电 子 科 技 大 学

UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA

# 硕士学位论文

**MASTER THESIS**


论文题目 **机场出租车运力需求预测技术研究**

学 科 专 业 **通信与信息系统**

学　　号 **201521010526**

作 者 姓 名 **林　思　睿**

指 导 教 师 **段景山　　高级工程师**

分类号 _____ 密级 _____

UDC [注1] _____

# 学 位 论 文

## 机场出租车运力需求预测技术研究


林思睿

指导教师 段景山 高级工程师

电子科技大学 成 都


申请学位级别 硕士 学科专业 通信与信息系统

提交论文日期 2018.04.02 论文答辩日期 2018.05.23

学位授予单位和日期 电子科技大学 2018 年 6 月

答辩委员会主席 _____

评阅人 _____

注 1：注明《国际十进分类法 UDC》的类号。

# RESEARCH ON FORECAST TECHNOLOGY OF AIRPORT TAXI CAPACITY

A Master Thesis Submitted to

University of Electronic Science and Technology of China


**Discipline :** **Communication and Information Systems**

**Author:** **Lin Sirui**

**Supervisor:** **Senior Engineer Duan Jingshan**

**School:** **School of Information and Communication Engineering**

# 独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得电子科技大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

作者签名：林思睿 日期：2018年4月2日

# 论文使用授权

本学位论文作者完全了解电子科技大学有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权电子科技大学可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后应遵守此规定）

作者签名：林思睿 导师签名：段高山

日期：2018年4月2日

# 摘 要

整体经济的增长及居民消费水平的提升刺激了航空客运需求的急速增长，更加大众化和更加普及的飞机出行方式加速了机场旅客吞吐量的增加，这为大型机场的发展提供了重要的人群基础，也给机场多元化交通方式的运力带来了更大的压力。出租车作为一种具有较好灵活性的交通工具，便于进行全天候的客流数量分析。另外，研究出租车运力需求预测技术可以推广到其他交通工具的需求分析中，为机场运力需求提供参考。因此，本文将主要针对出租车的运力需求情况进行建模分析，并用首都机场实时统计的数据检验模型的有效性。

本文首先进行客流计数统计设备数据的分析，并进行数据清洗，即重新审查和校验客流计数统计设备采集到的数据，发现并纠正数据文件中的错误。接着，本文研究并选取了机场出租车客流模型特征，即通过分析机场出租车客流的规律，从各种形式的航班、旅客数据中挑选出最能刻画运力需求预测模型的特征供机器学习的算法和模型使用，以进行运力需求预测。在随后运力需求预测算法的研究与设计中，本文根据待分析的机场出租车客流数据的特点，构建适当的模型，设计合适的数据挖掘算法，并根据当前数据库中的实时信息，预测出租车运力需求量。最后对研究成果进行了验证，本文在模拟环境下进行仿真，并使用首都机场实际采集得到的客流及运力调度数据对算法的准确率和预测效率进行验证，并根据验证结果，对预测模型进行了改进。验证结果表明：基于机场出租车客流特点的趋势季节性去除算法在各个时间点上的预测有更好的表现，能够有效预测候车区客流数量。

模型得到的出租车运力情况，可以为机场出租车调度中心潜在的出租车运力不足进行预警，以合理分配调度出租车资源，解决旅客滞留，提高机场的客流运输能力。

**关键词：**首都机场，交通流预测，数据清洗

# ABSTRACT

The growth of the overall economy and the increase in the level of resident consumption have stimulated the rapid growth in the demand for air passenger transport. More popular aircraft travel modes have accelerated the increase in airport passenger throughput, which provides an important population base for the development of large airports. It also exerted greater pressure on the transportation capacity of the airport's diversified modes of transportation. Taxi, as a kind of flexible vehicle, is convenient for the analysis of the number of passengers around the clock. In addition, the research on the demand forecasting technology of taxi capacity can be extended to the needs analysis of other modes of transport and provide reference for the demand for airport transportation capacity. Therefore, this paper will mainly focus on the analysis of the capacity requirements of taxis, and use the real-time statistical data of Capital Airport to test the effectiveness of the model.

This paper first analyzes the data of the passenger flow counting statistics equipment and performs data cleaning, that is, re-reviews and checks the data collected by the passenger flow counting statistics equipment, finds and corrects errors in the data file. Then, this paper studies and selects the characteristics of airport taxi passenger flow model, including analyzing the law of airport taxi passenger flow, and selecting the features for machine learning that best characterize the capacity demand forecasting model from various forms of flight and passenger data. Models are used to forecast capacity requirements. In the research and design of subsequent capacity demand forecasting algorithms, this paper builds an appropriate model based on the characteristics of airport taxi passenger flow data to be analyzed, designs appropriate data mining algorithms, and predicts taxis capacity demand based on real-time information in current databases. Finally, the research results are verified. This paper simulates the simulation environment, and uses the passenger flow and capacity scheduling data collected by the Capital Airport to verify the algorithm's accuracy and prediction efficiency. Based on the verification results, the prediction model is improved. The verification results show that the seasonality and trend removal algorithm based on the characteristics of passenger flow in airport taxis has better performance at various time points and can effectively predict the number of passengers in the waiting area.

The taxi capacity obtained by the model can provide an early warning for the potential shortage of taxi capacity at the airport taxi dispatch center, allocate and dispatch taxi resources reasonably, solve the passenger detention, and improve the passenger flow capacity of the airport.

**Key words:** Capital Airport, Traffic Flow Prediction, Data Cleaning

# 目 录

# 图目录

# 表目录

# 第一章 绪 论

## 1.1 研究工作的背景与意义

整体经济的增长及居民消费水平的提升刺激了航空客运需求的急速增长，国内外的机场，特别是大型枢纽机场逐渐向中枢辐射式发展，正经历新一轮改扩建的新时期。分析近年民航旅客运输市场构成情况，以首都机场为例，2016 年的旅客吞吐量为 94,393,454 人次，上年同期为 89,939,049 人次，同比增长了 5.0%。起降架次为 606,081，上年同期为 590,199，同比增长了 2.7%。更加大众化和更加普及的飞机出行方式使得旅客吞吐量快速增长，这为大型机场的发展提供了重要的人群基础，也给机场多元化交通方式的运力带来了更大的压力。

大型枢纽机场拥有巨大的旅客吞吐量，同时人员流动呈现高动态、高密度可变、时间分布不均匀的特点，随之而来的是巨大的服务压力，也因此，需要一个与机场旅客吞吐量相匹配的交通运力系统才能充分发挥机场的作用和潜力。从现状来看，首都机场作为当今最具发展潜力的航空枢纽之一，承担着国内主要的航空业务，但其运力供应能力并不完善。由于首都机场地处远离市区的位置，导致机场周边交通干道一旦拥堵，将会有大量旅客滞留机场，这会大大降低旅客对机场的体验度。随着进出港客流量的持续增长，首都机场运力供应以及周边道路通行保障能力将承受巨大的考验。

出租车作为一种具有较好灵活性的交通工具，可以进行全天候的客流数量分析。另外，研究出租车运力需求预测技术可以推广到其他交通工具的需求分析中，如轨道交通、机场巴士、出租车等等，为机场运力需求提供参考。因此，本文将主要针对出租车的运力需求情况进行建模分析，并用首都机场实时统计的数据检验模型的有效性。模型得到的出租车运力情况，可以为机场出租车调度中心预警潜在的出租车运力不足的情况，合理分配调度出租车资源，解决旅客滞留，提高机场的客流运输能力。对于旅客来说，可以提供诸如出租车候车区预计队伍长度和打车时间等可靠的信息。

## 1.2 运力需求预测的国内外研究现状

运力需求预测，即通过分析机场出租车运力需求数据，充分考虑客流数据的特性，有机结合机器学习的各种理论方法，研究其变化趋势，并据此对短期的需求进行预测。同时深入分析单个事件对运力需求的影响，以最大限度提升算法的预测效

果。进行客流数据分析的目的，是通过识别客流数据中的趋势和变化，加深对运力数据的理解并进行建模。

国外针对机场运力需求预测问题有很多相关的研究。

Senay Solak 等人[1]讨论了对于一个机场客流场景，很难建立一个整体模型，一方面是因为航班的交错会使得客流的分析变得复杂，另一方面，由于天气等客观因素，模型之后会掺杂很多偶然性。这种偶然性在 Wai Hong Kan Tsui 等人[2]的研究中有所体现，Wai Hong Kan Tsui 等人用 ARIMA 和 SARIMA 进行仿真，利用香港机场的数据进行验证，讨论了由于 SARS 病毒这种偶然事件使得机场客流模型受到的影响。Xia Liu 等人[3]用三亚机场的数据讨论了使用时间序列预测模型对机场人数进行评估的几种方案。这些研究启发了本文在后续的研究中，采用时间序列分析和数据挖掘的方式，而不是建立仿真模型。

Antony Evans 等人[4]的研究和 Abdulla Al-Dhaheri 等人[5]的研究分别以仿真的形式和分析实际数据的方式讨论了机场容量的发展并提出了一些优化机场客流服务的设想。之前提到的 Wai Hong Kan Tsui 等人[2]指出了机场的发展和周围的交通建设互相制约的关系，机场需要接续运输保障体系的持续完善才能充分发挥作用和潜力。这些研究对于本课题后续的应用场景有重要的启示。

Shin-Lai (Alex) Tien 等人[6]讨论了天气因素对于机场客流量的影响。Günther Yves 等人[7]深入讨论了对于机场天气因素的考虑，对优化机场体系的作用。这些研究启发了本文在之后的预测中考虑天气这个重要的因素。

Sung Wook Yoon 等人[8]则讨论了构建一个乘客下机时间和取行李行为关系的模型，其中的一些数据，对于本文后续根据机场排班估计乘客的数量有很大的帮助。

Chamath Malinda Ariyawansa 等人[9]分析了不同的数据挖掘方式的特点，并且提到了几个国外的数据集，这些数据集对于侧面验证本文模型的准确度有重要的作用。

Hai Yang 等人[10]通过数学方式解释了出租车和乘客之间的平衡关系。Junghoon Lee 等人[11]基于出租车公司所记录的历史数据分析了出租车的载客模式，并利用 K 聚类方法对乘客上车地点进行研究。Harold Nikoue 等人[12]讲述了如何利用排队论对排队乘客进行仿真。Wenbo Ma 等人[13]研究了如何对飞机乘客的数量进行估计。Duilio Curcio 等人[14]和 P. Fonseca i Casas 等人[15]的研究涉及候机乘客数量的仿真。这些论文中的数据对于后续的候车区客流人数验证都是非常重要的参考。

在国内，陈玉宝，曾刚[16]采用 Holt-Winter 季节模型、ARMA 和线性回归模型对三亚机场 2016-2017 年的客流量进行了预测。王磊[17]运用 Logit 模型对机场轨道交通的分担率进行客观地预测，并用西安咸阳国际机场的数据验证上述理论。上述

研究中的模型和预测方法对于本文的工作有重要的启发，其中采用的一些数据可以用来验证本文预测的准确度。

## 1.3 研究内容与章节安排

基于上述研究背景以及本课题的研究目标，结合国内外研究现状，有以下研究内容：

（1）客流计数统计设备数据分析

首都国际机场在出租车候车区的入口和出口分别安装了一套客流计数统计设备进行出租车客流数量统计，然而统计设备统计得到的数据中有很多的错误数据，并不能直接应用，需要进行数据清洗。数据清洗，即重新审查和校验客流计数统计设备采集到的数据，发现并纠正数据文件中的错误。客流数据充满各种错误、噪声、异常值会使得运力需求预测算法无法正常运行。在运力需求预测模型中，需要将错误数据进行剔除，用算法进行清洗，以提供更为可靠的数据。

（2）机场出租车客流模型特征选取与研究

因本课题采用机器学习的方式进行运力需求预测，故需要结合相关的模型预测需求，分析机场出租车客流的规律，并基于这些规律，选择有意义的特征输入机器学习的算法和模型进行训练。

客流模型特征是在运力需求预测中的一种独立、可测量的属性，客流计数数据的可靠性和选取的特征决定了运力需求预测的上限。所谓特征选择，即从已有的各种形式的航班、旅客数据中挑选出最能刻画运力需求预测模型的特征。研究并有效选取模型的特征，才能最大限度从原始数据中提取特征供算法和模型使用。

（3）运力需求预测算法的研究与设计

机场出租车候车区客流运输数据是按观测时间排列、随着时间变化且观测数据之间相互关联的数据序列。客流运输数据分析，本质上就是一个时间序列的趋势分析，而且是一个季节时间序列的分析。

本课题根据待分析的机场出租车客流数据的特点，基于机器学习的相关理论，借鉴现有的研究成果，构建适当的模型，设计合适的数据挖掘算法。并结合数据库操作，根据当前数据库中的实时信息，计算对应的出租车参考数量。

（4）机场出租车运力需求预测的验证

通过在模拟环境下进行仿真以及对比在首都机场采集得到的客流及运力调度数据和预测数据，对算法的准确率和预测效率进行验证。最后根据验证结果，对预测模型进行改进。本文实现了算法并对比分析了几种模型的预测效果。

论文的章节安排如下：

第一章绪论部分，论述了出租车运力需求预测技术的相关研究背景、进展和研究的意义，阐述了首都机场的规划发展情况和相应的运力需求预测研究现状，最后指出本文的研究内容。

第二章运力需求预测方法综述部分，结合预测的目标，介绍了不同的预测模型，这些模型也是后续研究的重要基础。

第三章运力需求预测模型结构部分，首先介绍了运力需求预测技术及其目的，接着介绍了课题重要的相关概念和方法，最后介绍了运力需求预测的几个要素，包括运力需求数据的分析、数据清洗、特征工程和模型设计。

第四章首都机场出租车候车区场景分析部分，首先介绍首都机场实际的出租车接续运输模型，描述了客流数据统计出现的问题、出错的原因和数据清洗的方法，最后结合机场统计的旅客信息、旅客流量在各个时段的分布和旅客到港的行为特点对这些数据进行分析。

第五章基于机场出租车运力需求特点的预测算法，在第二章、第三章的基础上，实现了 ARIMA 模型和 LSTM 模型并使用这些模型对出租车客流的运力需求进行预测。最后对预测方法进行优化，使得预测模型有更好的表现，并和之前实现的模型作对比，对模型的提高有更直观和更量化的认识。

第六章是全文的总结和展望，同时阐述了本课题下一步的研究方向。

## 1.4 本章小结

本章主要讨论了出租车运力需求预测技术的相关背景和研究意义，阐述了首都机场的发展情况以及国内外相应的理论研究成果，如短时交通流预测和相应的影响因素。最后，指出本文的主要研究内容和相应的论文结构。

# 第二章 机场出租车运力需求预测方法综述

机场出租车运力需求因为航班的安排、天气因素的影响以及突然的爆发性事件例如第一章提到的 SARS 病毒等具有不确定性和动态性，同时也因为人们出行的总体规律性和交通条件的约束具有周期性和相关性。机场出租车运力需求预测受到很多随机因素的影响，具有很大的时变性和不确定性。

运力需求预测中有三种应用广泛的预测模型：均值模型、ARIMA 时间序列模型和 LSTM 模型。本课题先采用 ARIMA 模型进行运力需求预测。在北京交通委提供了更多具体的航班及旅客数据之后，整理得到更多客流特征，使用 LSTM 模型解决运力需求预测问题，并在之后对 LSTM 模型进行优化，达到更准确的预测效果。本章是之后提出基于机场出租车客流特点的趋势性季节性去除算法的重要基础。

## 2.1 均值模型

对于机场出租车候车区客流数据，历史统计值的加权平均值是一个较好的估计值。用公式表示就是：

$$F_{t+1} = \sum_{i=1}^{n} \alpha_i F_i \tag{2-1}$$

其中，$F_i$表示每个时间点统计得到的出租车候车区客流人数，$F_t$为需要预测的客流人数，$\alpha_i$为每个统计值的权重，当$\alpha_i = \frac{1}{n}$，即取历史均值。

均值模型是一个静态预测模型，在常规有规律的环境中有较好的表现，但是应用到客流量预测中可能会出现悬崖式跳变的情况，例如前一分钟的历史均值是 30 人，下一分钟的历史均值是 2 人，考虑到出租车候车区的客流具有连续性的特征，出现这种断点的情况是不合理的。另外，在预测客流量这种环境中，均值模型难以适应突发事件引起的不确定性和非线性，例如第一章提到的因为 SARS 病毒爆发事件使得机场客流受到严重影响。因此，均值模型需要结合其他模型进行修正。

## 2.2 ARIMA 时间序列模型

### 2.2.1 AR 模型

一个 $m$ 阶的自回归模型可以表示为：

$$x_{t+1} = \varphi_1 x_{t-d} + \cdots + \varphi_m x_{t-d-m+1} + \varepsilon_{t+1} \tag{2-2}$$

记为 $AR(m)$，其中 $m$ 是自回归阶数，$\varepsilon_t$为白噪声，代表当前值所受到的外部扰

动。

AR 模型的理论基础是，假设时间序列的任意一个元素$x_t$和之前的若干元素$x_{t-1}$，$x_{t-2}$ ...之间存在一定的数学关系，于是可以根据过去的观测值进行预测。对于机场出租车候车区客流模型，*AR*(*m*)模型假定当前的客流状态只由过去 *m* 个客流状态决定，利用之前 *m* 个观测值的线性组合作为回归变量描述之后的某个时刻的随机变量，是一个基于多元回归分析的模型。

### 2.2.2 MA 模型

一个 *m* 阶的移动平均模型可以表示为：

$$y_{t+1} = \varphi_1 \varepsilon_t + \cdots + \varphi_m \varepsilon_{t-m+1} + \varepsilon_{t+1} \tag{2-3}$$

对于机场出租车候车区客流模型，*MA*(*m*)模型假定当前的客流状态只由过去 *m* 个环境的扰动决定，利用之前 *m* 个时刻的误差项的线性组合作为回归变量描述之后的某个时刻的客流状态。当 MA 模型是有限阶时，这个移动平均过程总是平稳的，因此当平稳性条件无法满足，不能使用 AR 模型时，可以考虑使用 MA 模型。

### 2.2.3 ARMA 模型

自回归移动平均模型 ARMA，将 AR 模型和 MA 模型结合起来产生一个新的模型。

可以表示为：

$$y_{t+1} = \varphi_1 y_t + \cdots + \varphi_p y_{t-p+1} + \varepsilon_{t+1} + \theta_1 \varepsilon_t + \cdots + \theta_q \varepsilon_{t-q+1} \tag{2-4}$$

显然 AR 模型和 MA 模型是 ARMA 模型的特殊情况。

ARMA 模型表明，当前的状态受到过去的状态和环境的扰动影响，因此既有 AR 项又有 MA 项。

### 2.2.4 ARMA 求解的前提

理论上，通过合理的函数变换之后，任何时间序列模型，都可以分解为趋势项，周期项和随机噪声项：

$$y(\mathrm{t}) = g(t) + s(t) + h(t) + \epsilon_t \tag{2-5}$$

其中，*g*(*t*)表示趋势的变化，*s*(*t*)是时间序列中的周期性部分，*h*(*t*)代表了节日的影响，$\epsilon_t$代表了一些突发事件。

对于机场客流来说，趋势项，即随着航空客运的发展，每年机场客流会呈现一个增长的趋势；周期项，呈现为航空客流有旺季淡季之分，夏季和早春是旺季，客流量会比较多，另外每年的八月份都是当年客流量最多的月份；同时每天的客流量

都有相似的趋势，会在一些固定的时段出现客流高峰的情况。

趋势的变化和周期性部分是导致客流序列非平稳的主要因素。对于这种非平稳的序列，需要转换为平稳序列进行处理，建立合适的平稳模型之后，再通过逆变换，得到非平稳序列的统计模型。

ARMA 是一个线性系统模型，在推导 ARMA 的求解方程时认为系统的输入是高斯白噪声序列。

高斯白噪声的功率谱是一个常数$\sigma^2$，此时 ARMA 系统的传输函数和输出等效：

$$H_{out}(z) = H_{in}(z) \cdot H_{arma}(z) = \sigma^2 \cdot H_{arma}(z) \quad (2\text{-}6)$$

即当高斯白噪声序列通过一个线性系统，会产生一个平稳序列，也就是当前已知的时间序列。如果$H_{in}(z)$不是常数，那么就需要先验地知道输入序列的特征才能求解 ARMA 方程，对于机场出租车运力需求预测模型来说，客流数据的特征太多，先验地知道输入序列的特征是做不到的。

### 2.2.5 ARIMA 模型

AR 模型、MA 模型和 ARMA 模型要求客流序列是平稳的，非平稳序列不能用 *ARMA*(*p*,*q*)模型来描述。

差分自回归移动平均模型，*ARIMA*(*d*,*p*,*q*)模型是 ARMA 模型的延伸，如果机场客流候车区客流运输数据经过 $d$ 阶差分之后能够转换为一个平稳的时间序列，那么就可以对差分之后的数据建立 ARMA 模型。其中 $p$ 是自回归项，$q$ 为移动平均项数。

## 2.3 LSTM 模型

### 2.3.1 传统神经网络模型

如图 2-1 所示是传统神经网络模型的结构。

同一层的节点相互独立，不同层的节点相互连接，信息在层和层之间直接传递。这种模型对于数据各个时刻的信息的处理是相互独立的，即，$t$ 时刻的输出$y_t$只和该时刻的输入$x_t$有关。对于机场出租车候车区客流模型，各个客流信息的时间顺序是十分重要的，他们之间具有相关性，因此传统神经网络模型不适合用来对本模型进行建模。

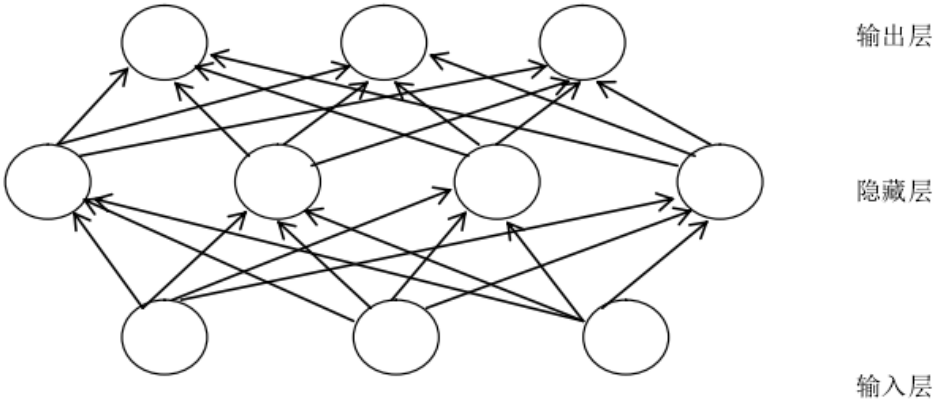


图 2-1 传统神经网络模型的结构

### 2.3.2 RNN 模型

RNN，循环神经网络，可以解决各个时刻的信息需要相互关联的问题。

如图 2-2 所示是循环神经网络模型的结构，RNN 模型中，隐藏层相邻节点之间也互相连接，即，同一层节点上的信息也会互相影响。

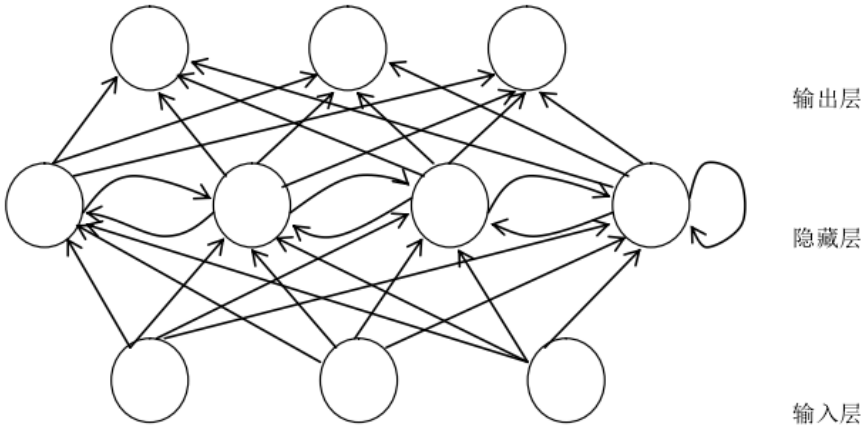


图 2-2 循环神经网络模型的结构

如图 2-3 所示是循环神经网络模型状态更新示意图。

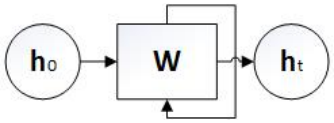


图 2-3 循环神经网络模型状态更新示意图

RNN 模型中，权重参数是在各个时刻共享的，于是模型向后传递信息的过程就是将初始状态的信息多次乘以连接隐藏层前后的权重矩阵 $W$ 的过程，如图 2-3 所示，计算 $t$ 时刻的状态信息，即是：

$$h_t = W^t h_0 \tag{2-7}$$

当矩阵的特征项小于 1，则$W^t$会向 0 矩阵逼近，即在隐藏层传递信息的过程中，信息消失，反之，当特征项大于 1，$W^t$逼近∞矩阵，则在传递信息的过程中出现信息爆炸的情况。RNN 在处理相隔较远的信息时有自身的局限性。因此对于机场出租车运力模型来说，RNN 模型可以将过去状态的信息传递到当前状态的计算中，但是难以传递距离较远的信息。

### 2.3.3 LSTM 模型

LSTM 是一种时间递归神经网络，标准 LSTM 单元如图 2-4 所示。

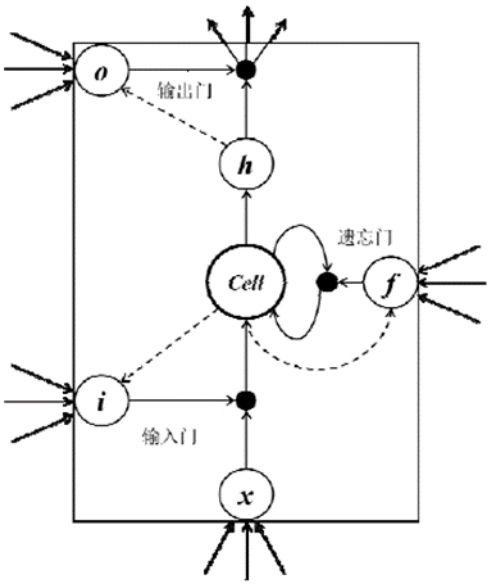


图 2-4 标准 LSTM 单元

和传统 RNN 相比，增加了输入门、遗忘门和输出门来控制和保存历史信息。LSTM 主要是为了解决传统 RNN 网络随着层数增多出现的梯度消失或爆炸的问题，相比传统 RNN，将每一层神经元设计成具有多个门的结构，使得误差在传播时有些可以通过“门”直接到下一层，保证梯度不会因为计算的原因完全消失，使得权值在每次训练时都有更新变化，因此有很好的收敛性。因此对于机场出租车运力模型来说，LSTM 的层数可以更多，可以有效利用长距离的历史客流信息，捕捉长期依赖关系的特性。

