引用格式:刘新民,王琪,孙秋霞. 基于 Fisher 聚类的公交客流量时间序列预测及对比[J]. 山东科技大学学报(自然科学版),2019,38(2):73-81.

LIU Xinmin, WANG Qi, SUN Qiuxia. Prediction and comparison of bus passenger flow time series based on Fisher cluster [J]. Journal of Shandong University of Science and Technology (Natural Science), 2019, 38(2): 73-81.

基于 Fisher 聚类的公交客流量时间序列预测及对比

刘新民[1],王 琪[2],孙秋霞[2]

(1. 山东科技大学 经济与管理学院,山东 青岛 266590;2. 山东科技大学 数学与系统科学学院,山东 青岛 266590)

摘 要:公交客流量预测是城市公共交通管理的基础,科学的客流量预测能够为公交系统管理和路线调整提供可靠依据。考虑到公交客流量的波动差异性以及预测的复杂性,首先利用 Fisher 算法对原始数据聚类,并依时段划分为六种类型;然后选择自回归差分移动平均模型以及季节性自回归差分移动平均模型两种方法开展公交客流量的预测,并以广州市公交客流量数据进行实证分析,最后计算两种模型的平均绝对误差和平均绝对百分比误差,对比分析基于聚类数据的两模型预测效果的优劣。结果发现:基于 Fisher 聚类数据,季节性自回归差分移动平均模型的预测效果较好,且比数据未聚类前对应模型预测的效果更优。

关键词:时间序列模型;Fisher 聚类;公交客流量;预测;对比

中图分类号:U491.1 **文献标志码**:A **文章编号**:1672-3767(2019)02-0073-09

DOI:10.16452/j.cnki.sdkjzk.2019.02.009

Prediction and comparison of bus passenger flow time series based on Fisher cluster

LIU Xinmin[1], WANG Qi[2], SUN Qiuxia[2]

(1. College of Economics and Management, Shandong University of Science and Technology, Qingdao, Shandong 266590, China; 2. College of Mathematics and Systems Science, Shandong University of Science and Technology, Qingdao, Shandong 266590, China)

Abstract: The prediction of bus passenger flow is the basis of urban public transport management. Scientific prediction of passenger flow can provide a reliable basis for the management of bus systems and route adjustment. Considering the dissimilarity of fluctuation and the complexity of prediction of bus passenger flow, the original data was firstly clustered by using Fisher's algorithm and was divided into six types according to time interval. Then, the autoregressive differential moving average model and seasonal autoregressive differential moving average model were selected to make predictions of public traffic passenger flow, and empirical analysis was conducted based on Guangzhou public traffic flow data. Finally, the average absolute errors and the average absolute percentage errors of the two models were calculated, and the prediction effect of the two models based on clustering data were compared and analyzed. The results show that based on the Fisher clustering data, the seasonal autoregressive differential moving average model has a better prediction effect, which is much better than that with data without clustering.

收稿日期:2018-03-22

基金项目:国家自然科学基金项目(71371111);国家自然科学基金青年科学基金项目(71501114);山东科技大学科研创新团队支持计划项目(2015TDJH103)

作者简介:刘新民(1965—),男,山东莒南人,教授,博士生导师,研究方向为管理科学与工程、企业家理论等.
E-mail:Liu-xinmin@163.com

Key words: time series model Fisher; bus cluster; passenger flow; prediction; comparison

随着大数据技术的不断发展,数据挖掘的形式在不断变化,而数据预测也成为数据挖掘的一部分。常用的数据预测方法有很多,而时间序列模型因其对预测领域背景要求不高而被广泛使用,除了常见的经济问题预测以及医疗问题预测外,时间序列模型在交通预测方面应用也比较广泛。

在交通预测方面,Liu 等[1-2]采用时间序列模型、灰色模型和改进后的时间序列模型三种模型对机场客流量进行预测;Vasantha 等[3]利用季节性自回归差分移动平均模型(seasonal autoregressive differential moving average,SARIMA)对印度钦奈三条主干路车流量进行了预测;成诚等[4]考虑了节假日效应,利用多元季节性时间序列模型对交通枢纽客流量进行预测;蔡昌俊等[5-6]采用乘积自回归差分移动平均模型(autoregressive differential moving average ARIMA)模型对北京市地铁站进出站客流量进行了预测,马佳羽等[7]采用傅里叶级数与 ARIMA 结合的方式对铁路客流量进行预测;在数据处理方面,周辉宇[8]利用时间序列的关联规则数据挖掘算法验证预测结果,徐晓伟等[9]通过聚类出行数据更为详细地分析了居民出行特征,闫伟等[10]利用模糊聚类的数据挖掘方式对交通流进行预测,取得较好的预测精度。

