Operations Research and Fuzziology 运筹与模糊学, 2016, 6(1), 15-26
Published Online February 2016 in Hans. http://www.hanspub.org/journal/orf
http://dx.doi.org/10.12677/orf.2016.61003

Hans汉斯

Prediction of Shanghai Metro Line 16 Passenger Flow Based on Time Series Analysis

—with Lingang Avenue Station as a Study Case

Yanli Chen[1], Yuwu Sha[2], Xiaolin Zhu[1], Xiaohong Zhang[1]

[1]College of Arts and Sciences, Shanghai Maritime University, Shanghai
[2]Ministry of Operational Scheduling, Shanghai Maglev Transportation Development Co. Ltd., Shanghai
Email: chenyl_1119@foxmail.com

Received: Jan. 27th, 2016; accepted: Feb. 17th, 2016; published: Feb. 23rd, 2016

Abstract

Problems emerge along with the continuous development of urban rail transit, and how to predict the passenger flow to improve the efficiency of the rail transit operation by the scientific method has caused widely public concern. Time series analysis is the mainstream of forecasting method. And ARIMA model acts on all kinds of sequences, so it is the most common time series prediction method by far. This study proposes Autoregressive Integrated Moving Average Model (ARIMA model) to predict the passenger flow data of the line 16 Lingang Avenue Station based on the historical datum through time series analysis in order to improve the operational efficiency of the urban rail transit and effective cohesion with buses in Lingang area. We utilize the autocorrelation and partial auto-correlation function to preliminarily judge and identify the parameters of ARIMA model.

Keywords

Urban Rail Transit, Time Series Analysis, ARIMA Model

基于时间序列分析的上海地铁16号线客流预测

—以临港大道站为例

陈彦莉[1]，沙玉五[2]，朱小林[1]，张小红[1]

文章引用: 陈彦莉, 沙玉五, 朱小林, 张小红. 基于时间序列分析的上海地铁 16 号线客流预测[J]. 运筹与模糊学, 2016, 6(1): 15-26. http://dx.doi.org/10.12677/orf.2016.61003

[1]上海海事大学文理学院，上海
[2]上海磁浮交通发展有限公司客运服务部，上海
Email: chenyl_1119@foxmail.com

收稿日期：2016年1月27日；录用日期：2016年2月17日；发布日期：2016年2月23日

摘　要

随着不断发展城市轨道交通建设也不断涌现诸多问题，如何以科学手段来预测客流，从而提高轨道交通运营的效率引起广泛关注。时间序列分析是主流的预测方法，其中ARIMA模型适用于各类的序列，是迄今最通用的时间序列预测法。本文将以上海地铁16号线临港大道站为例，对其日客流通过时间序列分析方法，建立差分自回归移动平均模型(ARIMA模型)，利用自相关函数和偏自相关函数来初步判断和识别ARIMA模型各个参数，并根据所建立的模型来预测16号线临港大道站后两周客流数据，以此为提高临港地区城市轨交运营效率，改善临港地区地铁与公交高效衔接建立基础。

关键词

城市轨交，时间序列，ARIMA模型

1. 引言

目前，我国的城市化进程已经进入到城市加速发展阶段，城市人口急剧增加，城市中心区的高密度开发和人口的高度集中，使得交通出行总量剧增。其中由于城市边缘和远郊城市化地区的发展，将出现大量新的长距离的出行需求。地铁，作为绝大多数的城市轨道交通系统，都会逐渐成为城市交通的骨干。因此，由公交运营系统过渡到城市轨道交通而产生的诸多问题也应运而生。

上海轨道交通 16 号线，北起龙阳路站，南至滴水湖站。全长 58.96 公里，其中地下线长约 13.74 公里，高架线长约 45.22 公里，共设车站 13 座。16 号线选择目前国内最快的 120 公里时速的技术路线。2012 年 11 月底，轨道 15-2 标上下行全部贯通，由龙阳路站至滴水湖站[1]。2014 年 12 月 28 日，龙阳路——罗山路区段开通。而自从 16 号线开通以来，虽然在一定程度上缓解近郊的交通压力，但也有新问题不断暴露出来：

16 号线运行后不久，周边原先的公交线路都有所调整，使其周边沿线绝大多数的客流都转移至 16 号线。不仅如此，由于原先设计方案为观光线路，因此对于仅有 3 节车厢，且每节车厢座位类似于公交车的设计安排，使 16 号线更不堪重负。

举实例说，浦东新区、南汇地区许多高校学生原先大多乘坐公交回校以及进市区，但由于 16 号线的建成，一些公交线路相继取消或剧减班次，学生客流很大程度都转移到地铁线路上。除了学生客流，周边有固定工勤时间规定的上班族一来由于公交线路的改变，二来由于地铁的准时快捷，主观上也倾向乘坐地铁。这无疑对于 16 号线是一个重大考验。有反映 16 号线要室外排队 40 分钟，排队长龙在道路上颇为震撼，被不少市民称为上海目前“最挤的地铁”。

