计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200300099

使用 ARIMA 模型预测公园绿地面积

闫祥祥
对外经济贸易大学统计学院　北京 100029

摘　要　在时间序列中使用 ARIMA 模型是常见的分析预测方式之一。为了预测公园绿地面积,在其他预测模型优势不明显的情况下,最终选择 ARIMA 模型作为预测方法。文中调研并选取了北京市 1978—2017 年园林绿化及森林情况数据,在 SPSS 系统中,通过数据选择、描述性统计分析、自相关图平稳性检验、数据平稳性处理、模型检验等步骤最终确定适合采集数据的 ARIMA 模型,并在该模型上对 2018—2020 年的公园绿地面积进行预测。可视化和模型统计量等实验结果表明,该模型的拟合及预测效果良好。

关键词:ARIMA;SPSS;时间序列模型;预测;平稳性检验
中图法分类号　O211.61

Using ARIMA Model to Predict Green Area of Park

YAN Xiang-xiang
School of Statistics,University of International Business and Economics,Beijing 100029,China

Abstract　Using ARIMA model in time series is one of the common analysis and prediction methods. In order to predict the green area of the park,in the case where the advantages of other prediction models are not obvious,the ARIMA model is finally selected as the prediction method. The data of landscaping and forestry in Beijing from 1978 to 2017 are surveyed and collected. In the SPSS system,through the steps of data selection,descriptive statistical analysis,autocorrelation graph stationarity test,data stationarity processing,model test,etc. ,the ARIMA model suitable for data collection is finally determined,and to predict the green area of the park. Experimental results such as visualization and model statistics show that the model fits and predicts well.

Keywords　ARIMA,SPSS,Time series model,Prediction,Stationarity test

1　引言

随着城市经济的发展和城市居民生活水平的不断提高,人们的环保意识趋于成熟,对生活环境的期望和要求也越来越高。众所周知,城市绿化无论对于自然环境还是对于人们的生活都至关重要,它不仅可以改善城市的健康状况,产生良好的经济效益,而且有利于人们的身心健康。

但是,城市绿化所消耗的资金量巨大,如果在建设之前没有进行科学的预测分析,不仅会造成资金浪费,规划也会缺乏合理性和科学性[1]。因此本文探索的内容在生活方面不仅有实际意义,而且具有一定的前沿性。

ARIMA 模型作为时间序列分析的常见方法之一,主要是从时序数据自相关的角度反映其本身的发展规律,它本身十分简单,只与内生变量有关而无须借助外部变量。目前,ARIMA 模型由于其成熟的理论而成为时间序列分析中的经典模型,在实际的企业生产和生活中得到了广泛的应用[2]。笔者检索多篇文献资料,发现 ARIMA 模型在经济[3]、工业[4]、医学[5]等行业领域应用较多;在生活环境方面的应用相对偏少[6],因此本文研究也具有一定探索意义。

SPSS Statistics(SPSS)作为一款历史悠久、专注商业数据分析的软件,具有易学易用、分析功能多样等优点,目前被更多地应用在统计、预测等分析场景中。因此,结合数据情况和软件的特点等,本文选择在 SPSS 系统中对采集数据进行 ARIMA 模型的探索研究。

2　基本概念

2.1　时间序列

时间序列是指某一现象的统计指标值按时间顺序所排列的数据。其主要目的是基于随机过程理论(Stochastic process theory)和数理统计方法(Mathematical statistics)研究随机时序数据所遵从的统计规律,对未来进行预测,用于解决实际问题。

2.2　平稳性

分析任何问题之前都要做一定的基本假设。在时间序列分析中,保证数据的平稳性是在进行后续时序分析步骤之前的必要前提。对于非平稳时间序列来说,统计分析会受到方法和理论的局限性,由于时间序列不稳定,从历史数据中获得的统计特性对未来没有意义。

平稳性的统计性质包括均值、方差和协方差。在时序分析当中数据具有平稳性意味着以上统计量不发生明显变化。

如果一个时间序列的均值为常数、自协方差只与时间间隔有关而与时间起点无关,则其为宽平稳;严平稳的一维分布都是相同的,二维及二维以上的联合分布只与时间间隔有关而与时间起点无关,条件更为严格[7]。因此从理论上和实际上来看,弱平稳相对应用更广泛,故进行平稳性分析时基本上使用弱平稳条件进行判别。

本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
通信作者:闫祥祥(519141272@qq.com)