ARIMA 模型的缺点在于，在预测模型建立之后，因为参数固定了，无法随着时间的前进实时更新，模型的预测没有考虑最新的观测信息。一种解决方法是，预测模型重建，但这无疑会增大计算量。LSTM 是 ARIMA 模型的自然延伸，但是更加灵活，是一种非参数模型，这简化了学习的过程，同时可以通过滚动预测的方式

反映最新的客流观测信息。

LSTM 模型的另一个优势在于，ARIMA 模型假设变量之间呈现一个线性的关系，但是 LSTM 模型避免了这种事先的指定。ARIMA 模型比较适合稳定的客流状态，对于机场出租车候车区客流这种变化频繁的场景难以很好应对。

另外，LSTM 可以处理多个输入变量的问题，其他外在的特征能够很简单地插入到模型中。基于 ARIMA 的时间序列模型预测可以为预测提供一个比较好的基准值，但很多波动无法刻画，这些波动往往都是有航班到达事件引起的，需要通过航班特征来刻画。LSTM 可以通过增加输入变量，灵活地加入有价值的假设或经验，更好地刻画时间序列的变化。

## 2.4 本章小结

本章介绍了出租车候车区客流的一些运力需求预测算法：均值模型，ARIMA 时间序列模型、LSTM 模型并将这些模型应用到实际的预测之中，其中均值模型和 ARIMA 时间序列模型属于参数回归模型。本章希望通过充分挖掘不同运力需求预测模型所获取的不同方面的有效信息，将多个模型融合成一个整体，以达到对运力需求预测准确度、模型复杂度以及模型响应时间等多方面的综合性能提高的目的。

# 第三章 机场出租车运力需求预测模型结构

本文第二章主要阐述了相关的运力需求预测算法，本章将以此为基础，针对运力需求预测模型设计与实现中的关键技术进行分析和论述。

运力需求预测模型的结构如图 3-1 所示。机场运力需求预测有几个要素：第一个是机场出租车运力需求数据的分析，运用时间序列的理论方法，分析客流数据的统计规律。第二个是数据清洗，利用客流数据的时空相关性，剔除错误的数据，填补缺失的数据，排除数据中各种错误、噪声、异常值的干扰，有助于提高预测的准确性。第三个是特征工程，用一个好的状态向量描述客流数据的大部分信息，这有助于去掉冗余的信息，降低模型训练的时间，提高模型的推广能力。第四个是运力需求预测模型的研究，并使用首都机场真实的出租车客流数据对模型的有效性进行检验。

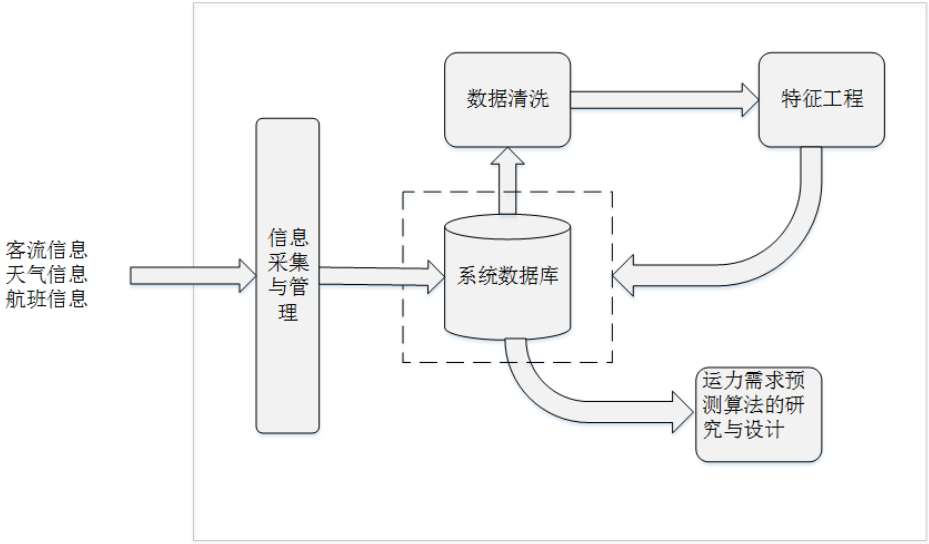


图 3-1 运力需求预测技术总体设计

## 3.1 时间序列研究

时间序列是指按照时间先后顺序排列的各个观测记录的有序的离散的集合，数据包含了所观测的系统在各个时间点的客观记录。

机场出租车候车区的客流数量序列就是一个典型的时间序列，客流量的改变可以看作是之前机场出租车候车区旅客数量受到机场航班等因素的影响而发生变化的结果，观测值随着时间推移不断变化并且观测值之间互相关联，数据源不断录入更新，是一个动态的时间序列。

时间序列中包含着丰富的信息，包括显式的直观信息和隐式的内在信息。对于机场出租车候车区客流数量的时间序列而言，显式的直观信息是指过去时期出租车候车区的客流量数量，隐式的内在信息则是出租车客流长期变化的趋势。另外，随着时间的增长，收集的机场客流量信息不断增多，时间序列的信息量也在不断扩充，可以渐渐从把握客流量的日流量趋势到分析客流量的年增长变化。进行机场运力需求分析时，光是客流数据就有超过 1000 万条，海量数据对算法的储存空间和性能也提出了更高的要求。

机场出租车候车区的客流数据有以下的特性：

- 运力需求变化规律与人的生产和生活规律相关，随着人的工作和休息的相互交替，具有一定的周期性。具体来说，每天机场的工作人员和出租车司机的生产和生活安排有一定的规律，因此，每天的运力需求曲线有相似性；对于每周来说，相同的工作日的安排比较相似，客流的趋势也比较相似，而不同工作日之间的到港旅客人数有较大的差异，经过统计，周四、周六的日进港旅客人数最多，而周二最少；对于每年来说，夏季是旅行的最佳季节，因此夏季的到港旅客人数最多。另外，法定节假日中因为出行游客增多，到港旅客激增，长假日和正常日有很大差别，但是长假日之间有相似的曲线趋势。
- 在机场出租车候车区客流数量的观测中，相邻观测值间具有相互依赖的特性，不会出现相邻的两个观测时间客流数量差别巨大的情况，两个值之间遵循着客流的连续性，而普通静态数据没有时间相关性。
- 在机场运力数据的演变中，随着时间的推移，之前统计的客流数据对于当前运力需求预测的影响越来越小，越接近当前时段，越反映最新的客流信息，参考价值越大。
- 由于客流的进出比较频繁，体现在数据上就是短期波动频繁，趋势变化特征不明显。

## 3.2 数据清洗

在本课题中，客流计数统计设备采集到的数据受到各种意外因素的影响，有很多的错误数据，利用这些数据计算得到的客流人数和实际统计人数相差较大，并不能直接应用，需要滤除冗余的干扰数据，进行数据清洗。

所谓数据清洗，即重新审查和校验客流计数统计设备采集到的数据，发现并纠正数据文件中可识别的错误。客流数据充满各种错误、噪声、异常值会使得运力需求预测算法无法正常运行。在运力需求预测模型中，需要将错误数据进行剔除，用算法进行清洗，以提供更为可靠的数据。

数据清洗的内容包括：

● 修正人为的错误，例如在之前时间序列的采集中，是通过直接在高清网络球机的录像中统计该时刻的客流人数的，这样的统计方式难免会因为视觉的角度等原因，导致统计的人数和实际人数有出入。

● 检查数据一致性，不同系统产生的数据进行合并的时候，也可能导致错误，例如在机场的表项中，有些表项中的时间是格林威治时间，有些表项的时间是北京时间，在两份数据进行合并时，如果没有进行处理就会产生错误，这种错误属于表示方式不统一导致的不一致。

● 数据统计方式会产生异常，要替代、修改和删除错误数据，例如在机场的客流统计场景中，就因为光照、工作人员的影响使得数据产生异常。机场工作人员在特定时间的一些工作，例如收集手推车在出入口走动，会给实际数据的统计带来一定程度的干扰，使得这段时间的客流量变动和其他时间相比特别反常。在机场的表项中，有些延误的航班在人工修改到达时间时，也可能因为工作人员的疏忽而登记错误。另外，在机场的出租车标注车牌的时候，出现了 O 和 I 两个现实中不会出现的车牌字母，即使不考虑现实中车牌被设计成没有 O 和 I，这种分类的不同也显然会带来问题。

● 处理无效值、缺失值等，客流设备会因为偶然的断电缺失一整段数据，缺失过多的数据直接丢弃。

通过数据清洗，可以使得原始数据变得完整、正确、一致。

国外对数据清洗的研究最早出现在美国，是从对全美的社会保险号错误的纠正开始。美国信息业和商业的发展，极大地刺激了对数据清洗技术的研究[18]。

Jiawei Han 等人[19]的工作讲述了数据清洗的方法，其中关于时间数据序列数据清洗的方法对于本文客流计数统计设备数据的处理有重要的参考意义，在运用这些方法对数据进行预处理后，能够有效提高预测的准确率。

## 3.3 特征工程

“数据决定了机器学习的上限，而算法只是尽可能逼近这个上限”，这里提到的数据指的是经过特征工程处理得到的数据。因本课题采用机器学习的方式进行客流数量预测，故需要结合相关的模型预测需求，分析机场出租车客流的规律，并基于这些规律，选择有意义的特征输入机器学习的算法和模型进行训练。从某种层面来说，使用的特征越好，得到的效果就会越好。在机场出租车运力需求预测模型中，需要在众多的属性中提取最相关的属性进行分析，与目标相关性高的特征，应当优先选择。

所谓特征工程，指的是把原始数据转变为模型的训练数据的过程，时间序列的维数一般很高，而特征工程根据特定的评估标准，从当前统计得到的变量中，精选出数量较少的变量作为特征子集，并且这个特征子集能够有效反应当前时间序列的主要信息，从而有效节省储存和计算的资源[20]。

进行特征选择的原因有：

- 特征数量越多，模型越复杂，会降低这个模型的推广能力。本课题使用首都机场出租车候车区客流运输数据验证模型的有效性，推广能力下降，也就难以应用到其他运输工具的需求预测中。
- 原始数据之中存在冗余信息，特征之间可能存在相互依赖。另外，机场运力需求预测涉及的表项有很多，包含的特征也有很多，如未来一小时到港航班数量、未来一小时到港旅客数量、附近的交通路况信息、航班起飞到达等，并不是所有的特征都和当前运力需求预测的目标相关。
- 特征数量过多，容易导致分析特征、训练模型的时间过长，特征矩阵过大会导致巨大的计算量。

特征选择的方法有以下几种：

- 过滤法：计算各个特征的方差/相关系数/互信息，设定阈值选择特征。
- 包装法：根据目标函数，每次选择或排除若干特征。
- 集成法：使用某些机器学习算法进行训练，得到各个特征的权值系数，根据系数的计算值选择特征。

## 3.4 运力需求预测算法的研究与设计

机场客流数据分析，本质上就是一个时间序列的趋势分析。

Erich Fuchs 等人[21]提供了两种研究时间序列的方法。一种是通过正交多项式拟合研究短期时间序列，另一种是通过神经网络研究长期时间序列。现阶段数据相对较少，对于一周的数据乃至一个月数据的拟合，考虑参考第一种方法，对时间序列的趋势有更深入的了解，也方便规律的提取。这种方法的使用在 Erich Fuchs 等人的研究中[22]有讲述。在项目后期，随着数据的增多，可以考虑参考第二种方法，达到分析和预测乘客流量和出租车流量规律的研究目的，这种方法的使用在 Georg Dorner 的研究[23]中有讲述。

Guido Perboli 等人[24]讨论了如何对取行李的乘客的行为进行仿真。Gabriel Aguilera-Venegas 等人[25]利用元胞自动机和神经网络仿真，估计取行李乘客的数量。这些论文对于机场项目中的短时运力实时预测有着重要参考。在分析旅客的出站时间时，假如将取行李的行为考虑进来，得到的结果会更加准确。

另外，由于机场客流服从某种季节模式，会在某个特定时期呈现增长或下降的趋势。具体来说，旅行通常都会在夏季达到最高峰，而第二高峰出现在春季，这些特点都会体现在机场客流数量上。因此，机场客流序列，更精确地讲，是一个季节时间序列。时间序列的相关研究[26]讲述了几种季节模型的构建方式，以及如何通过加入简化算子和建立乘积模型对模型进行简化。

本课题在数据学习方向上，打算采用集成学习的方式。集成学习使用某种规则把各个模型进行整合，从而获得比单个模型更好的预测效果。Thomas G. Dietteric[27]介绍了几种集成学习的方法，并针对不同的问题，给出了不同的集成学习的方式，最后对不同方式的特点进行总结。本文最后将通过动态的到港离港航班信息和出租车候车区实时统计的客流信息情况进行分析计算得到预测结果。

## 3.5 本章小结

本章首先对机场出租车运力需求预测在实际过程中的需求进行分析，并主要介绍了研究中所需要的理论基础和方法基础，包括机场出租车运力需求数据的分析、进行客流计数统计设备数据清洗、客流模型特征选取与研究以及最重要的运力需求预测算法的研究与设计，为后续工作提供研究基础。

# 第四章 首都机场出租车候车区场景分析

本文之前两章分别对运力需求预测模型算法和模型结构的关键技术进行研究和分析。本章将依据运力需求预测技术的需求，结合相应的运力需求预测算法，对首都机场出租车候车区场景中运力需求预测的关键内容进行详细设计。

## 4.1 机场出租车候车区客流运输模型

如图 4-1 所示，乘客在候车厅等候上车，需要预测的就是出租车候车区的客流数量，从而为机场调度部门的决策提供数据支持。

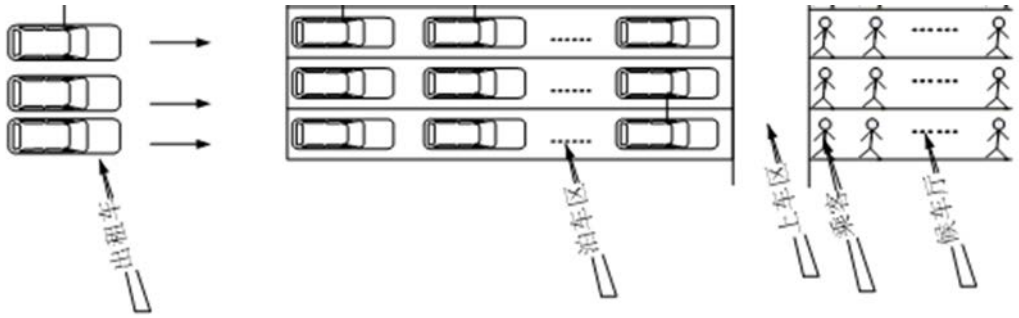


图 4-1 机场出租车候车区客流接续运输场景示意图

以首都机场 T1、T2 航站楼为例，如图 4-2 所示，这两个航站楼共用一个出租车蓄车池，接到调度指令后出租车到 T1 航站楼或 T2 航站楼。


图 4-2 T1、T2 航站楼出租车待运区车流采集示意图

客流数据主要采集自首都机场 T2 和 T3 航站楼出租车候车区，为了获取高质量的客流数据，通过客流计数摄像机采集数据。另有首都机场信息服务部门提供的 4 万余条旅客相关记录。

以 T3 航站楼为例，客流计数摄像机的分布如图 4-3 所示，安装位置共有 7 个，

为 5 个出口和 2 个入口，分别统计 T3 航站楼出租车候车区出口离开的客流数量和入口进入的客流数量，采集时间为每分钟一次。另外还有两台高清网络球机可以查看候车区的现场实时视频，可以进行人为的统计，和通过客流计数摄像机的数据计算得到的人数进行对比校准。

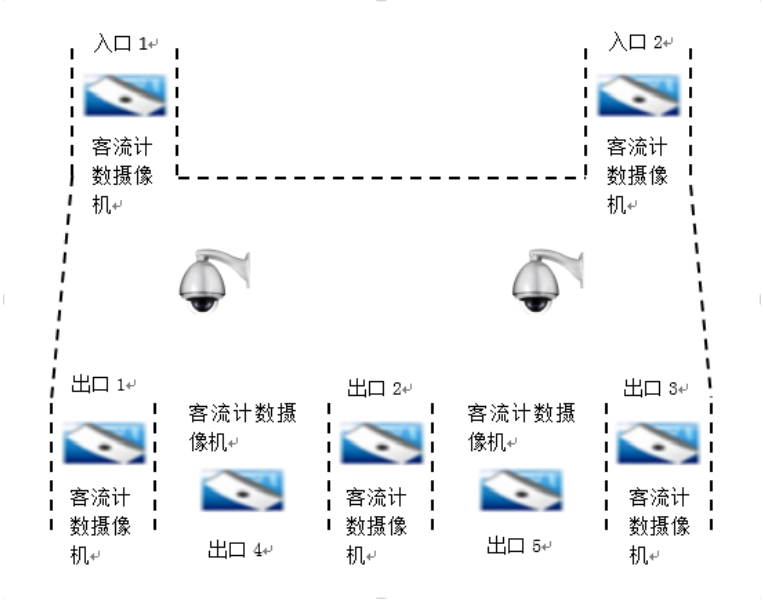


图 4-3 T3 航站楼传感器分布图

T3 航站楼出租车候车区每分钟的旅客变化数为：

$$P_{\text{旅客每分钟变化人数}} = \sum P_{\text{入口 i}} - \sum P_{\text{出口 i}} \tag{4-1}$$

于是 T3 航站楼出租车候车区的旅客计算方式为：

$$P_{\text{排队旅客数}} = P_{\text{排队区域原有旅客数}} + P_{\text{旅客每分钟变化数}} \tag{4-2}$$

T1 航站楼和 T2 航站楼旅客排队数量统计示意图如图 4-4 和图 4-5 所示。

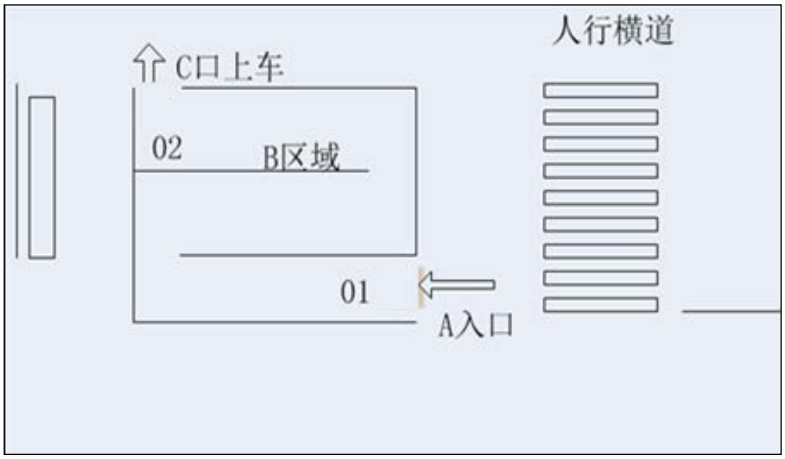


图 4-4 T1 航站楼旅客排队数量统计示意图

T1 航站楼和 T2 航站楼这两个航站楼同样安装了客流计数摄像机统计出租车候车区出口离开的客流数量和入口进入的客流数量。

这两个区域中的出租车候车区旅客的数量可以用类似的方法计算。

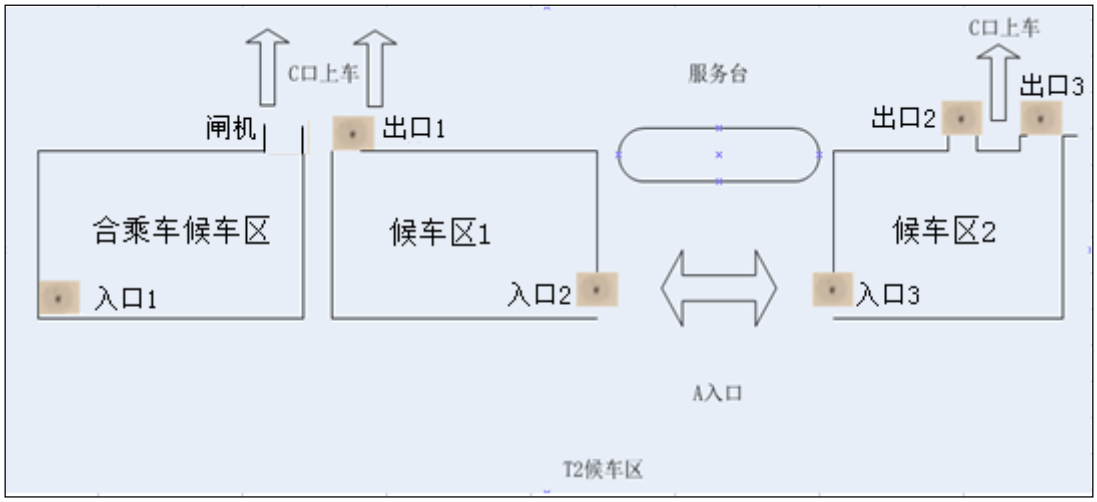


图 4-5 T2 航站楼旅客排队数量统计示意图

T3 航站楼中的客流计数摄像机的安装位置和覆盖区域如图 4-6 所示。

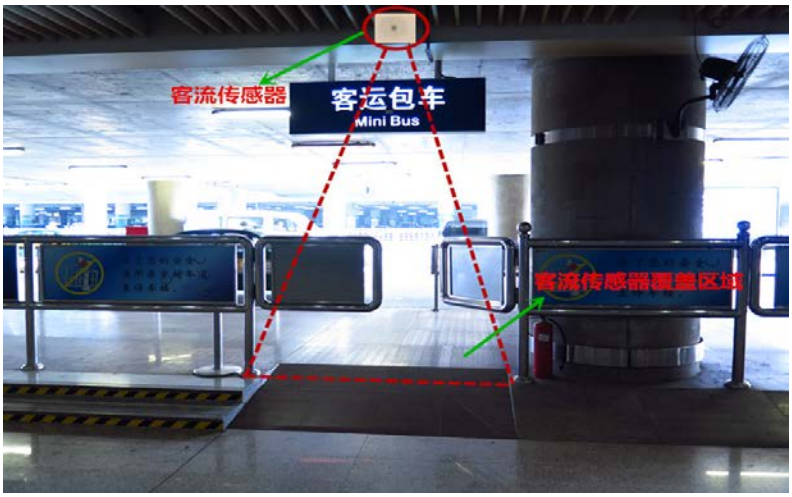


图 4-6 客流传感器安装位置和覆盖区域

## 4.2 客流数据统计

### 4.2.1 客流数据统计中出现的问题

如图 4-7 和 4-8 所示，X 轴代表一天的时间，统计的时间是从凌晨 4 点开始到第二天的凌晨 4 点，Y 轴是出租车候车区排队人数。

图 4-7 是使用客流计数摄像机统计的原始数据计算得到的机场出租车候车区排队人数，可以看到，会出现人数为负数的情况，统计人数出现了明显的错误。

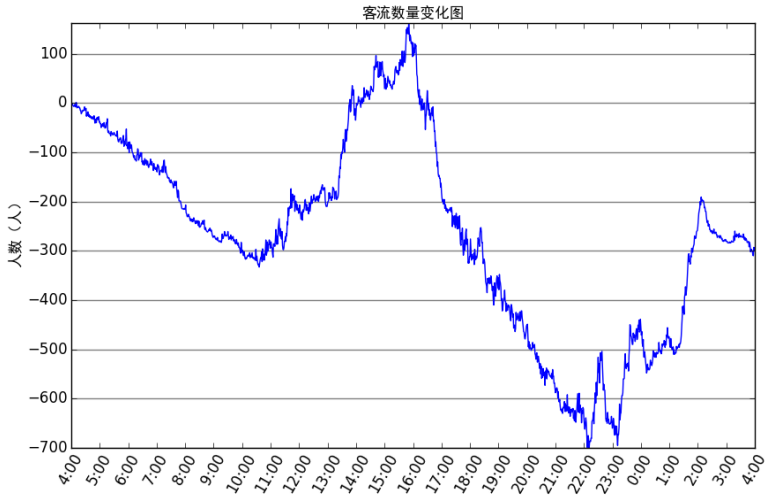


图 4-7 利用原始数据直接进行计算得到的客流数量变化图

图 4-8 代表修正错误数据之后计算得到的机场出租车候车区排队人数，在体现了原来数据的增长趋势的情况下，又修正了统计的误差。

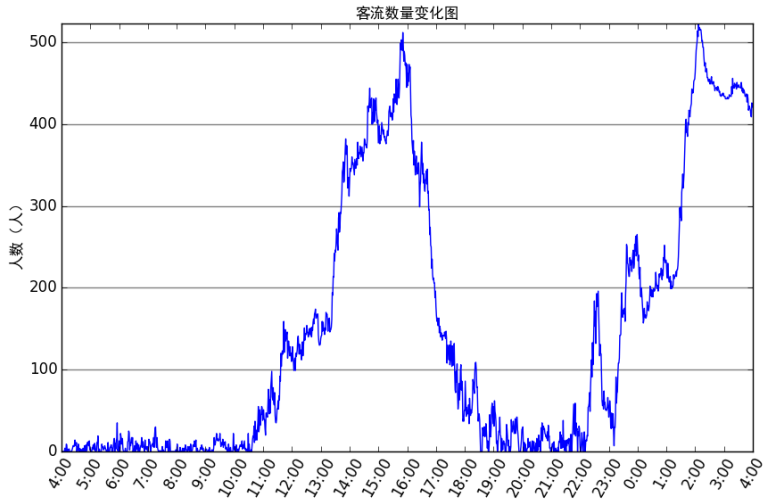


图 4-8 修正错误数据后计算得到的客流数量变化图

计算的方式见式（4-2）。

修正错误数据的效果就是调整图 4-7 中曲线的走势，使得其在受到各种工作人员的影响以及本身测量误差的情况下，能够一定程度上保持原来的走势，同时解决误差的累积，如图 4-8 所示。解决误差累积，就是在出现负数的情况下，及时调整回正值。