而临港地区高校学生也是排队长龙中的一员，因此本文将针对临港新城区域高校学生出行活动为研究对象，对现行的 16 号线客流进行研究，我们发现临港地区学生出行所选择的地铁站为临港大道和滴水湖，由于滴水湖站点会有许多出游乘客造成干扰，故本文选取临港大道站作为研究站点，并希望对其站点每日客流数据进行预测。由于现主流的客流预测手段为进行时间序列分析，建立 ARIMA 模型[2]-[4]，

故本文将借以这些手段，对临港大道站点客流进行预测，为 16 号线运营以及周边公交接驳配套运营方案的改进奠定基础。

2.ARIMA 模型

所谓的时间序列，是指一个依时间顺序组成的观察数据的集合。而时间序列分析，是将预测对象随时间推移而形成的数据序列视为一个随机序列，用一定的数学模型来近似描述这个序列。这个模型一旦被识别后就可以从时间序列的过去值及现在值来预测未来值。序列的变化趋势可以分为平稳的和非平稳的。从直观上来讲，平稳性时间序列是序列观测值围绕平均值上下小范围波动的序列。对于非平稳的时间序列，首先看其序列图的变化趋势，如果呈近似直线上升，则一阶差分可令其平稳，若是呈波动上升变化，一般二次差分即可，若呈指数型上升，则对其对数差分也可使其平稳。当把非平稳过程转成平稳过程后，即可按照平稳时间序列的分析方法处理该序列。

移动平均法、指数平滑法师早期时间序列分析的主流方法，随着计算机科技的发展与普及，差分自回归移动平均模型(Autoregressive Integrated Moving Average Model, ARIMA 模型)被广泛应用于时间序列分析之中。

ARMA 模型是由自回归模型(简称 AR 模型)与滑动平均模型(简称 MA 模型)为基础“综合”构成。传统的趋势模型外推预测方法只适合于具有某种典型趋势性变化现象的预测，然而在现实中，许多现象的序列资料并不总是具有这种典型趋势特征，依此方法建立的模型所产生的误差项不一定完全是具有随机性质的，从而影响了预测效果．ARIMA 模型是将序列先进行差分从而转化为 ARMA 模型，根据序列识别一个试用模型，再加以诊断，做出必要调整，反复进行识别、估计、诊断，直到适合的模型，因此它适用于各类的序列，是迄今最通用的时间序列预测法[5]。

在介绍 ARIMA 模型之前，本文首先要介绍一个特殊的序列——白噪声序列。白噪声序列是一种特殊的平稳序列。我们定义：若随机序列 $\{X_t\}$ 由互不相关的随机变量构成，在不同时点上的随机变量的协方差为 0，即对所有 $s \neq t$， $Con(X_t, X_s) = 0$，则称其为白噪声序列。有该性质的时间序列意味着人们无法根据其过去的特点推测其未来的走向，其变化没有规律可循。当模型的残差序列成为白噪声序列时，可认为模型达到了较好的效果，即剩余残差中已经没有可以识别的信息。

ARMA [6]模型的一般形式如下：

$$y_t - \varphi_1 y_{t-1} - \varphi_2 y_{t-2} - \cdots - \varphi_p y_{t-p} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

式中：$y_t - \varphi_1 y_{t-1} - \varphi_2 y_{t-2} - \cdots - \varphi_p y_{t-p}$ 表示模型的自回归部分；$\varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$ 表示模型的移动平均部分； y_t 称为 ARMA (p,q) 序列，非负整数 p,q 分别称为自回归阶数和移动平均阶数，参数 $\varphi_1, \varphi_2, \cdots, \varphi_p$ 称为自回归系数， $\theta_1, \theta_2, \cdots, \theta_q$ 称为移动平均系数。

当 $p = 0$ 时，则 ARMA $(0,q)$ 模型

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

称为 q 阶移动平均模型，记为 MA(q)。

当 $q = 0$ 时，则 ARMA $(p,0)$ 模型

$$y_t - \varphi_1 y_{t-1} - \varphi_2 y_{t-2} - \cdots - \varphi_p y_{t-p} = \varepsilon_t$$