2.3 差分

差分，又名差分函数和差分运算，从计算的角度来说就是相邻两个变量值的差，即后一个值减去前一个值，它反映了离散变量之间的变化。在时序数据中，根据时间逆序将每一个变量值减去前一个变量值即为一阶差分，对应差分阶数为 1；再次重复差分过程，阶数为 2，以此类推。

在 ARIMA 模型中差分是进行时序数据平稳化的唯一方式，阶数一般取 1 或 2，其余高阶差分使用较少。

3 ARIMA 模型

3.1 ARIMA 模型概况

自回归移动平均模型(Autoregressive Integrated Moving Average Model，ARIMA)是由 Box 和 Jenkins 在 20 世纪 70 年代初提出的时间序列检测方法，又称为 Box-jenkins 模型[6]。它在统计学和计量经济学等多种学科当中都有所应用，是最常见的一种用于进行时间序列预测的方法。

ARIMA 模型是几种情况的综合表述形式，其具体取决于 3 个方面：是否由时序数据本身的回归构成；是否由随机误差项的回归构成；是否对序列做了差分。时序数据在平稳性的基础上才能对具体模型形式进行判别。因此 ARIMA 模型根据实际情况具体表现为：滑动平均模型(MA)、自回归模型(AR)、自回归-滑动平均模型(ARMA)、差分后的滑动平均自回归模型(ARIMA)。

3.2 ARIMA 模型参数说明

ARIMA 模型由 3 个重要参数(p,d,q)决定：

p 即自回归系数(Auto-Regressive)，表示序列值滞后 p 阶；

d 对应 ARIMA 模型中的“I”，表示时序数据变为平稳序列需要最少进行 d 次差分；

q 即滑动平均系数(Moving Average)，表示误差项滞后 q 阶。

3.3 ARIMA 模型的数学表示

3.3.1 自回归模型(AR)

p 阶自回归模型的公式定义为：

$$y_t=\mu+\sum_{i=1}^{p}\gamma_i y_{t-i}+\varepsilon_t \tag{1}$$

其中，y_t 为当前值，μ 为常数项，γ_i 为自相关系数，ε_t 为误差。

3.3.2 滑动平均模型(MA)

q 阶滑动平均模型的公式定义为：

$$y_t=\mu+\sum_{i=1}^{q}\theta_i\varepsilon_{t-i}+\varepsilon_t \tag{2}$$

其中，θ_i 为滑动平均系数。

3.3.3 自回归-滑动平均模型(ARMA)

自回归-滑动平均模型的公式定义为：

$$y_t=\mu+\sum_{i=1}^{p}\gamma_i y_{t-i}+\varepsilon_t+\sum_{i=1}^{q}\theta_i\varepsilon_{t-i} \tag{3}$$

3.4 ARIMA 模型参数的确认

3.4.1 自相关函数介绍

自相关函数(Autocorrelation function，ACF)体现了时序数据中相邻观测点的相关性，其公式为：

$$\rho_k=\frac{\mathrm{cov}(y_t,y_{t+k})}{\sqrt{\mathrm{var}(y_t)\mathrm{var}(y_{t+k})}} \tag{4}$$

偏自相关函数(Partial autocorrelation funciton，PACF)是在随机变量去除中间 $k-1$ 个值的影响之后衡量 y_t 和 y_{t-k} 之间的相关性，其公式为：

$$\varphi_{kk}=\begin{cases}\rho_1, & k=1\\ \dfrac{\rho_k-\sum_{j=1}^{k-1}\varphi_{k-1,j}\varphi_{k-j}}{1-\sum_{j=1}^{k-1}\varphi_{k-1,j}\varphi_{k-j}}, & k>1\end{cases} \tag{5}$$

3.4.2 参数 d 的确定

将时序数据做自相关图平稳检验，分别查看 ACF 图和 PACF 图的截尾性，二者不是拖尾就是截尾，否则为非平稳序列。非平稳序列在进行差分后进行平稳性检验，直到满足平稳性条件，则 d 值确定。

