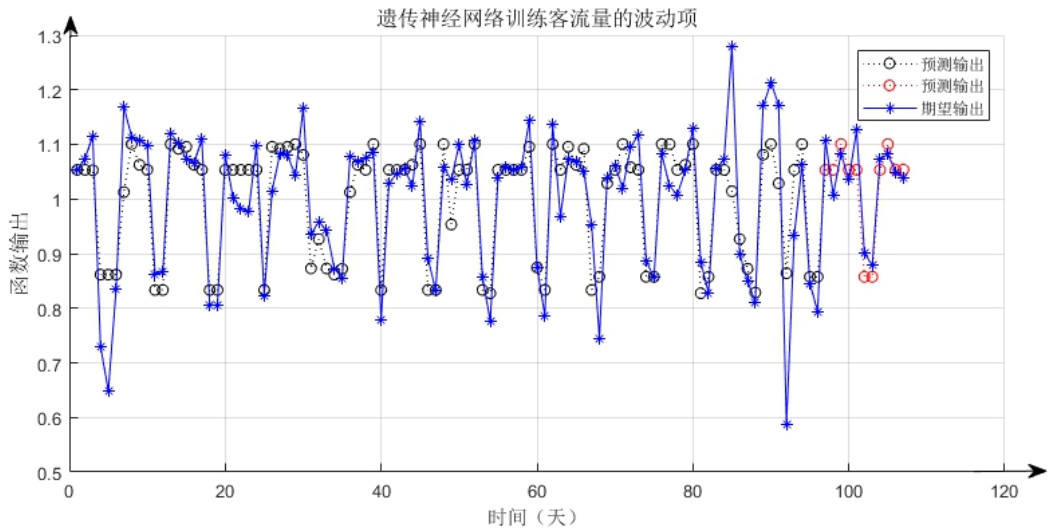


图 57：遗传神经网络训练客流量的波动项

图中，由*连接而成的蓝色线条表示实际的客流量波动项，由“o”连接而成的黑色和红色虚线分别表示 97 个训练集样本和 10 个测试集样本的预测结果，在该网络中，训练集的平均误差（即每个预测结果与实际值的平均偏差）为 4.831%，测试集的平均误差为 2.747%；训练集的 R^2 为 0.72485，测试集的 R^2 为 0.80217。可以看到，训练的网络有着比较好的拟合效果，能够得到较为准确的预测结果。

通过该网络，预测未来 7 天的客流量波动项的结果，如下所示：

表 12：未来七天客流量波动项预测结果表

天数	第一天	第二天	第三天	第四天	第五天	第六天	第七天
客流量波动项	1.0287	0.8575	0.8747	1.0534	1.0534	1.1005	1.0626

4.2.6 客流量的预测

将神经网络预测的客流量波动项和时间序列预测的客流量趋势项相乘，即可得到未来预测的客流量，如下图所示：

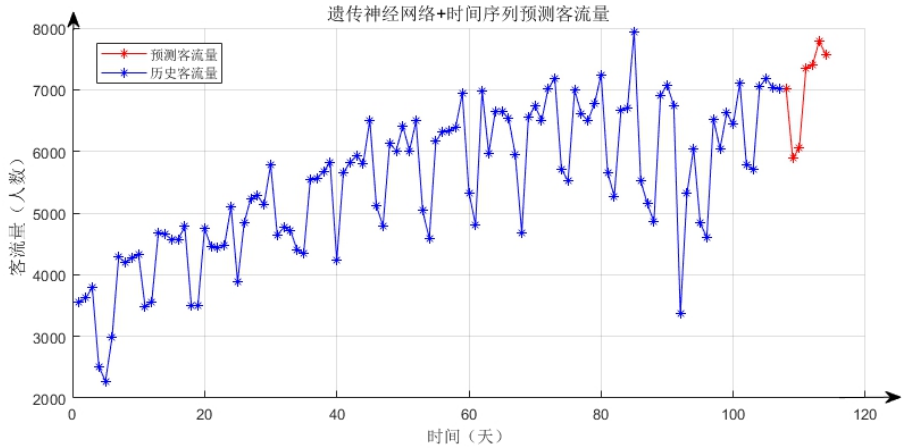


图 58：遗传神经网络+时间序列预测未来客流量

预测未来 7 天的客流量数据如下表所示：

表 13：未来七天客流量预测结果表

天数	第一天	第二天	第三天	第四天	第五天	第六天	第七天
客流量	7009.1	5889.4	6054.2	7346.1	7400.1	7786.5	7570.5

在上文的预测中，我们是以所有站点的总客流量为例，进行客流量的预测的，接下来，我们将从不同角度入手，根据上文所建立的模型，分别对未来单个站点的客流量、未来所有站点在一天中不同时段的客流量进行预测。

4.3 未来单个站点客流量的预测

我们提取出每个站点 107 天来的客流量数据，以每个站点的客流量数据作为基础数据，以同样的方法，对未来 7 天的客流量进行预测。

以站点 9 为例（其他站点的预测结果详见附录一），通过时间序列预测客流量趋势项的结果如下图所示：

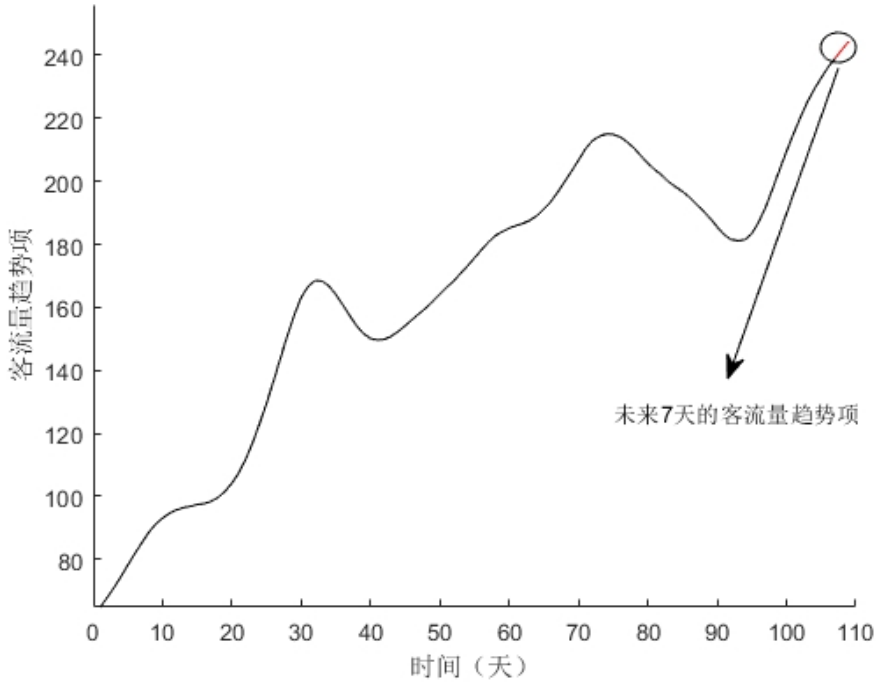


图 59：时间序列预测客流量趋势项

通过遗传神经网络，各项参数及输入层不变，输出层变为站点 9 客流量数据的波动项，得到训练好的网络，训练集和测试集的预测结果和实际值的对比如下：

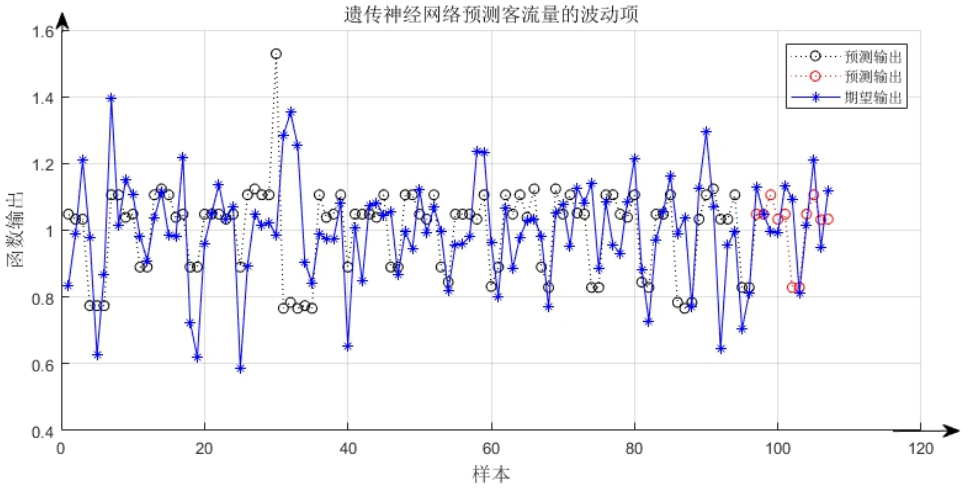


图 60：训练集和测试集的预测结果和实际值的对比

在该网络中，训练集的平均误差（即每个预测结果与实际值的平均偏差）为 10.6051%，测试集的平均误差为 8.8976%；训练集的 R2 为 0.48547，测试集的 R2 为 0.51269。可以发现，在第 30 天左右，训练集的预测值与实际值产生了较大的偏差，事实上，在对 168 个站点的网络进行训练的过程中，网络的预测误差均在 10%上下，相比于总客流量波动项的预测，准确度会有一定的下降。