称为 p 阶自回归模型，AR(p)。

为了方便地表示时间序列的滞后项，以下定义刻画其性质的表示方法即延迟算子。延迟算子类似于一个时间指针，当前序列值乘以一个延迟算子，就相当于把当前序列值的时间向过去拨了一个时刻，记 B^k 为 k 步延迟算子，即 $B^k y_t = y_{t-k}$， $B^k \varepsilon_t = \varepsilon_{t-k}$， $B^k c = c$ (c 为常数)，并令

$$\varphi(B)=1-\varphi_1 B-\varphi_2 B^2-\cdots-\varphi_p B^p$$

$$\theta(B)=1-\theta_1 B-\theta_2 B^2-\cdots-\theta_p B^p$$

则 ARMA(p,q) 模型简记为 $\varphi(B)y_t=\theta(B)\varepsilon_t$

当时间序列数据存在趋势性，我们需要通过差分处理使该序列趋于平稳化，这样的时间序列被称为准平稳序列，相应的模型为 ARIMA(p,d,q) 模型，其形式为：

$$\begin{cases}\varphi(B)\nabla^d y_t=\theta(B)\varepsilon_t\\ E(\varepsilon_t)=0, Var(\varepsilon_t)=\sigma_\varepsilon^2\\ E(\varepsilon_t\varepsilon_s)=0, s\neq t\\ Ey_s\varepsilon_t=0, (\forall s<t)\end{cases}$$

其中：

$$\varphi(B)=1-\varphi_1 B-\varphi_2 B^2-\cdots-\varphi_p B^p$$

$$\theta(B)=1-\theta_1 B-\theta_2 B^2-\cdots-\theta_p B^p$$

当时间序列数据既存在趋势性又存在周期性时，可通过逐期差分和季节差分使序列趋于平稳化，其可以采用 ARIMA(p,d,q) (P,D,Q) 模型。其中 P,D,Q 分别表示以 S 为间距的自回归、差分、和移动平均阶数，S 为季节周期，模型公式如下：

$$\Phi(B)\Phi_S(B)\nabla^d\nabla_S^D y_t=\Theta(B)\Theta_S(B)\varepsilon_t$$

3. 临港大道站点客流预测模型

3.1. 平日客流预测模型

本文以上海地铁 16 号线 2014 年 12 月 28 日至 2015 年 8 月 31 日的每日进站客流量进行分析，刻画出其序列图，见图 1。

从图中我们可以明显看出，其数据变动趋势在 2015 年 2 月 7 日至 2015 年 3 月 4 日、7 月 23 日至 8 月 31 日，两个时间段出现了明显的低谷值，而在 12 月 31 日、4 月 3 日、4 月 30 日出现峰值，经调查研究本文发现其低谷值为临港大学城寒暑假时间，出现峰值的时刻都是国定假日前一天，则本站点的客流趋势完全受到临港大学城区高校学生生活的影响。故本文将选取 2015 年 3 月 6 日至 2015 年 4 月 2 日这段时间段的客流来进行临港大道站平日客流量的研究。

图 2 和图 3 为临港大道 3 月 6 日至 4 月 2 日日客流量 $\{x\}$ 的自相关和偏自相关函数图，从此图中我们 l 可以确定此序列的变化趋势，并且大致确定 ARIMA 模型中的参数设定。我们可以看出序列的 ACF 拖尾衰减，且呈周期性，周期为 7，PACF 为拖尾衰减，故此序列为平稳序列，识别为混合模型，即 ARIMA 模型。进行反复尝试，我们可以判断对其进行 1 阶季节性差分可得平稳序列。

图 4 和图 5 为一阶季节性差分后所得到的自相关函数和偏自相关函数图，从图中我们可以看出序列 $\{x\}$的自相关系数和偏自相关系数都不具有统计学意义。

由此我们尝试拟合 ARIMA$(0,0,0)(0,1,0)$ 模型，通过 spss [7]中分析－预测－创建模型进行模型的建立，由此得到模型统计量以及残差的检验结果，如表 1 和图 6。

表 1 中 Ljung-Box 是用于检验某个时间段内观测值是否是随机、独立的，当 Ljung-Box 检验值大于 0.05，即说明其观测值为独立的，本模型的检验的结果我们从表中可知为 0.981 远远大于 0.05，则认为在 95%的置信水平下无法拒绝原假设，即不能显著拒绝原序列为纯随机序列(白噪声)的假定。我们可以得知

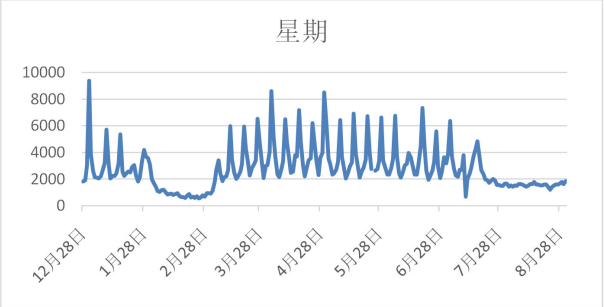

Figure 1. The sequence diagram of total passenger flow at Lingang Avenue station (2015.12.28-2016.8.31)
图 1. 临港总客流序列图(2015.12.28~2016.8.31)