已有文献主要集中于对原始数据预测模型的探索,并多采用组合多个预测方法或改进预测模型的方式进行交通预测,而对交通数据本身的特征提取或挖掘关注较少。同时,又由于数据聚类的某些方法能够有效提高预测模型的预测精度,结合公交数据本身具有的有序性和周期性,选择恰当的数据聚类算法提取和挖掘数据特征,而后再运用适当的时间序列模型进行预测,在一定程度上可以取得更好的预测效果。基于此,本文先利用 Fisher 有序聚类算法划分公交客流量时段类型,并针对各时段类型数据运用时间序列模型进行公交客流量预测,对比分析不同模型预测效果的优劣,为城市公共交通系统的管理优化和线路调整提供指导。

1 理论基础

1.1 Fisher 聚类算法

Fisher 有序聚类算法是用离差平方和来表示同类样本之间的差异程度。通过简便的计算和作图,确定最优分类数,使同类样本间的差异最小,各类别样本间的差异最大,并用 F 检验法检验最优分类数的合理性[11]。Fisher 聚类算法通常应用在公交客流分析中,通过对有序样本的聚类分析,能够有效地划分公交客流量的相似时段,降低预测的复杂程度,减少预测的重复性。

设有序变量依次为 $x_1,x_2,\cdots,x_n$,如果 $b(n,k)$ 表示对样本 $x_1,x_2,\cdots,x_n$ 的最优 k 分割,则 $b(n,k)$ 一定是在某一个截尾子段的最优 $k-1$ 分割 $b(n,k-1)$ 之后再添加一段形成的。这样就可以从各个截尾子段的最优二分段出发,建立一种递推公式,求出各种 k 值下的最优划分,从而使得求最优划分的精确解得以实现。

1.2 时间序列模型

时间序列预测方法是通过时间序列的历史数据揭示现象随时间变化的规律,将这种规律延伸到未来,从而对该现象的未来做出预测。常用的时间序列模型有两种,一种是自回归移动平均模型(autoregressive and moving average,ARMA),一种是自回归差分移动平均模型(ARIMA)[12]。当原始序列具有平稳性时,不存在差分阶数,被认为是第一种模型,即自回归移动平均模型(ARMA);当原始序列具有不平稳性时,需要进行差分处理,模型被认为是第二种模型,即自回归差分移动平均模型(ARIMA)。特殊的,当某些原始序列具有季节性时,可以提出季节因子,构造季节性自回归移动平均模型(SARMA)或者是季节性自回归差分移动平均模型(SARIMA)。

1.3 数据收集及预处理

公交客流量数据主要来源于乘客的刷卡数据,根据广州市交通部门提供的公交 IC 卡数据绘制原始数据图(如图 1),发现数据具有间断点和异常点,这是由于公交 IC 卡数据在采集、传输和储存的过程中,不可避免会出现一些错误。本文通过数据清洗和数据检查等方式对异常数据进行处理,最终选择了 2014 年 10 月 11 日至 12 月 31 日的数据进行分析,其中 10 月 11 日至 12 月 26 日的客流量数据为模型的实验阶段,12 月 27 日至 12 月 31 日的客流量数据为模型验证阶段。