表 4-1 和表 4-2 是 T2 航站楼和 T3 航站楼检测到的进出人数差值。

表 4-1 T2 航站楼检测到的进出人数差值

| 序号 | 航站楼 | 检测时间 | 进出人数差值 |
|---|---|---|---|
| 1 | T2 | 2017-1-1 04:00 | 53 |
| 2 | T2 | 2017-1-1 05:00 | 62 |
| 3 | T2 | 2017-1-1 06:00 | 61 |
| 4 | T2 | 2017-1-1 07:00 | 61 |
| 5 | T2 | 2017-1-1 08:00 | 99 |
| 6 | T2 | 2017-1-1 09:00 | 173 |
| 7 | T2 | 2017-1-1 10:00 | 238 |
| 8 | T2 | 2017-1-1 11:00 | 286 |
| 9 | T2 | 2017-1-1 12:00 | 347 |
| 10 | T2 | 2017-1-1 13:00 | 384 |
| 11 | T2 | 2017-1-1 14:00 | 430 |
| 12 | T2 | 2017-1-1 15:00 | 496 |
| 13 | T2 | 2017-1-1 16:00 | 577 |
| 14 | T2 | 2017-1-1 17:00 | 613 |
| 15 | T2 | 2017-1-1 18:00 | 665 |
| 16 | T2 | 2017-1-1 19:00 | 742 |
| 17 | T2 | 2017-1-1 20:00 | 851 |
| 18 | T2 | 2017-1-1 21:00 | 965 |
| 19 | T2 | 2017-1-1 22:00 | 1023 |
| 20 | T2 | 2017-1-1 23:00 | 1023 |
| 21 | T2 | 2017-1-2 00:00 | 1099 |
| 22 | T2 | 2017-1-2 01:00 | 970 |
| 23 | T2 | 2017-1-2 02:00 | 881 |
| 24 | T2 | 2017-1-2 03:00 | 875 |

利用客流计数摄像机采集到的数据，计算一天 24 小时的总进站人数和总出站人数进行比较，每经过一个小时计算一次进出人数的差值，结果如表 4-1 和表 4-2 所示，分别是 T2 航站楼和 T3 航站楼 2017 年 1 月 1 日凌晨 4 点到第二天凌晨 4 点统计的进出人数差值。

容易发现，在原始数据的基础上进行客流人数的计算，会出现进入和出去人数相差过大的情况，这显然是不符合实际的。进入人数超过出去人数，说明旅客一直

滞留在出租车候车区中，出去人数超过进入人数则更不合理。随着时间的增加，客流计数摄像机统计人数和实际人数的差距有逐渐增大的趋势。

表 4-2 T3 航站楼检测到的进出人数差值

| 序号 | 航站楼 | 检测日期 | 进出人数差值 |
|---|---|---|---|
| 1 | T3 | 2017-1-1 04:00 | 1 |
| 2 | T3 | 2017-1-1 05:00 | -35 |
| 3 | T3 | 2017-1-1 06:00 | -86 |
| 4 | T3 | 2017-1-1 07:00 | -181 |
| 5 | T3 | 2017-1-1 08:00 | -278 |
| 6 | T3 | 2017-1-1 09:00 | -340 |
| 7 | T3 | 2017-1-1 10:00 | -462 |
| 8 | T3 | 2017-1-1 11:00 | -519 |
| 9 | T3 | 2017-1-1 12:00 | -637 |
| 10 | T3 | 2017-1-1 13:00 | -733 |
| 11 | T3 | 2017-1-1 14:00 | -856 |
| 12 | T3 | 2017-1-1 15:00 | -956 |
| 13 | T3 | 2017-1-1 16:00 | -1145 |
| 14 | T3 | 2017-1-1 17:00 | -1235 |
| 15 | T3 | 2017-1-1 18:00 | -948 |
| 16 | T3 | 2017-1-1 19:00 | -1161 |
| 17 | T3 | 2017-1-1 20:00 | -1469 |
| 18 | T3 | 2017-1-1 21:00 | -1685 |
| 19 | T3 | 2017-1-1 22:00 | -1825 |
| 20 | T3 | 2017-1-1 23:00 | -1841 |
| 21 | T3 | 2017-1-2 00:00 | -1822 |
| 22 | T3 | 2017-1-2 01:00 | -1646 |
| 23 | T3 | 2017-1-2 02:00 | -1743 |
| 24 | T3 | 2017-1-2 03:00 | -1796 |

### 4.2.2 客流数据统计出错的原因

客流数据统计出错可以分为多检和漏检两种，多检是因为客流计数摄像机将影子或大件物品等干扰项识别成旅客，使得这段时间统计的旅客数量偏多，漏检则

是没有将经过客流计数摄像机的旅客统计到数据中。

客流数据统计出错的原因主要有以下几个，以T3航站楼的录像为例进行说明：

（1）人流密度大、速度快，出现漏检现象。

如图 4-9 所示，统计到出去的人数有三个，用黄色标记，蓝色的就是漏检的旅客。

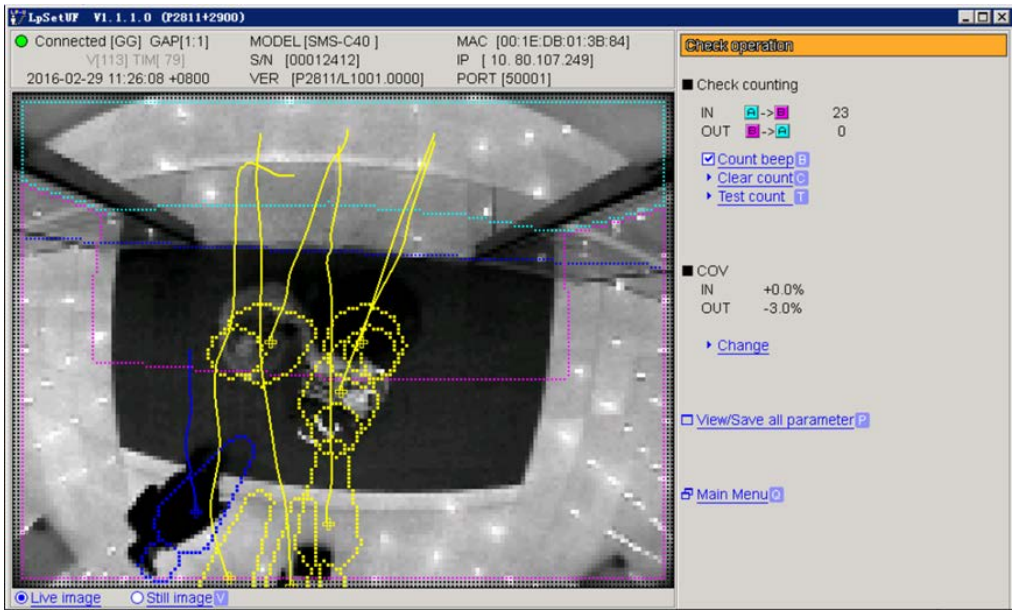


图 4-9 人流密度大、速度快，出现漏检现象

（2）在出口处等待的时间过长，导致漏检，如图 4-10 所示。

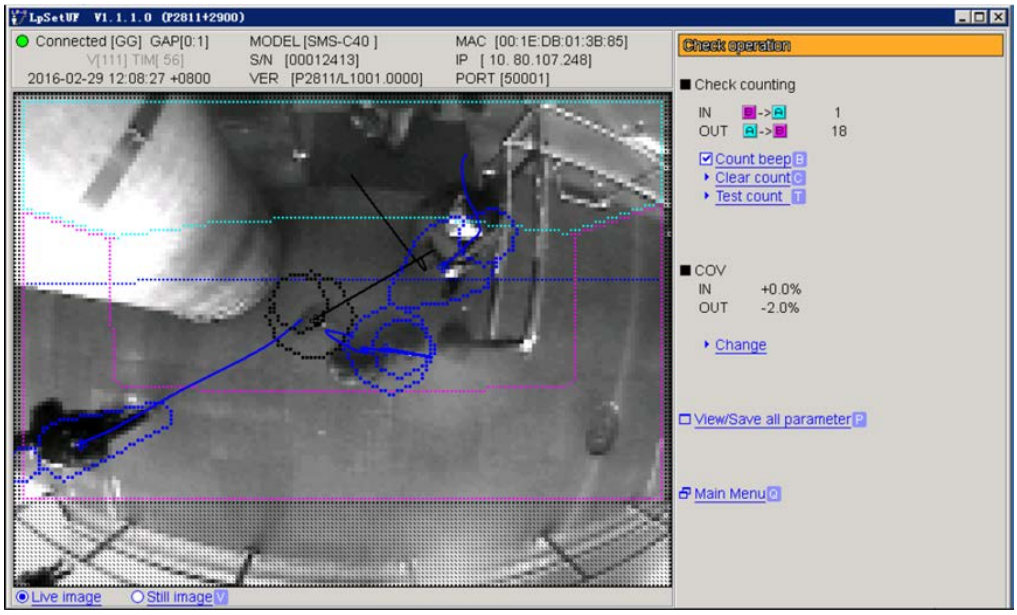


图 4-10 在出口处等待的时间过长，导致漏检

实际出去的旅客有三个，但只检测到两个，有一个旅客在检测区处停留超过 1

分钟，导致并没有被统计到客流数据中。

（3）较大的行李箱经过，出现多检现象。

如图 4-11 所示，在实际统计中，会出现将行李箱识别成旅客的现象，这种现象相对影响较小，因为入口和出口都会出现类似的错误，不会使得计算得到的统计值变成负值。

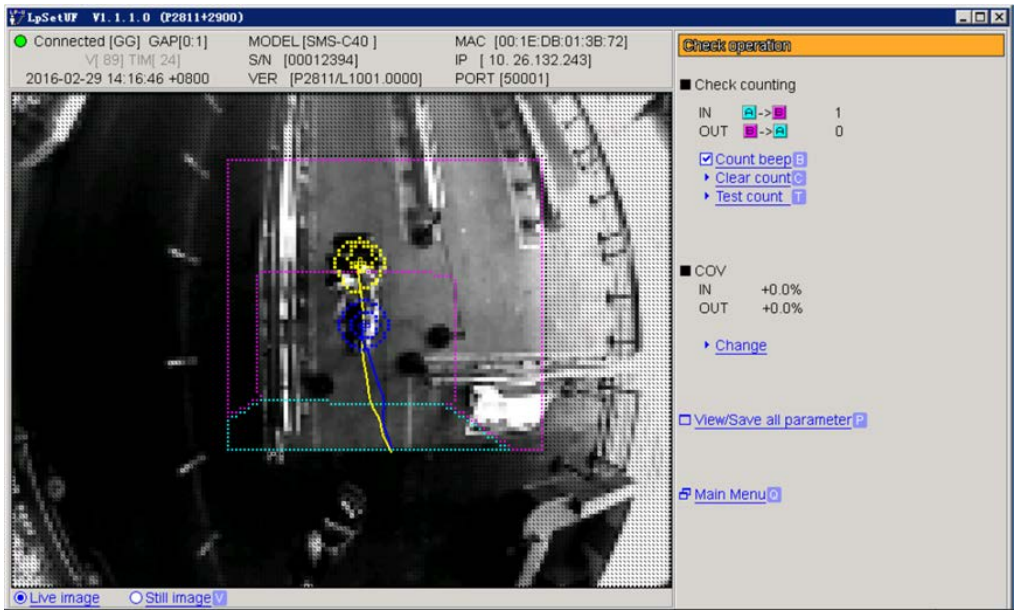


图 4-11 较大的行李箱经过，导致多检

（4）工作人员收集长串行李推车或在出口徘徊，会出现多检现象。如图 4-12 所示。

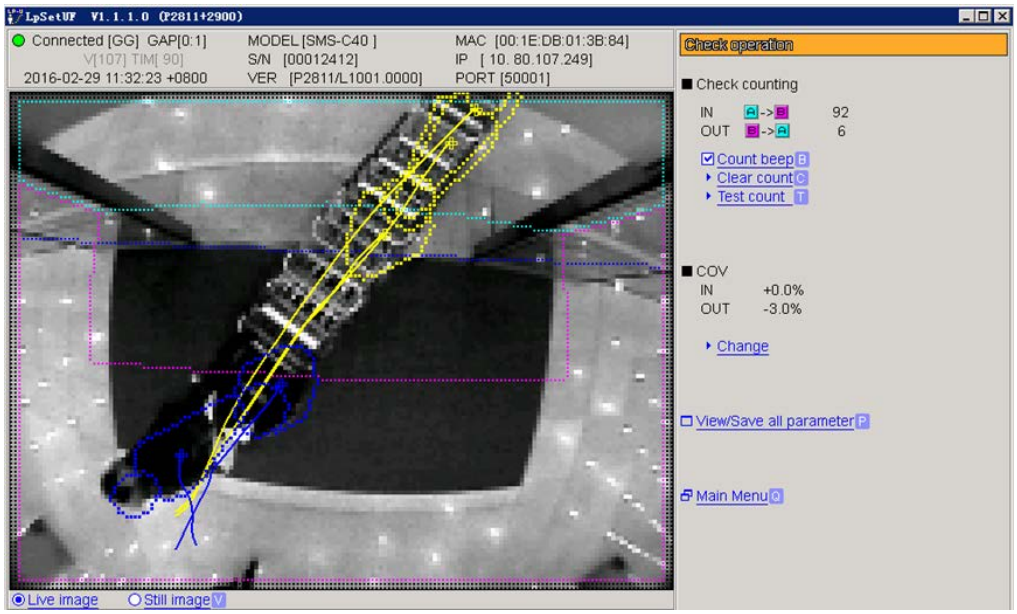


图 4-12 工作人员收集长串行李推车或在出口徘徊，导致多检

在图 4-12 中，实际只有这个工作人员出入，但是因为传感器将手推车识别成旅客，统计数据记录为 3 个人。

（5）如图 4-13 所示，大人领着小孩，检测出大人，小孩漏检。

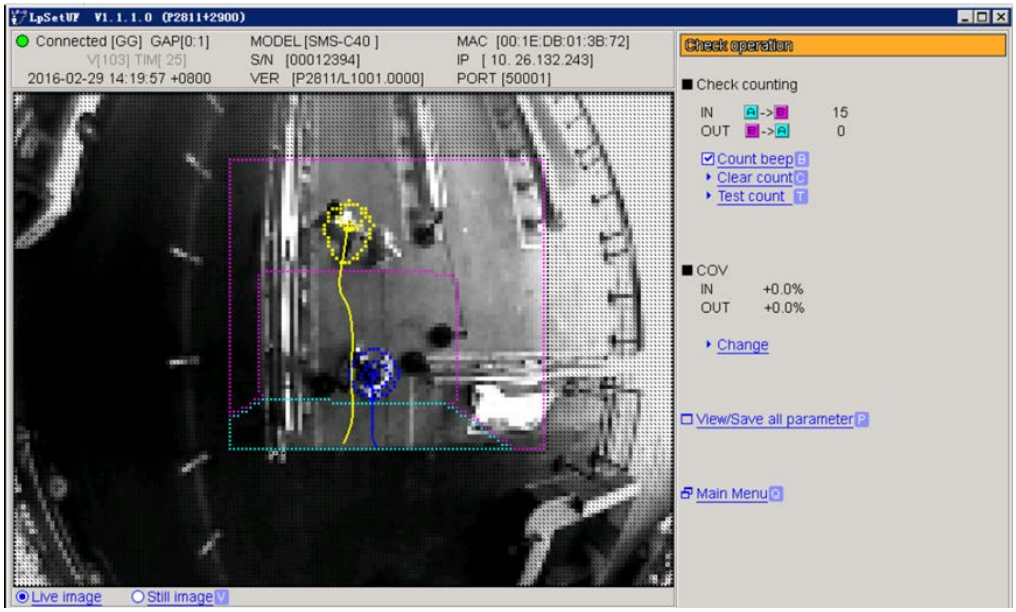


图 4-13 大人领着小孩，检测出大人，小孩漏检

（6）如图 4-14 所示，因为光照原因，客流计数摄像机将影子检测为一个人，出现多检现象。

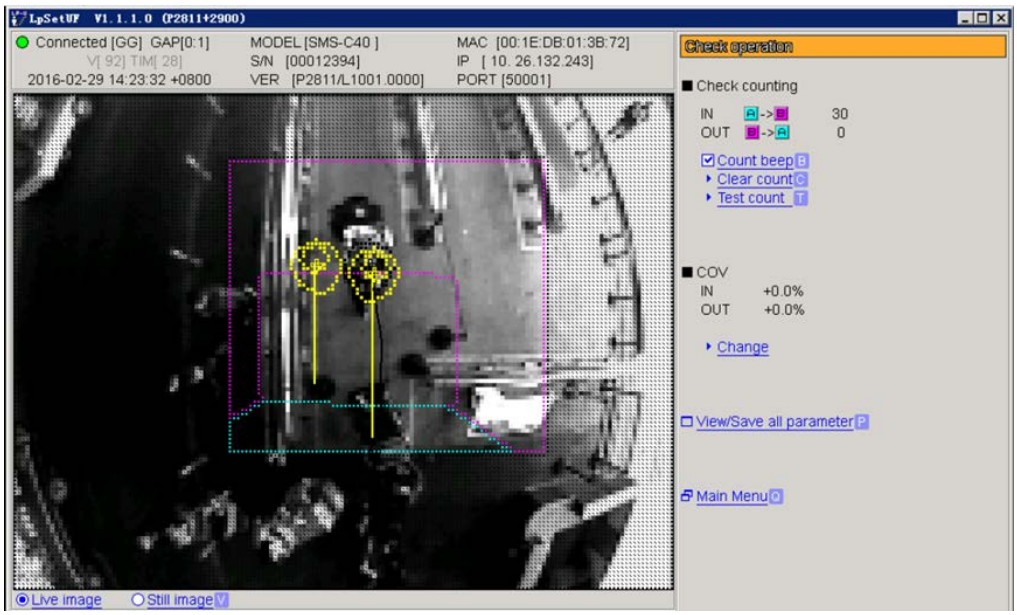


图 4-14 因为光照原因，将影子检测为一个人，出现多检现象

如图 4-15 所示，是 T2 航站楼的旅客上车情况，旅客从红线排队上车，机场工作人员从绿线口收集手推车，经过入口 1 或入口 2 将手推车送到楼上。在此过程

中，出口不计数，入口离开人数增加。其他工作人员也会从绿线口进出。


图 4-15 T2 航站楼错误检测示意图

在实际数据的分析中，T2 工作人员收集长串行李推车或徘徊导致的多检现象对乘客数量的统计造成最大的影响，但是这样的数据在样本中的数量不多，且多发生在凌晨这种乘客较少的时段，于是对这些统计数据进行剔除。

本文将通过多元数据协同修正的方式，对客流统计数据进行处理，使得通过客流传感器的数据计算得到的机场出租车候车区客流人数和实际统计得到的人数的误差减小，为接下来的机场出租车运力需求预测模型打下基础。

### 4.2.3 客流数据统计数据处理

#### 4.2.3.1 数据清洗

所谓数据清洗，即重新审查和校验客流计数统计设备采集到的数据，发现并纠正数据文件中可识别的错误。

通过在机场候车区的实地考察，设定一系列的人工规则，可以滤除一部分的错误数据，设定的人工规则如下：

（1）T3 航站楼出口 4 和出口 5 是比较特殊的点位，一般在排队人数不多的情况下不会轻易开启，如果排队人数不多，但是那两个口依旧有人数变动，那应该是工作人员出入导致；

（2）出口点位一般是等待的旅客离开候车区，很少出现旅客反向进入候车区的现象，主要是工作人员因指挥人流需要，在传感器周围徘徊造成误判，在进出口人数都很少的时候，从出口进入排队候车区的干扰数据可以剔除；

（3）对于入口点位，外出方向突然增大的数值可能是工作人员推动行李车列反向通过入口时误判，针对此类现象，统计出口和入口的变化人数的分布，以统计人数的$\mu \pm 2\sigma$作为限制，其中$\mu$为均值，$\sigma$为均方根，以排除异常的人数变化；

（4）在没有旅客进来的情况下，在一定的时间内，人数应该有相应程度的减少，不应该长时间不变；

（5）近期机场出租车候车区的统计均值具有比较高的参考价值，所以明显高于或低于均值水平的数据应该进行适当平滑。

实际得到的数据因为设备维修等原因，存在缺失的情况。对于个别时间段的数据缺失，利用前后数周同一天同一时刻的加权均值来进行填充，通常取近两到三周的数据，离缺失值时间越近，权重越大。这种填补策略能够最大程度结合机场客流量的周期性。公式为：

$$\widehat{y_t} = \sum \alpha_i y_i \tag{4-3}$$

其中$\alpha_i$是数据的权重，$y_i$是客流计数设备统计得到的实际数据，$\widehat{y_t}$是填补值。

对于大段时间段的数据丢失，考虑这些数据应该剔除。

另外，输入的时间周期中可能会有假期，但是预测的时间段没有。例如对于机场到港人数，春运就是一个重要的的影响事件，假期的流量相比正常的流量是异常值，对正常预测帮助不大，在训练的时候应该进行剔除。

#### 4.2.3.2 数据拟合修正

在实际数据的处理中，经过数据清洗之后的数据依然会在某些节点上出现悬崖式跳变的现象，本文解决的方法是对一段时间中的数据进行回归平滑，使得数据更接近实际观测值，拟合的方法是正交多项式回归。

正交多项式回归分析方法是一种利用最小二乘法实现拟合误差最小的时间序列特征表示方法。Fuchs 等人[21]通过把时间序列映射到多项式系数所形成的特征空间，选取部分反映时间序列整体特性的主要形态特征来近似表示时间序列，实现原时间序列的特征表示和数据降维。短期趋势可以通过正交多项式进行拟合，一个低阶的近似多项式足够描述均值、增长减少的趋势和相应的曲率，只需要几个值就足够描述一个时间窗口中的时间序列。

朱晓东等人的工作[28]讲述了为何使用正交多项式曲线拟合，是对以上论文的补充。自变量或者因变量两者中必须有一个量是没有误差的精确值才能应用最小

二乘法来处理。当自变量或因变量的误差较大时，在曲线拟合中就不能忽略，应该在曲线拟合的时候考虑进来。在机场项目中，由于乘客密度大、速度快以及工作人员对统计设备的干扰等原因，使得测量数据，即因变量出现误差；而数据中的时间测量，又可能因为设备断电的原因产生错误记录，不符合使用最小二乘法进行拟合的前提。在这种情况下，使用正交多项式曲线拟合是比较准确的。

本文使用正交多项式拟合解决悬崖式跳变的情况。

考虑到选取过多客流数据进行拟合会影响拟合的准确性，因为异常点的增多会使得拟合极大偏离客流变化的趋势，同时，过多拟合数据会影响算法的速度。本文通过对一段时间中的数据进行回归平滑，分段拟合，最后结合所有的拟合曲线得到客流的变化趋势，目的是用绝大部分正确的客流数据进行拟合去除少数的异常点。

假设实际统计得到的真实值是$y_i$，对$y_i$的变化曲线进行拟合。本文中选取的时间窗口长度为 1 个小时，也就是使用 60 个观测数据分段拟合。

分段拟合得到的客流变化曲线如图 4-16 所示。红色的曲线是分段拟合的曲线，标记为*的是客流计数摄像机数据计算得到的值。

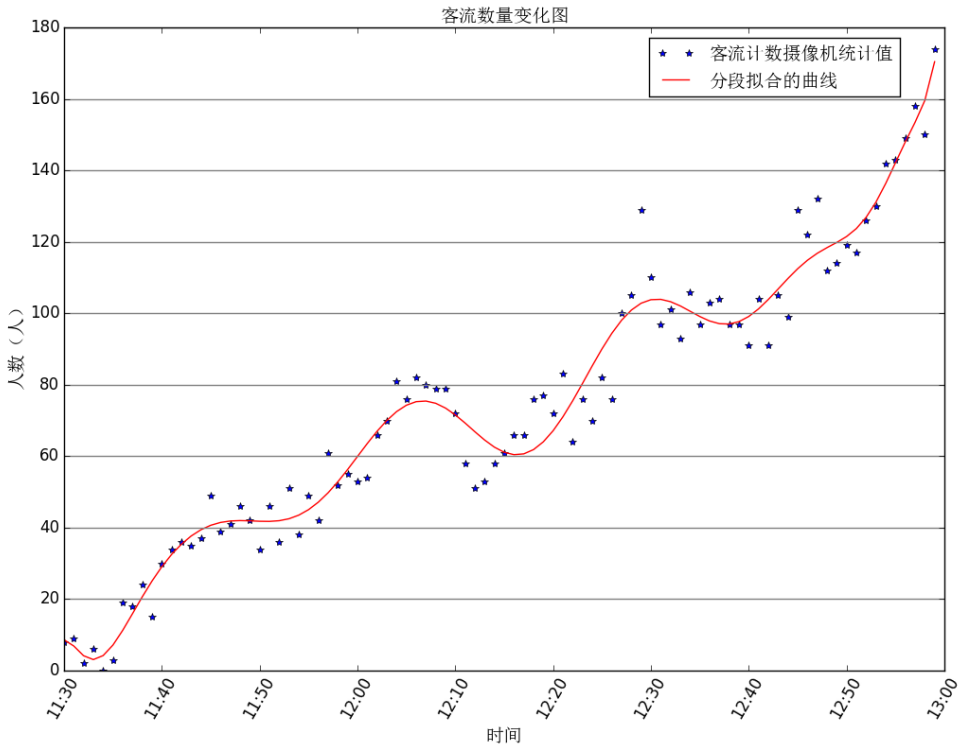


图 4-16 客流数量变化示意图

在图 4-16 中，偏离较大的点可能出现了悬崖式跳变，客流计算值和实际统计

值相差变大，通过拟合进行处理后，变化的幅度更加贴近机场出租车候车区的客流速度，可以使得计算值更贴近实际统计值。

#### 4.2.3.3 数据归一化

在实际的应用中，需要简化数据，将数据进行归一化处理，数据的归一化使得不同来源的数据具有相同量纲和相同数量级，让每个数据都有很规范的格式，即每一个输入源必须均值是 0，方差是 1，从而客观比较每个输入的作用。

具体就是，假设时间序列是$X_1, X_2 \ldots X_t$。

它们的均值是：

$$\mu = \frac{\sum_{i=1}^{t} X_i}{t} \tag{4-4}$$

它们的方差是：

$$\sigma^2 = \frac{\sum_{i=1}^{t}(X_i - \mu)^2}{t} \tag{4-5}$$

归一化后得到的新变量是：

$$Y_i = \frac{(X_i - \mu)}{\sigma} \tag{4-6}$$

#### 4.2.3.4 数据相似性度量

在机场出租车客流检测区，在一定时间段内，人们的活动反映在首都机场客流及运力调度数据上，以一天时间为周期，不同天但对应的同一个时刻的出租车客流量具有相似的趋势；以一周时间为周期，不同星期但对应的同一个星期几的出租车客流量具有相似的趋势；以一年为周期，不同年但对应的同一个月份或节假日的出租车客流量具有相似的趋势。

首都机场客流及运力调度数据的相似度和重复度定义如公式(4-7)所示：

$$\mathrm{MSD}\ (X, Y) = \frac{1}{n_x \times n_y} \sum_{\forall x \in X, \forall y \in Y} s(x, y, T) \tag{4-7}$$

其中，

$$\mathrm{s(x, y, T)} = e^{-\frac{\lambda}{m_T}\sum_{k=1}^{m_T}|x_k - y_k|} \tag{4-8}$$