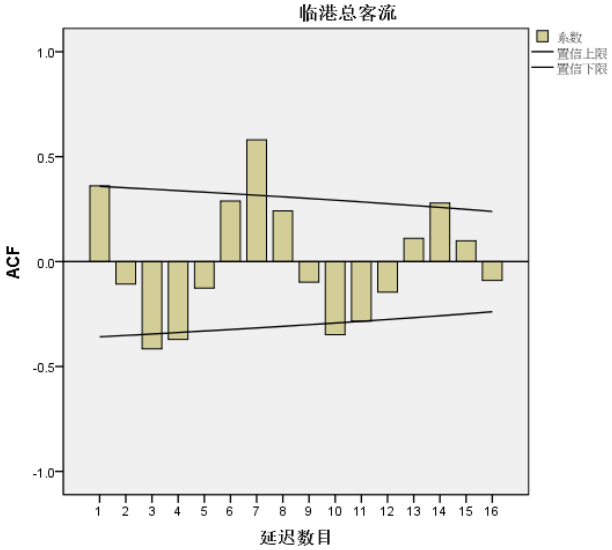

Figure 2. Autocorrelation Function of $\{x\}$
图 2. $\{x\}$的自相关函数

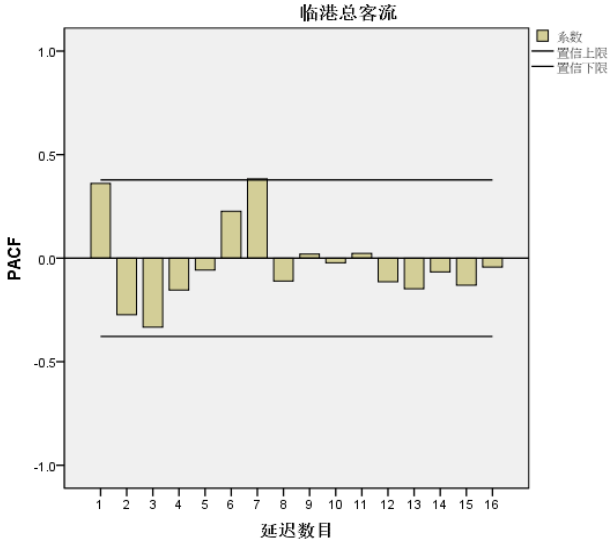

Figure 3. The partial autocorrelation function of $\{x\}$
图 3. $\{x\}$的偏自相关函数

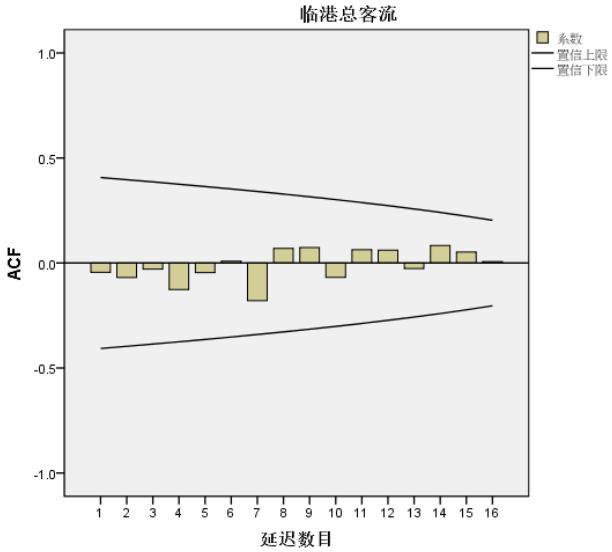

Figure 4. The autocorrelation function of $\{x\}$ first order seasonal difference
图 4. 一阶季节差分后$\{x\}$的自相关函数

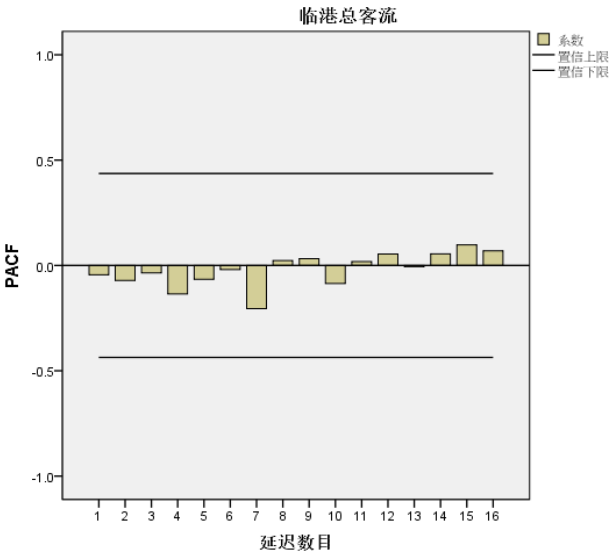

Figure 5. The partial autocorrelation function of $\{x\}$ first order seasonal difference
图 5. 一阶季节差分后$\{x\}$偏自相关函数