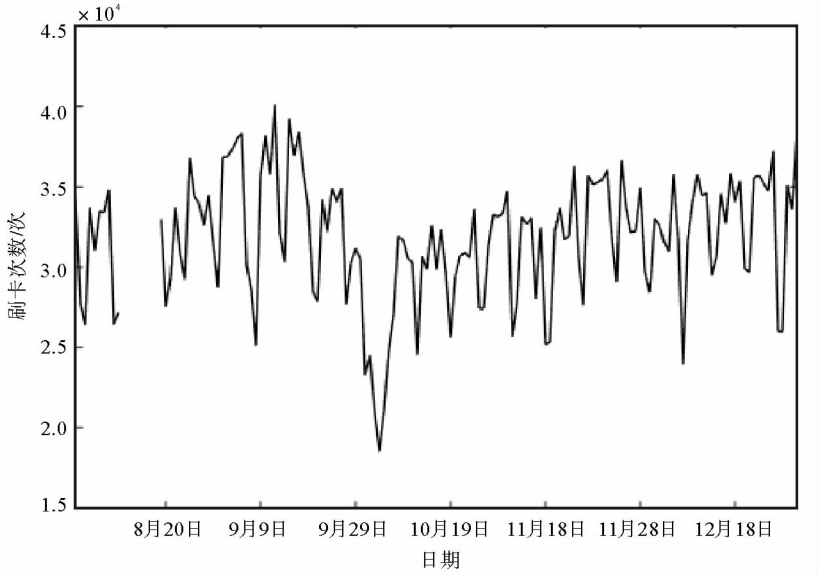

图 1　原始公交客流量数据序列

Fig. 1　Original bus traffic data sequence

为了更细致更直观地展现公交客流量的变化特征,利用统计学中均值的思想,将原始数据整合为单日平均客流量数据(如图 2 所示)。根据整合后的数据图能够发现各时刻客流量差异大小不一,乘客出行具有明显的早晚高峰的特质,在早晚阶段还具有一定的低峰现象。对不同时段的客流量进行预测能够极大提高预测精度,但由于涉及的出行时刻较多,预测工作量较大。采用聚类算法划分相似时段在一定程度上能够减轻预测繁琐性,也可以提高预测精度。

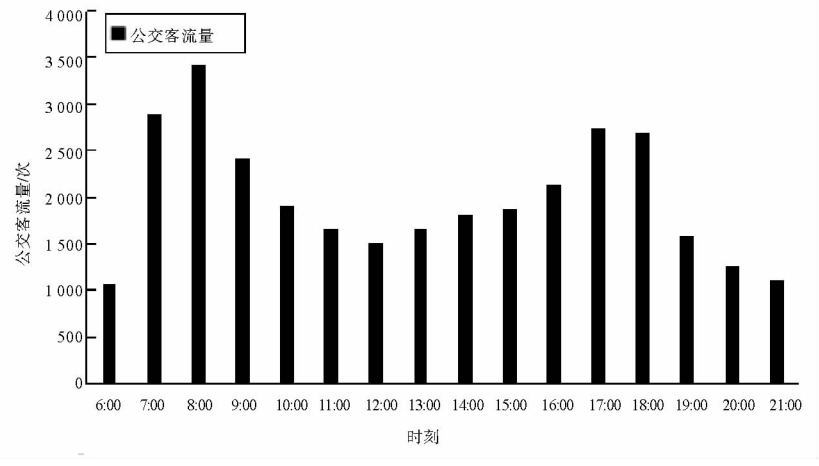

图 2　单日平均客流量数据图

Fig. 2　Data graph of daily bus passenger flow sequence

2　实例分析

2.1　基于 Fisher 算法的聚类分析

2.1.1　定义类的直径

设时段点为 $i,i=1,2,\cdots,n$,公交客流区间划分时段表示为 $G_{ij}=\{i,i+1,\cdots,j\}$,$i\leqslant j$,其均值向量元

素值

$$\overline{X}_{G_{ij}} = \frac{1}{j-i+1}\sum_{t=i}^{j} X_{(t)} 。 \tag{1}$$

式中 $X_{(t)}$ 表示 t 时段客流量所占百分比。用类的直径 $D(i,j)$ 表示 t 时段的公交客流量百分比 $X_{(t)}$ 到平均客流量百分比 $\overline{X}_{G_{ij}}$ 的距离,定义为

$$D(i,j)=\sum_{k=i}^{j}(X_{(t)}-\overline{X}_{G_{ij}})^2 。 \tag{2}$$

$D(i,j)$ 表示公交客流量不同时间段 $i,i+1,\cdots,j$ 内部变量之间的差异,值越小表示不同时段内公交客流量之间差异越小;反之,则表示不同时段内客流量差异越大,即公交出行的时间段越分散。根据公式(1)和(2)计算得到的相似时段类直径如表 1 所示。