3.4.3 参数 p 和 q 的确定

序列平稳之后，参数 p,q 和模型的确定方式如表 1 所列。

表 1 参数 p,q 和模型的确定条件

Table 1 Parameters p,q and determination conditions of model

模型	ACF	PACF
$AR(p)$	拖尾	p 阶后截尾
$MA(q)$	q 阶后截尾	拖尾
$ARMA(p,q)$	q 阶后拖尾	p 阶后拖尾

3.5 ARIMA 模型的建模步骤

ARIMA 模型的建模步骤[6,8-9]如下：

(1)对原始时序数据进行平稳性检验。如果序列不满足平稳性条件则进行差分平稳化处理。

(2)模型识别。对平稳序列分别进行 ACF 和 PACF 图分析，综合差分阶数，得到最佳参数 p,q 和 d。

(3)模型估计。通过可视化观察拟合效果并检验参数显著性。

(4)模型检验。结合 SPSS 软件功能的特性，使用不同配置方式进行建模，对比拟合参数结果并选定最终模型。

(5)残差白噪声检验。根据检验结果决定是否继续建模。

(6)数据预测。

4 数据选择

本实验所使用的数据是园林绿化及森林情况(1978－2017 年)，共 1 张表格，包括年份、年末公园绿地面积(公顷)、人均公园绿地面积(平方米/人)、城市绿化覆盖率(%)、林木绿化率(%)等若干字段，除年份、年末公园绿地面积、人均公园绿地面积、城市绿化覆盖率 4 个字段之外，其余数据均不完整，并且有些变量随年份的增加变化不大，有些变量数据不多，因此考虑到数据完整性、数据量和实际意义等因素，最终选择年末公园绿地面积(公顷)作为被观测数据。

5 实际建模过程

5.1 实验环境

本实验所有运行步骤均在 PC 上执行，系统为 Windows10，SPSS 软件版本为 20.0。

5.2 平稳性检验

将表格数据复制到 SPSS 系统中，变量属性如表 2 所列。

表 2 数据基本属性

Table 2 Basic data attributes

名称	类型	宽度	小数
年份	数值(N)	8	0
年末公园绿地面积	数值(N)	8	0

原始时间序列如图 1 所示。通过原始时序图可知序列有长期递增趋势，进行 ACF、PACF 图检验，结果如图 2、图 3 所示。

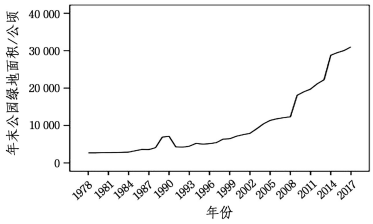

图 1 原始时间序列

Fig. 1 Original time series

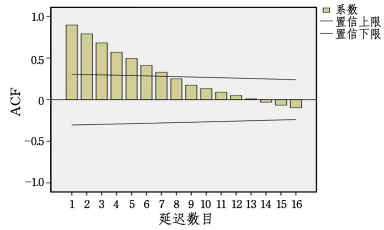

图 2 原始时间序列 ACF 图

Fig. 2 Original time series ACF graph

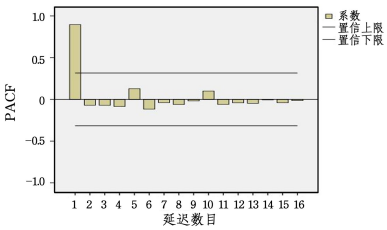

图 3 原始时间序列 PACF 图

Fig. 3 Original time series PACF graph

由图 2、图 3 可知,ACF 表现很像拖尾,但是后面的数据并没有收敛,反而有增大的趋势,并没有呈现波动现象,这说明序列具有单调趋势;而 PACF 为 1 阶截尾,系数围绕零轴上下小范围波动。因此判定原始时间序列为非平稳序列。

