Hans Journal of Data Mining 数据挖掘, 2017, 7(1), 26-36
Published Online January 2017 in Hans. http://www.hanspub.org/journal/hjdm
http://dx.doi.org/10.12677/hjdm.2017.71003

Hans汉斯

Passenger Flow Analysis under the Big Data

—A Case Study of Guiyang to Chengdu Direction

Qimei Zhang[1,3], Yumei Liao[1,2,3], Yongcheng Ren[1], Peng Huang[1]

[1]Department of Mathematics and Computer Science, Guizhou Normal College, Guiyang Guizhou
[2]Internet + Innovation and Entrepreneurship Center of Guizhou Normal College, Guiyang Guizhou
[3]Research Center of Industrial-IoT Engineering Technology of the Higher Education Institutions in Guizhou Province, Guizhou Normal College, Guiyang Guizhou
Email: 1433211664@qq.com, liaoyumei-1999@163.com, 776089711@qq.com, 2541377241@qq.com

Received: Jan. 10th, 2017; accepted: Jan. 21st, 2017; published: Jan. 24th, 2017

Abstract

By analyzing travelers' seat rates during some holidays, we can draw a conclusion that the passenger flow has a big fluctuation during holidays, taking Guiyang to Chengdu direction as an example. We analyze the trend of the rate with time order method, exam the smoothly of data, and translate the order into the ARIMA model. Then, the data are fitted by Hlot two-parameter exponential smoothing. We quantify the qualitative data and analyze the data based on the theoretical knowledge in the analysis of regression. We regard the beginning time, duration, and the comfort level of the seats as indexes and we obtain the relationship between passenger flow and the above indexes.

Keywords

Time Series, Regression Analysis, ARIMA Model, Hlot Two-Parameter Exponential Smoothing

基于大数据下的旅客流量分析

—以贵阳到成都方向为例

张啟梅[1,2]，廖玉梅[1,2,3]，任永成[1]，黄　鹏[1]

[1]贵州师范学院数学与计算机科学学院，贵州 贵阳

文章引用: 张啟梅, 廖玉梅, 任永成, 黄鹏. 基于大数据下的旅客流量分析[J]. 数据挖掘, 2017, 7(1): 26-36.
http://dx.doi.org/10.12677/hjdm.2017.71003

[2]贵州师范学院大学生互联网+创新创业训练中心，贵州 贵阳
[3]贵州师范学院贵州省高校工业物联网工程技术研究中心，贵州 贵阳
Email: 1433211664@qq.com, liaoyumei-1999@163.com, 776089711@qq.com, 2541377241@qq.com

收稿日期：2017年1月10日；录用日期：2017年1月21日；发布日期：2017年1月24日

摘 要

本文以贵阳到成都方向为实例，经过对部分节假日客座率进行分析，可知在节假日前后旅客流量波动较大；用时间序列法对趋势客流量进行分析，对数据进行平稳性检验，对原序列拟合ARIMA模型，然后利用Hlot两参数指数平滑法对数据进行拟合；在回归分析中，将定性数据数量化，根据理论知识对数据进行分析，以起始时间、时长、座位的舒适度为指标，得出客流量与这几个指标之间的关系。

关键词

时间序列，回归分析，ARIMA模型，Hlot两参数指数平滑法

1. 引言

随着发改委发布的《关于改革完善高铁动车组旅客票价政策的通知》，铁路总公司根据市场情况自行对火车票价进行定价的政策出台。铁路部门为了保持市场竞争力，实现利润最大化，需要了解日常铁路客运流量、淡旺季的变动指数等具体情况，所以对客流量的充分了解和预测是准确把握市场的首要条件，因此有关于铁路客流预测的研究也成了铁路客运服务需要重点研究的对象。

然而铁路客流量受诸多因数的影响，比如：节假日期间铁路客流量骤增，造成人多车少的情况，导致铁路客运量无法满足客户乘车需求，同时也给铁路局带来巨大压力，在非节假日期间，造成车多人少的情况，一些冷门线路区间上客座率不足，这样就造成铁路车辆资源的浪费，因此客流量进行预测，可以为之制定合理的价格，改善火车站运营方式、优化铁路资源配置、促进城市间之间的发展，从而更好的带动各城市间餐饮、住宿等服务业的发展。