此模型的残差彼此独立，由此可以判断此模型的残差通过了白噪声的检验。图 6 可以更为直观地观察到此检验结果。

但是我们观察模型的参数检验发现，如表 2 所示，此模型并不是最为适合的模型。因此我们判定选取的观测值太少，并不能清晰准确的建立合适的模型，故后文我们选取了 5 月 2 日至 7 月 3 日两个月的时间序列$\{x'\}$。我们观察数据发现，由于 16 号线数据平台的记录问题造成 5 月 25 日那天缺数据，故本文将首先对数据进行预处理。本文利用 spss 中转换－替换缺失值的功能，对于此处数据进行了序列均值的替换，并基于此基础上再次进行 ARIMA 模型的建立，经过多次尝试，最终所得模型为 ARIMA(2,0,2)(0,1,1)，我们对模型的拟合程度进行检验，平稳的 R 方 = 0.579，正态化的 BIC = 12.472，

并从表 3 可知模型的决定系数为 0.181 > 0.05，故此模型较好的解释了原序列。

表 4 是针对 5 月至 7 月的序列建立季节性 ARIMA 模型进行模型参数检验的结果，从表中我们可以发现 AR 滞后 1 阶、滞后 2 阶以及 MA 滞后 2 阶的系数均通过了参数检验，因此此模型具有现实意义。此外，从图 7 我们可以直观的发现，此模型的残差通过了白噪声检验。

因为其模型结果较好，故本文利用 ARIMA(2,0,2)(0,1,1)对 16 号线客流数据进行了拟合和预测，从图 8 我们可以看出拟合值与原序列值拟合程度较好，能较为真实地反映原序列所表达的信息。

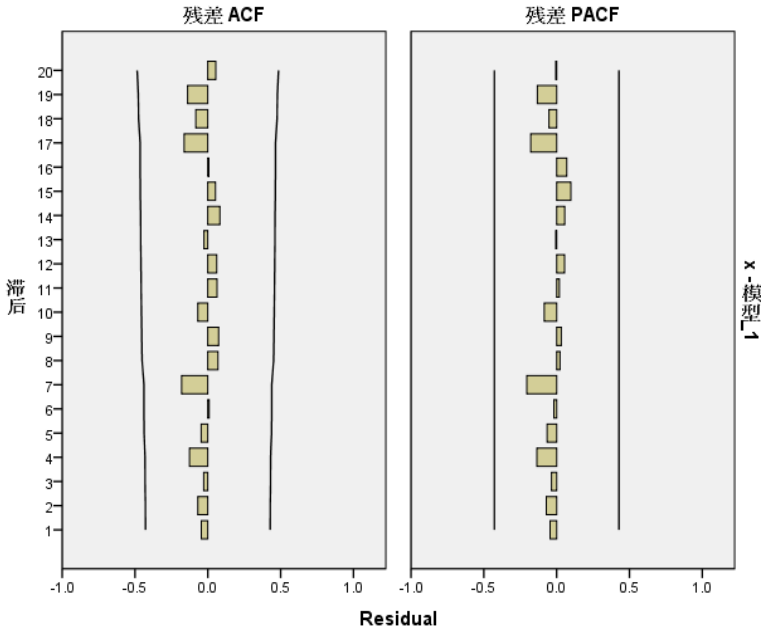

Figure 6. The test of ARIMA(0,0,0)(0,1,0) residual error by ACF and PACF
图 6. ARIMA(0,0,0)(0,1,0)残差 ACF 和 PACF 检验图

Table 1. The test of ARIMA(0,0,0)(0,1,0) model statistics
表 1. ARIMA(0,0,0)(0,1,0)模型统计量

模型	预测变量数	模型拟合统计量		Ljung-Box Q(18)			离群值数
		平稳的 R 方	MAE	统计量	DF	Sig.	
临港总客流-模型_1	0	1.105E-016	379.619	7.865	18	.981	0

Table 2. The test of ARIMA(0,0,0)(0,1,0) model parameters
表 2. ARIMA(0,0,0)(0,1,0)模型参数

				估计	SE	t	Sig.
临港总客流-模型_1	临港总客流	无转换	常数	455.571	151.383	3.009	.007
			季节性差分	1			

Table 3. The test of ARIMA(2,0,2)(0,1,1) model statistics (from May to July)
表 3. ARIMA(2,0,2)(0,1,1)模型统计量(5 月至 7 月)

模型	预测变量数	模型拟合统计量	Ljung-Box Q(18)			离群值数
		平稳的R方	统计量	DF	Sig.	
SMEAN(VAR00001)-模型_1	0	0.579	17.407	13	0.181	0

Table 4. The test of ARIMA(2,0,2)(0,1,1) model parameters (from May to July)
表 4. ARIMA(2,0,2)(0,1,1)模型参数(5 月至 7 月)