表 1 相似时段类直径

Tab. 1 Similar hour diameter

i/j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2	1.65														
3	3.02	0.14													
4	3.02	0.50	0.50												
5	3.26	1.25	1.18	0.13											
6	3.65	2.05	1.82	0.30	0.03										
7	4.10	2.81	2.39	0.48	0.08	0.01									
8	4.29	3.17	2.63	0.52	0.09	0.02	0.01								
9	4.35	3.31	2.70	0.52	0.10	0.04	0.04	0.01							
10	4.38	3.40	2.73	0.52	0.12	0.08	0.08	0.02	0.00						
11	4.39	3.38	2.74	0.60	0.25	0.24	0.22	0.12	0.70	0.03					
12	4.86	3.73	3.19	1.29	1.06	1.06	0.98	0.74	0.56	0.41	0.19				
13	5.18	3.96	3.48	1.75	1.58	1.58	1.44	1.08	0.79	0.55	0.23	0.00			
14	5.47	4.34	3.79	1.94	1.74	1.74	1.63	1.36	1.16	1.03	0.89	0.86	0.60		
15	6.12	5.12	4.47	2.44	2.18	2.18	2.12	1.93	1.82	1.77	1.73	1.73	1.11	0.05	
16	6.92	6.04	5.28	3.05	2.73	2.72	2.69	2.57	2.52	2.51	2.51	2.45	1.50	0.11	0.01

2.1.2 定义损失函数

用 $b(n,k)$ 表示将 n 个样本分为 k 类的某一种分法,即将不同公交客流量划分为 k 类:

$$\begin{aligned} G_1 &= \{i_1, i_1+1, \cdots, i_2-1\}, \\ G_2 &= \{i_2, i_2+1, \cdots, i_3-1\}, \\ &\cdots \\ G_k &= \{i_k, i_k+1, \cdots, n\}。 \end{aligned} \tag{3}$$

定义上述分类法的损失函数为

$$e[b(n,k)]=\sum_{t=1}^{k} D(i_t, i_{t+1}-1) 。 \tag{4}$$

其中 $n=i_{k+1}-1$,当样本数 n 和划分数 k 给定时,总离差平方和一定;当类内平方和越小、类间平方和越大,分类越合理。Fisher 聚类算法的目的在于找到一种划分使误差函数 $e[b(n,k)]$ 达到最小,即求解 $\min\limits_{1=i_1<i_2<\cdots<i_k<n} e[b(n,k)]$。公式如下[11]:

$$e[b(n,2)]=\min_{2\leqslant j\leqslant n}\{D(1,j-1)+D(j,n)\},\tag{5}$$

$$e[b(n,k)]=\min_{k\leqslant j\leqslant n}\{e[b(j-1,k-1)]+D(j,n)\}。\tag{6}$$

将表 1 中的直径数据代入上述公式,求得最小的损失函数如表 2 所示。以表 2 中第三行第二列数据 2.70(4)为例,其意义为:有 5 个样本需划分为 3 种类型,存在 {1},{2},{3,4,5} 、{1},{2,3},{4,5} 、{1,2},{3},{4,5} 、{1},{2,3,4},{5} 、{1,2},{3,4},{5} 、{1,2,3},{4},{5} 共 6 种划分情况,计算 6 种情况下的损失函数为分别为 11.77、2.70、2,70、4.99、4.99、4.99,取损失函数最小的划分情况(即第二、第三种情况),最小值在 $j=4$ 时取得,记为 2.70(4)。

表 2 损失函数 $e[b(n,k)]$

Tab.2 The value of loss function $e[b(n,k)]$

k/n	2	3	4	5	6	7	8	9	10	11	12	13	14	15
3	1.40 (2)													
4	5.00 (2)	1.40 (4)												
5	12.48 (2)	2.70 (4)	1.30 (4)											
6	20.50 (2)	4.41 (4)	1.72 (5)	0.32 (5)										
7	28.10 (2)	5.84 (5)	2.25 (5)	0.84 (5)	0.12 (6)									
8	31.10 (2)	5.85 (5)	2.26 (5)	0.85 (5)	0.16 (6)	0.11 (7)								
9	31.21 (5)	5.97 (5)	2.37 (5)	0.97 (5)	0.45 (6)	0.16 (9)	0.11 (8)							
10	31.44 (5)	6.20 (5)	2.60 (5)	1.20 (5)	0.82 (6)	0.18 (9)	0.40 (9)	0.02 (9)						
11	32.78 (5)	7.40 (4)	3.94 (5)	2.53 (5)	0.97 (11)	0.78 (10)	0.18 (11)	0.14 (11)	0.02 (11)					
12	37.35 (2)	14.27 (4)	7.40 (12)	3.94 (12)	1.20 (12)	0.97 (12)	0.78 (12)	0.18 (12)	0.14 (12)	0.02 (12)				
13	39.62 (2)	18.86 (4)	7.43 (12)	3.96 (12)	1.22 (12)	0.99 (12)	0.80 (12)	0.20 (12)	0.16 (12)	0.05 (12)	0.02 (12)			
14	43.44 (2)	20.76 (4)	15.11 (11)	7.43 (14)	3.94 (14)	1.20 (14)	0.99 (14)	0.80 (14)	0.20 (14)	0.16 (14)	0.05 (14)	0.02 (14)		
15	51.24 (2)	25.75 (4)	19.38 (14)	7.95 (14)	3.96 (15)	1.74 (14)	1.22 (15)	0.99 (15)	0.72 (15)	0.20 (15)	0.16 (15)	0.05 (15)	0.02 (15)	
16	52.94 (14)	31.94 (4)	19.99 (14)	8.56 (14)	4.06 (15)	2.35 (14)	1.31 (15)	1.09 (15)	0.90 (15)	0.30 (15)	0.20 (15)	0.15 (15)	0.05 (15)	0.02 (16)