其中，$\lambda$是大于 0 的调节系数，用于调节相似度在[0,1]之间的分布。$T$ 为采样时间，$m_T$为采样点数。

计算得到每周各天的客流数据的相似度如表 4-3 所示：

对于客流数据来说，可以根据数据相似性，以一周七天对数据进行划分，同时去除特殊的节假日的数据。

表 4-3 客流数据相似度

| 时间 | 周一 | 周二 | 周三 | 周四 | 周五 | 周六 | 周日 |
|---|---|---|---|---|---|---|---|
| 周一 | 0.8959 | 0.7858 | 0.7832 | 0.7820 | 0.7810 | 0.7806 | 0.7954 |
| 周二 | | 0.8748 | 0.7722 | 0.7716 | 0.7744 | 0.7746 | 0.7919 |
| 周三 | | | 0.8631 | 0.7557 | 0.7696 | 0.7644 | 0.7459 |
| 周四 | | | | 0.8787 | 0.7665 | 0.7560 | 0.7697 |
| 周五 | | | | | 0.8714 | 0.7714 | 0.7607 |
| 周六 | | | | | | 0.8685 | 0.7792 |
| 周日 | | | | | | | 0.8815 |

## 4.3 客流数据分析

### 4.3.1 出租车候车区客流量时段分析

如图 4-17 和图 4-18 所示，是出租车候车区客流量随时间的变化趋势。

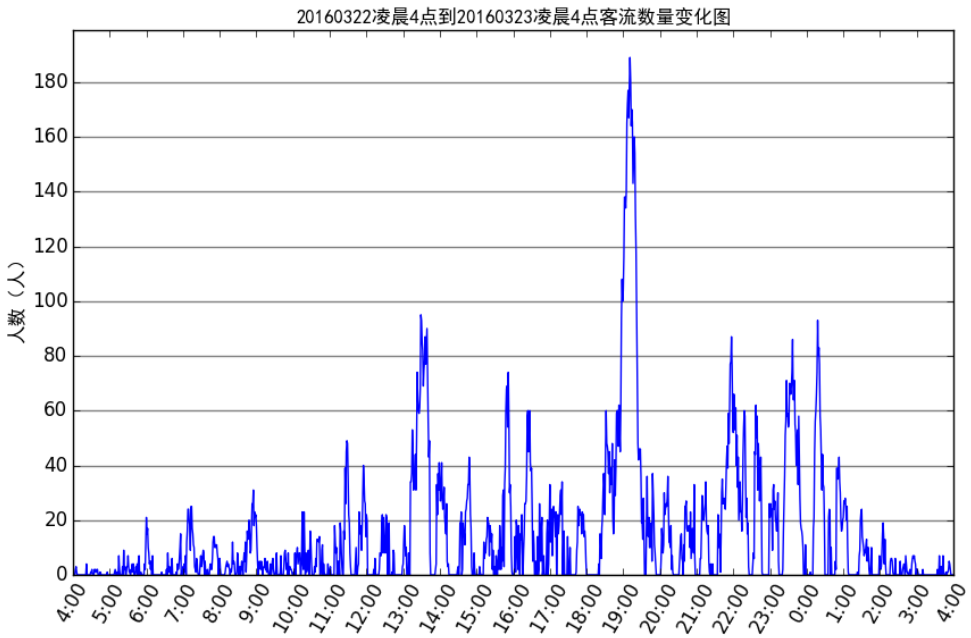


图 4-17 20160322 凌晨 4 点到 20160323 凌晨 4 点客流数量变化图

对于每一天，都可以生成类似的趋势图。这样对一天中人数的高峰情况，以及每天的不同变化有直观的理解。由图中可以看出，虽然高峰的人数有所不同，但每天出租车候车区客流量的大致趋势是一致的，在 8 点到 9 点的时间段会出现一个小高峰，13:00、16:00、19：00、22:00 的时候是机场出租车候车区客流量的高峰期。

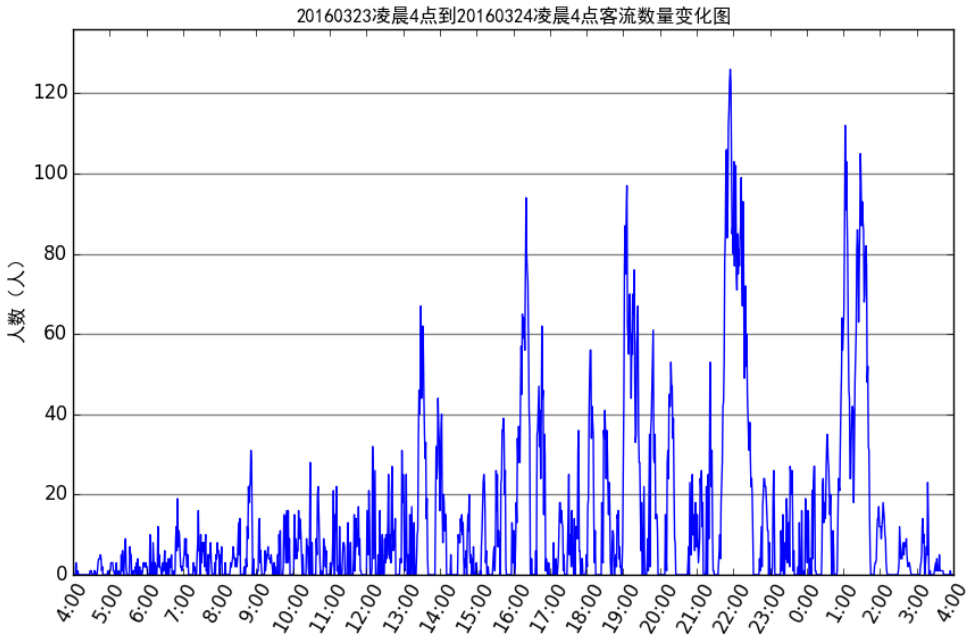


图 4-18 20160323 凌晨 4 点到 20160324 凌晨 4 点客流数量变化图

如图 4-19 所示是各个时间点统计的机场出租车客流数量平均值变化图。

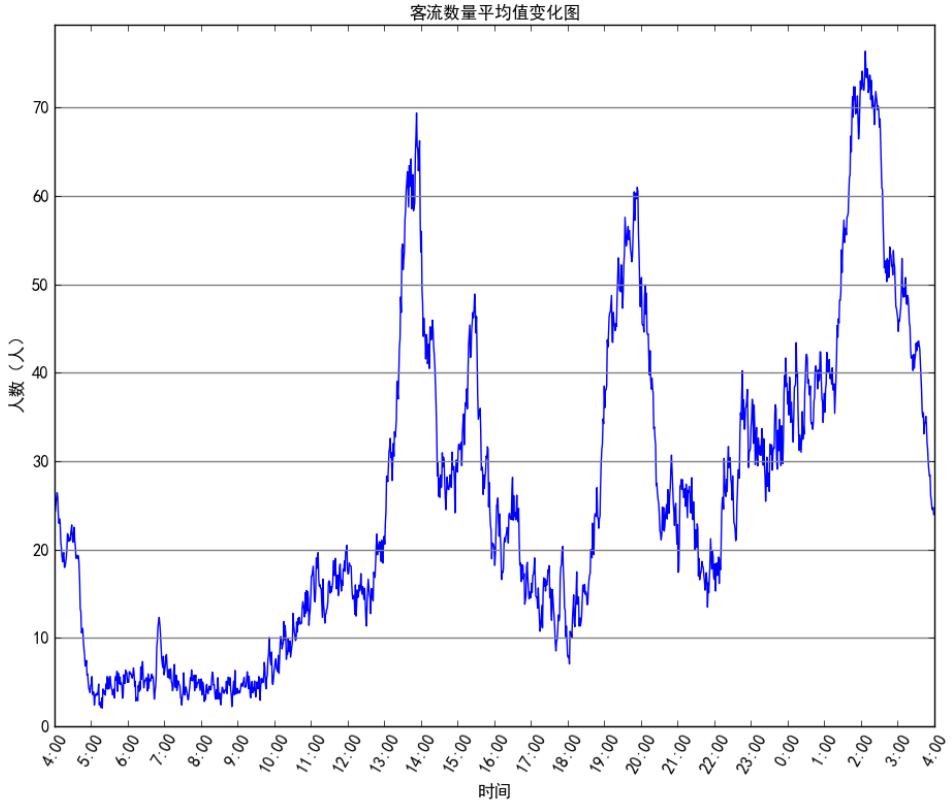


图 4-19 机场出租车候车区客流数量平均值变化图

可以看到，高峰期和上述提到的几个时间段是比较吻合的，说明这几个时间段的客流人数出现高峰不是一两天的偶然现象，而是因为机场每天的航班排班基本稳定和用户在机场内的的行走模型比较固定导致的，所以历史均值是一个良好的估计值。

航班计划是影响旅客动向的主要因素，在航班降落之后，出租车候车区会比较容易汇聚较多的乘客。因此，机场出租车候车区的旅客数量可以结合机场旅客到港情况进行分析。如图 4-20 所示，机场出租车客流高峰和到港航班数量的高峰相对应，结合到港航班数量的情况，可以有效对机场运力需求做出预测。

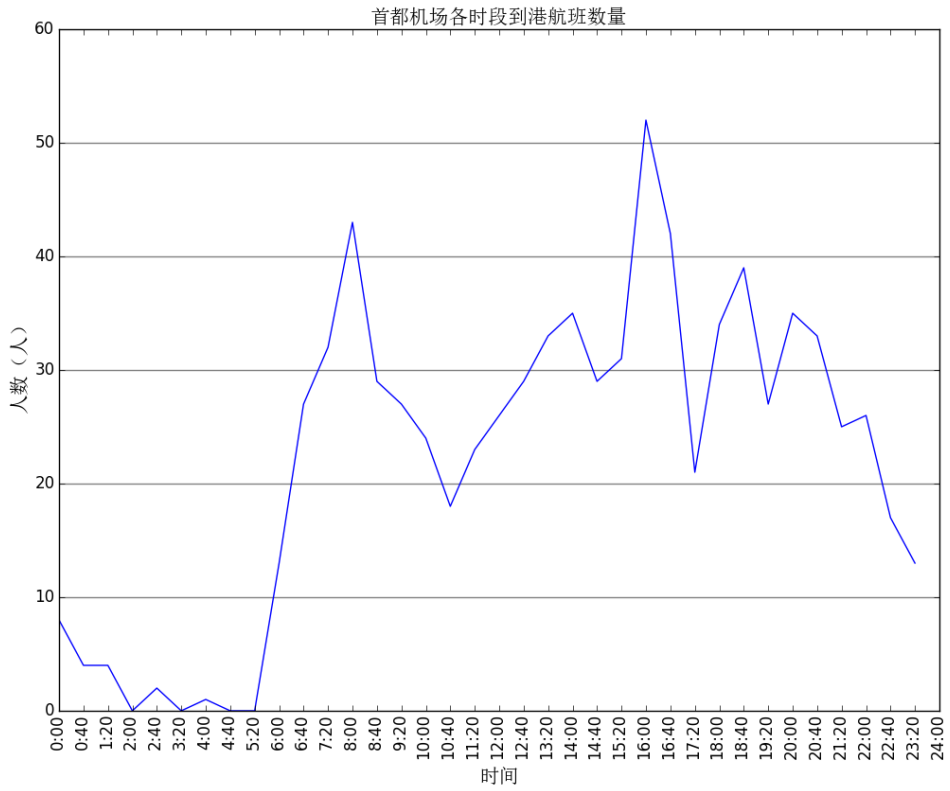


图 4-20 首都机场各时段到港航班数量

如图 4-21 所示是首都机场各时段旅客进港情况，结合首都机场各时段到港航班数量，可以分析各个时段的旅客进港情况。

在图 4-21 中，从 01:00 到 06:00 凌晨这个时间段，很少甚至没有旅客到达机场；06:00 开始，由图 4-20 可以看出，到港航班数量开始增多，在 11:00 到 12:00 之间到港旅客数出现第一个高峰期，每半小时就有超过 3000 名旅客抵港，这和之前的到港航班高峰是相对应的，出租车候车区客流量在这段时间也达到高峰；接下来到港航班减少，到港旅客数量略有下降，这段时间出租车候车区的旅客数也相应

有所下降；到下午 16:00 左右到港航班数量开始增加，到港人数和出租车候车区的旅客数达到第二个高峰期，之后开始慢慢下降；22:00 的时候到港航班、到港人数和出租车候车区的旅客数又有开始上升的趋势，并在 23:30 的时候达到又一个高峰期。可以看出，首都机场候车区客流数量和到港航班数量、到港人数是紧密相关的。

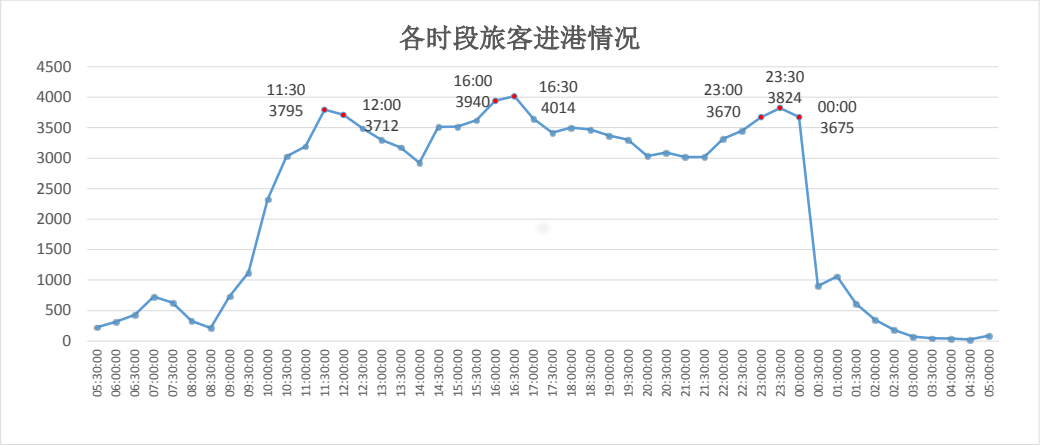


图 4-21 首都机场各时段旅客进港情况

### 4.3.2 出租车候车区客流量分布

出租车候车区客流量分布如图 4-22 所示，图中显示了对应各个出租车候车区客流人数，对应的出现次数。

关注数据分布的原因在于，希望通过训练集训练得到的模型能够合理地应用在测试集上。另外，可以通过数据的分布去除明显错误的数据，例如凌晨两点之后，出入口一分钟超过 20 个旅客是不合理的，可能是工作人员推车导致的错误统计。虽然不能纠正所有错误，但是能够去除大部分的异常数据。

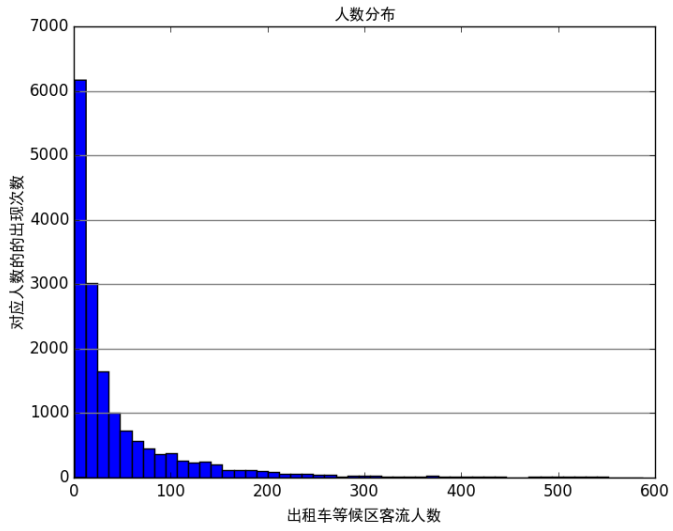


图 4-22 出租车候车区客流量分布

### 4.3.3 出租车候车区客流序列频谱分析

时间序列也可以用频谱来表示。从众多的实践应用中来看，离散傅里叶变换要求序列具有平稳性、周期性。经过一次差分去除趋势性之后的机场候车区客流数据是一个稳定的时间序列，可以将其视为一个离散的信号，通过离散傅里叶变换将时间序列从时间域上映射到频域空间。对序列的抽样频率为 1000Hz，频谱图如图 4-23 所示。

从频谱图中可以看出，序列中的主要频率成分是 0.7Hz，存在周期性特征，以 1424 个数据为一个周期。

对于输入的序列，一天有 1440 个数据，即客流量的周期大概为 1 天，这也对上面的猜测进行了验证。常识上认为，每周也会有对应的变化。机场客流时间序列数据波动的动态变化过程本身具有一定的周期性、季节性、阶段性。

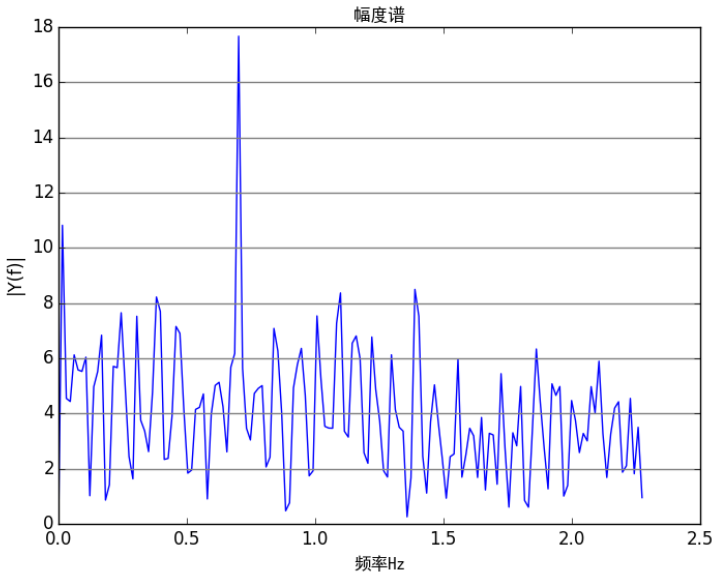


图 4-23 机场出租车候车区客流时间序列的频谱

### 4.3.4 旅客到港的出站时间分布

Kim Wonkyu 等人[29]讨论了一个飞机到港之后，旅客出站人数关于出站时间的概率分布，本文将机场的到港信息和这个概率分布结合起来，对每个时刻的到港人数进行评估，并且这篇论文得到的期望值和之后论文中通过实际统计求出的平均值相差不大，也从侧面应证了这个模型的可靠程度。SW Yoon 等人[8]则讨论了构建一个乘客下机时间和取行李行为关系的模型，其中的一些数据和张泉峰的数据[30]对于本文后续根据机场排班估计乘客的数量有很大的帮助。

根据大部分旅客步行速度相相同，特别慢和特别快的旅客相对较少的特点，可

以假设旅客从飞机落地起到达出站口所需时间为正态分布。

通过实际调查统计分析，得到旅客到港后各环节所需的时间值。如图 4-24 所示。

根据公式(4-9)和公式(4-10)

$$\mu = \sum_{旅客各个出港的过程} \mu_{相应的过程} \tag{4-9}$$

$$\sigma^2 = \sum_{旅客各个出港的过程} \left(\frac{maxTime_{相应的过程} - minTime_{相应的过程}}{6}\right)^2 \tag{4-10}$$

令不取行李乘客所需时间分布变量为$X_1$，均值为$\mu_1$，方差为${\sigma_1}^2$的正态分布函数，概率密度函数为$f_1(x)$。由以上数据计算可以得到，$\mu_1 = 41$，$\sigma_1 = 3.50$。

令取行李乘客所需时间分布变量为$X_2$，均值为$\mu_2$，方差为${\sigma_2}^2$的正态分布函数，概率密度函数为$f_2(x)$。由以上数据计算可以得到，$\mu_2 = 54$，$\sigma_2 = 4.15$。

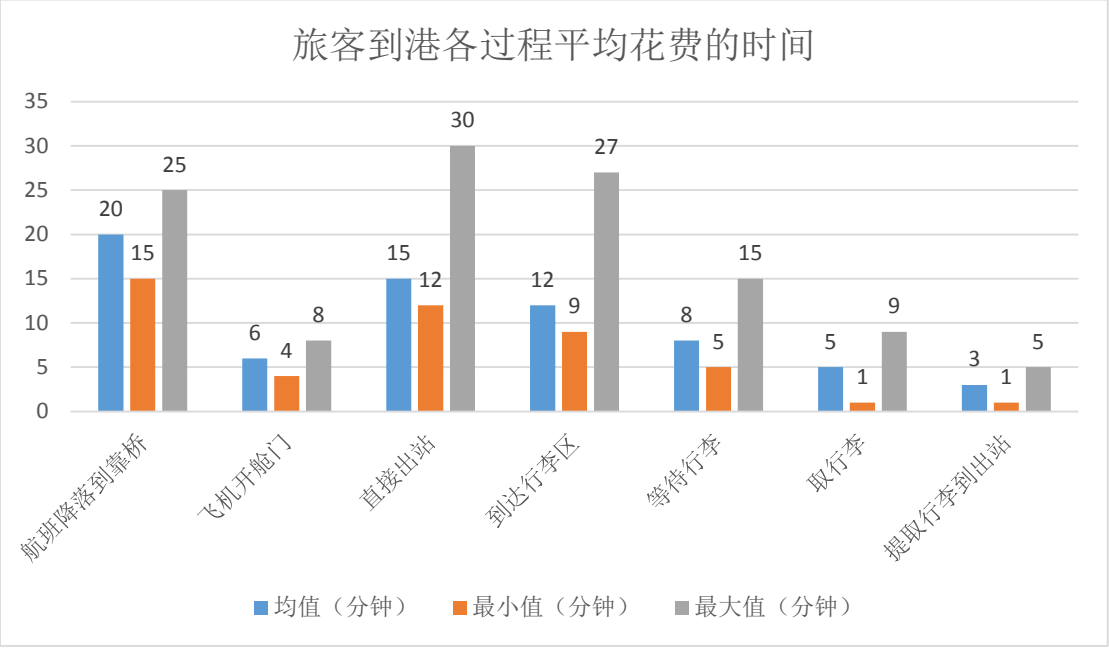


图 4-24 旅客到港各过程平均花费的时间

经过统计得到，不取行李的乘客和取行李的乘客的比例约为 7:3，则所有旅客到达出站口所需时间 X 的概率密度函数为：

$$\mathrm{f(x)} = 0.7f_1(x) + 0.3f_2(x) \tag{4-11}$$

通过统计安检历史数据，得到每次航班的平均值作为预测航班的载客量。将航班到达时间和人数分布结合起来，可以得到到港人数的时间序列曲线。

如图 4-25 所示，虚线是每个航班旅客到达出站口的时间分布，实线是所有航班旅客到达出站口的时间分布，通过计算当前时刻前一个小时所有实际到达的航

班和后一个小时计划到达航班的旅客走到出站口时间的分布情况，则总旅客到达是每个航班旅客到达之和。为了更直观展示到港航班和到港人数的变化，本文选择从 0 点到 4 点这段旅客人数较少的时间段。

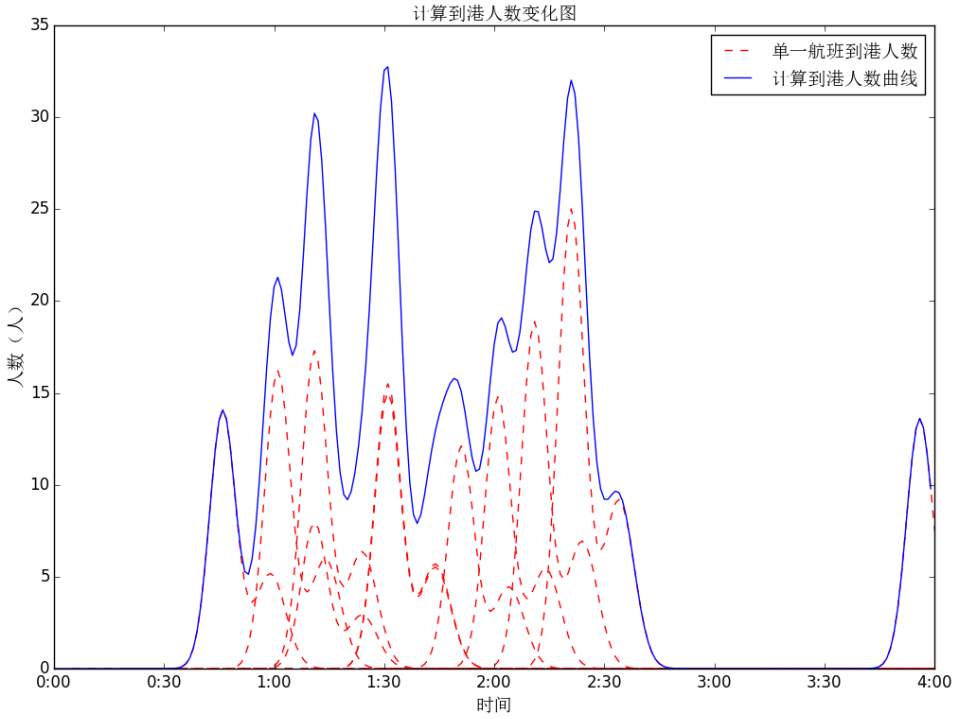


图 4-25 计算到港人数变化图

### 4.3.5 天气数据分析

考虑到飞机飞行的各个阶段会受到天气状况很大的影响，天气数据是对机场客流重要的影响因素。

可以将常见的天气类型按照对机场客流影响程度的大小分为以下几类，如图 4-26 所示。

统计显示，B 类天气、C 类天气的日平均到港旅客数是比较接近的，A 类天气下的日平均到港旅客数要比 B、C 两类天气高出 35%。D 类天气影响较大，不做参考。天气因素是运力需求预测中一个重要的影响因素。

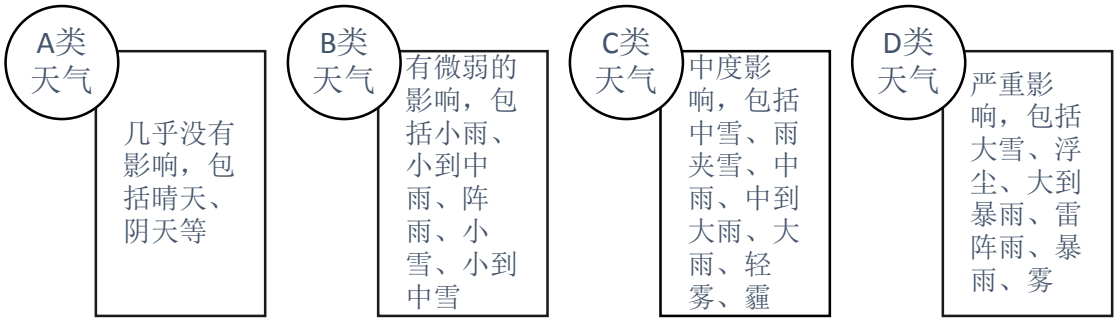


图 4-26 机场天气类型

如图 4-27 所示是通过 wunderground 网站采集得到的天气信息。记录了包含气温，露点，湿度，气压，能见度，风速，瞬时风速，降水量，天气状况等信息。

```
Port,Date,Time,Temp,Bodytemp,Dew,Humidity,Pressure,Visibility,Wind_dir,Wind_speed,Gust_speed,Event,Condition
ZBAA,2017-01-01,12:00 AM,-6.0,-,-7.0,93%,1026,1.5,East,3.6,-,,Mist
ZBAA,2017-01-01,12:30 AM,-6.0,-9.4,-7.0,93%,1026,1.4,East,7.2,-,,Mist
ZBAA,2017-01-01,1:00 AM,-6.0,-9.4,-7.0,93%,1026,0.9,SSE,7.2,-,Fog,Light Freezing Fog
ZBAA,2017-01-01,1:30 AM,-6.0,-,-6.0,100%,1027,0.8,Variable,3.6,-,Fog,Light Freezing Fog
```