2. 趋势客流量分析

趋势客流量预测的方法较多，如时间序列法、回归分析法、灰色预测法、BP 人工神经网络模型、重力模型等。这些方法在各自的领域都有各自的优缺点，因为是研究贵阳到成都方向的铁路客流量，且数据是由时间顺序记录的，结合铁路客流量增长特点，以铁路历史客流量为基础，建立时间序列分析模型对趋势客流量进行预测[1]。

2.1. 数据处理

选取样本数据，在贵州省统计局的统计月报中得出数据如下表所示：

数据缺失原因：由于数据来源是贵州省统计局统计月报，在数据收集过程中，数据缺失的情况是无法避免的。因此，在大多数情况下，信息系统是不完备的，而处理不完备数据的方法有以下三类[2]：

一：删除元组：也就是将存在遗漏信息的值进行删除，进而得到一个完备的信息表；

二：数据补齐：这类方法是用一定的值去填充空值，从而使信息完备化。通常基于统计学原理，根据决策表中其余值的分布情况来对一个空值进行填充。在数据挖掘中又有多种补齐方法：人工填写、特

殊值填充、平均值填充、热卡填充等；

三：不处理：在进行数据挖掘时，也可以包含空值，这类方法出现在贝叶斯网络和人工神经网络中[3]。

在上述数据中，有缺失值的出现，我们采用的是均值填充的方法：数据的属性又分为定性型数据与定量型数据，如果数据为定性型数据，就以该数据的众数来补齐缺失的值；如果数据为定量型数据，就以该数据存在值得平均值来插补缺失的值；可知我们的数据为定量型数据，所以将已知的数据全部相加求和，最后取均值即可，整理后的数据如表 1 所示。

由图 1 可知：2015 年 9 月到 2016 年 7 月这一期间，其中 2 月、7 月的人数分别为 468.53 万、443.25 万人，是这 11 个月当中最多的，且 2 月是春节期间，大量务工人员需要回家与家人团聚，人员迁徙量有所增加，7 月是在夏季，天气好，旅游量会增加，而且又是暑假，也增加了火车的客流量。

下面用模拟数据的方法对数据进行处理，以下时间序列模型所用数据来源于第四届泰迪杯官方网站的 B 题部分数据。

2.2. 模型总体流程

ARIMA 模型建立流程图(图 2)。

2.3. 具体流程

对抽取列车进行数据分析，并通过残差分析来判断模型的拟合度，建立简单模型来反映客流规律，经过分析，我们发现在过去一年里节假日最多的是周六，所以我们对周六的客流量进行统计分析，得出节假日对客流量的影响规律。并对抽中的客流量数据进行统计分析以及构建 ARIMA 模型。

每个月的客流量都在随着很多因素(比如：节假日、周末、寒暑假等)的影响而变化，因此我们初步判断这些数据所构成的序列是非平稳的。

1：对 K17 次列车提取出来的序列值进行平稳性检验：根据时序图观察是否平稳

由图 3 可知：k17 次列车总人数的时序图是非平稳序列。

2：差分运算与 Hlot 两参数指数平滑

Table 1. The passenger flow of railway, road and water transport in Guizhou province from September 2015 to July 2016
表 1. 2015 年 09 月至 2016 年 07 月贵州省在铁路、公路、水路的旅客运输量

	旅客运输总量(万人)	铁路(万人)	公路(万人)	水路(万人)
2015.09	7560.33	383.33	6998	179
2015.10	7826.27	387.27	7234	205
2015.11	7234.63	309.63	6717	208
2015.12	7274.43	287.43	6762	225
2016.01	7412.13	380.53	6855.61	176
2016.02	8022.53	468.53	7365	189
2016.03	7532.84	395.84	7001	136
2016.04	6905.36	392.36	6383	130
2016.05	6994.24	366.24	6472	156
2016.06	6902.38	371.38	6363.09	168
2016.07	7868.25	443.25	7261	164