2.1.3 确定最优分类数 k

由于 Fisher 聚类算法本身并未给出合适的划分类数 k，目前常用的 k 值确定方法是曲线法和 β 检验法，本研究采用两种方法结合的方式确定 k 值[13]。

1) 曲线法。通过绘制最小误差函数 e 随划分类数 k 变化的曲线，取该曲线拐弯处或开始变平处对应的划分类数 k 作为最适宜的分类数，根据图 3 确定分类数为 6 或者 7 较为合适。

2) β 检验法。通过计算 β 值确定分类数，

$$\beta = \frac{e[b(n,k)]}{e[b(n,k+1)]} 。 \tag{7}$$

当 β 值较大时，说明划分为 $k+1$ 类比 k 类要好；β 值接近于 1 时，则可以认为该划分数类 k 为最优分类数。通过计算得到的不同 k 值对应的 β 值如表 4 所示。当 $k=6$ 时对应的 β 值最小且接近于 1，因此，确定分类数为 6。

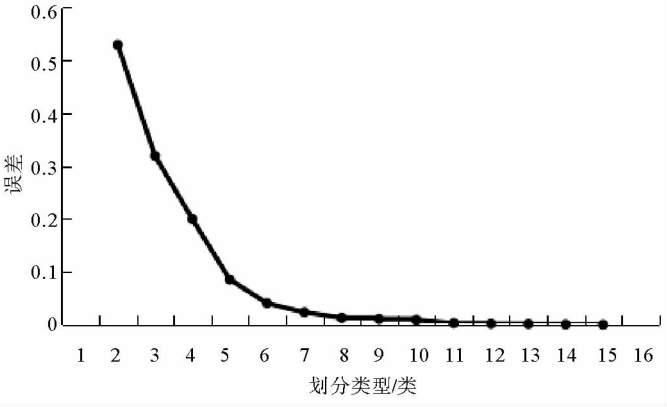

图 3 损失函数变化曲线图

Fig. 3 Variation curve of loss function

表 4 不同 k 值对应的 β 值

Tab. 4 The value of β corresponding to different k values

k	4	5	6	7
β	2.335	2.108	1.727	1.793

在损失函数 e 和 β 值都变小的情况下，综合两种方法的分类结果，最终确定最优分类数 $k=6$。

2.1.4 最优分类的确定

将样本划成 k 类，首先确定 j_k 使得 $e[b(n,k)]$ 达到极小值，即满足

$$e[b(n,k)] = e[b(j_k-1,k-1)] + D(j_k,n) 。 \tag{8}$$

得到第 k 类 $G_k = \{i_k, i_k+1, \cdots, n\}$。进而找到 j_k-1，使其满足

$$e[b(j_{k-1},k-1)] = e[b(j_{k-1}-1,k-2)] + D(j_{k-1},k-1) 。 \tag{9}$$

得到第 $k-1$ 类 $G_{k-1} = \{i_{k-1}, i_{k-1}+1, \cdots, i_k-1\}$。依次可得到所有类，最终得到最优分类集合 $b(n,k) = \{G_1, G_2, \cdots, G_k\}$。

根据表 2 结果，将公交客流量时刻划分为 6 时、7 时至 8 时、9 时至 16 时、17 时至 18 时、19 时、20 时至 21 时共 6 种时段类型。聚类算法的意义在于降低相似时段公交客流量的差异性，认为划分时段内的公交客流量是相等的，因此，初始不平稳的公交客流量图转为划分后的流量图(如图 4)，根据图 4 所示的数据流量进行预测。