图 4-27 天气信息采集示意图

wunderground 网站提供了世界各地在机场附近检测到的气象信息。历史气象信息的采样间隔为 30 分钟。

北京首都国际机场 2017 年 1 月 1 日到 2017 年 1 月 2 日气温条件的时间序列分布如图 4-28 所示。

对于天气情况这个信息，需要将各种天气情况数值化，进行 one hot 编码。one hot 编码能让特征之间的距离计算更加合理。在这个场景下，首都机场的天气情况有 Mist，Light Freezing Fog，Heavy Fog，Light Snow 等状态，如果不使用 one-hot 编码，表示分别是$x_{\text{Mist}}=0, x_{\text{Light Freezing Fog}}=1, x_{\text{Heavy Fog}}=2, x_{\text{Light Snow}}=3\cdots$，则距离$d\left(x_{\text{Mist}}, x_{\text{Light Freezing Fog}}\right)=1$，$d\left(x_{\text{Mist}}, x_{\text{Light Snow}}\right)=3$，这样的特征距离是不合理的，因为$x_{\text{Light Freezing Fog}}$和$x_{\text{Light Snow}}$是两个平等的状态。如果使用 one-hot 编码，即 $x_{\text{Mist}}=(1,0,0,0\ldots)$，$x_{\text{Light Freezing Fog}}=(0,1,0,0\ldots)$，$x_{\text{Heavy Fog}}=(0,0,1,0\ldots)$，$d\left(x_{\text{Mist}}, d_{\text{Light Freezing Fog}}\right)=d\left(x_{\text{Mist}}, x_{\text{Light Snow}}\right)=\sqrt{2}$，即每两个状态的距离是一样的，显得更合理。

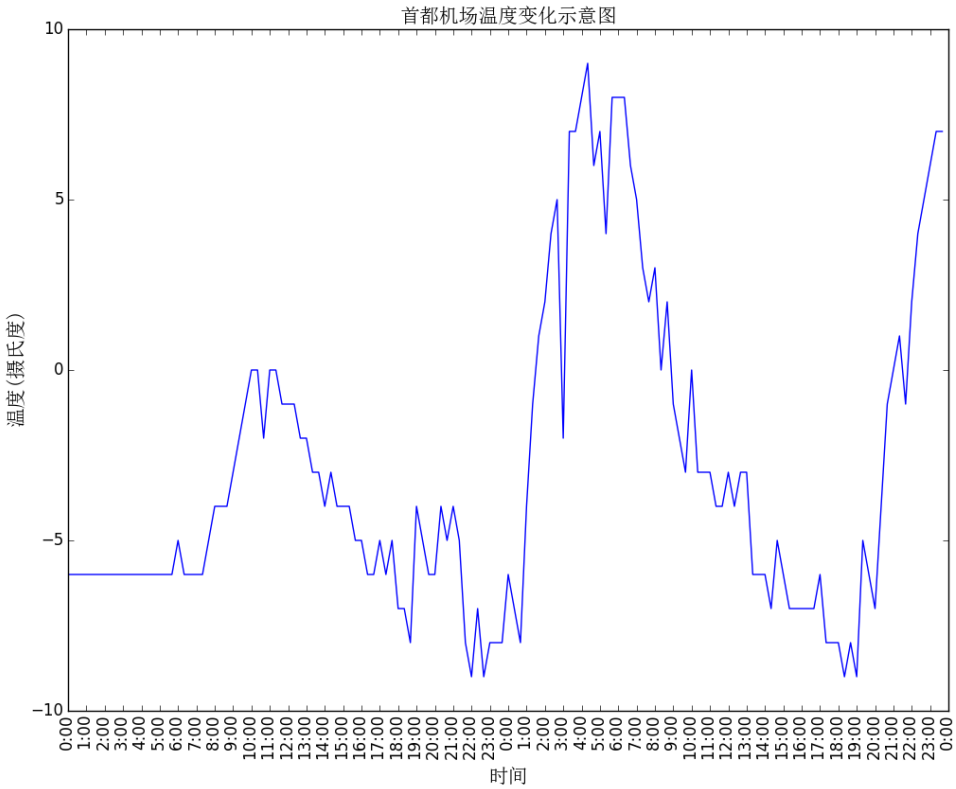


图 4-28 北京首都机场 2017 年 1 月 1 日到 2017 年 1 月 2 日气温变化

## 4.4 数据库建立

系统数据库中存储着大量的客流到港、航班安排、出租车运力等重要的信息，需要高效稳定的数据库系统为机场出租车运力需求预测提供数据支持。在本课题中，为了实现流量分析和建立预测方法，需要整理、加工每个时间间隔里机场公共区以及出租车站台上客流计数统计设备收集到的大量数据，产生新的数据，以便反映机场出租车客流的规律和特征之间的内在联系。

建立数据库需要满足：

- 方便地定义数据和操作数据，满足数据容量的需求。
- 在预期的时间和相应的数据规模下，满足更新和查询的性能要求。
- 系统涉及旅客数据和机场周边交通等大量敏感的交通流数据，需要保证系统的安全性，方便机场工作人员、旅客等多用户对数据的并发使用。在故障恢复方面，机场的系统采用框架结构，在不改变原程序的情况下，可以方便的进行维护和升级。

数据库中还需要方便数据的二次加工，例如：

- 对于 1000w 的客流统计数据，根据采集的摄像头的唯一标识号，可以提取

到该摄像头的所有客流记录，并且按照时间进行降序排列，可以得到该入口/出口一天的客流变化情况。在此基础上，在数据库中提取对应日期的航班信息进行分析处理，和对应的客流信息进行关联。

- 天气因素数据需要与机场天气预报系统协调，通过爬虫方式获取机场的天气实况和预报信息加入数据库中并和客流数据航班数据一一对应。
- 与机场出租车管理部门合作，在机场已有的人工发牌统计的基础上，开发并向机场出租车管理值班岗亭提供定时记录出租车发/收牌数量的软件，软件将记录的信息实时录入数据库。

在海量数据的基础上，进行数据的二次梳理，对于后面模型的建立和分析有着重要的作用。

数据库相关的研究[31]中深入分析了 Oracle 数据库体系结构，利用具体示例讨论了一些重要的数据库主题，同时分析了数据库中的物理结构，并介绍采用哪些技术能最优地使用这些物理结构，是本文数据库操作的重要参考。

## 4.5 本章小结

首都机场出租车候车区客流数据的清洗和建立运力需求预测模型是相辅相成的，本章主要对机场出租车候车区客流数据进行清洗工作。

首先，本章描述了客流数据错误的现状，同时通过观察数据采集的录像，找到客流数据统计出错的原因，并提出了数据清洗的方案。数据清洗的方式是根据客流连续性和各个时间段的客流人数分布特点，去除干扰值，并用针对性的正交多项式拟合进行平滑，解决悬崖式跳变的问题。最后通过统计学的方式和频谱分析的方法，挖掘机场出租车候车区客流数据的特点，并分析了影响机场客流量的航班排班和天气因素。本章为下文的机场运力需求预测算法打下了坚实的基础。

# 第五章 基于机场出租车客流特点的趋势性季节性去除算法

在前面的章节中，详细地介绍了运力需求预测以及相应的模型结构，并以首都机场出租车候车区为研究背景进行分析。本章实现了 ARIMA 模型和 LSTM 模型，在这些模型预测的基础上，考虑通过更好地灵活拟合趋势成分和更精确地建模季节性，并从原始客流数据中将这些成分移除，最后再添加到预测中，达到更准确的预测结果。这里提出基于机场出租车客流特点的趋势性季节性去除算法（Trend and Seasonality Removal Algorithm Based on Passenger Flow Characteristics of Airport Taxi），简称 TSR 算法，对机场出租车运力需求进行预测，并在本章的最后使用首都机场候车区的真实数据对各个模型的预测结果进行验证。

## 5.1 ARIMA 模型算法实现

### 5.1.1 平稳性检验

时间序列建模建立在大数定理和中心极限定理的前提下，即满足样本同分布的前提，也就是需要满足时间序列的平稳性。

平稳性检验的方法如下：

（1）时间序列图检验

从机场候车区客流时序图中，可以比较直观地看到随着时间的变动数据的整体变化趋势。平稳的时间序列波动的幅度很小，可以看出没有明显的周期性变化和趋势，即这个序列在一个常数值附近随机波动。通过观察时序图可以比较直观地检验时间序列的平稳性，但不够精确。

（2）单位根检验

单位根检验是一种常用的检验方法，原假设为序列具有单位根，即非平稳，要检验当前的机场候车区客流数据平稳，需要在给定的置信水平上显著拒绝原假设。

（3）自相关图检验

平稳时间序列一般具备短期相关性。如果当前的机场候车区客流数据平稳，则随着延迟期数的增加，序列的自相关系数会快速衰减为 0。

### 5.1.2 时间序列的差分 *d*

ARIMA 模型对时间序列的要求是平稳型。在机场客流的场景中，采样得到的机场客流数据是不平稳的，每年随着客运的发展，伴随着系统性的上升趋势，并且

具有明显的季节性和周期性的波动。对于这样一个非平稳的时间序列，首先要做的即是做时间序列的差分，直到得到一个平稳时间序列，之后再利用 ARMA 模型模拟随机过程。

对于机场出租车候车区客流运输数据，进行 ADF 检验，原假设是存在单位根，即客流运输数据是非平稳的。支持原假设的检验值为 0.144453。说明在 85%的置信区间可以拒绝原假设，即，现有的客流数据在 85%的置信区间是平稳的。当进行一次差分之后，支持原假设的检验值约等于 0，说明进行一次差分之后，可以得到一个平稳的时间序列。这与直观的判断较为吻合，机场的客流数据呈现一个大致以天为周期的情况，不是一个平稳序列，做了一次差分之后，可以去除这个趋势性。

### 5.1.3 选择合适的 $p$ 和 $q$

经过一次差分之后的机场出租车候车区客流数据是一个平稳的时间序列，之后需选择合适的 ARMA 模型，即通过检查平稳时间序列的自相关图和偏自相关图，为 ARMA 模型选择合适的 $p$ 和 $q$。

各个观测记录的自相关性的计算方式如下：

$$r_k = \frac{\sum_{t=1}^{n-p}(y_t - \bar{y})(y_{t+k} - \bar{y})}{\sum_{t=1}^{n}(y_t - \bar{y})^2} \tag{5-1}$$

其中$\bar{y}$表示观测值的平均值，$n$ 表示观测的时间序列的长度，$k$ 表示之后的阶数。

偏相关系数的计算方式如下：

$$p_{kk} = \begin{cases} r_1, & k = 1 \\ \dfrac{r_k - \sum_{j=1}^{k-1} p_{k-1,j} \cdot r_{k-j}}{1 - \sum_{j=1}^{k-1} p_{k-1,j} \cdot r_j} & k = 2,3,\dots \end{cases} \tag{5-2}$$

其中，

$$p_{k,j} = p_{k-1,j} - p_{kk} p_{k-1,k-j} (j = 1,2,\dots,k-1) \tag{5-3}$$

计算经过一次差分的客流运输数据的自相关系数和偏相关系数，如图 5-1 所示，自相关图显示在滞后 3 阶时的自相关系数超出了置信边界；偏相关图显示在滞后 1 至 3 阶时的偏自相关系数超出了置信边界，从滞后 3 阶之后偏自相关系数值缩小至 0，则有 *ARMA*(0,3)，*ARMA*(3,0)，*ARMA*(3,3)等模型可供选择。

为了更准确地确定 $p$ 和 $q$ 的具体值，采用准则函数定阶法确定 $p$ 和 $q$。

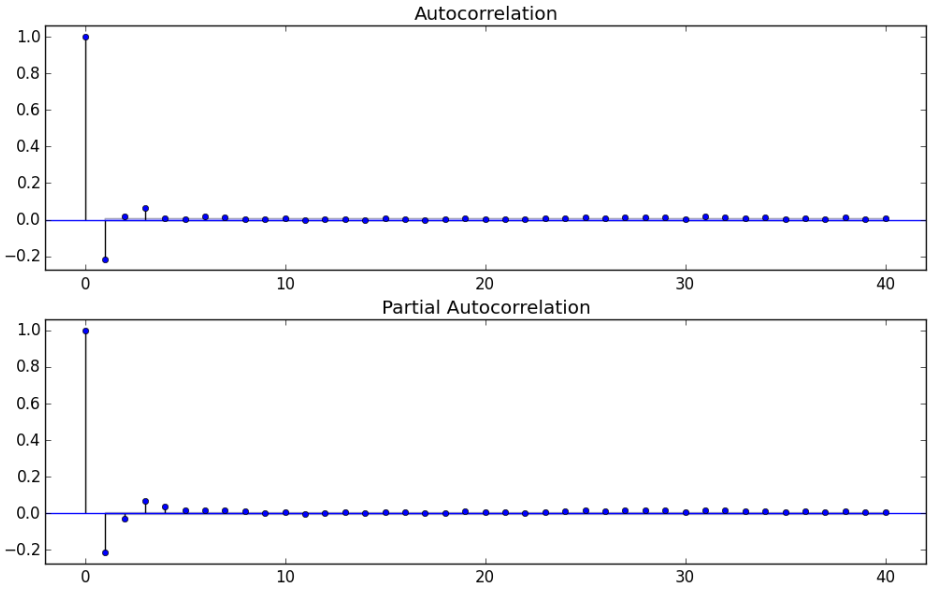


图 5-1 观测数据的自相关图和偏相关图

- AIC 法则，即赤池信息准则的定义是：

$$AIC = 2k - 2\,ln(L) \tag{5-4}$$

- BIC 法则，即贝叶斯信息准则的定义是：

$$BIC = ln(n) * k - 2\,ln(L) \tag{5-5}$$

- HQIC 法则，即汉南和奎因准则的定义是：

$$HQIC = ln(ln(n)) * k - 2ln(L) \tag{5-6}$$

在上述公式中，$k$ 是模型参数个数，$L$ 是似然函数，$n$ 为样本数量。

构造这些统计量所遵循的统计思想是一致的，就是在使得拟合残差尽可能小的同时，使得模型参数个数尽可能少。对于机场出租车候车区客流模型来说，参数数量过多，模型精度过高，会造成模型复杂度过高，也会出现过拟合的情况，无法推广到更普遍的场景。本课题使用首都机场出租车候车区客流运输数据验证模型的有效性，推广能力下降，也就难以应用到其他运输工具的需求预测中。客流模型需要很好地解释客流数据，同时尽量包含最少的参数，鼓励数据拟合的优良性。

经过上述准则计算之后，优先考虑的客流模型应该是值较小的那一个。

如表 5-1 所示，可以看到 *ARMA*(3,3)的 AIC，BIC，HQIC 均最小，因此是最佳模型。

表 5-1 模型使用不同准则的评价值

| 选择模型 / 准则 | AIC | BIC | HQIC |
|---|---|---|---|
| *ARMA*(1,1) | 795247.475 | 795285.781 | 795259.067 |
| *ARMA*(1,3) | 794576.525 | 794633.984 | 794593.913 |
| *ARMA*(3,1) | 794498.192 | 794555.650 | 794515.579 |
| *ARMA*(3,3) | 794355.057 | 794431.669 | 794378.241 |

### 5.1.4 模型检验

ARIMA 是一种线性相关模型，模型的残差被假定为高斯白噪声序列，在拟合残差项存在自相关的情况下，系数估计量将有相当大的方差，模型的预测功能失效，所以需要对回归分析中的拟合残差项是否存在自相关进行检验。

#### 5.1.4.1 Durbin-Watson 检验

德宾-沃森检验，简称 D-W 检验，可以用来检测回归分析中的残差项是否存在自相关。

原假设 $H_0$ 表示时间序列是平稳的，随机误差项不存在一阶序列相关。

假设残差项$U_t$可以表述为：

$$U_t = \rho U_{t-1} + \varepsilon \tag{5-7}$$

其中$\rho$是自相关系数。$\rho$的值介于-1 和 1 之间。当$\rho = 0$，表明没有自相关性。

D-W 统计量的计算公式如下：

$$DW = \frac{\sum (U_t - U_{t-1})^2}{\sum {U_t}^2} \approx 2(1-\rho) \tag{5-8}$$

所以$0 \le DW \le 4$。

$DW = 0$ 即$\rho = 1$，即存在正自相关性。

$DW = 4$ 即$\rho = -1$，即存在负自相关性。

$DW = 2$ 即$\rho = 0$，即不存在一阶自相关性。

通过蒙特卡罗模拟得到 D-W 统计量的概率分布，当 $DW$ 值在给定的显著水平下接近于 0 或 4 时，则存在自相关性，而在给定的显著水平下接近于 2 时，则不存在一阶自相关性，即可以根据临界值的位置对时间序列是否平稳进行检验。检验结果是 1.998，说明不存在自相关性。

#### 5.1.4.2 Ljung-Box 检验

计算 ACF 和 PCAF 并观察其图像和 Ljung-Box 检验都可以对滞后相关进行检验。两者的不同在于，计算 ACF 和 PCAF 只考虑是否存在某一特定滞后阶数的相

关，而 Ljung-Box 检验考虑在某一滞后数之前时间序列都是随机和独立的，从而判断序列总体的相关性是否存在，是一种统计检验。

Ljung-Box 检验的原假设是相关系数为零，即没有相关性。

如果检验概率大于特定临界值，则一个或多个滞后的自相关可能显著不等于零。

检验的结果如表 5-2 所示，取显著性水平为 0.05，最后一列前十二行的检验概率，即滞后 1~12 阶的时候，小于给定的显著性水平 0.05，说明相关系数与 0 没有显著差异，即 ARIMA 模型的残差为白噪声序列，ARIMA 模型是一个适合机场客流候车区客流运输数据的模型。

表 5-2 Ljung-Box 检验结果

| lag | AC | Q | Prob(>Q) |
|---|---|---|---|
| 1.0 | 0.000863 | 0.079287 | 7.782660e-01 |
| 2.0 | -0.000224 | 0.084612 | 9.585765e-01 |
| 3.0 | 0.006366 | 4.403688 | 2.210436e-01 |
| 4.0 | 0.002399 | 5.017017 | 2.855559e-01 |
| 5.0 | 0.011367 | 18.785343 | 2.107330e-03 |
| 6.0 | 0.012638 | 35.806965 | 3.005271e-06 |
| 7.0 | 0.003164 | 36.873856 | 4.955902e-06 |
| 8.0 | -0.004905 | 39.437684 | 4.076453e-06 |
| 9.0 | -0.006651 | 44.151887 | 1.323318e-06 |
| 10.0 | -0.005045 | 46.864791 | 9.992702e-07 |
| 11.0 | -0.008552 | 75.796927 | 2.596590e-11 |
| 12.0 | -0.006218 | 79.916874 | 1.143738e-11 |

### 5.1.5 ARIMA 时间序列模型预测

将进行了一次差分之后的机场出租车候车区客流数据输入 *ARMA*(3,3)模型中进行训练，并得到预测值。

如图 5-2 所示，得到的值是差分序列的预测值，所以还需要将预测值恢复到没有差分的原始区间中，如图 5-3 所示。

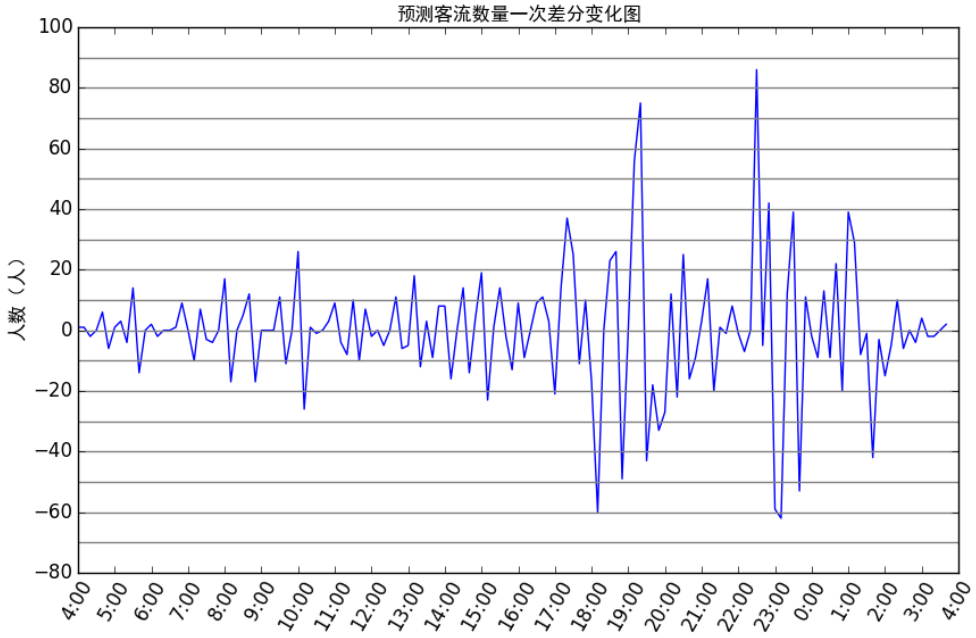


图 5-2 预测客流数量一次差分变化图

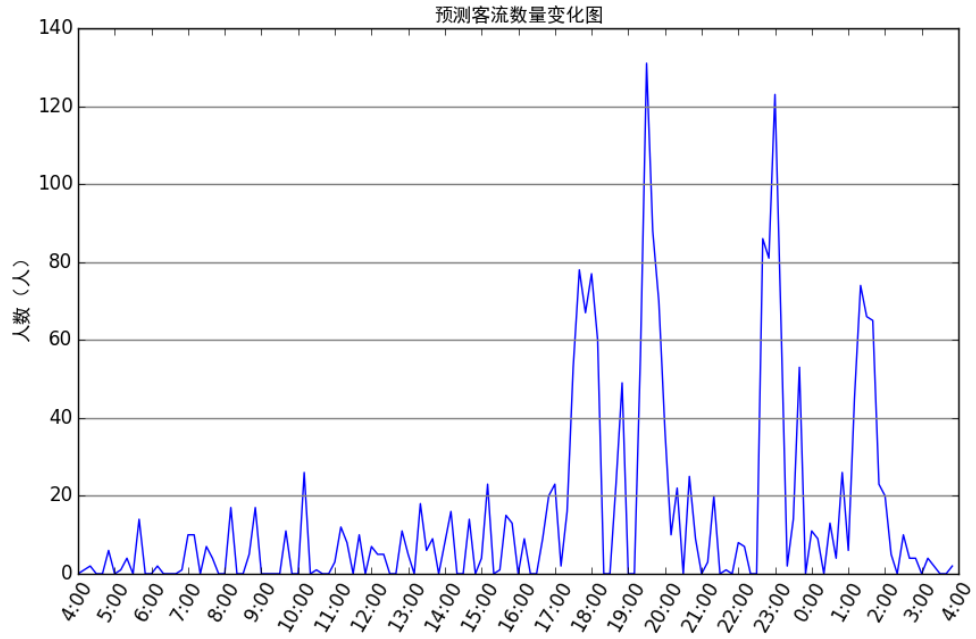


图 5-3 预测客流数量变化图

## 5.2 LSTM 模型实现

### 5.2.1 模型转换

机场出租车运力需求的预测可以转化为一个监督学习问题，输入是客流数据

和相应的天气和航班信息的组合，输出是当前时间的预测值，通过多变量输入数据拟合 LSTM。

现在固定一个一定长度的窗口对这些数据进行采用，作为输入的一部分。即用$x_{t-d}, x_{t-d+1}, \ldots, x_{t-1}$进行模型训练，用$x_t$对预测的结果进行验证。使用这些窗口后，就可以通过最小化损失函数，如均方根误差来训练神经网络。

如图 5-4 所示，输入和输出验证窗口都是以单个增量进行滑动对训练数据和测试数据进行处理。

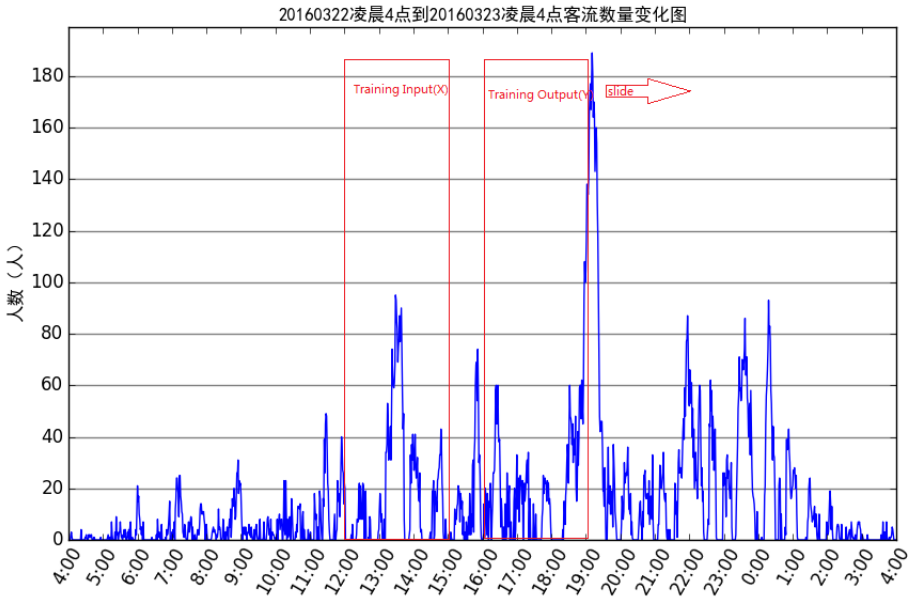


图 5-4 数据处理示意图

整理后的数据形式如图 5-5 所示：

| data1 | data2 | data3 | … | dataT | weather | passenger_flow | time |
|---|---|---|---|---|---|---|---|
| 6 | 0 | 0 | | 0 | 0001 | 135 | 0 |
| 0 | 0 | 0 | | 3 | 0001 | 137 | 1 |
| 0 | 0 | 0 | | 0 | 0001 | 128 | 2 |
| 0 | 0 | 2 | | 1 | 0001 | 121 | 3 |
| 0 | 2 | 0 | | 0 | 0001 | 170 | 4 |
| 2 | 0 | 0 | | 1 | 0001 | 171 | 5 |
| 0 | 0 | 3 | | 1 | 0001 | 175 | 6 |
| 0 | 3 | 0 | | 0 | 0001 | 155 | 7 |
| 3 | 0 | 0 | | 7 | 0001 | 230 | 8 |
| 0 | 0 | 9 | | 3 | 0001 | 189 | 9 |
| 0 | 9 | 0 | | 0 | 0001 | 186 | 10 |
| 9 | 0 | 0 | | 7 | 0001 | 197 | 11 |