注：数据来源于省交通运输厅、成都铁路局、省机场集团有限公司。

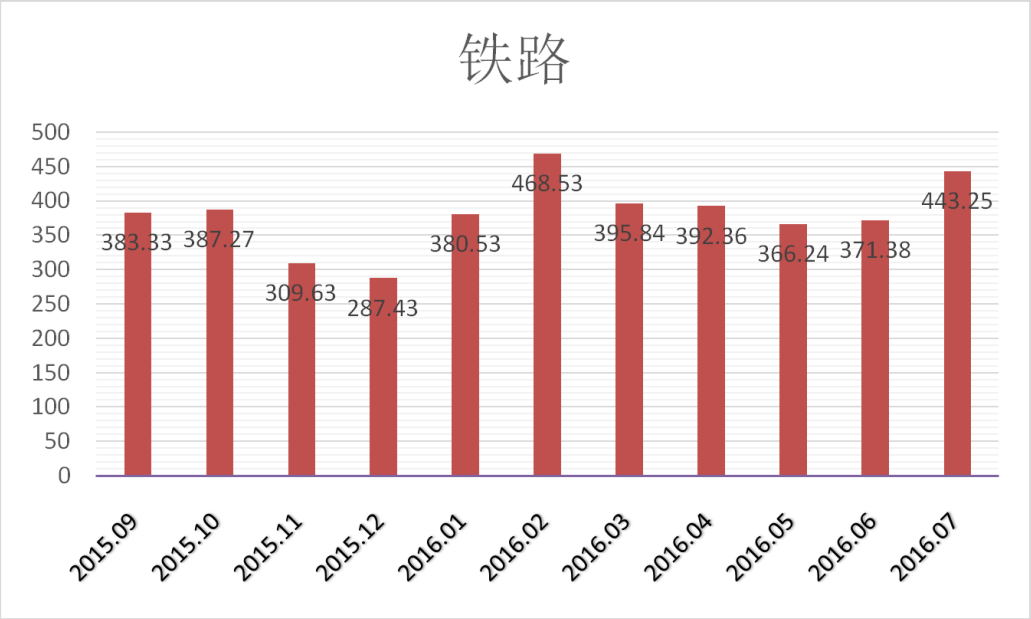

Figure 1. The number of the passengers of railway from September 2015 to July 2016
图 1. 2015 年 9 月到 2016 年 7 月火车客流量人数

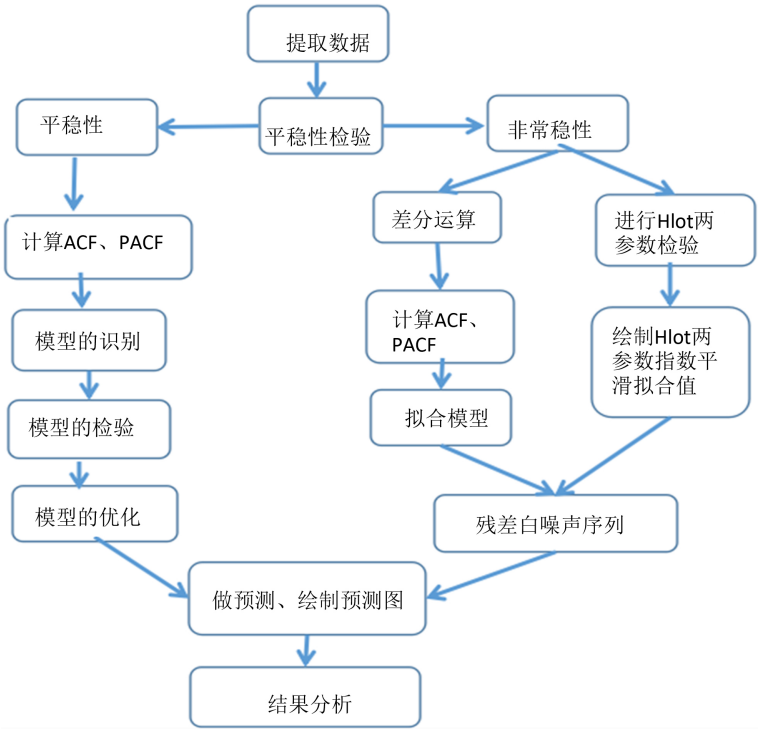

Figure 2. The flow chart of ARIMA model
图 2. ARIMA 模型建立流程图

a. 差分运算：

由图 3 时序图的确定性信息提取的情况来看，他们对确定性信息的提取不够充分，所以我们又采用差分运算，对此序列进行一阶差分处理，从图 4 可知差分后的序列是平稳的。