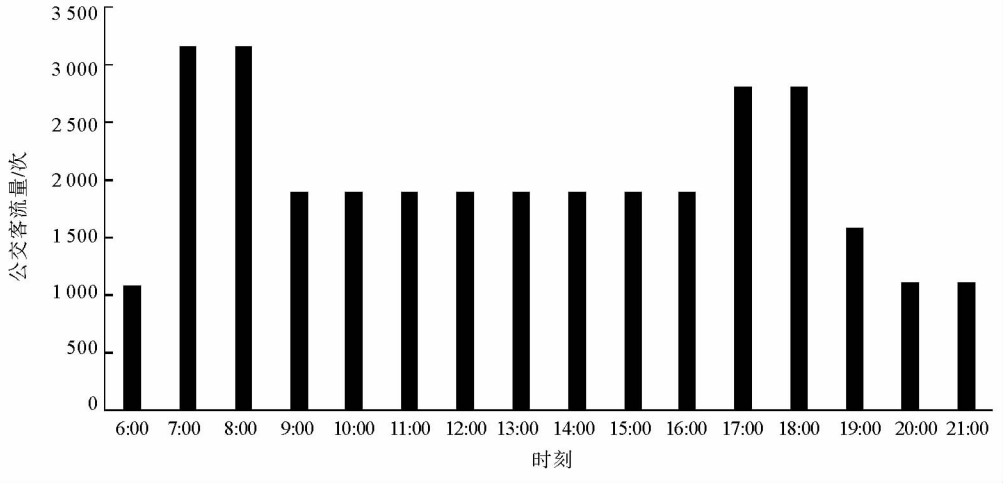

图 4 划分后的公交客流量时序图

Fig. 4 Time series of bus traffic flow after it is divided

2.2 基于时间序列模型的预测实证

2.2.1 自回归差分移动平均模型(ARIMA)

自回归移动平均模型(ARMA)模型的基本思想是:将预测对象随时间推移形成的序列视为一个随机序列,用数学模型来近似描述[14],该模型描述的序列为平稳序列。根据公交客流量原始数据图得知,该序列不具有平稳性,通过一阶差分后该序列具有平稳性,因此,差分阶数 $d=1$。假设随机变数 y_t 是在时间 t 的一个观测值,则 y_t 所构成的数列就成为随机序列,一般标准的 $ARIMA(p,1,q)$ 模型可以记为 $y_t \sim ARIMA(p,1,q)$,定义为

$$\varphi_p(B)W_t = \theta_q(B)a_t \text{。} \tag{10}$$

式中:$\varphi_p(B)=1-\varphi_1 B-\cdots\varphi_p B^p$,$W_t=(1-B)y_t$,$\theta_q(B)=1-\theta_1 B-\theta_2 B^2-\cdots-\theta_q B^q$,$a_t$ 为白噪音,p、q 为非负整数,B 为后移分算子,即 $By_t=y_{t-1}$,$\varphi_1,\varphi_2,\cdots,\varphi_p$ 为自回归参数,$\theta_1,\theta_2,\cdots,\theta_q$ 为移动平均参数。

运用 Eviews 软件对公交客流量进行预测,具体步骤如下:首先,构造原始序列的变化图像;其次,根据图像的变化趋势,对序列进行平稳化处理;再次,通过判断序列自相关系数和偏自相关系数以及模型的特点来确定阶数 p,q,从而得到 $ARIMA(p,d,q)$,值得注意的是,在模型识别的过程中可能会存在不同阶数的模型,一般的,采用 AIC 最小准则来选择模型阶数;最后,运用最小二乘法对其进行参数估计,为确保所选模型符合实际要求,需要进行模型检验,一方面检验参数的估计值是否具有显著性,另一方面检验模型残差是否为白噪声,检验后本文确定的模型为 $ARIMA(2,1,2)$。模型预测结果对比如表 5 所示。

表 5 ARIMA 模型预测结果对比

Tab. 5 Comparison of ARIMA model prediction results

日期	实际数据	预测值
12 月 27 日	1 779.375	2 207.522
12 月 28 日	1 884.000	2 093.287
12 月 29 日	2 107.750	1 894.661
12 月 30 日	1 902.500	1 963.805
12 月 31 日	2 336.500	1 981.117

2.2.2 季节性自回归差分移动平均模型(SARIMA)

当时间序列数据既存在趋势性又存在周期性时,可通过逐期差分使序列趋于平稳化。因此,选择季节性自回归差分移动平均模型,即 $SARIMA(p,d,q)(P,D,Q)_S$ 形式进行预测,其中 D 表示季节差分的阶数,P、Q 分别表示季节性的自回归和移动平均阶数,S 表示季节周期[15]。模型公式如下:

$$\varphi_p(B)\Phi_P(B)\nabla^d\nabla_S^D y_t = \theta_q(B)\Theta_Q(B)a_t \text{。} \tag{11}$$