未来七天的客流量如下图所示：

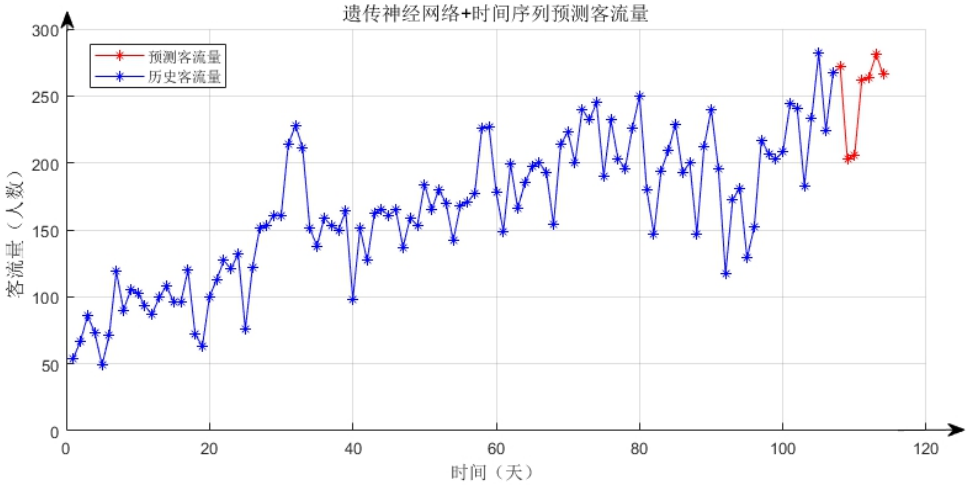


图 61：遗传神经网络+时间序列预测未来 7 天客流量

未来 7 天客流量的趋势项、波动项和客流量的预测值附表如下：

表 14：未来 7 天客流量的趋势项、波动项和客流量的预测值表

天数	第一天	第二天	第三天	第四天	第五天	第六天	第七天
客流量波动项	1.1239	0.8286	0.8315	1.0482	1.0482	1.1062	1.0387
客流量趋势项	241.818	244.463	247.001	249.435	251.768	254.001	256.136
客流量	271.790	202.563	205.392	261.461	263.906	280.976	266.054

4.4 未来所有站点在一天中不同时段客流量的预测

我们以 1 小时为间隔，统计出 6：00 到 24：00 每个时段的总客流量数据，对未来 7 天的客流量进行预测。具体操作步骤及图片与上文相似，在此不再赘述。使用神经网络预测的平均误差也在 10%左右。所有预测数据见**附录 1**。

此外，我们也考虑过对未来单个站点在一天中不同时段客流量的预测，但由于所给数据样本量太小，单个站点在一个小时内的客流量往往只有个位数，预测的偏差会比较大，因此最终没有对此进行预测。

4.5 客流预警系统的构建

对于预测出的未来的客流量，具有实际意义的就是**建立客流预警系统**，方便轨交业主对突发客流事件提前预知、提前部署应急方案；同时使未来的地铁拥堵情况具有可预见性，便于用户及时调整其出行方案。在这里，我们将对每个站点未来的客流量进行评估，并对可能存在的拥堵情况进行预警。

一般而言，一个站点的拥堵情况是站点实际客流与站点容量相比较而产生的。由于题目所给的信息不足以让我们知道各个站点的容量，因此我们选择行程数据范围内，站点单日客流量最大值作为我们的站点容量的代替值。

我们定义预警值如下：

预警值 = 站点实际客流 / 站点单日客流量最大值

将 4 月 1 日~7 月 16 日的各个站点的预警值计算出来，计算结果做频数分析，如下图所示：

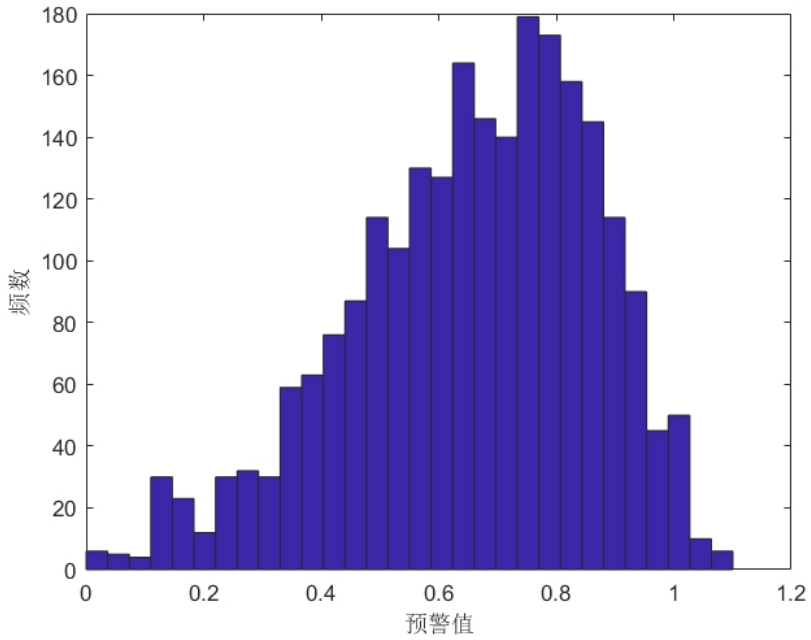


图 62：预警值频数分布图

可以看到，预警值在 0.8 左右的频数最多，超过 0.8，频数迅速下降，因此我们把 0.9 作为判断该站点当天是否拥堵的阈值。这样，我们就可以对未来的站点客流量作出预警。

我们对未来 7 天的各个站点进行客流预测分析，预测各站点每天的拥堵程度，并将结果中最拥堵的 10 种情况展示出来：

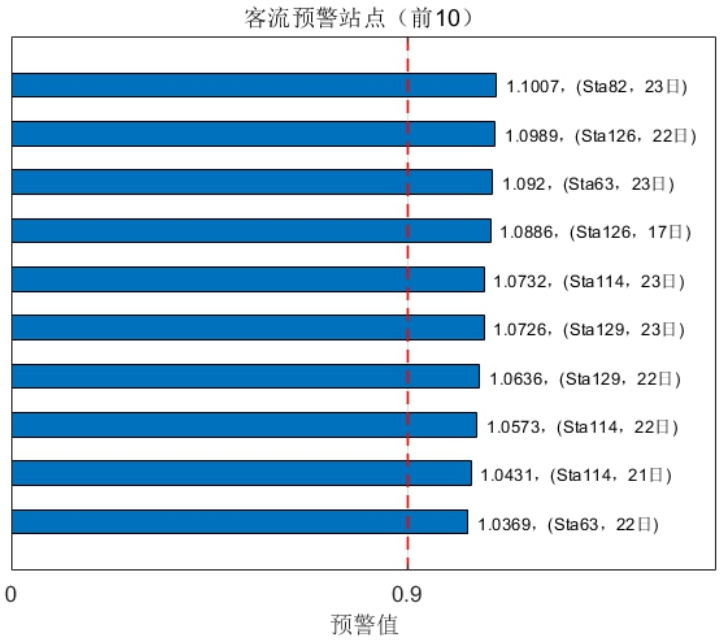


图 63：客流预警站点(前 10)图

对于各个站点未来 7 天的预警值的具体变化，可以参见**附录 2**。

实际上，对于现实中的情况，站点容量是已知的，在站点建造时就已经确定，因此如果用我们的方案解决现实问题，只需要将站点容量用实际值去代替，就能达到符合实际的效果。此外，也可以增加阈值的数量，将站点客流量情况划分为多个等级，从而更好地描述站点当天客流量的相对拥堵情况。

S2B 5.可交互界面的开发与设计

5.1 可交互界面介绍

为了使所有用户都能够对现有的客流、用户、站点数据进行数据分析与客流预测，我们团队利用 MATLAB 图形用户接口开发环境（GUIDE）建立了 GUI 对象，作为用户的交互界面，只需进行简单操作就能完成分析与预测，大大减少了用户的学习成本。

5.2 可交互界面开发与设计过程

该 GUI 对象开发过程主要有以下三个步骤：

Step1：制作图形界面

GUI 的图形界面由一系列的 GUI 对象构成，有各种 uicontrol 对象包括: Push Button、Slider、Radio Button、Check Box、Edit Text、Static Text、Pop-up Menu、Listbox 和 Toggle Button 对象，以及 Panel 和 Button Group 对象，还有用于数据可视化处理的 Table 和 axes 对象，和使 MATLAB 界面更美观的 ActiveX Control 对象。利用这些 GUI 对象，可以制作出满足功能需求的图形界面，为用户提供一个可以操作的界面外形。

Step2：编写回调函数

如果说图形界面是用户操作界面的外形，那么一系列的回调函数就是用户操作界面的灵魂。通过这些回调函数，就可以实现用户打开或关闭界面、点击 GUI 对象等操作发生时，图形界面能做出相应反馈，形成一种交互，使 GUI 对象能够真正实现一些我们需要的功能。回调函数的类型有 ButtonDownFcn、Callback、CellEditCallback、CellSelectionCallback、ClickedCallback、CloseRequestFcn、CreateFcn、DeleteFcn、KeyPressFcn、KeyReleaseFcn、OffCallback、OnCallback、ResizeFcn 等，常用的 CreateFcn 函数就可以实现图形界面打开时实现一些功能，而 Callback 函数能够实现用户点击一些 GUI 对象时，实现一些功能。