图 5-5 输入 LSTM 模型的数据示意图

窗口长度是 *T*，即采用 *t*-*T*…*t*-2，*t*-1 次的出租车客流人数作为输入，配合天气情况，当前首都机场旅客进港数量和对应的一天的时间等信息，进行模型训练，用

*t* 次的出租车客流人数对模型的预测结果进行验证。*t*=0 时，对应的是 4:00，*t*=1 对应的是 4:10，依次类推，每隔 144 个数据，又会从 0 开始编号，从而将一天的时间映射为相应的整数。

将数据集中提取的训练样本按照 4:1 的比例随机拆分，其中的四份用作训练和评估，另外的一份用来对模型进行纠正和调参，防止过度拟合。

### 5.2.2 实验平台

实验的平台是Google的开源项目平台Tensorflow，一个开放软件库。Tensorflow数值计算的方式是数据流图，流图中的节点表示数值操作，边线表示进行数值通信的多维数据数组，这些数组成为 tensors。

对于当前的 LSTM 模型，利用 Tensorflow，能够有效简化搭建过程，同时能够最大化利用 Tensorflow 内部的代码优化机制。

## 5.3 TSR 算法实现

结合之前分析得到的首都机场候车区客流特点以及相关的时间序列研究[32]，本文提出并实现了 TSR 算法。

### 5.3.1 模型趋势性拟合

机场出租车运力需求预测模型的趋势性，是指候车区客流运输数据在较少的时间表示出来的总趋势，即受益于整体经济的增长及居民消费水平的提升，表现出来的不断递增的基本趋势。

在进行拟合之前，对数据进行预处理，使用 Logistic 增长模型对机场客流人数的趋势项进行拟合。Logistic 增长模型，又称为自我抑制性方程。

Logistic 具有以下的形式：

$$N_{\mathrm{PassengerFlow}} = \frac{C}{1 + e^{-k(t-m)}} \tag{5-9}$$

对于当前的模型，$C$ 代表了机场的承载能力，$k$ 是增长速率，$m$ 是一个补偿值。

使用 Logistic 增长模型对模型进行拟合的原因是，这个模型比较符合机场客流人数的客观规律。在预测的过程中需要考虑到机场的实际情况，机场的出租车候车区的最大容量为 400 人，预测应该有一个上限值。这称为模型的承载能力。

Logistic 增长模型通常分为 5 个时期，能够和机场客流日增长，年度增长的几个阶段相对应。

对于机场客流的日增长：

- 开始期：属于机场工作人员开始工作的一段时间，这段时间也是航班安排

比较少的一段时间，这段时间的客流增长很缓慢。

● 加速期：随着航班安排的增多，机场工作有序展开，这段时间的客流增长开始加快。

● 转折期：当客流数达到机场承载能力的一半时，这时候的客流增长速度最快。

● 减速期：客流数超过机场承载能力的一半之后，这段时间航班安排开始减少，客流增长速度开始变慢。

● 饱和期：最后没有航班安排，客流数饱和。

对于机场客流的年度增长：

● 开始期：机场配套的轨道交通工程等还未完善，这时候的客流增长还较为缓慢。

● 加速器：受益于整体经济的增长及居民消费水平的提升，能够维持较快的发展速度。

● 转折期：到机场客流吞吐量达到设计产能的一半时，这时候每年的客流吞吐量增长最快。

● 减速期：机场设施的能力逐渐无法满足发展需要，航线航班的增长变得缓慢，客流增长速度也随之放缓。

● 饱和期：机场达到设计产能的瓶颈，增长近乎停滞，旅客吞吐量也处于低个位数增长的稳态。

根据中国民用航空局的数据，2015 年到 2017 年的旅客运输量如图 5-6 所示，X 轴是统计的时间，Y 轴是相应的旅客运输量。

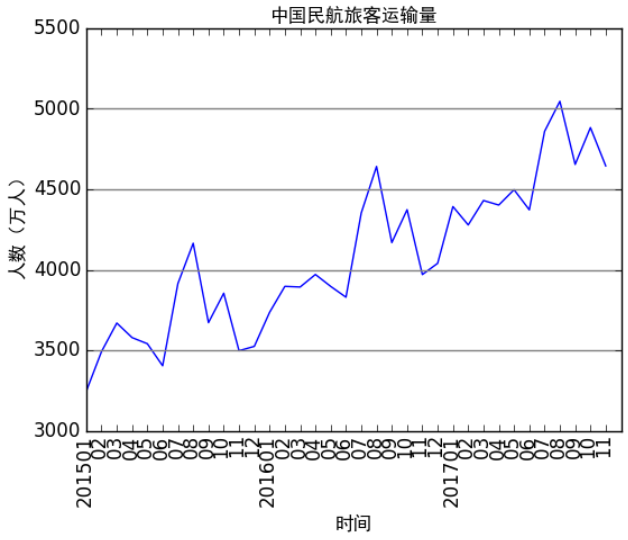


图 5-6 中国民航 2015 年到 2017 年旅客运输量

考虑模型的趋势性，能够有效提高模型的预测能力。

如图 5-7 所示，使用 2010 年到 2017 年的客流数据对模型的趋势性进行拟合。由于采集的数据数量有限，拟合的年度趋势也比较简单。

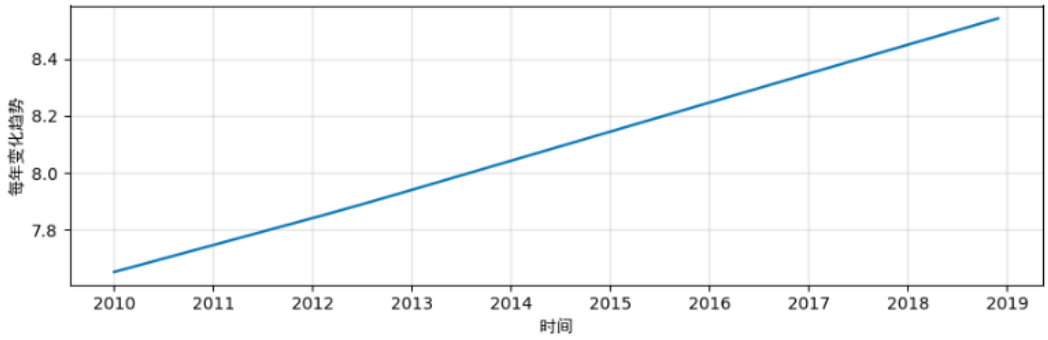


图 5-7 拟合的年度趋势

### 5.3.2 模型季节性和节日趋势拟合

模型的季节性和节日性趋势，是指和人的生产生活规律相关，具有一定季节和节日周期性变动的趋势。对于每年来说，夏季是旅行的最佳季节，因此夏季的到港旅客人数最多。另外，法定节假日中因为出行游客增多，到港旅客激增，长假日之间有相似的曲线趋势。

在预测的过程中可以将一些季节性和节日趋势进行单独处理。对于节日性趋势，比如春节和春节前夕，可以围绕这些特殊的时段单独建模。

拟合的方法是利用傅里叶系数[32]：

$$s(t) = \sum_{n=1}^{N}\left(a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin\left(\frac{2\pi nt}{P}\right)\right) \tag{5-10}$$

训练的时候，考虑每一个采样的时间，于是有：

$$s(t) = a_1\sum_{t}\cos\left(\frac{2\pi nt}{P}\right) + \cdots + a_n\sum_{t}\cos\left(\frac{2\pi nt}{P}\right) + \cdots + b_1\sum_{t}\sin\left(\frac{2\pi nt}{P}\right) + \cdots + b_n\sum_{t}\sin\left(\frac{2\pi nt}{P}\right) \tag{5-11}$$

可以把$a_1, \ldots, a_n, b_1, \ldots, b_n$作为属性放在模型中进行求解。

对于季节性趋势，$P$ 取 365.25，于是周期为：$\frac{2\pi}{\frac{2\pi}{P}} = P = 365.25$

对于节日和一周的趋势，$P$ 取 7，于是周期为：$\frac{2\pi}{\frac{2\pi}{P}} = P = 7$

将季节性趋势和节日的趋势像上式一样展开时，其实相当于一个低通滤波器，只保留了 $N$ 个低频成分，当 $N$ 增大，保留的频率成分越多，拟合越贴近数据，但可能出现过拟合的情况，实际测试的时候，对于季节性趋势，取 $N$=10，对于节日

和一周的趋势，取 $N$=3。

如图 5-8 所示，检测出来的明显的人数上趋势就是春节前夕的回乡高峰导致的。

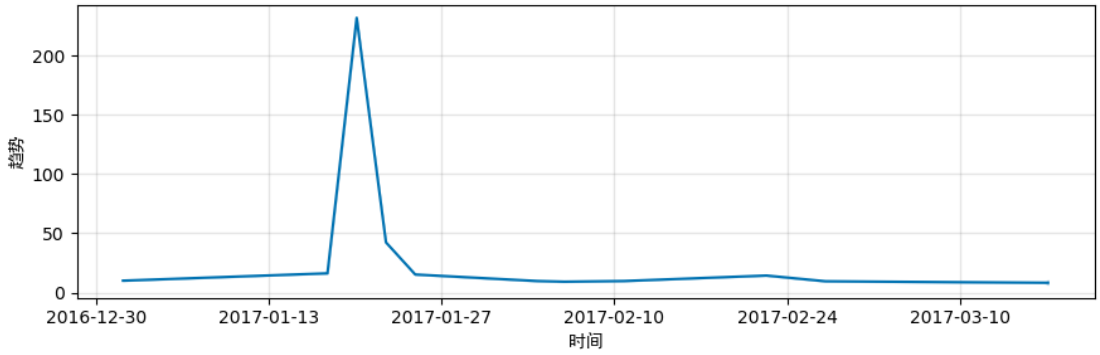


图 5-8 机场运力需求模型月趋势

对于一周的趋势，同样可以拟合出一周的变化趋势，如图 5-9 所示，检测出来在周日、周五有一个较大的趋势。

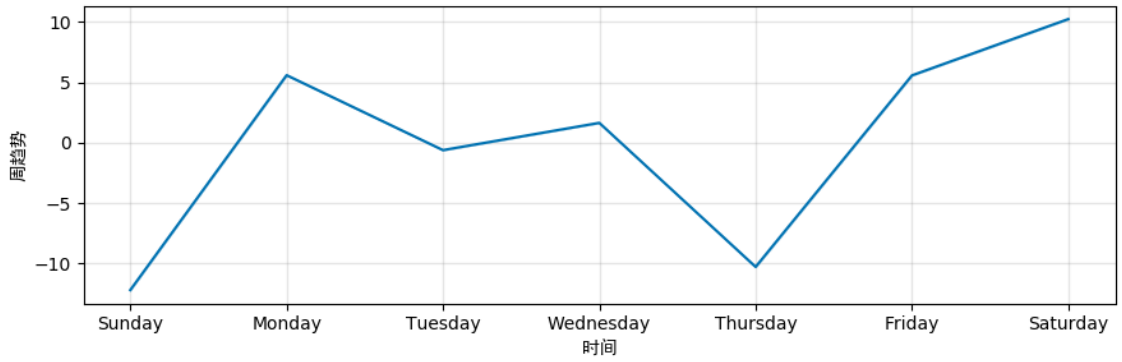


图 5-9 机场运力需求模型周趋势

对于一年中的每个月，可以拟合出如图 5-10 的趋势。如图所示，每天的六月份，旅客出行人数会迎来一个低谷，之后 8 月到 10 月是旅行的最佳时节，同时相对的假期天数较多，相应的旅客人数也是最多的。

整合一年的趋势和每个月客流量的变化趋势，可以得到如图 5-11 的拟合结果。从原始客流数据中将季节性成分和趋势性移除，最后再添加到预测中，可以达到更准确的预测结果。

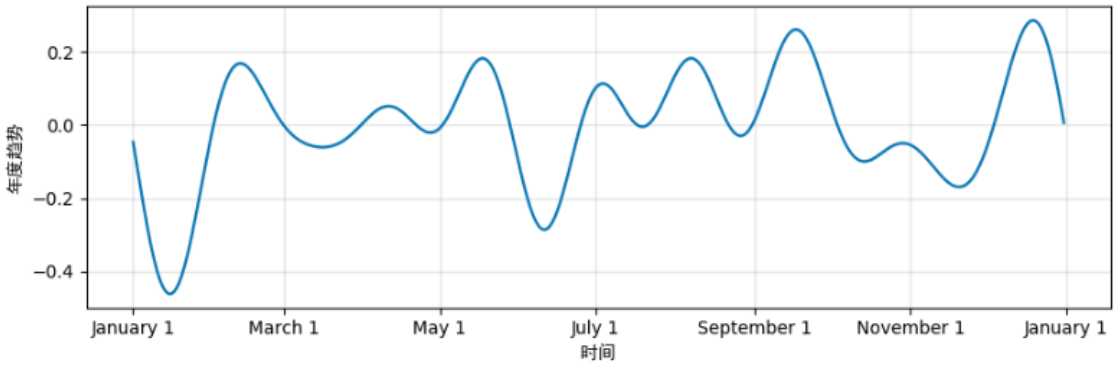


图 5-10 拟合一年中各个月客流量的变化趋势

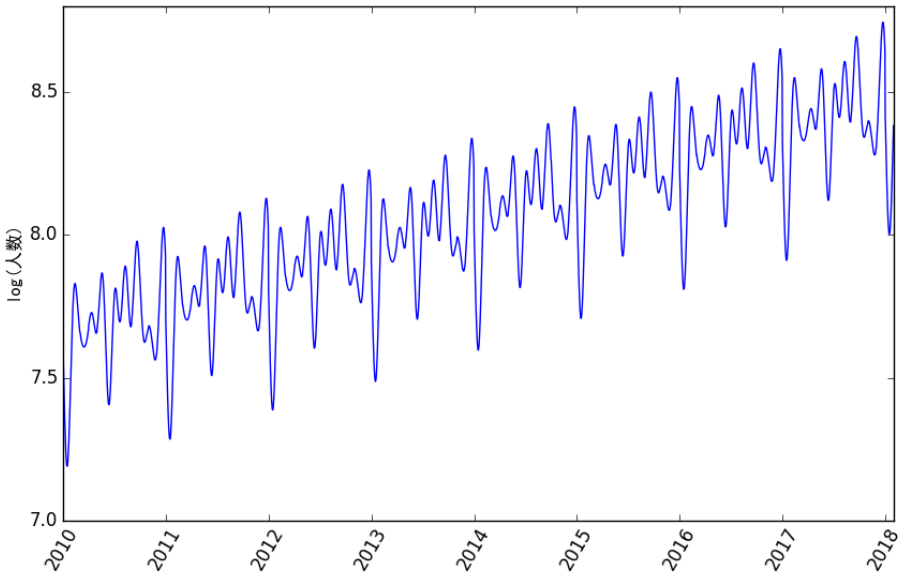


图 5-11 模型趋势性和季节性变化曲线

## 5.4 实验结果和分析

通过训练数据得出的机场运力需求预测模型需要利用测试数据进行模型预测效果的测试。

### 5.4.1 数据集描述

#### 5.4.1.1 机场出租车候车区客流数据

本文采集的数据是北京机场二号航站楼乘坐出租车的客流的统计数据。数据的采样间隔为 1 分钟。

在之后的实验中，考虑到计算的时间，数据统计的时间为 10 分钟一次，这样

每天统计的数据数量是 144。

实际采集数据的过程中，因为设备维修等原因，会导致某些日期的数据缺失。

系统数据库包括原始数据库和结算数据库，原始数据库存储的是机场中的原始信息，结算数据库存储的是预测算法计算得到的结果，并将结果提供给其他业务读取。

数据信息如图 5-12 所示，前 1-8 位表示设备编号。9-16 位表示日期。17-18 位，19-20 位分别是采集时刻的小时和分钟，21 位和 22 位分别是存储状态和设备状态，23-26 位是进入安装该传感器的门的人数，27-30 位是走出安装该传感器的门的人数。如图 5-12 所示，该设备的编号是 00005978，采集该数据的日期是 2015 年 12 月 3 日 0 时 7 分。


图 5-12 机场传感器统计信息

传感器统计的方向有两个，对于入口，实际统计的进入人数是进来的方向的统计人数减去出去的方向的统计人数；对于出口，实际统计的出去人数是出去的方向的统计人数减去进来的方向的统计人数。对于出口和入口的传感器，处理数据的方式是类似的。

图中统计的传感器是入口的传感器。进入该传感器的人是 2，走出该传感器的人是 3，也表明，在这分钟，有一个人走出该传感器。

其中在 T2 航站楼中，前八位标识为 00012160，00012394，00012405 的是入口，标识为 00012414，00012404，00012415 的是出口。

在 T3 航站楼中，前八位标识为 00012412，00005978 的是入口，标识为 00012413，00013281，00012423，00012411，00013282 的是出口。

整理后的数据如图 5-13 所示：

| DEVICEID | POSITION | POSITIOON_DV | INCOUNTER | OUTCOUNTER | DATACOLLECTTIME |
|---|---|---|---|---|---|
| 12415 | T2 | 出口3 | 0 | 19 | 201601230000 |
| 12415 | T2 | 出口3 | 0 | 13 | 201601230001 |
| 12415 | T2 | 出口3 | 0 | 6 | 201601230002 |
| 12415 | T2 | 出口3 | 0 | 9 | 201601230003 |
| 12415 | T2 | 出口3 | 0 | 1 | 201601230004 |
| 12415 | T2 | 出口3 | 0 | 9 | 201601230005 |
| 12415 | T2 | 出口3 | 2 | 11 | 201601230006 |
| 12415 | T2 | 出口3 | 0 | 11 | 201601230007 |
| 12415 | T2 | 出口3 | 0 | 2 | 201601230008 |

图 5-13 机场传感器采集信息示意图

在仿真实验中，将这些数据分为两部分，一部分作为训练集，一部分作为预测集，对于训练集，建立模型对这部分数据进行拟合，最后用模型预测得到的结果和预测集中的实际数值进行对比，计算误差的大小。

#### 5.4.1.2 机场航班数据

机场航班信息和客运信息都是通过访问北京市交通运行监测调度中心（TOCC）内部数据库获得。涉及的数据有每日进出港总客运量、进出港总航班架次、出租车待运车数量、出租车出发车数量和总共乘坐人数等等。

本系统涉及的表项有：

机场数据表 FLOP_ACTT 记录了机场的实际起飞、到达事件、航班 ID 和到达的实际时间。

机场数据表 FLOP_PAST_ACCOUNT 记录了起飞和到达的航班数。

机场数据表 FLOP_DELY 记录了飞机的延误数量以及航班取消事件。在这张表中，有飞机的延误的开始时刻和延误时间。这些信息在数据表 FLOP_ACTT 和数据表 FLOP_CNCL 中有更详细的记录。飞机排班的差异会带来旅客流量的大幅度变化，上述表项对于客流量分析有重要的作用。

机场数据表 SCHD 记录了航班的详细信息，如航班 ID、计划的到港或离港时间、实际的到港或离港时间、航班航站楼以及该航班的最大载客数。从这张表中可以挖掘出不同航班对应旅客的大致人数，结合机场数据表 FLOP_ACTT，可以有效预测未来人流量的变化。

机场数据表 MDKA01_ROUTE_LINK_INF 记录了首都机场附近的道路的拥堵情况。

### 5.4.2 实验结果分析

实验将 ARIMA 模型，LSTM 模型和 TSR 算法的预测结果进行比较，实验的主要目的是测试各个预测模型对于机场运力需求预测的效果。

由于机场不同的航站楼的航班公司有区别，因此按照不同航站楼对机场的出租车运力需求进行预测。以下是各个模型对机场运力需求的预测曲线图。

T2 航站楼的实验结果如图 5-14，5-16，5-18 所示，图中的数据分别是 2017 年 4 月 25 日 4 点到 2017 年 4 月 26 日 4 点采集到的真实出租车运力需求数据和各个模型的预测数据。

T3 航站楼的实验结果如图 5-15，5-17，5-19 所示，图中的数据分别是 2017 年 3 月 15 日 4 点到 2017 年 3 月 16 日 4 点采集到的真实出租车运力需求数据和各个模型的预测数据。

如图 5-14 和图 5-15 所示是传统 ARIMA 模型对于机场出租车运力需求的预测曲线图。

ARIMA 模型可以根据过往的数据捕捉到几个重要的峰值，但是在具体的时间

点上和实际值的差距较大，ARIMA 模型通过提升阶次已经很难满足数据预测的要求。

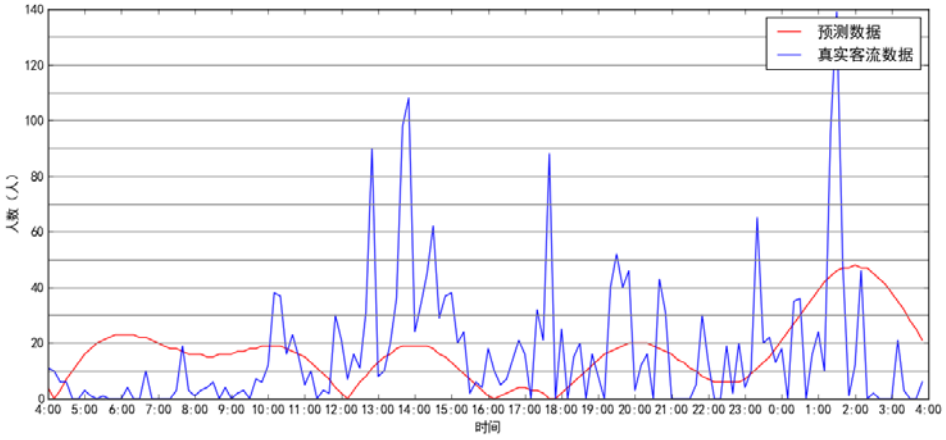


图 5-14 T2 航站楼机场出租车运力需求数据和 ARIMA 模型预测数据对比图

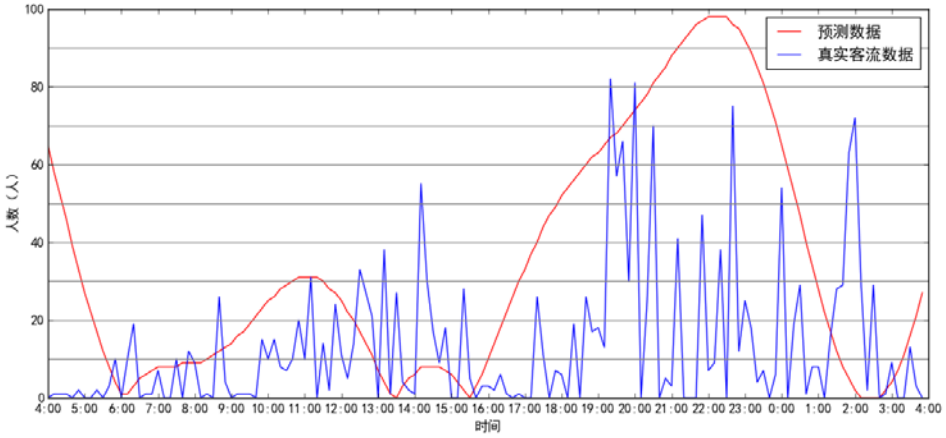


图 5-15 T3 航站楼机场出租车运力需求数据和 ARIMA 模型预测数据对比图

如图 5-16 和图 5-17 所示是 LSTM 模型对于机场出租车运力需求的预测曲线图。

LSTM 模型进行预测的效果较好，主要是在预测的过程中将一部分机场的相关信息运用到模型的预测中，所以预测能够比较贴近实际值，LSTM 模型的预测随着时间的推移，误差逐渐增多，这和模型本身的构建方式有关，模型不断根据一天中前半部分的预测值，预测之后的运力需求，误差不断累积，最后预测值偏差也会越来越大。同时随着时间的推移，机场的航班延误等突发情况增多也会导致误差的增大。

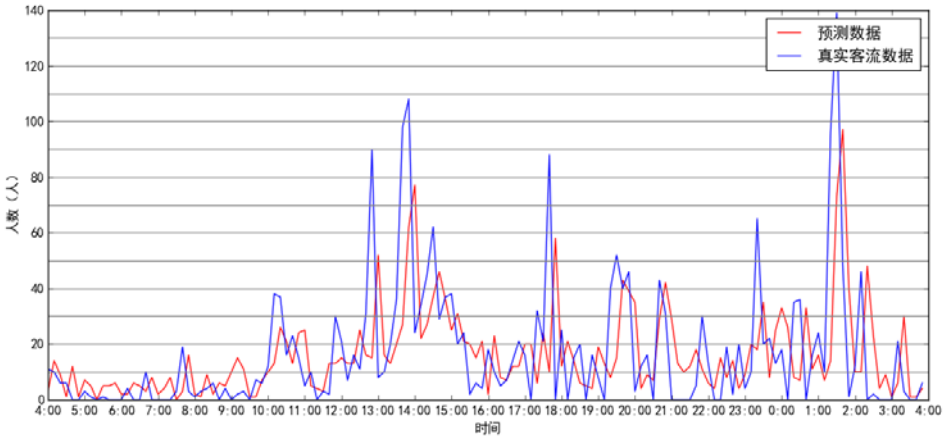


图 5-16 T2 航站楼机场出租车运力需求数据和 LSTM 模型预测数据对比图

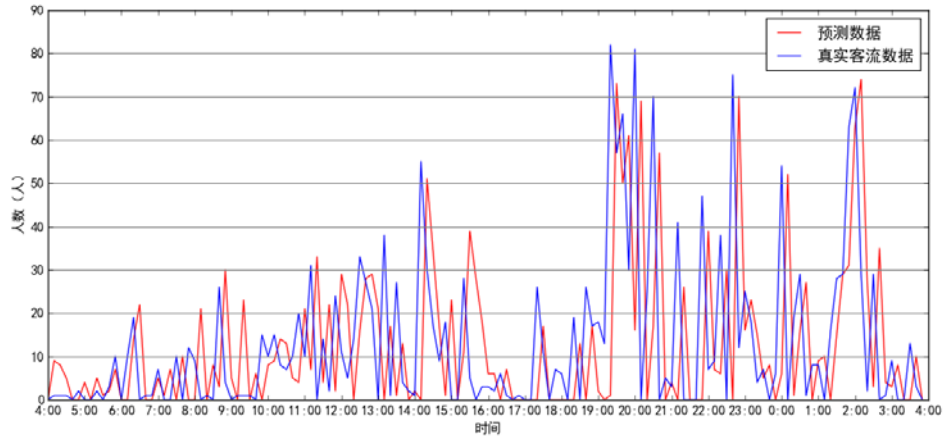


图 5-17 T3 航站楼机场出租车运力需求数据和 LSTM 模型预测数据对比图

如图 5-18 和图 5-19 是基于机场出租车运力需求特点的预测算法对于机场出租车运力需求的预测曲线图。

基于机场出租车运力需求特点的预测算法表现出良好的效果，算法对于一些比较高峰的时段预测已经比较准确，相比于前两个模型，预测效果更加理想。

为了更加准确地反应模型预测的准确程度，选取从 2017 年 3 月 7 日到 2017 年 3 月 16 日连续的 10 天计算模型的评估指标，以更加准确地衡量模型预测的准确度。这 10 天中没有假期的影响，可以比较在正常的工作日和周末这三个模型的预测效果。