一阶差分后序列时序图(图 3)在均值附近波动平稳，借助差分后序列自相关图(图 5)进一步考察差分序列的平稳性，除了一阶的自相关系数显著非零，其他阶的自相关系数有超过 2 倍标准差范围，可以知道它是拖尾的，而在偏自相关图中大多数值在负轴附近波动，且有明显的超过标准 2 倍差范围，也可以知道是拖尾，所以可以认为一阶差分后序列平稳，综合考察自相关图(图 5)和偏自相关图(图 6)的属性，可以认为自相关系数、偏自相关系数拖尾，所以对原序列拟合 ARIMA 模型[4]，得到的拟合模型：

$$x_t = 0.8782x_{t-1} + \varepsilon_t - 0.2538\varepsilon_{t-1}, \varepsilon_t \sim N(0,191804)$$

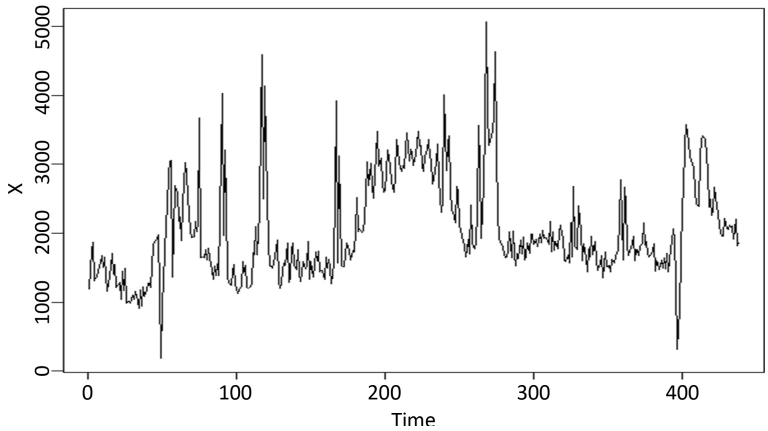

Figure 3. K17 train sequence diagram
图 3. K17 次列车时序图

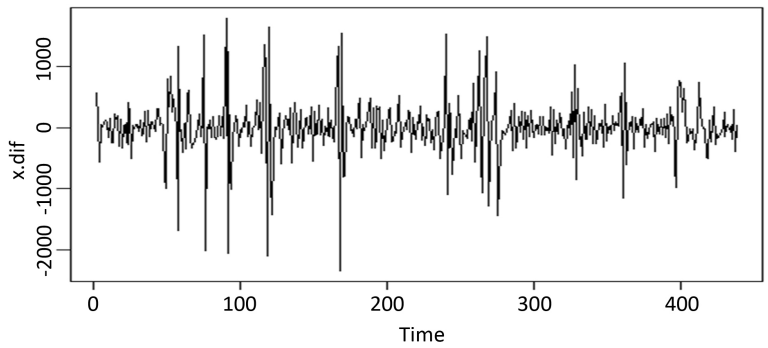

Figure 4. The sequence chart of K17 after difference
图 4. K17 次列车差分后的时序图

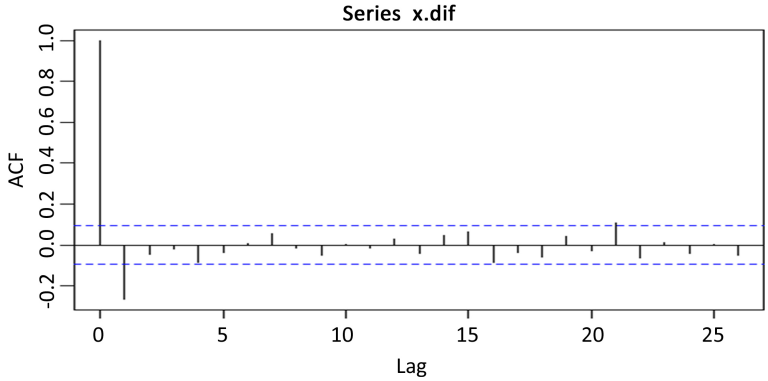

Figure 5. Autocorrelation figure of K17 train after the first order difference
图 5. K17 次列车一阶差分后的自相关图