Step3：编译程序：

最后为了实现脱离安装 MATLAB 的使用环境，实现在其他用户的使用 Windows 操作系统的计算机中使用我们团队的可交互界面，我们将 GUI 图形界面与相应的 m 文件（其中包括各种回调函数）进行编译，通过 MATLAB 的 deploytool 工具箱打包，产生可以在 Windows 操作系统中运行的可执行文件（exe 文件）。

5.3 可交互界面使用流程

5.3.1 安装说明

Windows 操作系统用户可以通过以下方式运行我们项目的可交互界面。

（1）对于计算机中**已经安装**了 Matlab 2019a 的用户，由于已经包含了该可交互界面的运行所需环境，可以直接打开 for_redistribution_files_only 文件夹，双击其中的 Analysis_and_Forecast_of_Rail_Transit_Passenger_Flow.exe 可执行文件，就能够运行我们项目的可交互界面。

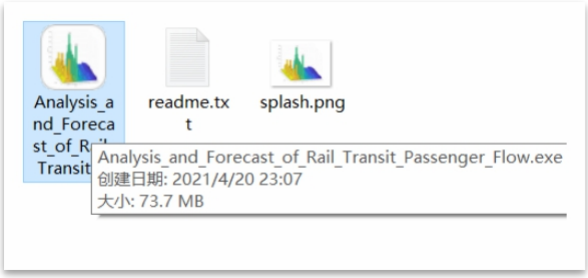


图 64：可执行文件

（2）对于**没有安装** Matlab 2019a 的用户，首先需要安装运行环境。打开 for_redistribution 文件夹，打开 MyAppInstaller_web.exe 可执行文件(打开后如下图所示)，按照页面中的操作步骤，下载安装程序所需要的运行环境 MATLAB Runtime R2019a。

图 65：打开可执行文件

值得注意的是，该过程需要联网进行，并且为了防止硬盘空间不足导致无法安装，用户的计算机必须有至少 1,604 MB 的剩余硬盘容量。安装完成后，即可按照（1）中的方法打开和运行软件。

5.3.2 操作说明

Step1：数据导入

打开软件后，首先会进入数据导入的页面。

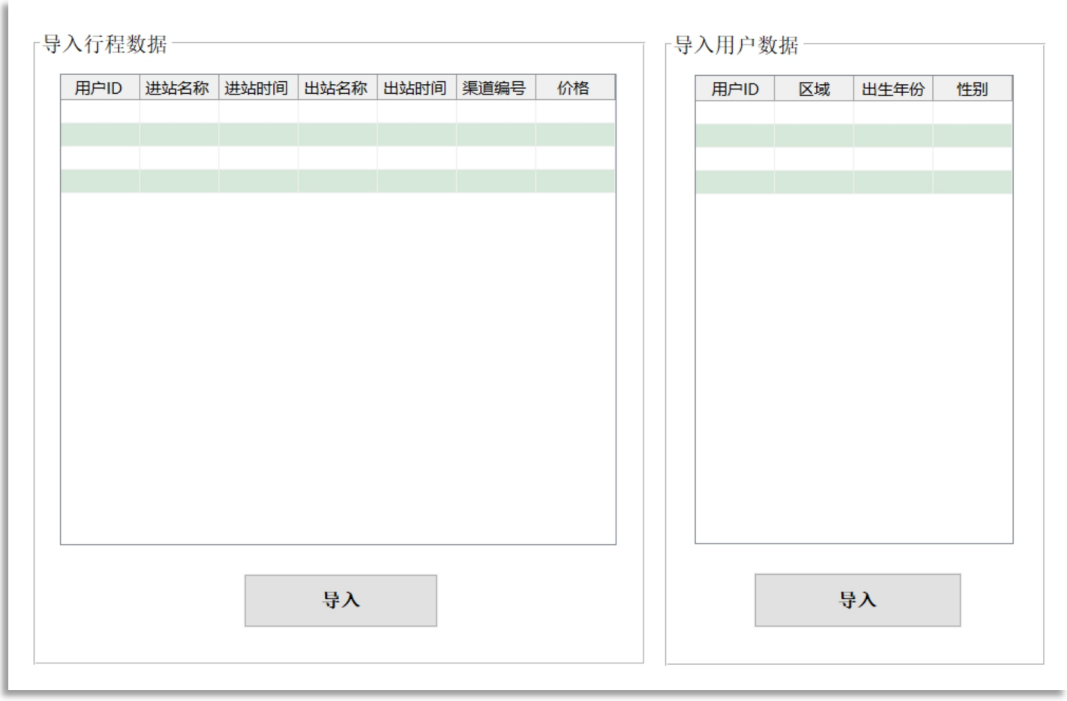


图 66：数据导入页面

点击“导入”按钮，即可自动查找到软件所在文件夹中名为“trips.csv”的文件，可选择并导入；亦可以自行更改路径，打开放在其他文件夹中的同名文件；若文件名不为“trips.csv”，可点击图中 A 处的菜单，选择“所有文件”模式，即可导入其他名称的数据文件（支持 xls、xlsx、csv 文件）。

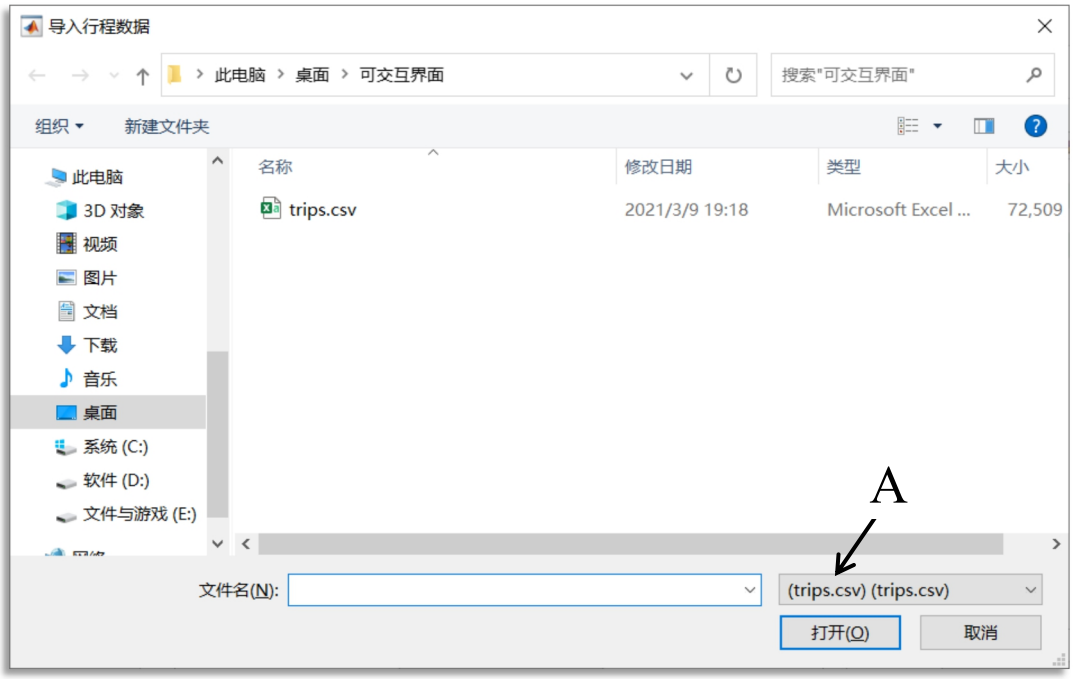


图 67：导入行程数据

完成“行程数据”与“用户数据”的导入后，等待 1 秒，可自动跳进入下一个数据导入页面，按照同样的方法，完成“工作日数据”、“站点数据”、“天气数据”的导入，再次等待 1 秒，自动跳转至“数据清洗”页面。