5.3 模型识别

对原始时间序列进行一阶差分,结果如图 4 所示。

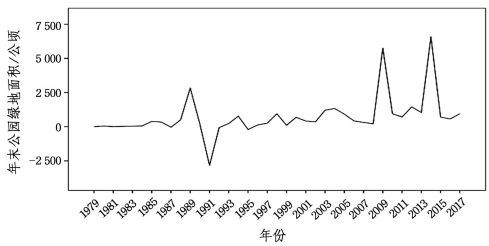

图 4 1 阶差分序列

Fig. 4 1st order difference sequence

然后进行 ACF 和 PACF 图检验,观察截尾性,结果如图 5、图 6 所示。

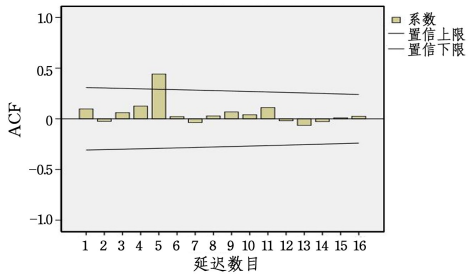

图 5 1 阶差分序列 ACF 图

Fig. 5 1st order differential sequence ACF graph

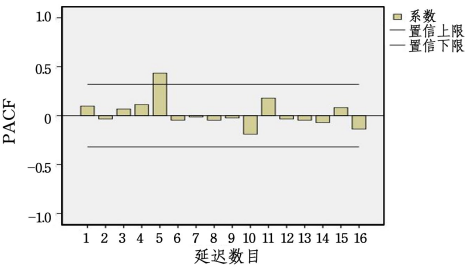

图 6 1 阶差分序列 PACF 图

Fig. 6 1st order differential sequence PACF graph

由图 5、图 6 可知,原始序列进行一阶差分后 ACF、PACF 图均为 0 阶拖尾,因此一阶差分后的序列为平稳序列,p 和 q 均等于 0。

5.4 模型估计

由于公园绿地面积是年末数值,故已消除季节性因素,不用做季节性分析。根据之前所确定的 p,d,q 3 个参数确定模型 ARIMA(0,1,0),在 SPSS 系统中进行建模,拟合效果如图 7 和表 3 所示。

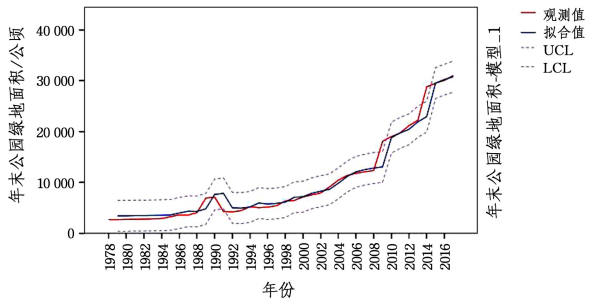

图 7 手动建模拟合效果

Fig. 7 Fitting effect of manual modeling

表 3 手动建模模型统计量

Table 3 Model statistics for manual modeling

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	正态化的 BIC	统计量	DF	Sig	
ARIMA(0,1,0)	0	-4.819×10^{-16}	14.722	17.292	18	0.503	0

本次拟合为手动创建 ARIMA 模型,没有选择“自动检测离群值”选项,无论从拟合图形还是模型统计量上来看效果差强人意。不仅拟合曲线有些滞后于原始时序曲线,并且平稳的 R 方数值不佳。

5.5 模型检验

在数据等其他因素不变的情况下,使用 SPSS 软件中的“专家建模期”功能进行建模,选择“自动检测离群值”选项,拟合效果如图 8 和表 4 所示。结合图表来看,模型拟合效果很完美,平稳的 R 方值很高,同时因为成功检测出 4 个离群值点,所以专家建模器给出的最佳模型为 ARIMA(0,2,0)。

Box 认为,大多数情况下 $p\leqslant2$ 和 $q\leqslant2$ 就可满足建模需要,通常使用 AIC 或 BIC 方法确定 p 和 q。

最小化信息准则(Akaike Information Criterion,AIC)的定义为:

$$AIC=2k-2\ln(L) \tag{6}$$

其中，k 值小表示模型简洁，L 值大表示模型精确。AIC 从预测角度来衡量模型的简洁性和精确性。当不同模型进行比较时，AIC 值最小的模型为最优模型。

贝叶斯信息准则(Bayesian Information Criterion，BIC)的定义为：

$$BIC=k\ln(n)-2\ln(L) \tag{7}$$

同样地，BIC 值越小越好。BIC 从模型拟合的角度来衡量模型对现有数据拟合的优劣性。

由于 SPSS 软件的特性，模型输出统计量时只输出 BIC 值，根据 BIC 准则，BIC 值越小，模型对数据的拟合效果越好，因此选取专家建模器模型作为最终模型。

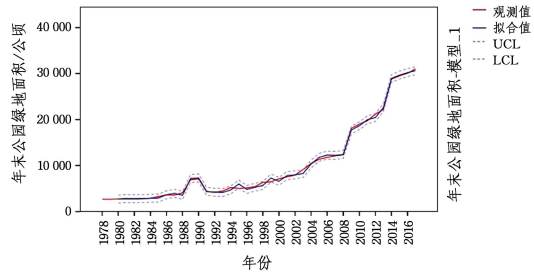

图 8 专家建模器拟合效果

Fig. 8 Fitting effect of expert modeler

表 4 专家建模器模型统计量

Table 4 Model Statistics for Expert Modeler

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	正态化的 BIC	统计量	DF	Sig	
ARIMA(0,2,0)	0	0.960	12.489	18.327	18	0.434	4