由残差序列白噪声检验中，不管是自由度为 6，还是自由度为 12 的 P 值明显都大于 0.05，即接受原假设，说明该模型显著成立，这说明 ARIMA 模型对该序列拟合成功。

b. Hlot 两参数指数平滑：适用于对含有线性趋势的序列进行修匀，假定序列有一个比较固定的线性趋势，每期的递增 r 或递减 r，那么第 t 期的估计值就应该等于第 $t-1$ 期的观察值加上每期固定的趋势变动值，对序列值进行指数平滑，绘制 Hlot 两参数指数平滑拟合效果图，即可进行预测，拟合效果图如图 7。

其中：黑色曲线为原始数据经 R 软件操作后的时序图，

红色曲线为 Hlot 两参数指数平滑拟合值。

由图 7 可知，因红色曲线为 Hlot 两参数指数平滑拟合值且与原始数据的时序图(黑色曲线)几乎重合，说明采用两参数指数平滑的方法得到的拟合值是正确的。

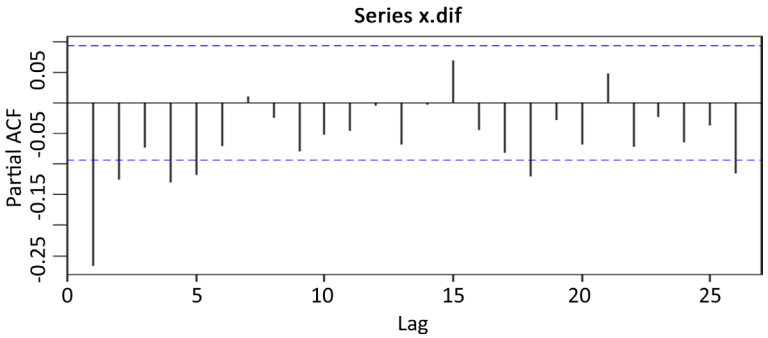

Figure 6. Partial Autocorrelation figure of K17 train after the first order difference
图 6. K17 次列车一阶差分后的偏自相关图

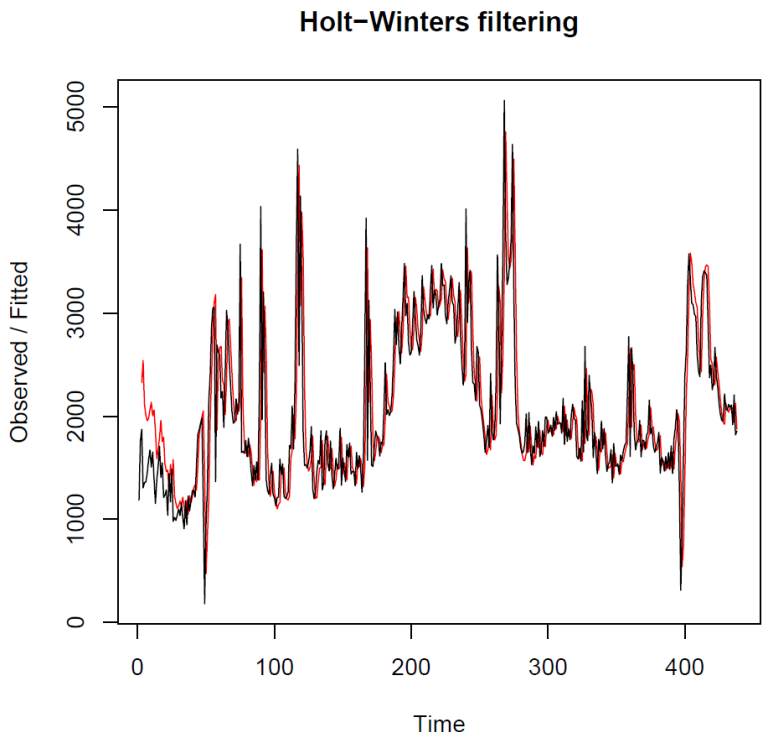

Figure 7. Hlot two-parameter exponential smoothing fitting values of K17 train
图 7. K17 次列车 Hlot 两参数指数平滑拟合值