					估计	SE	t	Sig.
SMEAN(VAR00001)-模型_1	SMEAN(VAR00001)	无转换	常数		5.039	33.265	0.151	0.880
			AR	滞后 1	0.658	0.238	2.769	0.008
				滞后 2	−0.626	0.198	−3.156	0.003
			MA	滞后 1	0.217	0.196	1.106	0.275
				滞后 2	−0.747	0.143	−5.214	0.000
			季节性差分		1			
			MA，季节性	滞后 1	0.912	0.654	1.395	0.170

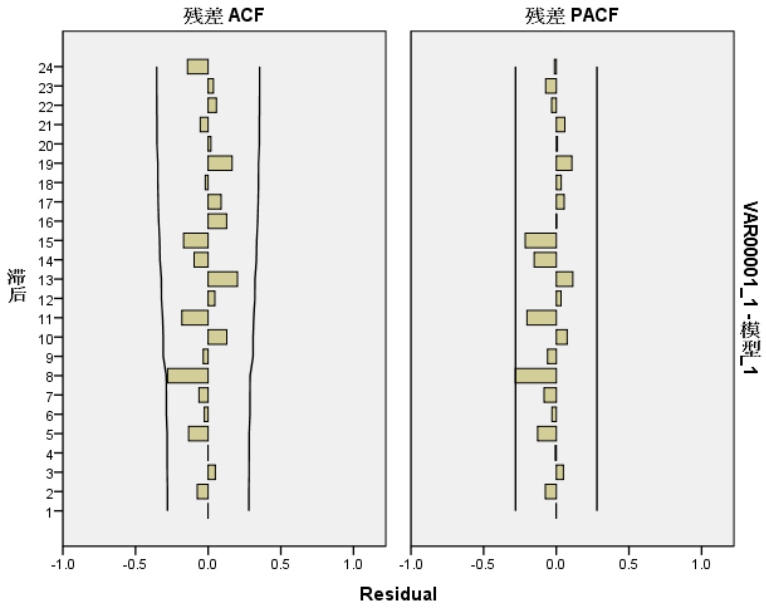

Figure 7. The test of ARIMA(2,0,2)(0,1,1) residual error by ACF and PACF (from May to July)
图 7. ARIMA(2,0,2)(0,1,1)残差 ACF 和 PACF 图

3.2. 只含有学生因素序列的 ARIMA 模型建立

从上文的分析，我们已知 16 号线临港大道站每日客流量变化趋势完全受临港地区周边大学生生活活动的影响，也就是说学生上课时期即可定义为含有学生因素以及周边客流因素序列，而暑期以及寒假期间，临港大道每日客流量变化只受周边地区居民影响。我们观察 16 号线临港大道站客流量时间序列图发现，暑期与寒假序列变化趋势稳定，故本文试想能否将平日客流剔除周边居民因素，以此研究只含有学生因素的 16 号线临港大道站客流量序列变化。后文将从此角度进行研究。

首先，我们选取临港大学城区学生上课的时间段：5 月 26 日至 6 月 29 日，以及暑假时期：7 月 28 日至 8 月 31 日。在选取时间段时，我们首先考虑到选取不涉及特殊情况(如跨度国定假期)的上学时间段，原因我们可以直观的从图 1 看出，在特殊情况下，临港大道站的客流量会出现激增，从而影响到后面的建模分析。其次，由于此站客流量变化趋势具有季节性，故选取的只含有周边居民影响的序列要与平时的时间序列所在周期节点一致，在随后的预处理后，才能得到更为合理的只含有学生因素的时间序列 $\{x_{revice}\}$。

要建立 ARIMA 模型，首先要对序列进行平稳性分析，本文利用 spss 对时间序列$\{x_{revice}\}$作其时间序列图，如图 9 所示。

通过直观的看序列的散点图并不能直接判断该序列是否平稳，但我们可以发现其序列具有周期性。因此本文通过作自相关和偏自相关图进行平稳性的进一步分析，所得结果如图 10 和图 11。

我们从图中可以发现自相关函数不拖尾，由此故此序列为非平稳序列，且表现出周期性变化，周期为 7。此外偏自相关函数也不截尾。故尝试对此序列进行平稳化，最终对此序列取自然对视并进行 1 阶季节差分，和 1 阶差分，从而得到 ARIMA(2,1,0)(1,1,0)模型。

从模型拟合程度来看，通常通过观察平稳的 R 方和正态化的 BIC 的值来进行拟合程度好坏的判断。平稳的 R 方用来比较模型中的固定成分与一个简单均值模型的差别，当原始序列中有季节成分时，其要优于 R 方统计量；而正态化的 BIC 用来度量模型拟合优度的同时还考虑了模型的复杂程度，每增加一个参数便会对其产生一个惩罚因子。此模型所得的平稳的 R 方为 0.51，正态化的 BIC 为 12.96。除此之外，我们从表 5 可以发现此模型解释了原序列 83.7%的信息，由此可知此模型拟合程度良好。