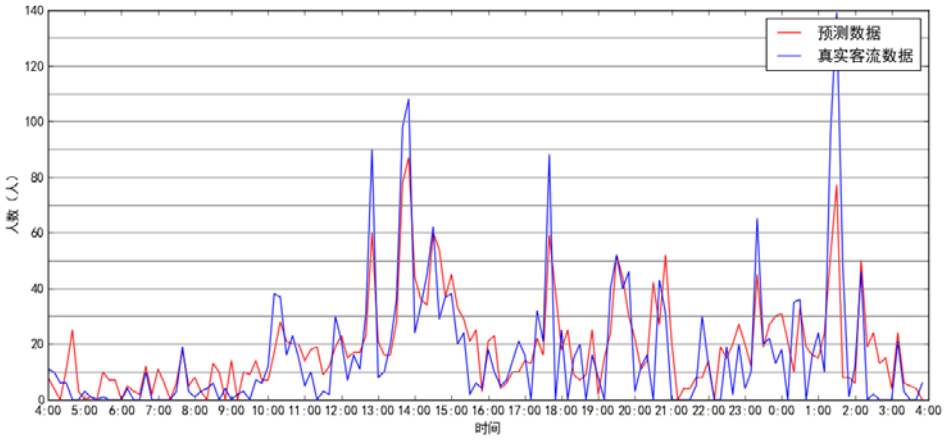


图 5-18 T2 航站楼机场出租车运力需求数据和 TSR 算法预测数据对比图

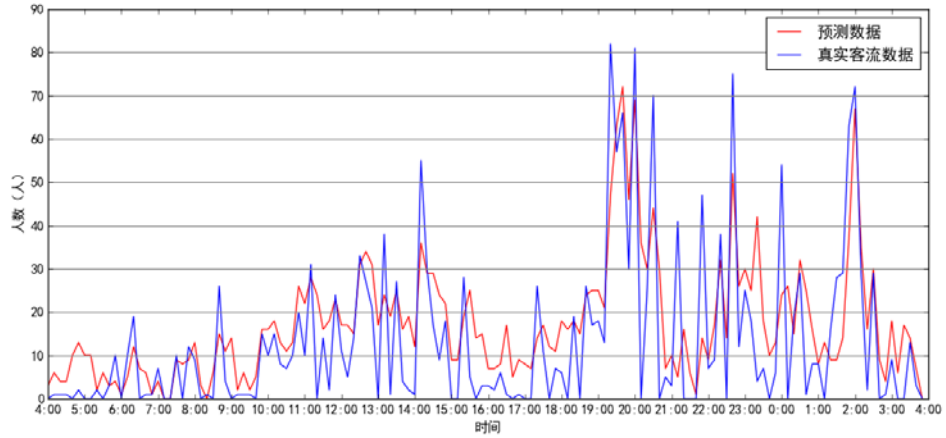


图 5-19 T3 航站楼机场出租车运力需求数据和 TSR 算法预测数据对比图

评估指标如下所示：

(1)均方根误差（Mean Absolute Error, MAE）

$$MAE = \sum_{passengerFlow,timeStamp} (P_{passengerFlow,timeStamp} - Q_{passengerFlow,timeStamp})^2 \quad (5\text{-}12)$$

其中$P_{passengerFlow,timestamp}$是系统的观测值，$Q_{passengerFlow,timestamp}$是模型的预测值。

这个指标能反映一个数据集的离散程度，同时对异常值特别敏感，也就是说预测值和实际值差距越大，误差函数越大，评估指标越差。在一些客流量大、客流量波动剧烈的时间点，预测的误差会比较大，这些时间点的预测对目标函数起决定性作用。这个预测值能够比较合适地反应模型预测的准确度。

类似的评估指标还有：

(2)平均绝对误差（Root Mean Square Error, RMSE）

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(P_{passengerFlow,timeStamp} - Q_{passengerFlow,timeStamp})^2}{n}} \quad (5\text{-}13)$$

其中，n 是预测的次数。

(3)相对绝对误差（Relative Absolute Error, RAE）

$$RAE = \frac{\sum_{t=1}^{n}|P_{passengerFlow,timeStamp} - Q_{passengerFlow,timeStamp}|}{\sum_{t=1}^{n}|P_{passengerFlow,timeStamp} - \overline{P_{passengerFlow,timeStamp}}|} \quad (5\text{-}14)$$

(4)相对平均根误差(Root Relative Square Error, RRSE)

$$RRSE = \sqrt{\frac{\sum_{t=1}^{n}(P_{passengerFlow,timeStamp} - Q_{passengerFlow,timeStamp})^2}{\sum_{t=1}^{n}(P_{passengerFlow,timeStamp} - \overline{P_{passengerFlow,timeStamp}})^2}} \quad (5\text{-}15)$$

T2 航站楼的实验结果如图 5-20，5-21，5-22，5-23 所示。可以看出 TSR 算法各项指标都比另外两个模型小，说明 TSR 算法预测更加准确。另外，2017 年 3 月 11 日是周六，在之前的趋势分析中是客流量较多的一天，在这一天，ARIMA 模型预测的各项指标都明显增大，准确度下降，但是 TSR 算法依然有比较好的表现。2017 年 3 月 15 日是周三，也是趋势分析中客流量较多的一天，这一天 ARIMA 模型的 MAE 和 RMSE 两个指标也有明显上升的趋势，表明 ARIMA 难以适应这种客流量较多，变化比较频繁的场景。另外两个模型在这种客流量较多的日期预测效果也变差。

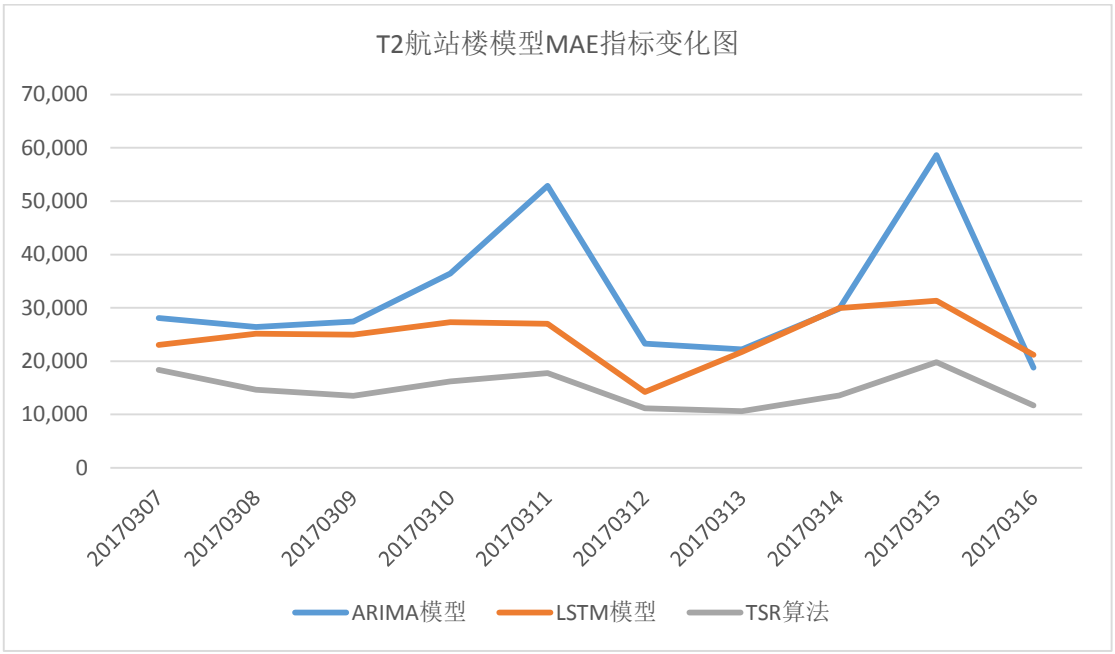


图 5-20 T2 航站楼模型 MAE 指标变化图

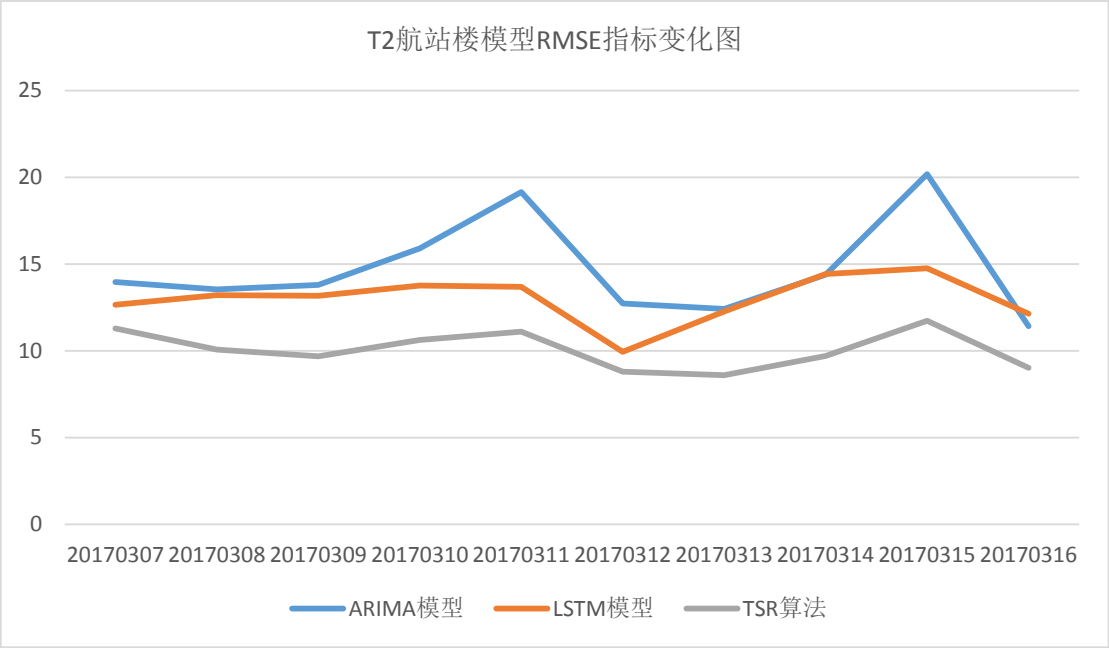


图 5-21 T2 航站楼模型 RMSE 指标变化图

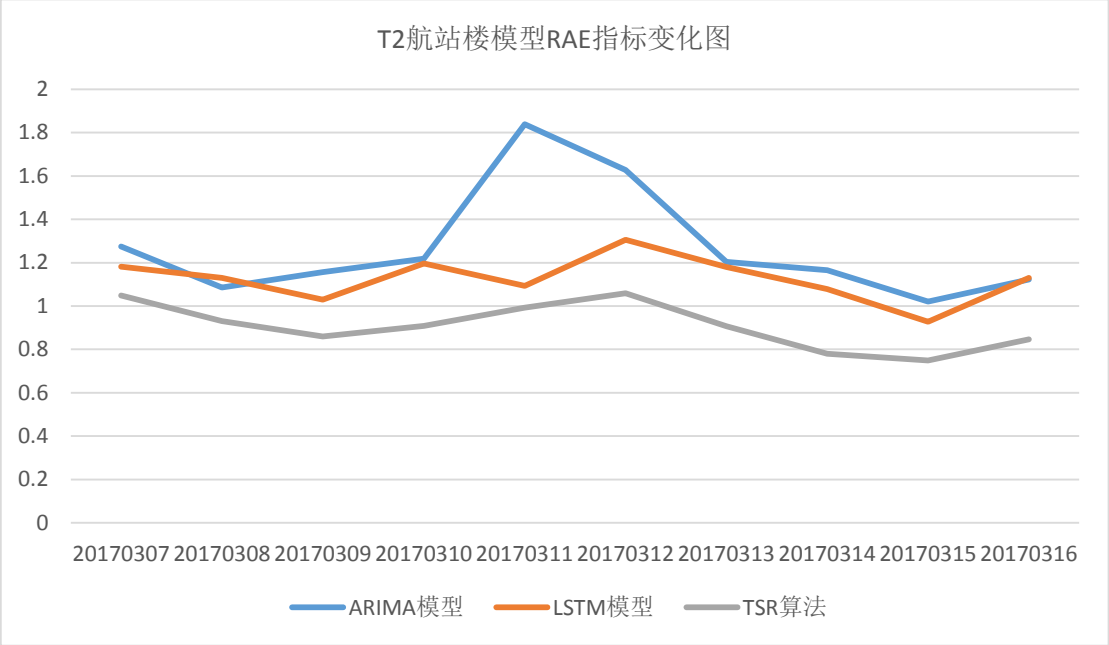


图 5-22 T2 航站楼模型 RAE 指标变化图

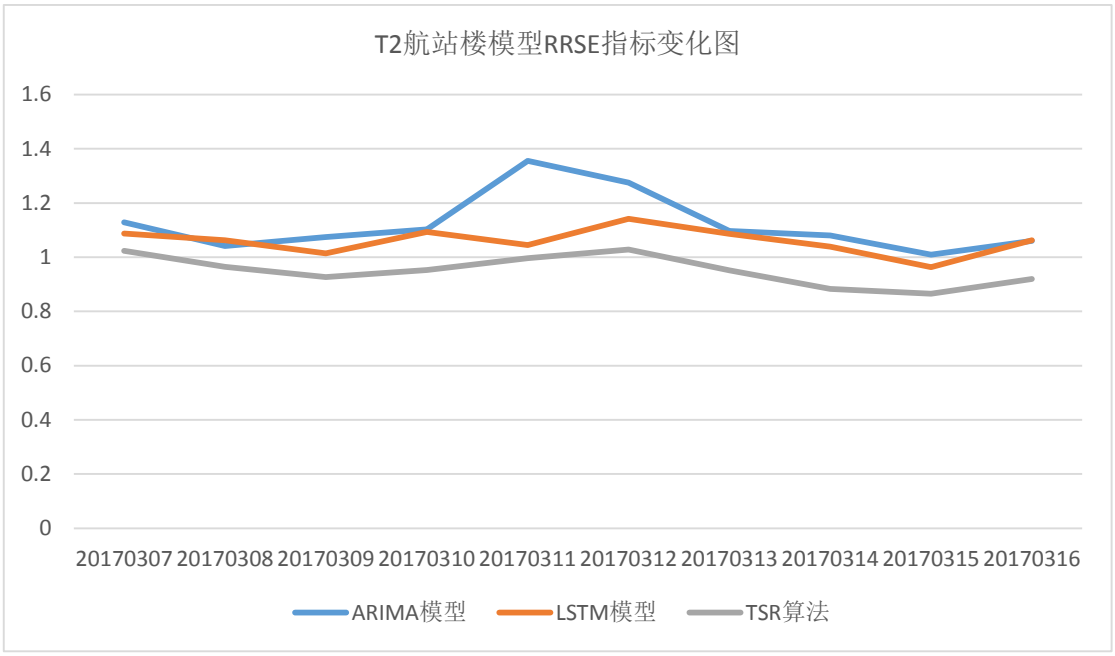


图 5-23 T2 航站楼模型 RRSE 指标变化图

T3 航站楼的实验结果如图 5-24，5-25，5-26，5-27 所示。2017 年 4 月 17 日是周一，2017 年 4 月 22 日是周六，同样是客流量较多的日期，各个模型的指标上升，预测效果下降。

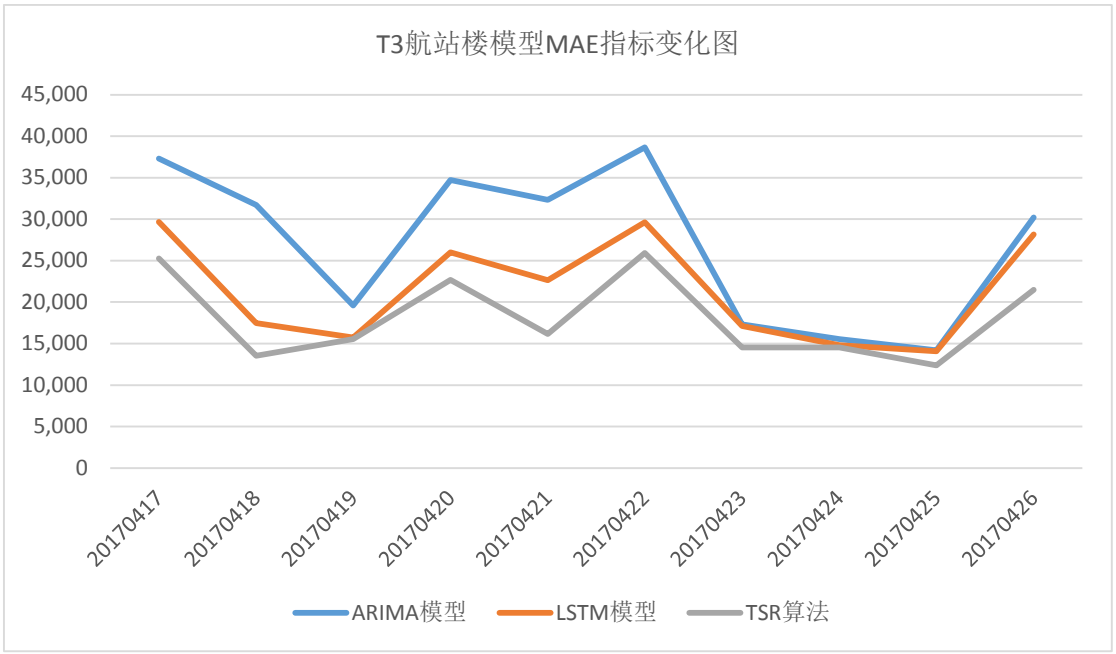


图 5-24 T3 航站楼模型 MAE 指标变化图

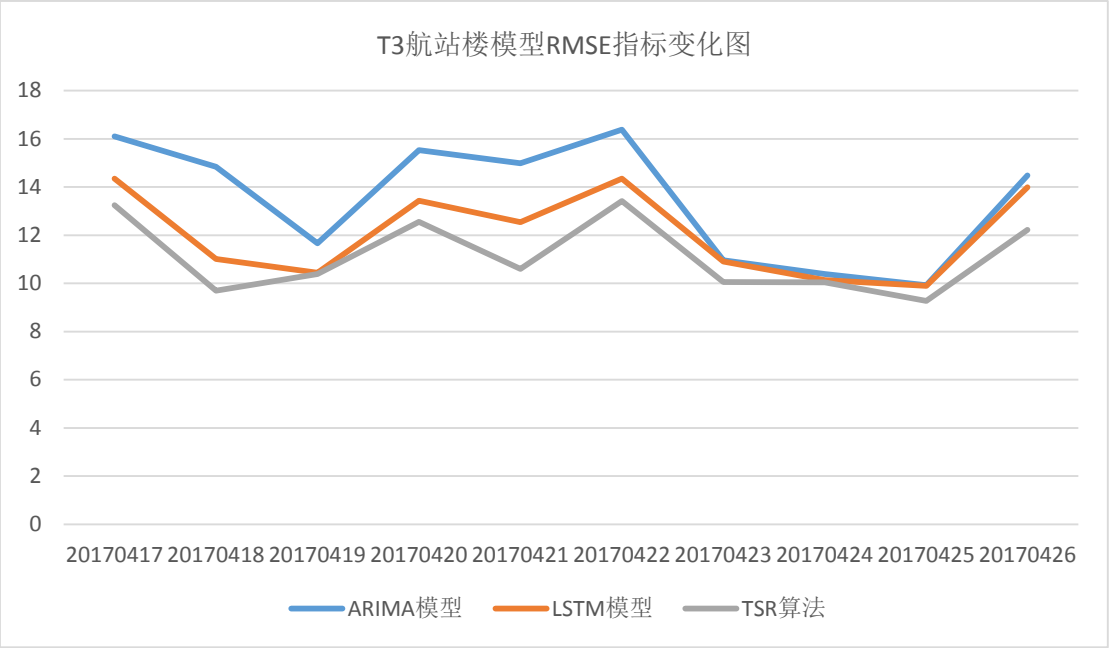


图 5-25 T3 航站楼模型 RMSE 指标变化图

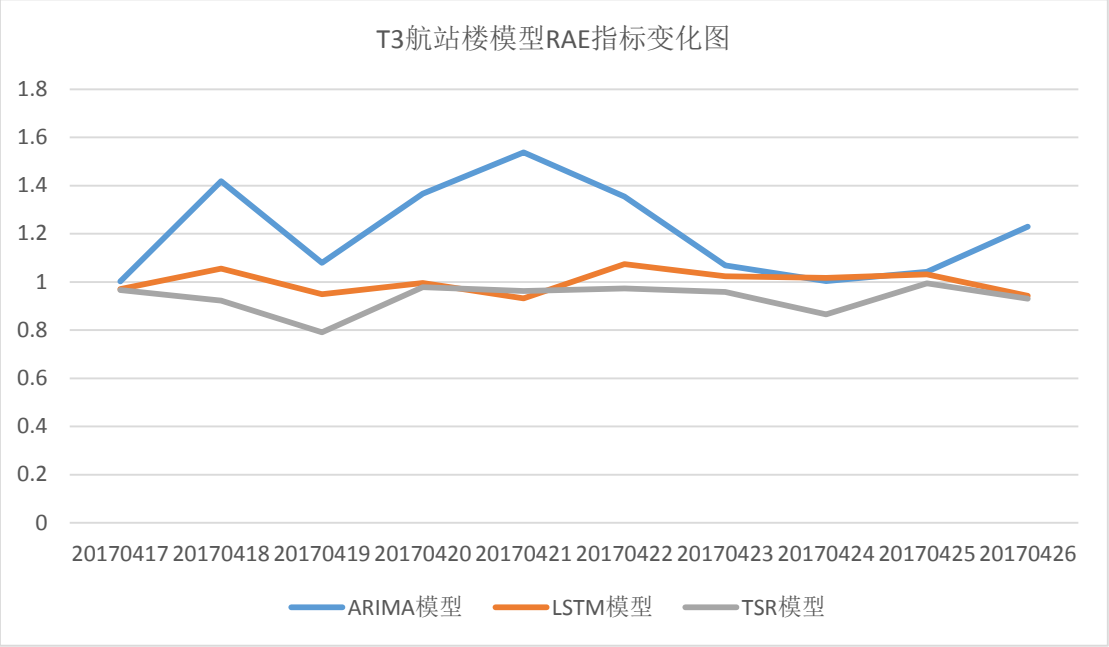


图 5-26 T3 航站楼模型 RAE 指标变化图

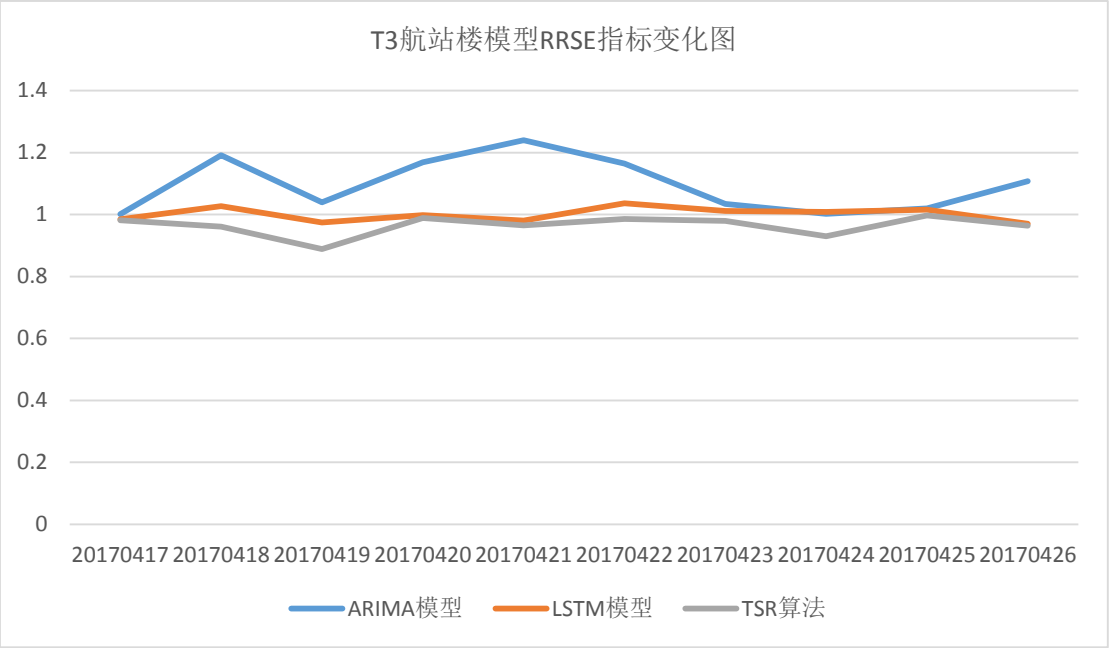


图 5-27 T3 航站楼模型 RRSE 指标变化图

指标值越小，表明预测值更接近真实值。可以看出，基于机场出租车运力需求特点的预测算法在各个指标下都有更好的表现，即基于机场出租车运力需求特点的预测算法的预测误差比其他参照模型小。

将 T2 航站楼的实验结果中的平均值的结果进行整合，如表 5-3 和表 5-4 所示：

表 5-3 T2 航站楼性能指标评价表

| 模型 | 性能指标 | | | |
|---|---|---|---|---|
| | MAE | RMSE | RAE | RRSE |
| ARIMA | 32,402 | 14.751 | 1.271 | 1.122 |
| LSTM | 24,592 | 13.002 | 1.125 | 1.059 |
| TSR 算法 | 14,729 | 10.059 | 0.908 | 0.951 |

表 5-4 T2 航站楼 TSR 算法性能提高百分比表

| 模型 | 性能提高百分比 | | | |
|---|---|---|---|---|
| | MAE | RMSE | RAE | RRSE |
| ARIMA | 54.54% | 31.80% | 28.56% | 15.24% |
| LSTM | 40.10% | 22.63% | 19.28% | 10.19% |

将 T3 航站楼的实验结果中的平均值的结果进行整合，如表 5-5 和表 5-6 所示：

表 5-5 T3 航站楼性能指标评价表

| 模型 | 性能指标 | | | |
|---|---|---|---|---|
| | MAE | RMSE | RAE | RRSE |
| ARIMA | 27,159 | 13.526 | 1.210 | 1.096 |
| LSTM | 21,527 | 12.104 | 0.999 | 1.000 |
| TSR 算法 | 18,210 | 11.148 | 0.934 | 0.964 |