3. 回归分析

回归分析研究的主要对象是客观事物变量间的统计关系，它是建立在对客观事物进行大量试验和观察的基础上，用来寻找隐藏在那些看上去是不确定的现象中的统计规律性的统计方法。

回归分析方法是通过建立统计模型研究变量间相互关系的密切程度、结构状态及进行模型预测的一种有效的工具[5]。

3.1. 数据处理

A：在时间的选取过程中，一天 24 小时，所以在起始时间的选取中，以小时为单位进行选取；

B：在选择时长的过程中，以分钟为单位对时长进行计算；

C：在回归分析中，对一些自变量是定性变量的情形给予数量化处理，处理方法是引进只取 0 和 1 两个值的虚拟自变量将定性数据数量化，由于贵阳到成都方向列车的座位分为：硬座、硬卧、软卧。所以在这里需要引进两个 0~1 变量，其中一个 0~1 变量为：1 表示硬座，0 表示软座；另外一个 0~1 变量为：1 表示卧铺，0 表示座位；

对数据进行以上处理，选取出来的数据如表 2 (注：数据来源于成都铁路局、携程官网)。

Table 2. Each index of Guiyang to Chengdu direction
表 2. 贵阳到成都方向各指标值

*x*1	*x*2	*x*3	*x*4	*y*
11	18.5	1	0	592
11	18.5	1	1	128
11	18.5	0	1	16
13	17	1	0	605
13	17	1	1	131
13	17	0	1	17
14	16.5	1	0	627
14	16.5	1	1	385
14	16.5	0	1	17
15	19.5	1	0	553
15	19.5	1	1	269
15	19.5	0	1	23
16	24.7	1	0	558
16	24.7	1	1	209
16	24.7	0	1	24
20	16.2	1	0	350
20	16.2	1	1	96
20	16.2	0	1	8

其中：因变量 y：列车的座位数；自变量 x_1：火车的起始时间；x_2：贵阳到成都所需要的时长；x_3：表示列车分为硬座、软座；x_4：列车分为座位、卧铺。

3.2. 设定理论模型

因为数据是非线性的，所以建立多项式回归模型，而多项式回归模型是一种重要的曲线回归模型，将多项式模型转化为一般线性回归模型，因而我们建立以下模型：

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i \tag{3.2.1}$$

其中 $i = 1,2,\cdots,n$，该数据中自变量的个数为 4，所以我们拟合了一个二阶多项式回归模型：

$$\begin{aligned} y = {} & \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_{11} x_1^2 \\ & + \beta_{12} x_{12} + \beta_{13} x_{13} + \beta_{14} x_{14} + \beta_{22} x_2^2 + \beta_{23} x_{23} \\ & + \beta_{24} x_{24} + \beta_{33} x_3^2 + \beta_{34} x_{34} + \beta_{44} x_4^2 + \varepsilon_i \end{aligned} \tag{3.2.2}$$

并打算检验是否有交互效应，在回归中，可以采用逐个引入自变量的方式，这样可以清楚的看到各项对回归的贡献，是显著性更加明确。在 spss 操作中引入自变量，并对自变量进行转化，其中 $x_1^2 = x_1 * x_1$；$x_{12} = x_1 * x_2$；$x_{13} = x_1 * x_3$；$x_{14} = x_1 * x_4$；$x_2^2 = x_2 * x_2$；$x_{23} = x_2 * x_3$；$x_{24} = x_2 * x_4$；$x_3^2 = x_3 * x_3$；$x_{34} = x_3 * x_4$；$x_4^2 = x_4 * x_4$；