图 68：导入页面

Step2：数据清洗

“数据清洗”页面仅有一个按钮，点击后即可根据我们设定的数据清洗规则，对异常数据进行剔除。

图 69：数据清洗页面

完成剔除后，中间的文本框中会显示剔除情况，并将在 1 秒后自动跳转至“数据分析、预测”页面。

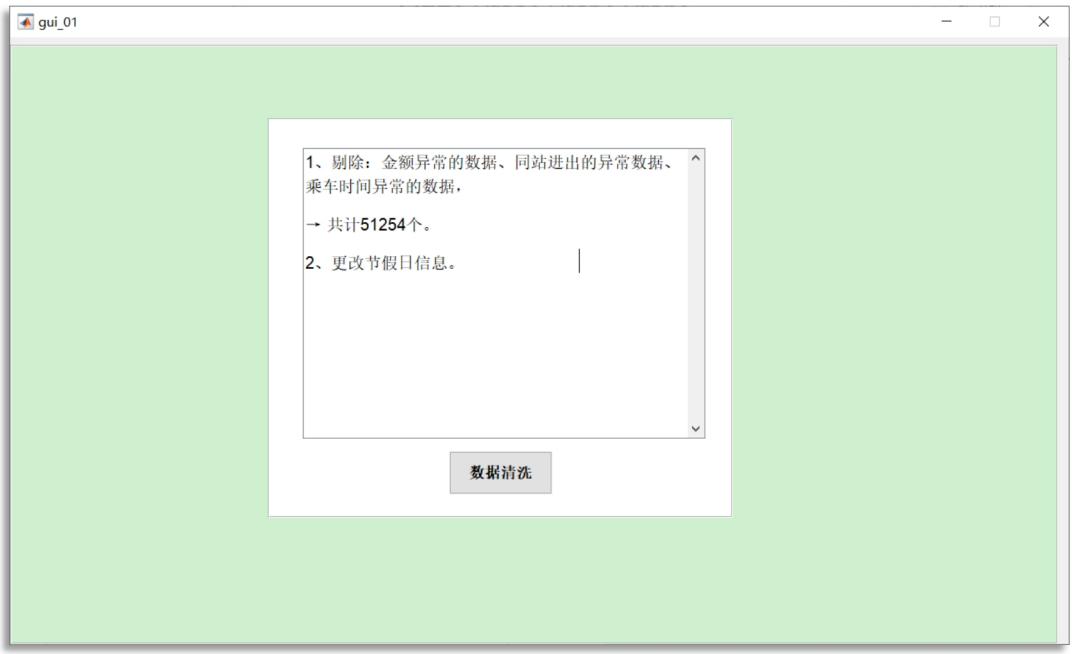


图 70：剔除情况

Step3：数据分析、预测

“数据分析、预测”页面由 4 个区域构成。

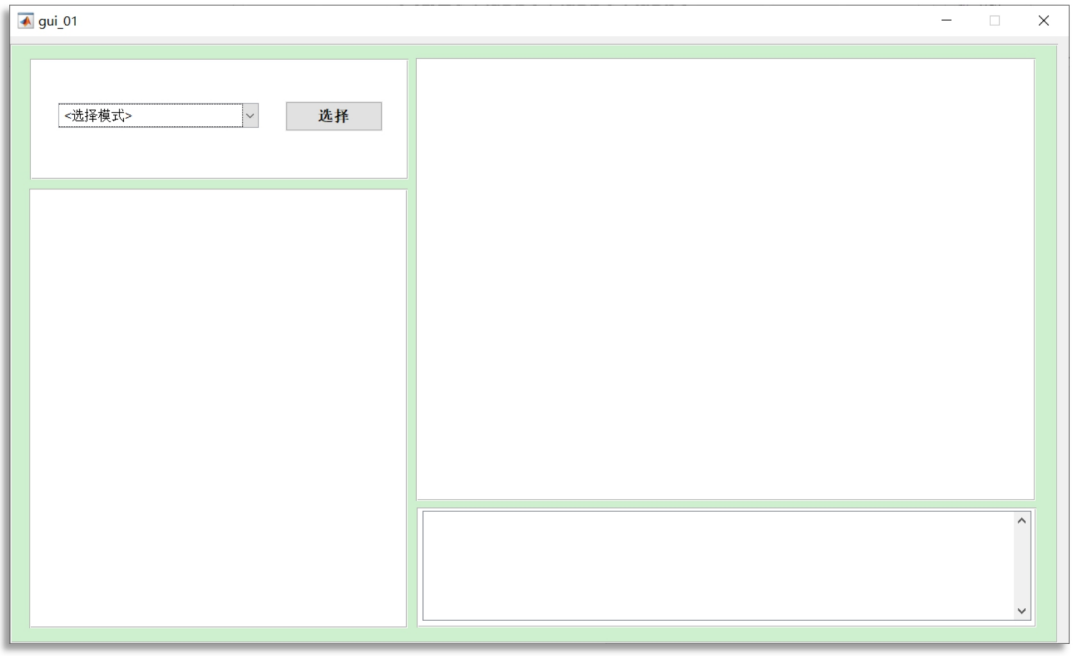


图 71：数据分析和预测的 4 个区块

位于左上方的是“模式选择”区域，可打开下拉式菜单，选择相应的分析模式，再点击“选择”按钮，即可进入相应的模式。

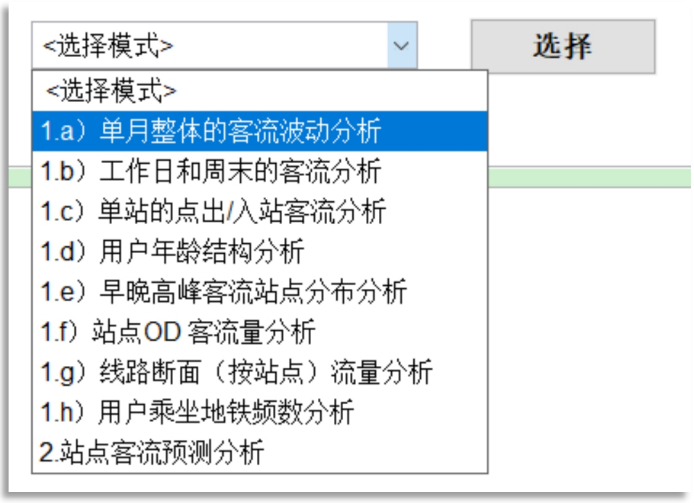


图 72：模式选择区域

1、单月整体的客流波动分析

点击“选择”按钮后，点击“计算”按钮，即可对行程数据进行分析，剔除不完整的月份，并给出可选择的月份，点击“绘图”按钮，给出相应结果。

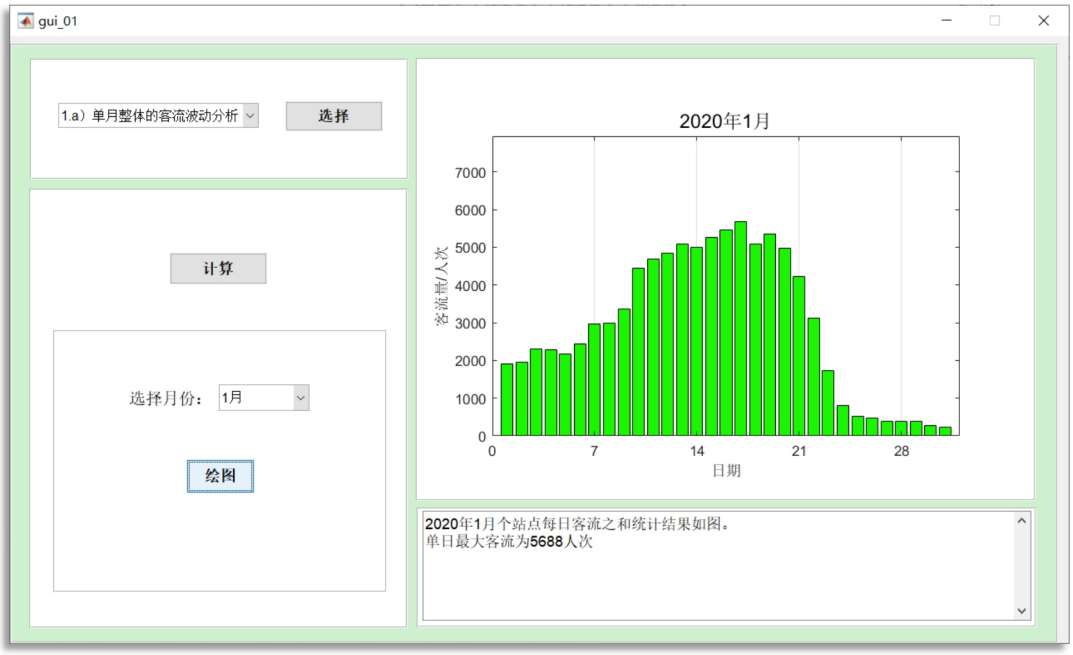


图 73：单月整体的客流波动分析

2、工作日和周末的客流分析

点击“选择”按钮后，点击“计算并绘图”按钮，给出相应结果。