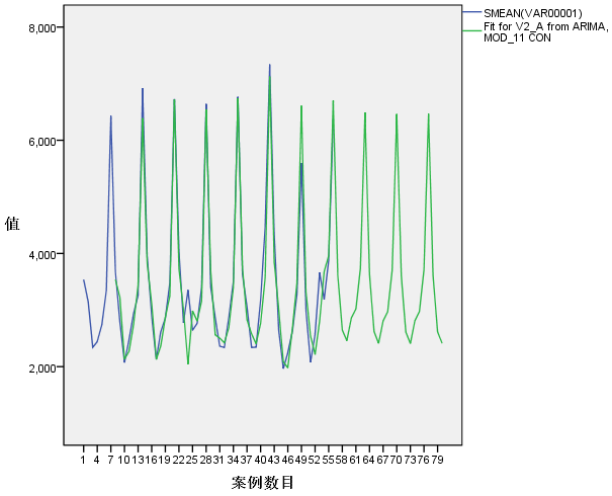

Figure 8. The fitting results
图 8. 拟合结果

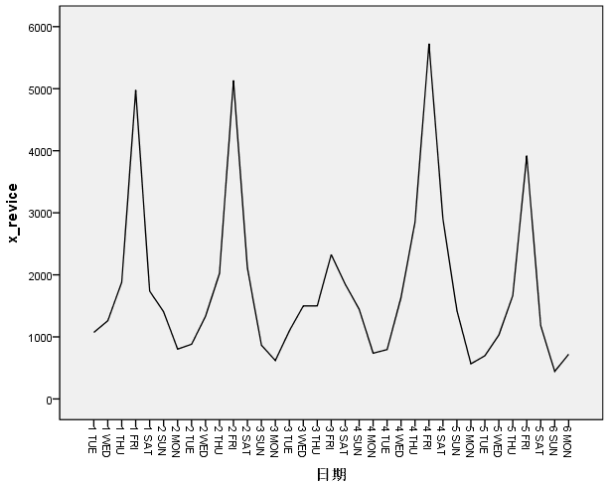

Figure 9. Time series plot of $\{x_{revice}\}$
图 9. $\{x_{revice}\}$时间序列图

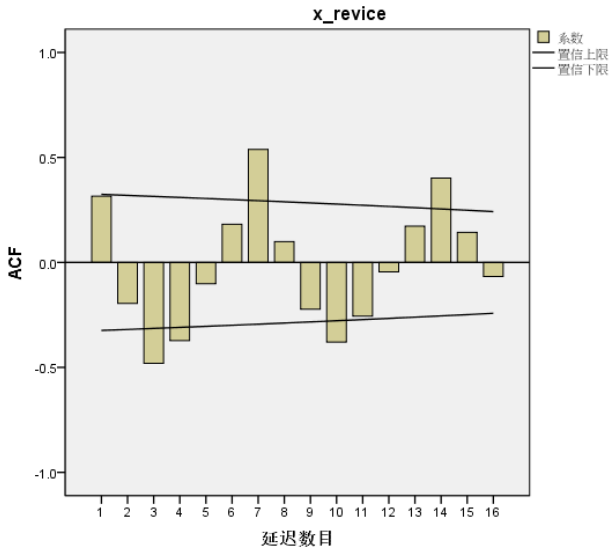

Figure 10. Autocorrelation function of $\{x_{revice}\}$
图 10. $\{x_{revice}\}$序列自相关图

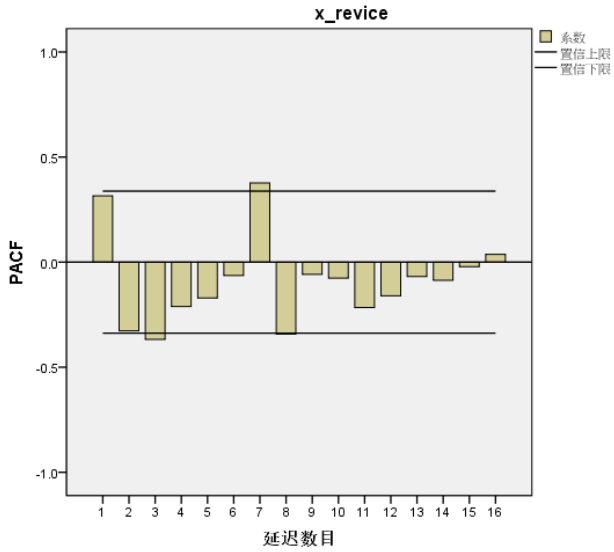

Figure 11. The partial autocorrelation function of $\{x_{revice}\}$
图 11. $\{x_{revice}\}$序列偏自相关图