对数据进行转化过后进行线性回归分析。

3.3. 参数估计

如果将曲线模型转化为线性模型，就可用普通最小二乘法估计未知参数，如果不能转化为线性模型，则参数的估计就要采用非线性最小二乘法，在该模型中，我们可以采用普通最小二乘估计[6]。

对每一个样本观测值 (x_i, y_i)，最小二乘法考虑观测值 y_i 与其回归值 $E(y_i) = \beta_0 + \beta_1 x_i$ 的离差越小越好，综合考虑 n 个离差值，其中离差平方和为：

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \left[y_i - E(y_i) \right]^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{3.3.1}$$

所谓最小二乘法，就是寻找参数 β_0，β_1 的估计值 $\hat{\beta}_0$，$\hat{\beta}_1$ 使公式(3.3.1)的离差平方和达到极小，即寻找 $\hat{\beta}_0$，$\hat{\beta}_1$ 满足：

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = \min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{3.3.2}$$

根据上式求出 $\hat{\beta}_0$，$\hat{\beta}_1$ 就称为回归参数 β_0，β_1 的最小二乘估计，称

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{3.3.3}$$

为 $y_i \ (i = 1,2,\cdots,n)$ 的回归拟合值，简称回归值或拟合值。

$$e_i = y_i - \hat{y}_i \tag{3.3.4}$$

为 $y_i \ (i = 1,2,\cdots,n)$ 的残差。

用一元线性回归方程拟合 n 个样本观测点 $(x_i, y_i)(i = 1,2,\cdots,n)$，就是要求回归直线 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ 位于这 N 个样本点中间，或者使这几个样本点靠近这条回归直线。

残差平方和

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2 \tag{3.3.5}$$

从整体上刻画了 n 个样本观测点到 $(x_i, y_i)(i = 1, 2, \cdots, n)$ 回归直线 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ 距离的长短。

对 β_0， β_1 进行求导，使得 β_0， β_1 满足下列方程：

$$\begin{cases} \left. \dfrac{\partial Q}{\partial \beta_0} \right|_{\beta_0 = \hat{\beta}_0} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0 \\ \left. \dfrac{\partial Q}{\partial \beta_1} \right|_{\beta_1 = \hat{\beta}_1} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) x_i = 0 \end{cases} \tag{3.3.6}$$

经过整理，后得到正规方程组：

$$\begin{cases} n\hat{\beta}_0 + \left(\sum_{i=1}^{n} x_i \right) \hat{\beta}_1 = \sum_{i=1}^{n} y_i \\ \left(\sum_{i=1}^{n} x_i \right) \hat{\beta}_0 + \left(\sum_{i=1}^{n} x_i^2 \right) \hat{\beta}_1 = \sum_{i=1}^{n} x_i y_i \end{cases} \tag{3.3.7}$$

用正规方程求解 β_0， β_1 的最小二乘估计：

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \end{cases} \tag{3.3.8}$$

利用上述公式就可以具体计算回归方程的参数[7]。

3.4. 结果分析

由上述理论知识对数据进行 spss 操作，先对数据进行多项式转化，将二次多项式转化为一次线性的式子，再对数据进行逐步回归操作，剔除不重要的变量，再对剩余数据进行回归分析，结果如下所示：

模型摘要(The model in this paper)

模型	R	R 平方	调整后的 R 平方	标准估算的错误
1	0.978	0.957	0.895	77.650

分析(ANOVA)

模型		平方和	自由度	均方	F	显著性
1	回归	936467.583	10	93646.758	15.531	0.001
	残差	42206.417	7	6029.488		
	总计	978674.000	17			

系数(The coefficient of)						
模型		非标准化系数		标准系数	t	显著性
		B	标准错误	贝塔		
1	(常量)	1447.846	1408.792		1.028	0.338
	x2	−72.227	136.672	−0.900	−0.528	0.614
	x11	−3.402	1.549	−1.282	−2.195	0.064
	x22	−0.110	3.485	−0.056	−0.031	0.976
	x44	−869.678	377.359	−1.758	−2.305	0.055
	x12	4.780	2.958	1.377	1.616	0.150
	x13	−4.714	16.045	−0.149	−0.294	0.777
	x14	23.379	16.045	0.737	1.457	0.188
	x23	0.466	15.429	0.018	0.030	0.977
	x24	−3.647	15.429	−0.143	−0.236	0.820
	x34	246.707	377.348	0.499	0.654	0.534