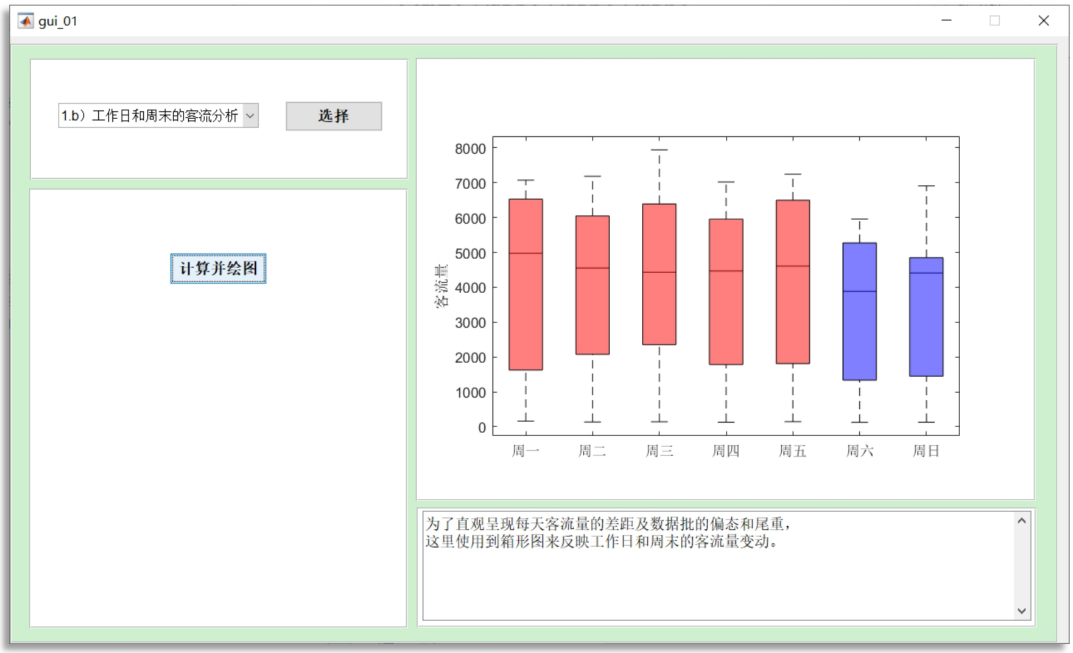


图 74：工作日和周末的客流分析

3、单站点的出/入客流分析

点击“选择”按钮后，点击“计算”按钮，即可对行程数据进行分析，输入相应站点后，点击“绘图”按钮，给出相应结果。

图 75：单站点的出/入站分析

4、用户年龄结构分析

点击“选择”按钮后，点击“计算并绘图”按钮，给出相应结果。

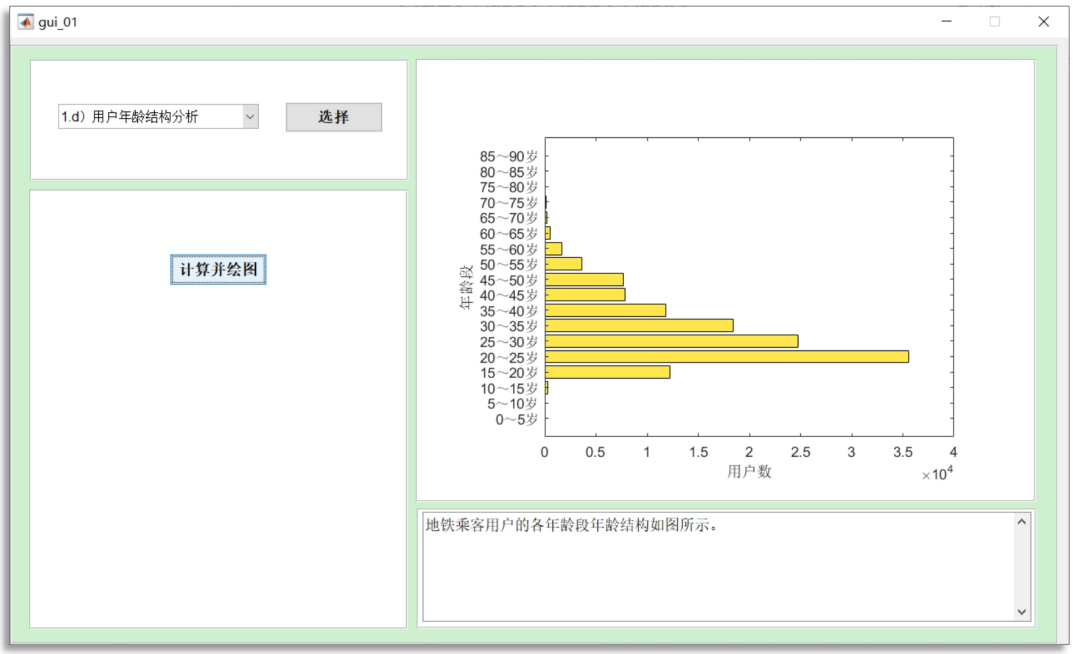


图 76：用户年龄结构分析

5、早晚高峰客流站点分布分析

点击“选择”按钮后，点击“计算”按钮，即可对行程数据进行分析，并给出可选择的线路，点击“绘图”按钮，给出所选线路上各站点相应结果。

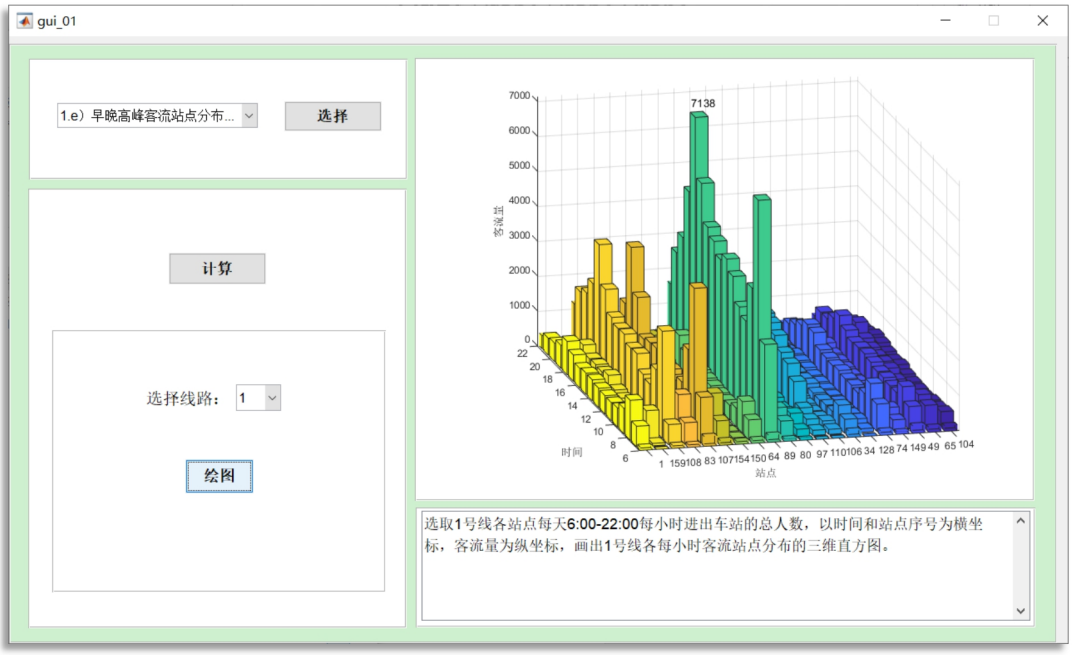


图 77：早晚高峰客流站点分布分析

6、站点 OD 客流量分析

点击“选择”按钮后，点击“计算”按钮，即可对行程数据进行分析，比较站点间、线路间的客流，并给出可选择的线路以及“线路间”的选项，点击“绘图”按钮，给出相应结果。

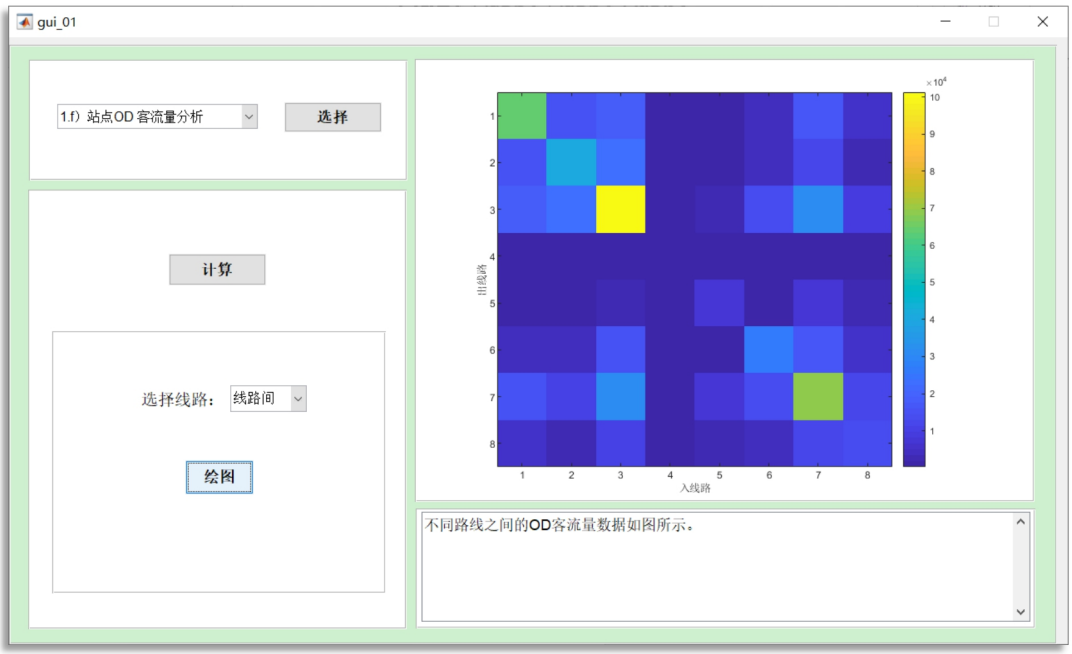


图 78：站点 OD 客流量分析

7、线路断面（按站点）流量分析