其中:$\varphi_p(B)$、$\Phi_P(B)$ 分别表示非季节与季节自回归多项式;$\theta_q(B)$、$\Theta_Q(B)$ 则表示非季节与季节移动平均系数多项式,y_t 为 t 时刻的观测值,a_t 为白噪声。季节性自回归差分移动平均模型的实证步骤与自回归差分移动平均模型的方法相类似,只是增加了季节性(周期性)因子以及差分阶数,在经过行图形的绘制、平稳化处理、模型识别、模型检验 4 个过程后,通过 Eviews 确定最终的预测模型为 $SARIMA(1,1,1)(1,1,1)_7$,模型预测结果对比如表 6 所示。

表 6 SARIMA 模型预测结果对比

Tab. 6 Comparison of SARIMA model prediction results

日期	实际数据	预测值
12 月 27 日	1 779.375	2 031.853
12 月 28 日	1 884.000	2 030.335
12 月 29 日	2 107.750	2 030.261
12 月 30 日	1 902.500	2 033.169
12 月 31 日	2 336.500	2 033.251

2.3 模型的预测误差

为了判断时间序列模型的预测效果,选用平均绝对误差(mean absolute error,MAE)和平均绝对百分比误差(mean absolute percent error,MAPE)来量化预测结果的好坏,误差计算公式如下:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i-\hat{Y}_i|\times 100\% \text{,} \tag{12}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| \times 100\% \text{ 。} \tag{13}$$

式中 Y_i 为 i 时刻的实际公交客流量值，$\hat{Y}_i$ 为 i 时刻的预测公交客流量值，n 为预测的小时数。各个模型的预测误差结果如表 7。

本研究运用平均绝对误差(MAE)和平均绝对百分比误差(MAPE)对上述两个模型的预测效果进行对比(如表 7 所示)能够发现：在方法的选择上，季节性自回归差分移动平均模型的预测效果最好，季节性自回归差分移动平均模型比自回归差分移动平均模型的平均绝对误差(MAE)降低了 4%左右，平均绝对相对误差降低了 1%左右，这是由于本文所选择的数据具有明显的周期性。因而，运用季节性自回归差分移动平均模型提取周期因子，进而对原始序列进行预测更为合适。

表 7 预测模型误差结果

Tab. 7 Predictive model error results %

时间序列模型	MAE	MAPE
自回归差分移动平均模型	15.84	7.63
季节性自回归差分移动平均模型	11.37	6.09

3 讨论

目前，运用时间序列模型对原始交通客流量数据进行预测较为常见，为了验证 Fisher 聚类算法的有效性，本研究又分别利用两种时间序列模型对未聚类的客流量数据进行预测，数据选择仍为 10 月 11 日至 12 月 31 日 6 时至 21 时的公交客流量数据，其模型的预测结果如表 8 所示。

表 8 聚类算法前后对比误差结果

Tab. 8 Results of comparison error before and after clustering algorithm %

时间序列模型	平均绝对误差(MAE)		平均绝对百分比误差(MAPE)	
	未进行聚类算法	聚类算法	未进行聚类算法	聚类算法
自回归差分移动平均模型	18.37	15.84	12.84	7.63
季节性自回归差分移动平均模型	15.72	11.37	9.12	6.09

结果表明：对原始序列进行聚类划分能够有效提高预测精度。从平均绝对误差(MAE)的结果来看，进行聚类划分后，每种预测模型的误差结果大约能降低 4%左右；从平均绝对百分比误差(MAPE)的结果来看，每个预测模型的精度能提高 3%至 6%不等，说明对原始序列进行数据处理，能够有效提高预测精度。

4 结论

为提高交通客流量的预测效果，对初始数据进行处理和分析显得尤为重要。本研究从原始数据着手，利用 Fisher 聚类算法，将具有周期差异的时间序列数据进行聚类划分，降低了原始数据的波动性以及预测的复杂性。在预测模型的方法选择上，通过比较平均绝对误差和平均绝对相对误差的大小，发现季节性自回归差分移动平均模型(SARIMA)的预测效果最好。对于具有周期波动性的数据序列，选择含有周期因子的季节性自回归差分移动平均模型进行预测，预测精度较高；当原始序列的周期性不显著时，运用自回归差分移动平均模型进行预测也能够达到很好的预测效果，这为以后运用时间序列模型进行预测提供简便的选择方法；在数据处理上，利用 Fisher 聚类算法能够有效划分数据时段类型，对聚类后的数据进行预测能够有效提高预测精度。

参考文献：

[1]LIU X, HUAN X. Prediction of passenger flow at Sanya Airport based on combined methods[J]. Data Science, 2017, 727: 729-740.