由模型摘要可看出回归方程中的 $R^2 = 0.957$ 远大于 0.9，说明拟合的模型是较好的。且是在方差分析表中可以看出：

$SSR = 936467.583$， $SSE = 42206.417$， $SST = 978674.000$， SSR 的自由度 $df = 10$， SSE 的自由度 $df = 7$， SST 的自由度 $df = 17$，且

$$SST = SSR + SSE = 936467.583 + 42206.417 = 978674.000$$

残差的平方和太没有上个模型的值大，说明拟合效果好。

所以得到模型为：

$$\begin{aligned} y_i = {} & 1447.846 - 72.227x_2 - 3.402x_1^2 - 0.100x_2^2 \\ & - 869.678x_4^2 + 4.78x_{12} - 4.714x_{13} + 23.379x_{14} \\ & + 0.466x_{23} - 3.647x_{24} + 246.707_{34} + \varepsilon_i. \end{aligned}$$

4. 结论

随着经济的快速发展，出行问题成为一重大问题，为了保持市场竞争力，实现利润最大化，我们采用时间序列、回归分析等方法对数据进行两参数指数平滑、多项式回归分析对数据进行整理、分析，获得精确结果。

本文以贵阳到成都方向为实例，经过部分节假日对客座率进行分析，可知在节假日前后旅客流量波动较大，铁路部门可根据节假日客流量波动对列车进行增减变动，而火车站附近餐饮业、服务业也可根据客流量为旅客提供便利；用时间序列法对趋势客流量进行分析，对数据进行平稳性检验，对原序列拟合 ARIMA 模型，后利用 Hlot 两参数指数平滑法对数据进行拟合；在回归分析中，将定性数据数量化，根据理论知识对数据进行分析，以起始时间、时长、座位的舒适度为指标，得出客流量与这几个指标之前的关系，这是关于客流量分析的一些方法，这对城市旅游管理部门根据客流量分析结果及时采取措施，有针对的应对淡旺季的不同需求，提高服务旅游质量，减少浪费，提高经济效益。

在多项式回归方法分析中，残差值经过处理依然较大，说明统计的数据尚不完备，给模型推广带来一定难度，初步断定影响客流量的原因还有季节因素、天气因素等，但随着旅游统计工作的进一步完善，

运用 ARIMA 模型进行更加准确的客流量分析。

基金项目

贵州师范学院校级学生科研项目(项目编号:2016DXS099);贵州省 2014 年省级本科教学工程项目“计算机科学与技术”专业综合改革(项目编号：黔教高发[2014]378 号)；卓越工程师教育培养计划项目(黔教高发[2013]446 号)；2015 年省级本科教学工程建设项目(黔教高发[2015]337 号)；2016 年大数据视角下的贵阳市交通优化配置问题研究(项目编号：201614223037)。

参考文献 (References)

[1] 王炜炜. 高速铁路影响下铁路客流量预测研究[J]. 铁道运输与经济, 2016, 38(4): 42-46.

[2] 金国. 客运专线条件下铁路运输通道运力资源优化配置模型研究[D]: [硕士学位论文]. 北京: 北京交通大学, 2009.

[3] 庞新生. 缺失数据处理方法的比较. 统计与决策, 2010(24): 152-155.

[4] 汤岩. 时间序列分析的研究与应用[D]: [硕士学位论文]. 哈尔滨: 东北农业大学, 2007.

[5] 付凌晖, 王惠文. 多项式回归的建模方法比较研究[J]. 数理统计与管理, 2004(1): 48-52.

[6] 索淑文. 自回归模型参数的普通最小二乘估计. 高等函授学报(自然科学版), 2009, 22(4): 3-4.

[7] 蔺焕泉. 普通最小二乘估计方差表达式的等价性. 长春大学学报, 2010, 20(2): 1-2.