点击“选择”按钮后，点击“计算”按钮，即可对行程数据进行分析，剔除不完整的月份，并给出可选择的月份，点击“绘图”按钮，给出相应结果。

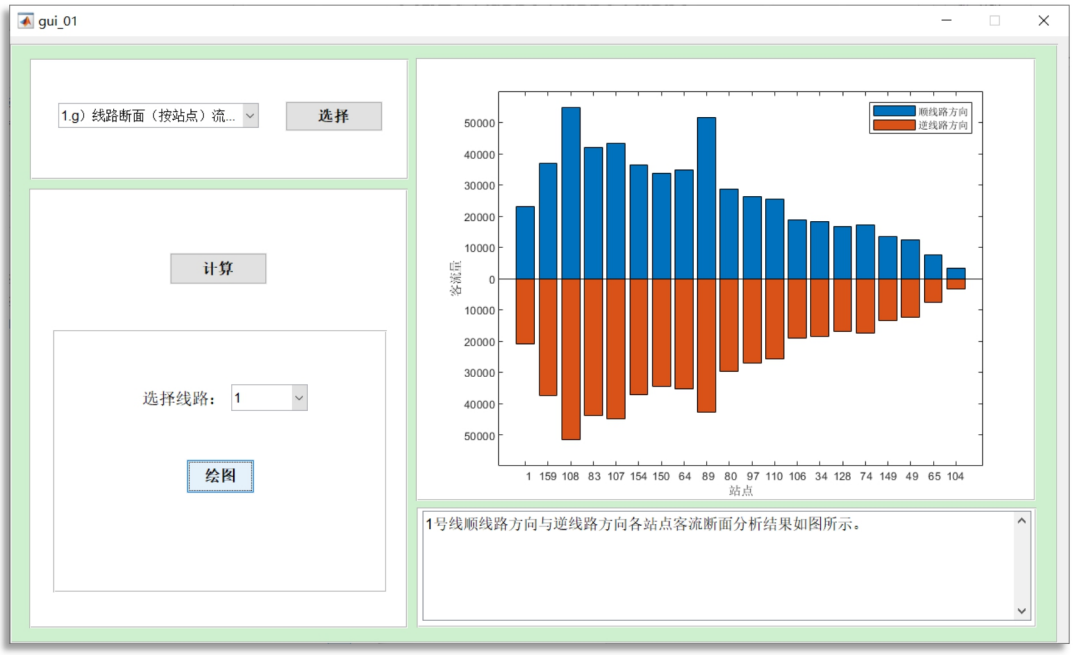


图 79：线路断面（按站点）流量分析

8、用户乘坐地铁频数分析

该分析模式为团队自愿补充的，点击“选择”按钮后，点击“计算并绘图”按钮，给出相应结果。

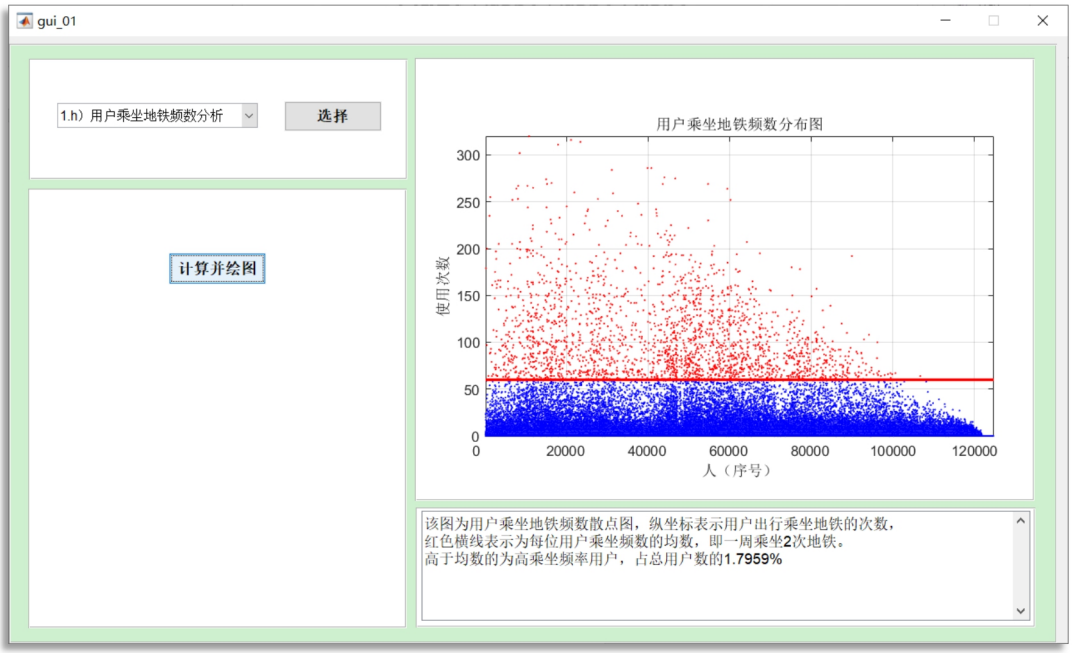


图 80：用户乘坐地铁频数分析

9、站点客流预测

点击“选择”按钮后，需要先预测模式的选择，共有 2 种预测模式可供选择。

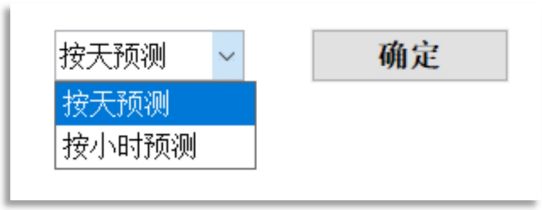


图 81：预测模式选择

(1) 按天预测

点击“确定”按钮后，出现 3 种预测的对象。依次为“站点客流预测”“总客流预测”以及“拥堵预警”。

(a) 站点客流预测

点击后，即可出现对所输入站点未来 7 天客流的预测，并以今天（数据中的最后一天）作为中心，以折线图的形式展示前 7 天的客流量与预测的未来 7 天的客流量，并给出拥挤的预警值，以红色水平虚线表示，客流量超过该水平虚线，该客流量在图像上的点将变为红色，预示当天该站点客流量较大，可能发生拥堵。