表 5-6 T3 航站楼 TSR 算法性能提高百分比表

| 模型 | 性能提高百分比 | | | |
|---|---|---|---|---|
| | MAE | RMSE | RAE | RRSE |
| ARIMA | 32.95% | 17.58% | 22.80% | 12.04% |
| LSTM | 15.40% | 7.89% | 6.50% | 3.60% |

其中，性能提高百分比公式计算如下：

$$P_{性能提高百分比} = \frac{|A_{基于机场出租车运力需求特点的预测算法} - A_{对比模型}|}{A_{对比模型}} \quad (5\text{-}8)$$

另外选取从 2017 年 10 月 1 日到 2017 年 5 月 7 日连续的 7 天进行计算。这 7 天中有五一假期的影响，可以比较在节假日的影响下这三个模型的预测效果。

实验的结果如图 5-28，5-29，5-30，5-31 所示：

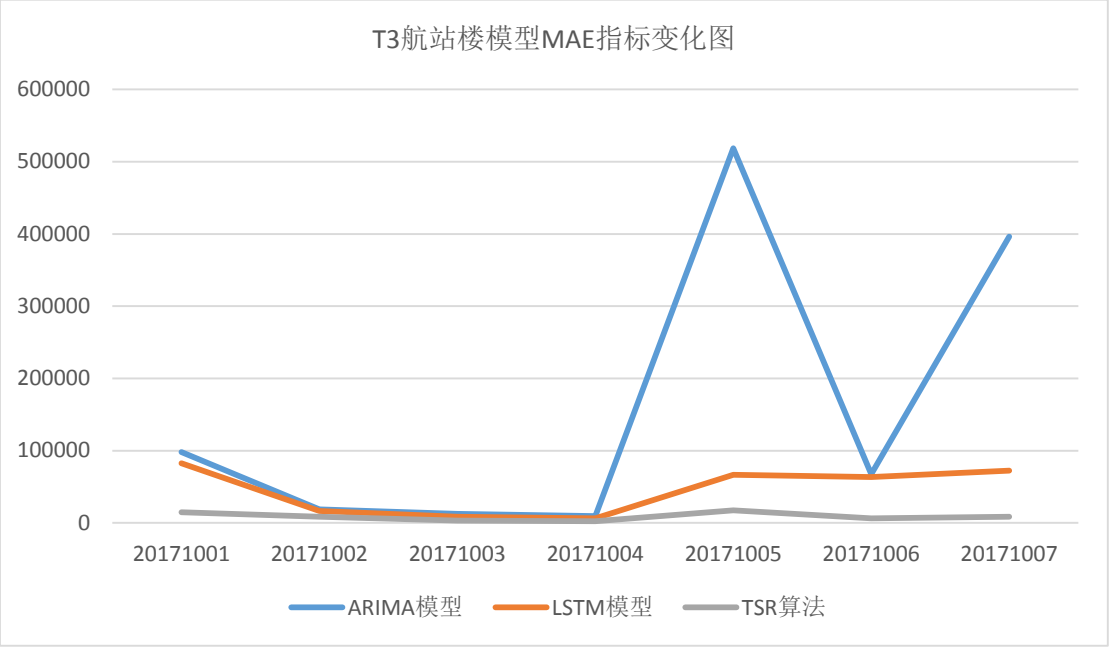


图 5-28 T3 航站楼 MAE 指标变化图

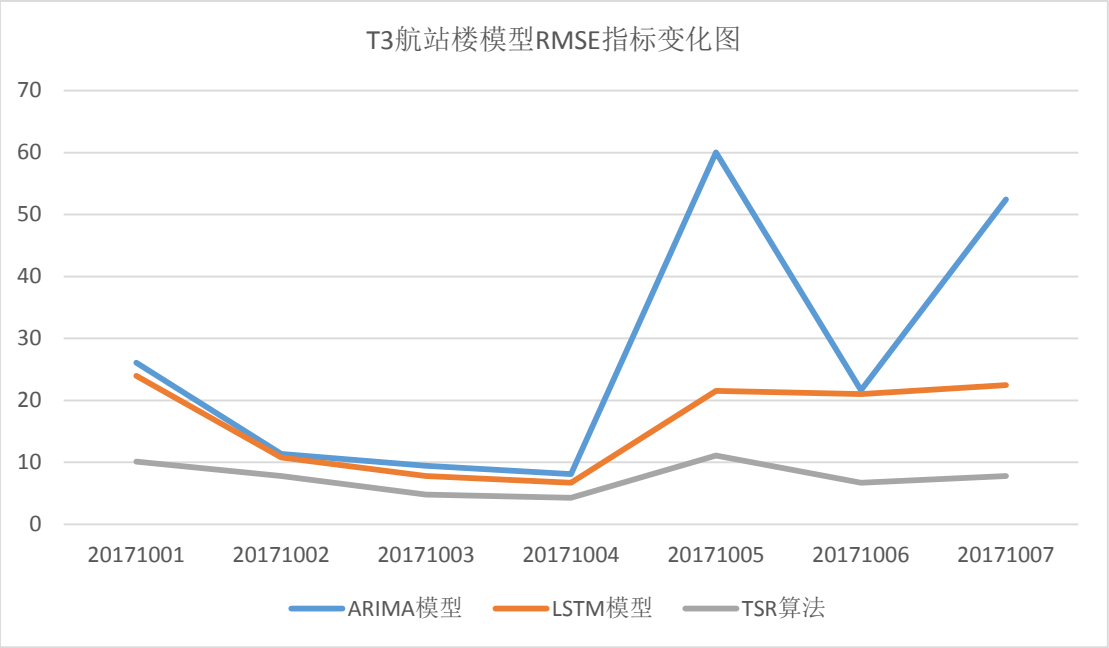


图 5-29 T3 航站楼 RMSE 指标变化图

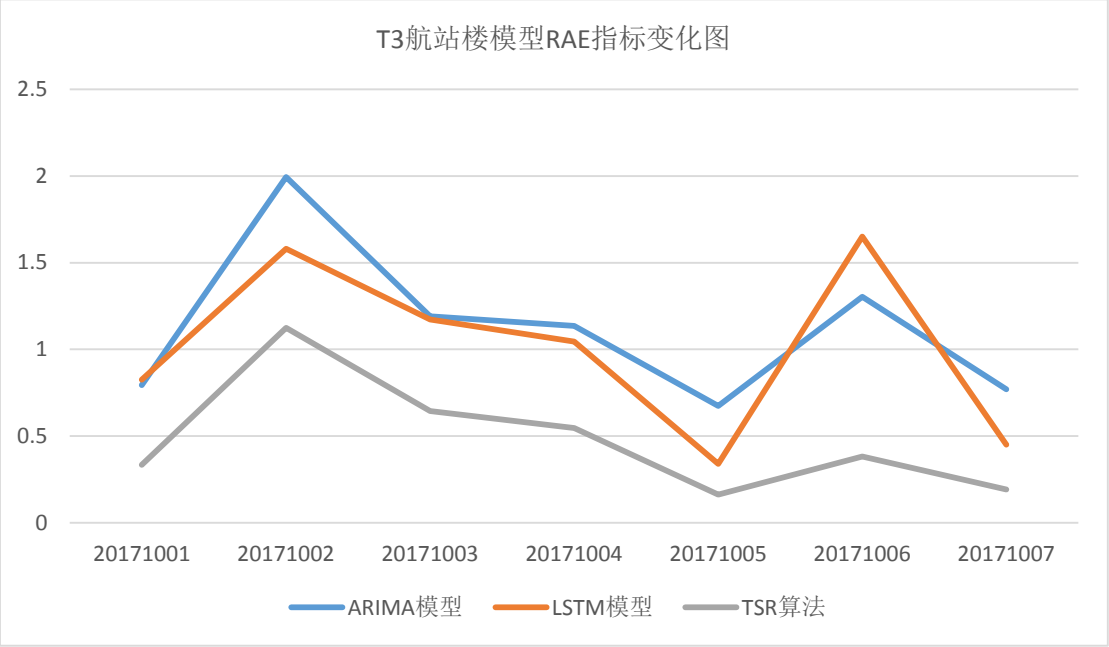


图 5-30 T3 航站楼模型 RAE 指标变化图

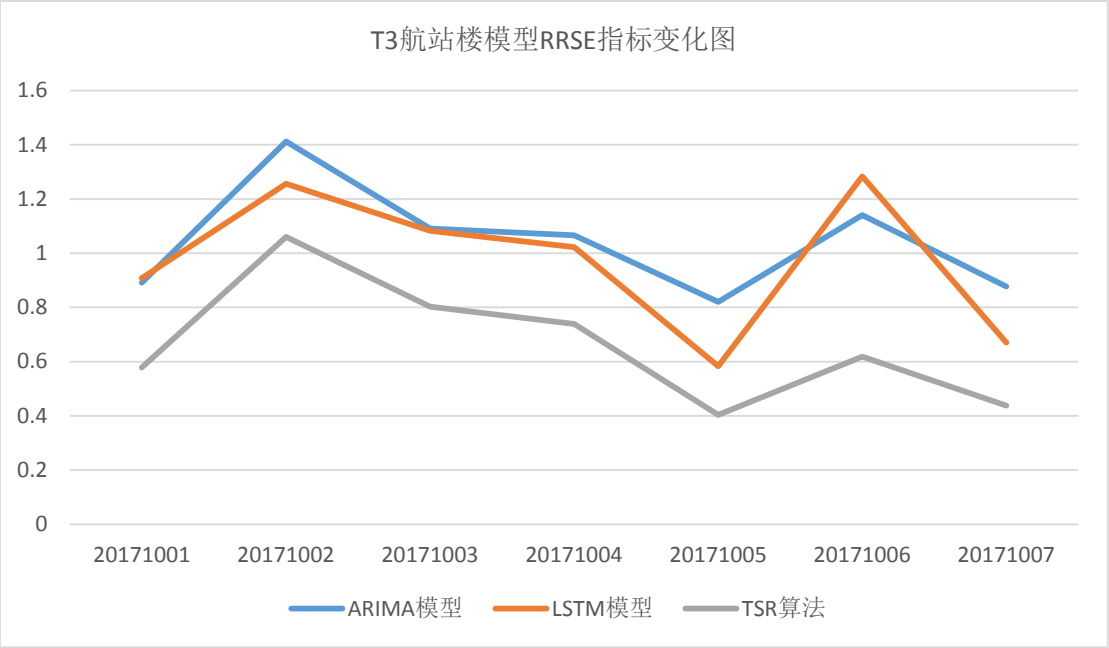


图 5-31 T3 航站楼模型 RRSE 指标变化图

将 T3 航站楼国庆假期的实验结果中的平均值的结果进行整合，如表 5-7 和表 5-8 所示：

表 5-7 T3 航站楼性能指标评价表

| 模型 | 性能指标 | | | |
|---|---|---|---|---|
| | MAE | RMSE | RAE | RRSE |
| ARIMA | 160,113 | 27.003 | 1.123 | 1.042 |
| LSTM | 45,367 | 16.319 | 1.008 | 0.972 |
| TSR 算法 | 8,871 | 7.491 | 0.483 | 0.662 |

表 5-8 T3 航站楼 TSR 算法性能提高百分比表

| 模型 | 性能提高百分比 | | | |
|---|---|---|---|---|
| | MAE | RMSE | RAE | RRSE |
| ARIMA | 94.45% | 72.25% | 56.99% | 36.46% |
| LSTM | 80.44% | 54.09% | 52.08% | 31.89% |

相比之前在普通工作日和双休日的预测情况，TSR 算法在节假日这种客流人数多的场景下预测准确度提高幅度更大，由此可以看出去除模型的趋势性和季节性再进行模型预测的重要性。

## 5.5 本章小结

进行机场出租车运力需求预测，对于提高首都机场接续运输保障能力以及周边道路通行保障能力有重要的意义。

本章基于机场数据的周期性和本身发展所具有的人数增长趋势特点，将这两者的影响在模型中进行考虑，并使用北京首都机场的数据对模型预测的准确度进行了验证，使用 ARIMA 模型、LSTM 模型和基于机场出租车运力需求特点的预测算法对机场出租车运力需求进行预测，实验结果以图形和图表的方式呈现出来，并对实验结果进行对比分析。基于机场出租车运力需求特点的预测算法表现出更理想的预测效果。

# 第六章 总结与展望

## 6.1 论文工作总结

机场客流量是制定航空发展规划的重要依据，也是进行航空建设项目经济评价与统计分析的重要基础。机场客流量的预测研究对于机场科学规划、提高运作效率和安全性、资源优化配置有着重要的意义。针对优化机场地区交通运输组织，满足日益上升的航空客流市内接续交通运输，提高机场地区的交通运输接续保障能力的需求，本课题通过分析机场客流在不同条件下的交通方式选择特点和规律，并根据机场动态客流和运力情况，研究一种运力需求数量预测算法，进行实时预测，最后利用在首都机场采集得到的客流及运力调度数据验证预测的准确率和预测效率。

本文的研究工作如下：

（1）结合首都机场的实际情况对机场客流计数摄像机的数据进行重新审查和校验，修正人为的错误，检查不同系统合并的过程中数据的一致性，修正因为数据统计方式产生的错误数据，处理因为各种异常情况导致的无效值、缺失值等，为下一步数据分析做准备。

（2）结合机场客流的特点、机场的客观环境条件等进行分析，充分发掘机场航班数据和客流数据的特征应用到机器学习中，对机场出租车运力需求预测模型进行训练。在此基础上，抓取诸如天气信息等更多的数据，从而提高预测的准确性。

（3）在 Tensorflow 上进行实验，实现了传统的 ARIMA 模型和 LSTM 模型，并在传统模型的基础上根据机场客流特点进行优化，实现了 TSR 算法。最后利用在首都机场采集得到的客流及运力调度数据评估了各种模型的性能和精度。实验证明，在传统模型的基础上充分考虑机场发展的趋势性、旅客出行的周期性会具有更高的预测精度，是有效可行的。

## 6.2 研究展望

本文在机场出租车运力预测方面取得了一定成果，但仍有部分工作有待进一步研究：

（1）分布式计算可以加速训练过程。

TensorFlow 可以使用 GPU 加速，有效缩短训练时间。

如图 6-1 和图 6-2 是在 GPU 板卡上进行训练的过程。由于时间和资金的关系

并没有在机场实际的预测环境中投入使用，但这对于机场逐渐增多的数据处理及预测无疑是有帮助的。

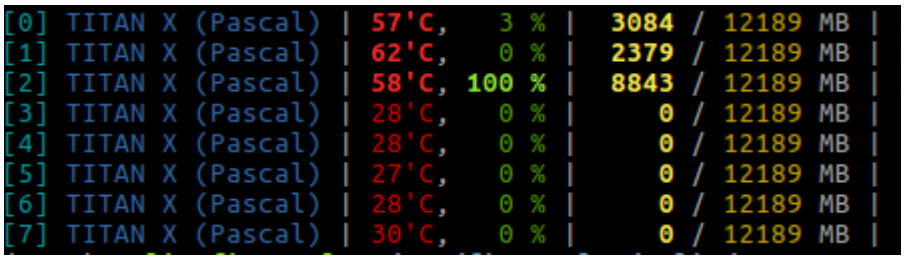


图 6-1 GPU 板卡占有率示意图

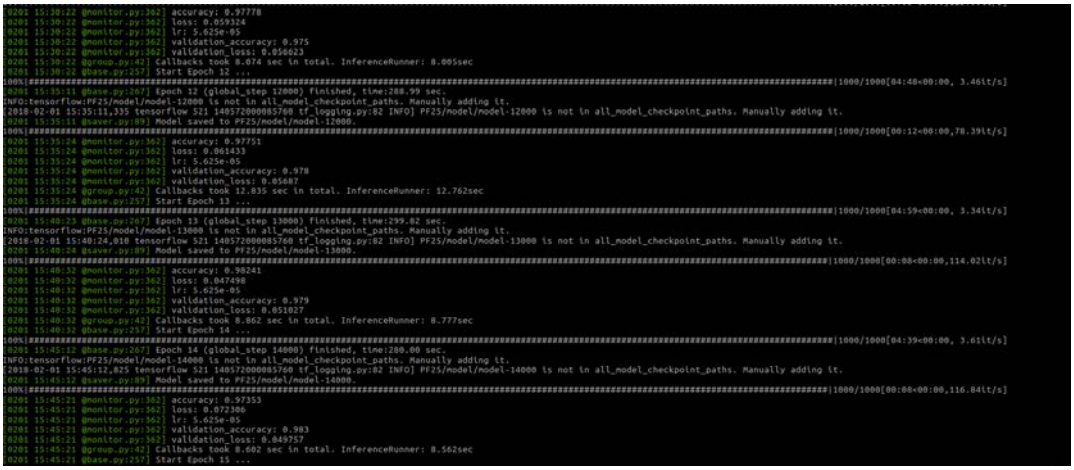

图 6-2 GPU 板卡上训练过程示意图

（2）如何有效结合其他智能算法进行机场出租车运力需求预测是未来的一个研究方向。

（3）做好机场运力需求预测还需要依赖于数据库的持续建设与完善。

项目从开展到我的研究告一段落，只有短短两年的时间，采集的数据数量还远远不够，现阶段只能通过少量数据对各种模型进行分析，同时机场出租车运力需求的统计还存在误差较大的问题，这些方面的改善都可以有效提高机场运力需求预测的准确度。另外，其他的一些指标例如机场周围的交通状况还没有有效利用到模型中，后续的研究中可以作为输入变量，并结合主成分分析算法等提取更多的特征。

（4）时间序列的相关研究可以应用到分子通信中。

分子通信基于生物启发的思想，使用分子作为信息载体来模仿微生物及生物界纳米机之间信息交互的方式，实现纳米机间信息的传输。在信息分子在传输过程中，可能会经历例如分子间碰撞、化学反应及热噪声等非确定因素的影响，这些影响因素即分子信道噪声[33]。

机场出租车运力需求预测中，需要对机场客流数据的各种错误原因进行分析，并且纠正或剔除错误的数据，针对分子信道噪声的抽象研究，同样需要个别化分析不同类型的噪声源，对数据进行整形。

由于时间有限，没有深入开展这部分的研究工作，但这会是以后时间序列研究的重要研究方向。

（5）时间序列的相关研究可以应用到网络异常检测中。

运力需求变化规律与人的生产和生活规律相关，随着人的工作和休息的相互交替，具有一定的周期性。具体来说，机场工作人员和出租车司机规律性的生产生活安排使得运力需求曲线之间有相似性。在内部网络中，网络流量主要由网络中的用户产生，受到的干扰较少，主要和用户的作息时间有关，因此也有着一定的规律性。长期来看，内部网络流量的规律性会因为网络结构的变化和调整出现较大程度的变化，但有比较明显的短期规律，呈现一种以一天或一周为单位的周期性的变化。本文之前提到的时间序列预测算法具有一般性，可以将上述算法应用到内部网络的场景中。

在实际的场景中，能够在防火墙捕捉到网络之间的交互信息，这种信息会形成一种流的形式。通过数据的变换，内部网络中的数据也可以转换为以时间序列形式表示的数据，可以借鉴分析时间序列的方式发掘内在的联系。

内部网络中，数据的各项指标在正常情况下是比较平稳的，有着一定的周期性，但是异常的网络活动，例如 DOS 攻击、病毒传播、网络故障等，会使网络数据产生一定程度的波动，剔除这些数据，对于研究内部网络的整体有重要的意义。另外，有些异常状态可能在出现的初期并不明显，无法检测，但是异常出现的前期已经有数据发生异变，如果能在异常初期检测到异常，就可以为网络故障处理争取大量的时间。

另外，在实际的场景中，源 IP 和目的 IP 的指派有时会因为网络信息中的错误产生误报，或者没有对完整的信息进行记录，类似地需要对这些信息进行预处理。

本文作者曾利用内部网络 1100 台主机和服务器 Cisco ASA 的防护数据日志对上述想法进行验证，是有效的，但在本文中不做具体展开。

# 致 谢

三年的硕士生活难以忘记，在此向所有帮助过我的老师、同学、亲人和朋友表示深深的谢意！

首先衷心感谢我的导师段景山老师，段老师严谨的治学态度和一丝不苟的工作作风使我在学业和人生中受益匪浅。这次论文研究从找准研究方向、确定研究思路到最后的实地采集数据，都得到段老师很多的帮助。在此特向段老师表示最诚挚的感谢！

在研究过程中，感谢北京交通委的支持，让我能够采集到机场的实时客流信息以及其他实时航班数据，让我的算法能够在实际的环境中得到验证。

感谢 2015 级的同窗好友以及各位师兄弟、师姐妹在我研究生期间对我的帮助。

还要特别感谢我的家人！感谢他们对我的学业的支持！感谢他们二十多年来的付出和关爱！

感谢在百忙之中抽出宝贵时间对本文进行评审的老师！

# 参考文献

[1] S Solak, J P B Clarke, E L Johnson. Airport terminal capacity planning[J]. Transportation Research Part B: Methodological, 2009, 43(6): 659-676

[2] W H K Tsui, H O Balli, A Gilbey, et al. Forecasting of Hong Kong airport's passenger throughput[J]. Tourism Management, 2014, 42: 62-76

[3] X Liu, X Huang, L Chen, et al. Prediction of Passenger Flow at Sanya Airport Based on Combined Methods[C]//International Conference of Pioneering Computer Scientists, Engineers and Educators. Springer, Singapore, 2017, 729-740

[4] A Evans, A W Schäfer. Simulating airline operational responses to airport capacity constraints[J]. Transport Policy, 2014, 34: 5-13

[5] A Al-Dhaheri, P S Kang. Lean Improvements to Passenger Departure Flow in Abu Dhabi Airport: Focus on Data from the Check-in Element[J]. Research & Technology, 2015:122-133

[6] S L Tien. Evaluation of an Airport Capacity Prediction Model for Strategizing Air Traffic Management[J]. Jul-2012, 2012:1-35

[7] Günther Y, Pick A, Kern S, et al. Improved airport operations planning by using tailored forecasts of severe weather[C]. Conferences in Air Transport & Operations. 2015

[8] S W Yoon, S J Jeong. An alternative methodology for planning baggage carousel capacity expansion: A case study of Incheon International Airport[J]. Journal of Air Transport Management, 2015, 42: 63-74

[9] C M Ariyawansa, A C Aponso. Review on state of art data mining and machine learning techniques for intelligent Airport systems[C] Information Management (ICIM), 134-138

[10] H Yang, C W Y Leung, S C Wong, et al. Equilibria of bilateral taxi–customer searching and meeting on networks[J]. Transportation Research Part B: Methodological, 2010, 44(8-9): 1067-1083

[11] J Lee, I Shin, G L Park. Analysis of the passenger pick-up pattern for taxi location recommendation[C] Networked Computing and Advanced Information Management, 2008. 1, 199-204

[12] H Nikoue, A Marzuoli, J P Clarke, et al. Passenger flow predictions at sydney international airport: a data-driven queuing approach[J]. arXiv preprint arXiv:1508.04839, 2015, 1-10

[13] W Ma, T Kleinschmidt, C Fookes, et al. Check-in processing: simulation of passengers with advanced traits[C] Proceedings of the Winter Simulation Conference. Winter Simulation

Conference, 2011, 1783-1794

[14]D Curcio, F Longo, G Mirabelli, et al. Passengers' flow analysis and security issues in airport terminals using modeling & simulation[C] European Conference on Modeling & Simulation, Praga-Repubblica Ceca. 2007, 4-6

[15]P F i Casas, J Casanovas, X Ferran. Passenger flow simulation in a hub airport: An application to the Barcelona International Airport[J]. Simulation Modelling Practice and Theory, 2014, 44: 78-94

[16]陈玉宝，曾刚. 基于组合预测方法的民航旅客吞吐量预测研究——以首都机场为例[J]. 中国民航大学学报, 2014, 32(2): 59-64

[17]王磊. 轨道交通机场线客流预测问题研究[D] 西安：长安大学, 2009, 5-10

[18]王曰芬，章成志，张蓓蓓，等. 数据清洗研究综述[J]. 现代图书情报技术, 2007, 2(12): 50-56

[19]J Han, J Pei, M Kamber. Data mining: concepts and techniques[M]. Elsevier, 2011, 88-92

[20]J Friedman, T Hastie, R Tibshirani. The elements of statistical learning[M]. Springer, Berlin: Springer series in statistics, 2001, 57-61

[21]E Fuchs, C Gruber, T Reitmaier, et al. Processing short-term and long-term information with a combination of polynomial approximation techniques and time-delay neural networks[J]. IEEE Transactions on Neural Networks, 2009, 20(9): 1450-1462

[22]E Fuchs, K Donner. Fast least-squares polynomial approximation in moving time windows[C] IEEE International Conference on. IEEE, 1997, 3, 1965-1968

[23]G Dorffner. Neural networks for time series processing[C] Neural Network World, 1996, 447-468

[24]G Perboli, S Musso, F Perfetti, et al. Simulation of new policies for the baggage check in the security gates of the airports: the logiscan case study[J]. Procedia-Social and Behavioral Sciences, 2014, 111: 58-67

[25]G Aguilera-Venegas, J L Galán-García, E Mérida-Casermeiro, et al. An accelerated-time simulation of baggage traffic in an airport terminal[J]. Mathematics and Computers in Simulation, 2014, 104: 58-66

[26]G E Box, G M Jenkins. Time series analysis: forecasting and control rev. ed[J]. Oakland, California, Holden-Day, 1976, 37(2):238-242

[27]T G Dietterich. Ensemble methods in machine learning[C] International workshop on multiple classifier systems. Springer Berlin Heidelberg, 2000, 1-15

[28]朱晓东,鲁铁定,陈西江. 正交多项式曲线拟合[J]. 东华理工大学学报(自然科学版), 2010,

04: 398-400

[29]Kim W, Park Y, Kim B J. Estimating hourly variations in passenger volume at airports using dwelling time distributions[J]. Journal of Air Transport Management, 2004, 10(6): 395-400

[30]张泉峰.首都机场接续运输协调保障技术研究与实现[D].电子科技大学,2015, 1-10

[31]Thomas Kyte. Oracle Database 9i/10g/11g 编程艺术[M]. (苏金国，王小振) 北京：人民邮电出版社, 2011, 311-380

[32]Taylor S J, Letham B. Forecasting at Scale[J]. The American Statistician, 2017, 44: 78-94

[33]黎作鹏,张菁,蔡绍滨,王勇,倪军.分子通信研究综述[J].通信学报, 2013, 34(05): 152-167

# 攻硕期间的研究成果

## 参与的主要科研项目

[1] 北京市交通行业科技项目“首都机场接续运输协调运行保障技术研究与示范应用”，项目主要研发人员。

[2] “基于 TD-LTE 的宽带移动专用通信网络总体方案与测试评估研究”，项目主要研发人员