[2]LIU X, HUANG X. Prediction for passenger flow at the airport based on different models[J]. Parallel Architecture, Algorithm and Programming, 2017, 729: 25-40.

[3]KUMAR S V, VANAJAKSHI L. Short-term traffic flow prediction using seasonal ARIMA model with limited input data [J]. European Transport Research Review, 2015, 7(3): 1-9.

[4]成诚,杜豫川,刘新. 考虑节假日效应的交通枢纽客流量预测模型[J]. 交通运输系统工程与信息,2015,15(5):202-207.
CHENG Cheng, DU Yuchuan, LIU Xin. A passenger volume prediction model of transportation hub considering holiday effects[J]. Journal of Transportation Systems Engineering and Information Technology, 2015, 15(5): 202-207.

[5]蔡昌俊,姚恩建,王梅英,等. 基于乘积 ARIMA 模型的城市轨道交通进出站客流量预测[J]. 北京交通大学学报,2014,38(2):135-140.
CAI Changjun, YAO Enjian, WANG Meiying, et al. Prediction of urban railway station's entrance and exit passenger flow based on multiply ARIMA model[J]. Journal of Beijing Jiaotong University, 2014, 38(2): 135-140.

[6]王莹,韩宝明,张琦,等. 基于 SARIMA 模型的北京地铁进站客流量预测[J]. 交通运输系统工程与信息,2015,15(6):205-211.
WANG Ying, HAN Baoming, ZHANG Qi, et al. Forecasting of entering passenger flow volume in Beijing subway based on SARIMA Model[J]. Journal of Transportation Systems Engineering and Information Technology, 2015, 15(6): 205-211.

[7]马佳羽,韩兆洲. 复杂季节时间序列模型研究[J]. 统计与决策,2017(6):27-30.
MA Jiayu, HAN Zhaozhou. Time series data with complex seasonal periods studies[J]. Statistics & Decision, 2017(6): 27-30.

[8]周辉宇. 基于大数据规则挖掘的交通拥堵治理研究[J]. 统计与信息论坛,2017,32(5):96-101.
ZHOU Huiyu. Study on traffic congestion based on big data rule mining[J]. Statistics & Information Forum, 2017, 32(5): 96-101.

[9]徐晓伟,杜一,周园春. 基于多源出行数据的居民行为模式分析方法[J]. 计算机应用,2017,37(8):2362-2367.
XU Xiaowei, DU Yi, ZHOU Yuanchun. Resident behavior model analysis method based on multisource travel data[J]. Journal of Computer Applications, 2017, 37(8): 2362-2367.

[10]闫伟,刘云岗,王桂华,等. 基于数据挖掘的交通流预测模型[J]. 系统工程理论与实践,2010,30(7):1320-1325.
YAN Wei, LIU Yungang, WANG Guihua, et al. Data mining using in a novel traffic flow forecasting model[J]. Systems Engineering: Theory & Practice, 2010, 30(7): 1320-1325.

[11]李林波,姜屿,王静,等. 基于数据融合的公交客流规模测算方法[J]. 城市交通,2016,14(1):43-50.
LI Linbo, JIANG Yu, WANG Jing, et al. Passenger volume estimation based on data fusion[J]. Urban Transport of China, 2016, 14(1): 43-50.

[12]刘自强,王效岳,白如江. 基于时间序列模型的研究热点分析预测方法研究[J]. 情报理论与实践,2016,39(5):27-33.
LIU Ziqiang, WANG Xiaoyue, BAI Rujiang. Analysis and forecasting method of research hotspots based on time series model[J]. Information Studies: Theory & Application, 2016, 39(5): 27-33.

[13]徐潇. 基于客流预测的公交调度优化研究[D]. 郑州:郑州大学,2017.

[14]刘建军. 几种时间序列模型在客流量预测上的比较[D]. 武汉:武汉邮电科学研究院,2015.

[15]乔占俊. 基于季节调整的区域中长期负荷预测[J]. 统计与决策,2014(21):83-85.
QIAO Zhanjun. Regional mid-and-long term load forecast based on seasonal adjustment[J]. Statistics & Decision, 2014(21): 83-85.

(责任编辑:傅 游)