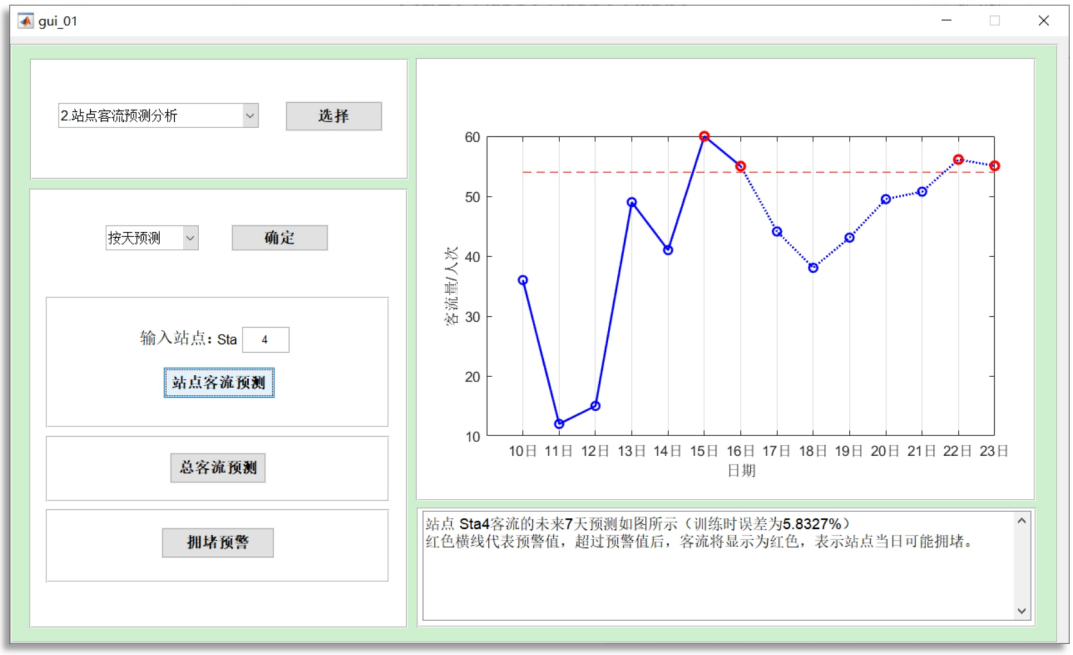


图 82：站点客流预测分析

(b) 总客流预测

点击后，即可出现对所未来 7 天总客流的预测结果，并以今天（数据中的最后一天）作为中心，以折线图的形式展示前 7 天的客流量与预测的未来 7 天的客流量。

图 83：总客流量预测分析

(c) 拥堵预警

点击后，即可对 7 天未来的各个站点进行客流预测分析，预测各站点每天的拥堵程度，并将结果中最拥堵的 10 种情况展示出来。

图 84：拥堵预警

(2) 按小时预测

点击“确定”按钮后，出现“总客流预测”按钮。点击后，即可预测未来 7 天客流逐小时变化情况。

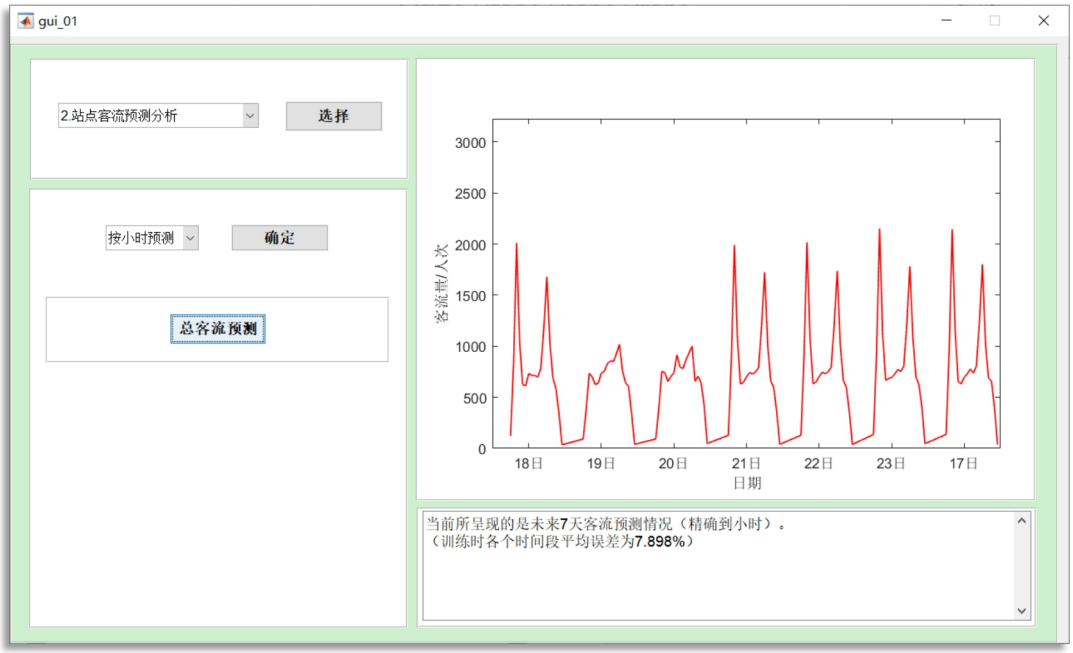


图 85：按小时预测

S2C 可行性分析

S2C 1.成本可行性

该项目主要成本为学习研究成本，本团队利用自学的方式掌握了大数据处理的方法，以及可交互界面开发的技术，以低成本高质量的方式完成了赛题的项目要求。

S2C 2.技术可行性

本团队为实现地铁客流量的精确统计分析以及未来客流的预测和预警，掌握了数据分析挖掘方法，统计学理论以及机器学习算法。同时，项目组成员能够熟练运用 Matlab、Python、SPSS 等软件，对项目的实现提供了技术支持。我们团队的技术储备已经足够完成这个项目了。

S2C 3.产品可行性

我们团队在客流分析预警系统全程程序的编写过程中，都紧紧围绕 Excel 所给数据展开，当 Excel 中的数据被更换时，亦能够对被更换的数据进行分析和预测，因此该程序的适用性很强。同时，我们开发的可交互界面虽由 Matlab 编写，但也基本提供了较为友好的与用户交互的方式。

S2D 项目管理与角色分配

S2D 1.角色分配及职责

我们团队共有 5 人，每个人都有一个角色，包括项目经理、技术总监、产品经理以及测试人员，每个角色的具体分工与职责如图 85 所示：

图 86：团队角色分工与职责图

S2D 2.成员介绍

1.指导老师：负责指导整个团队，及时发现并提出项目解决过程中的问题。

副教授，拥有资深的实力、丰富的教学经历，对数学、信息技术领域均有一定研究，指导过多次大数据的项目，指导学生多次获得数学建模、服务外包等国家级奖项，指导学生发表 SCI 一区 TOP 期刊多篇。

2.项目负责人 老陈： 项目主要负责人，对团队成员合理分工，规划项目整体宏观架构。

（1）荣获全国大学生数学建模竞赛国家二等奖，美国大学生数学建模竞赛二、三等奖。

（2）具有良好的项目规划能力，主持科研立项一项、挑战杯省一等奖。

（3）科研能力强，于《ACS Sensors》杂志(中科院一区 TOP，IF=7.33 分)发表论文《Effective and Robust Parameter Identification Procedure of a Two-site Langmuir Kinetics Model for a Gas Sensor Process》。

（4）具有较强的研发能力，以第一开发人员申请国家发明专利《一种量子点免疫荧光曲线的特征识别方法》，申请号: 202010795403.5。

以第一开发人员，获得生物免疫荧光试剂曲线特征识别软件著作专利 4 项。登记号：2020SR0318181、2020SR0318336、2020SR0515989、2020SR0991577。

3.技术总监 阿毛：负责可视交互界面开发、技术路线规划及视频的设计与制作，掌握本团队的主要功能应用。

（1）荣获全国大学生数学建模竞赛省级一等奖、三等奖，美国大学生数学建模竞赛二、三等奖。

（2）拥有丰富的实践应用经历，具有创新精神，熟练掌握本团队的应用。

4.产品经理 阿渣：负责模型的建立和技术路线的实施与完善。

荣获全国大学生数学建模竞赛省级二等奖，美国大学生数学建模竞赛三等奖。

5.客户经理 之渔：负责方案设计优化，技术路线的整理和文本的撰写。

（1）荣获美国大学生数学建模竞赛三等奖，校数学建模竞赛一等奖。
（2）有良好的文案撰写、排版能力。获得统计调查大赛省级三等奖等荣誉。
（3）科研能力较好，返修(minor revise)在 SCI 收录期刊论文 1 篇。

6.艺术总监 丁丁：负责 PPT 的设计与制作、团队图标 logo 和文本美化等。

（1）荣获美国大学生数学建模竞赛二等奖。
（2）熟练掌握 PR 和 vlog 小视频技术，多次参与学校多媒体竞赛、微信短视频竞赛等。
（3）有丰富的绘图、作图能力，熟练掌握 PS 技术。

S2D 3.团队管理

团队制定明确的开发管理制度，寒假期间线上讨论，每日一次任务派送，每周一次团队汇报。回校后聚在一起做题，探讨具体做题方法。一次大型会议，沟通时长为一个小时，小型会议的沟通时长为半小时。每日开发结束后，团队进行总结，进一步明确自己的任务和目标。每周汇报进展，确保大方向正确。

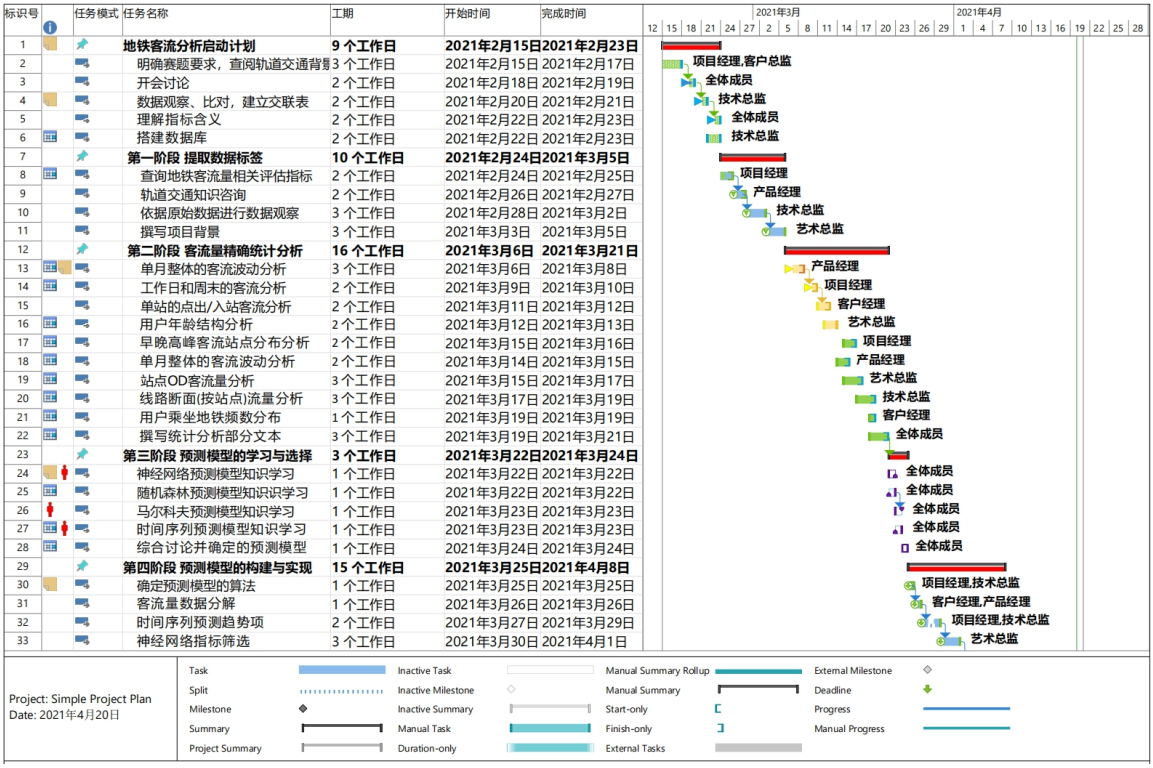
图87：开会做题

项目利用专业的研发环境，实施独立管理的项目团队，所有的任务都围绕一个共同的目标展开，团队成员相互独立，实行任务分解策略，每个成员有自己独自的任务，成员需要将任务精益求精；深入分析项目需求、理解项目目标、评估解决方案，包括技术平台、语言、交付进度等，保证项目的顺利执行。