Table 5. The test of ARIMA(2,1,0)(1,1,0) model statistics
表 5. ARIMA(2,1,0)(1,1,0)模型统计量

模型	预测变量数	模型拟合统计量	Ljung-Box Q(18)			离群值数
		平稳的R方	统计量	DF	Sig.	
x_revice-模型_1	0	0.508	10.536	16	0.837	0

从图 12 我们更可以直观的发现此模型的残差通过了检验，为白噪声序列。

表 6 给出了模型的参数检验，从表中我们可以发现其系数都通过了检验。由此我们可得模型为：

$$\left(1+0.753B^{2}\right)\left(1+0.563B^{7}\right)\nabla\nabla_{7}^{1}\log x_{revice}=\left(1+0.563B^{7}\right)\varepsilon_{t}$$

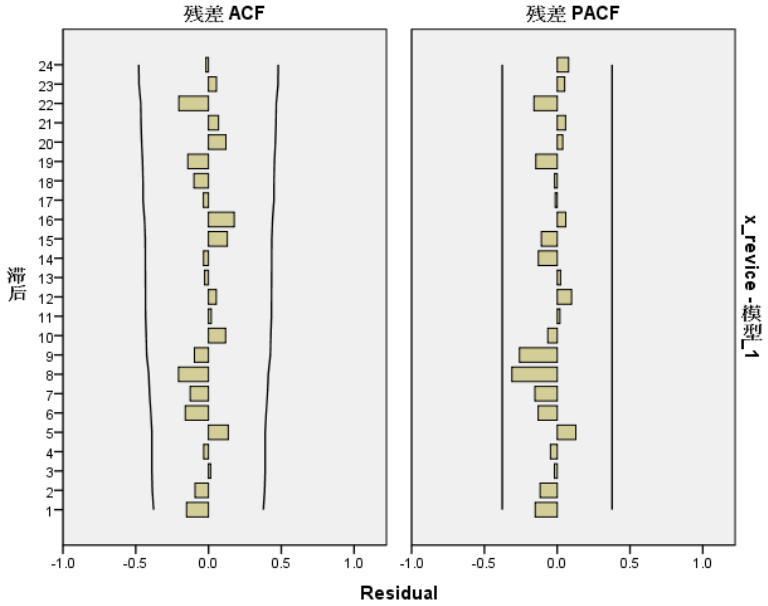

Figure 12. The test of ARIMA(2,1,0)(1,1,0) residual error by ACF and PACF
图 12. ARIMA(2,1,0)(1,1,0)残差 ACF 和 PACF 图

Table 6. The test of ARIMA(2,1,0)(1,1,0) model parameters
表 6. ARIMA(2,1,0)(1,1,0)模型参数

					估计	SE	t	Sig.
x_revice-模型_1	x_revice	自然对数	AR	滞后 2	−0.753	0.190	−3.960	0.001
			差分		1			
			AR，季节性	滞后 1	−0.563	0.220	−2.563	0.017
			季节性差分		1			

4. 结束语

城市轨道公交的建设随着社会的发展不断推进，但随着城市轨道交通的发展，不断出现的城市交通问题也不断涌现。本文通过对原有数据的选取，试图建立较为合理的 ARIMA 模型，并基于所构建的上海地铁 16 号线客流预测 ARIMA 模型，从而预测了后 2 周的客流数据，其次，由于临港大道站客流量的变化几乎受周围大学城区学生生活活动周期影响，故本文试图建立只含有学生因素的客流预测模型，此模型残差都通过了白噪声检验，并且解释了原时间序列 83.7%的信息，具有良好的适应性，从而对未来轨道交通规划以及周边配套公交安排具有较好的参考价值。

参考文献 (References)

[1] 百度百科: 上海地铁 16 号线[EB/OL]. http://baike.baidu.com/link?url=cMYPENKG_sTRZPvN6dhtD9CmzW0PMcTAAKf0ZHhiR3DcMZex9stMoL4ACt24LVCfkDLd5MXXe2u7rG10wyuuza

[2] 张杰, 刘小明, 贺玉龙, 陈永胜. 基于时间序列的我国铁路客流量预测[J]. 统计与咨询, 2008: 20-21.

[3] 裴武, 陈凤, 程立勤, 交通量时间序列 ARIMA 预测技术研究[J]. 山西科技, 2009(1): 75-79.

[4] 祁伟. 李晔, 汪作新. 季节性 ARIMA 模型在稀疏交通流下的预测方法[J]. 公路交通科技, 2014, 31(4): 130-135.

[5] 常国珍, 张前登. 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.

[6] 王燕. 应用时间序列分析[M]. 北京: 中国人民大学出版社, 2005: 64-68, 146.

[7] 杜强, 贾丽艳. SPSS 统计分析从入门到精通[M]. 北京: 人民邮电出版社, 2011.