5.6 残差白噪声检验

选择最终模型之后，还需要对残差项进行白噪声检验。如果残差存在自相关，则应考虑增加自回归或滑动平均的解释，重新建模并且进行模型评估，再对新模型的残差进行白噪声检验，如此往复，直到确定残差为白噪声为止。

白噪声检验仍然可以采用自相关图检验方式，结果如图 9、图 10 所示。

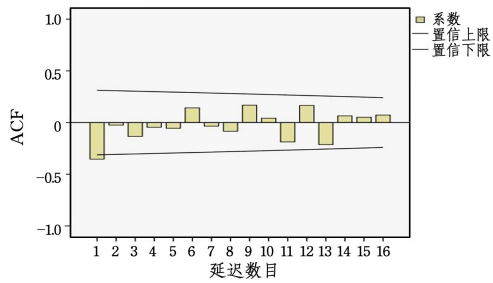

图 9 最终模型拟合残差 ACF 图

Fig. 9 ACF plot of final model fitting residuals

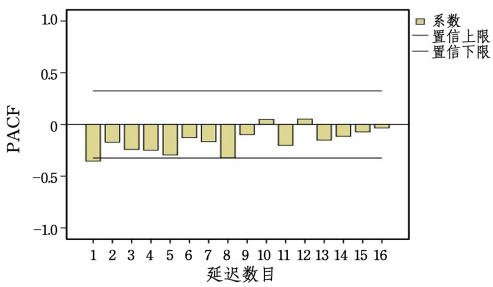

图 10 最终模型拟合残差 PACF 图

Fig. 10 PACF plot of final model fitting residuals

可以看出，在 1 阶差分之后所有自相关系数均落在置信区间之内，并且趋势逐渐趋近于 0。同样地，残差 PACF 系数也近似为 0。

残差 ACF 图统计量如表 5 所列，Ljung-Box 检验的 sig 值绝大部分大于 0.05，也就是滞后项不显著，无法拒绝各残差项不相关的原假设，时序数据有效信息基本已经被模型提取。

表 5 残差 ACF 图统计量

Table 5 Residual ACF graph statistics

滞后	自相关	标准误差	Ljung-Box 统计量		
			值	df	Sig.
1	−0.355	0.156	5.164	1	0.023
2	−0.025	0.154	5.190	2	0.075
3	−0.134	0.152	5.966	3	0.113
4	−0.045	0.150	6.059	4	0.195
5	−0.055	0.147	6.198	5	0.287
6	0.142	0.145	7.153	6	0.307
7	−0.036	0.143	7.215	7	0.407
8	−0.084	0.140	7.576	8	0.476
9	0.168	0.138	9.053	9	0.432
10	0.042	0.136	9.149	10	0.518
11	−0.187	0.133	11.119	11	0.433
12	0.166	0.131	12.732	12	0.389
13	−0.214	0.128	15.504	13	0.277
14	0.065	0.126	15.768	14	0.328
15	0.052	0.123	15.943	15	0.386
16	0.073	0.120	16.313	16	0.431

继续对残差项做正态性检验，在 SPSS 软件中绘制残差项的 Q-Q 图，结果如图 11 所示。

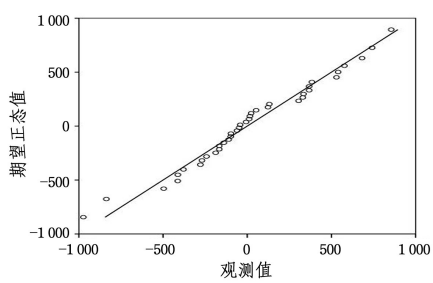

图 11 最终模型拟合残差 Q-Q 图

Fig. 11 Q-Q plot of final model fitting residuals

可以看出，最终模型拟合后的残差基本服从均值为 0 的正态分布。结合自相关图检验结果，确定最优模型拟合后的残差项为白噪声序列，无须继续建模。

5.7 数据预测

使用最佳模型对年末公园绿地面积后 3 年(2018－2020 年)进行预测，结果如表 6、图 12 所示。

(下转第 556 页)