S2D 4.项目监控

在项目研发过程中为了保证研发进度及研发质量，我们团队依据项目开发计划安排每位成员的任务，并对具体完成情况作出准确记录，最终形成项目主体开发过程实时进度表。

4.1 项目主体开发过程



标识号	任务模式	任务名称	工期	开始时间	完成时间
1		**地铁客流分析启动计划**	**9 个工作日**	**2021年2月15日**	**2021年2月23日**
2		明确赛题要求，查阅轨道交通背景	3 个工作日	2021年2月15日	2021年2月17日
3		开会讨论	2 个工作日	2021年2月18日	2021年2月19日
4		数据观察、比对，建立交联表	2 个工作日	2021年2月20日	2021年2月21日
5		理解指标含义	2 个工作日	2021年2月22日	2021年2月23日
6		搭建数据库	2 个工作日	2021年2月22日	2021年2月23日
7		**第一阶段 提取数据标签**	**10 个工作日**	**2021年2月24日**	**2021年3月5日**
8		查询地铁客流量相关评估指标	2 个工作日	2021年2月24日	2021年2月25日
9		轨道交通知识咨询	2 个工作日	2021年2月26日	2021年2月27日
10		依据原始数据进行数据观察	3 个工作日	2021年2月28日	2021年3月2日
11		撰写项目背景	3 个工作日	2021年3月3日	2021年3月5日
12		**第二阶段 客流量精确统计分析**	**16 个工作日**	**2021年3月6日**	**2021年3月21日**
13		单月整体的客流波动分析	3 个工作日	2021年3月6日	2021年3月8日
14		工作日和周末的客流分析	2 个工作日	2021年3月9日	2021年3月10日
15		单站的点出/入站客流分析	2 个工作日	2021年3月11日	2021年3月12日
16		用户年龄结构分析	2 个工作日	2021年3月12日	2021年3月13日
17		早晚高峰客流站点分布分析	2 个工作日	2021年3月15日	2021年3月16日
18		单月整体的客流波动分析	2 个工作日	2021年3月14日	2021年3月15日
19		站点OD客流量分析	3 个工作日	2021年3月15日	2021年3月17日
20		线路断面(按站点)流量分析	3 个工作日	2021年3月17日	2021年3月19日
21		用户乘坐地铁频数分布	1 个工作日	2021年3月19日	2021年3月19日
22		撰写统计分析部分文本	3 个工作日	2021年3月19日	2021年3月21日
23		**第三阶段 预测模型的学习与选择**	**3 个工作日**	**2021年3月22日**	**2021年3月24日**
24		神经网络预测模型知识学习	1 个工作日	2021年3月22日	2021年3月22日
25		随机森林预测模型知识学习	1 个工作日	2021年3月22日	2021年3月22日
26		马尔科夫预测模型知识学习	1 个工作日	2021年3月23日	2021年3月23日
27		时间序列预测模型知识学习	1 个工作日	2021年3月23日	2021年3月23日
28		综合讨论并确定的预测模型	1 个工作日	2021年3月24日	2021年3月24日
29		**第四阶段 预测模型的构建与实现**	**15 个工作日**	**2021年3月25日**	**2021年4月8日**
30		确定预测模型的算法	1 个工作日	2021年3月25日	2021年3月25日
31		客流量数据分解	1 个工作日	2021年3月26日	2021年3月26日
32		时间序列预测趋势项	2 个工作日	2021年3月27日	2021年3月29日
33		神经网络指标筛选	3 个工作日	2021年3月30日	2021年4月1日

图 88：进度安排甘特图-1

标识号	任务模式	任务名称	工期	开始时间	完成时间
34		神经网络算法编写与实践应用	5 个工作日	2021年4月2日	2021年4月6日
35		神经网络预测波动项	1 个工作日	2021年4月7日	2021年4月7日
36		撰写预测部分文本	2 个工作日	2021年4月7日	2021年4月8日
37		**第五阶段 整合调试**	**12 个工作日**	**2021年4月9日**	**2021年4月20日**
38		图形用户界面制作	7 个工作日	2021年4月9日	2021年4月15日
39		依据最佳方案，进行整合调试	1 个工作日	2021年4月16日	2021年4月16日
40		视频拍摄和剪辑	1 个工作日	2021年4月17日	2021年4月17日
41		PPT制作	1 个工作日	2021年4月18日	2021年4月18日
42		文本等所有材料整理	2 个工作日	2021年4月19日	2021年4月20日

图 89：进度安排甘特图-2

4.2 开发过程展示

本项目团队成员已多次组队参加各种比赛，相互之间较为熟悉，了解个体所擅长领域后，进行角色定位，明确各自目标、方向和职责，为团队协作打下基础。由于寒假期间团队不能面对面讨论，于是大家开展了线上语言沟通，一周一次大会，每天一次小会，每个参与发言，将自己的疑问与建议记录下来，并且进行总结，进一步明确自己的任务和目标。此外，团队在回校后积极在教室开展项目整体思路以及细节的探讨，在黑板上酣畅淋漓地互相交流想法。

S2D 5.项目总结

随着城市轨道交通网络化建设和运营的发展，乘客作为流动介质，数量在不断增长，而在一定时期内，城市轨道交通设施的内容量是有限的，势必会出现局部空间客流短时

间大量聚集的情况，从而形成大客流拥挤的问题。因此，在各时间段内对客流量的预测以及高峰时期对客流拥挤的预警显得非常有意义。

在互联网逐渐步入大数据时代后，用户轨道交通出行的站点信息、出行时间、乘坐票价、出行数量等多维度数据可以快速便捷地获得。从而对轨道交通的客流量进行多角度的统计和分析。本项目组基于 80 万条行程信息和 12 万条用户信息，在数据处理后进行了单月整体的客流波动、工作日和周末、单站的点出/入、用户年龄结构、早晚高峰客流站点分布、站点 OD 以及线路断面客流量分析；同时，也分析了地铁乘坐的频数分析。综上，得出了轨道交通客流分布在时间和空间上的一些规律。

在对客流进行预测和预警过程中，通过建立基于遗传神经网络的时间序列预测模型，对站点未来的客流情况进行分析。首先确定将 4 月 1 日至 7 月 16 日的 107 个客流量样本作为预测的基础数据，未来的 7 天作为预测的天数，利用时间序列分解将客流量数据分解为波动项和趋势项。然后对神经网络的输入层指标进行选取，经过天气、温度、节假日三方面的筛选，我们确定了 11 个指标。筛选完指标后，通过神经网络对输入层指标和输出层波动项进行训练，结果显示训练集的 R^2 为 0.72485，测试集的 R^2 为 0.80217，具有较好的拟合性。再用训练好的网络对未来客流量的波动项进行预测。通过 AR 时间序列模型可以预测客流量的趋势项，将预测的波动项和趋势项相乘即得到未来的客流量。通过上述方法，我们对未来所有站点的总客流量、未来单个站点的客流量以及未来所有站点在一天中不同时段客流量进行了预测。此外，我们还建立了客流预警系统，为企业的后置处理提供了一定的指导意见。最后，我们集成项目模型和方案进行了可交互界面的开发，用以相关从业人员的便捷使用。

参考文献

[1] 李飞羽. 城市轨道交通乘客行为特征分析及出行预测[D].华南理工大学,2020.

[2] 李欣盈. 基于客流时间序列数据的深圳与上海地铁站点比较研究[D].哈尔滨工业大学,2020.

[3] 郭文,肖为周,秦菲菲.基于支持向量机模型的地铁进站客流量预测[J].河北工业科技,2019,36(01):31-35.

[4] 杨永凯,宋瑞,李海荣. 地铁客流预测模型的分析与研究[A]. 中国运筹学会.第四届中国青年运筹与管理学者大会论文集[C].中国运筹学会:中国运筹学会,2001:7.

[5] 潘罗敏. 地铁短时客流量预测预警研究[D].首都经济贸易大学,2011.

[6] 苗沁,刘慧婷.城市轨道交通站点数据链客流预测方法研究[J].现代城市轨道交通,2021(03):64-67.

[7] 钱卫力,叶霞飞,陶志祥.东京、大阪、上海城市轨道交通高峰小时最大客流断面高峰系数对比分析[J].城市轨道交通研究,2012,15(02):50-53+58.

[8] 刘剑锋,罗铭,马毅林,王静,孙福亮,陈锋.北京轨道交通网络化客流特征分析与启示[J].都市快轨交通,2012,25(05):27-32.

[9] 周志华著.机器学习北京：清华大学出版社, 2016 年 1 月.

[10] 董升伟. 基于改进 BP 神经网络的轨道交通短时客流预测方法研究[D].北京交通大学,2013.