

文章编号: 1000-5013(2011)05-0511-04

基于 ARIMA 模型的福州市雷暴日趋势分析

刘隽, 张烨方, 黄岩彬

(福建省气象局, 福建 福州 350001)

摘要: 在分析 ARIMA(p, d, q) 预测模型的基础上, 以福州市 1961—2006 年的雷暴日为时间序列基础, 通过对该序列进行平稳性分析、差分处理、自相关、偏自相关系数计算与绘图、ARIMA 建模、参数估计、假设检验及模型预测, 将 ARIMA 模型运用在雷暴日的趋势分析上. 研究结果表明, ARIMA 能很好地拟合计算出未来短时段内的数据, 可以应用于实际的雷暴日分析.

关键词: 雷暴日; 差分自回归移动平均模型; 预测; 短期; 福州市

中图分类号: P 456; P 468. 028

文献标志码: A

雷暴日是反应一个地区雷电活动规律的重要气象参数, 是目前进行建筑物防雷类别及防雷装置设计的重要数据. 对雷暴日进行趋势分析是气象长期预报的一项重要内容. 一般而言, 雷暴日的年分布没有非常明显的趋势和规律, 目前对雷暴日的规律研究也还停留在研究和探索阶段^[1-3]. 自回归移动平均模型 (autoregressive integrated moving average model, 简记 ARIMA) 是由 Box 和 Jenkins 于 70 年代初提出的时间序列预测方法, 其中 ARIMA(p, d, q) 称为差分自回归移动平均模型. ARIMA 模型的基本思想: 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, 然后以时间序列的自相关分析为基础, 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, 就可从时间序列的过去值及现在值来预测未来值^[4-5]. 这一特点正好适合来预测、预报像雷暴日这样分布无明显规律, 具有时间序列的数据^[6]. 本文在建立差分自回归移动平均模型的基础上, 对福州市雷暴日趋势进行预测分析.

1 ARIMA 模型的原理

设 Y_t 为一个时间序列, 数据按照一定的时间段顺序排列, 以雷暴日为例, 该时间序列时间段为年. 进行 ARIMA 分析的数据必须满足两个条件, 一是平稳性的数据, 二是数据本身的随机过程是一个白噪声过程. 即当 Y_t 的数学期望为一固定值, 或者说系统受扰动后经过足够长的时间后可以趋于平稳的序列可认为是平稳序列; 而在满足平稳序列的前提下, 如果还满足 $\gamma_t = EX_t X_s = \begin{cases} \sigma^2, & t=s, \\ 0, & t \neq s, \end{cases}$ 则称序列 Y_t 为白噪声.

利用 Y_t 的前 p 个数据进行回归计算来预测 Y_t , 即 $Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p}$, 称为 p 阶 AR 自回归模型, 记为 AR(p); 利用 Y_t 的前 q 个预测残差值进行回归计算 Y_t , 即 $Y_t = \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$, 称为 q 阶 MA 滑动平均模型, 记为 MA(q).

如果将 AR(p), MA(q) 的预测方法相结合 (即 ARMA), 就可以对数据进行更加完整的预测. 由于很多数据本身并不是平稳的白噪声序列, 因此, 需要对数据进行一定的处理才能使其成为可以进行 ARMA 分析的序列. 常用的处理方法是对非平稳序列进行差分计算, 即按照差分算子 $\nabla Y_t = Y_t - Y_{t-1}$ 做数据的 d 阶差分, 这样组合起来的模型即称为 ARIMA(p, d, q).

引入一个后移算子 B , 使之满足 $B^k X_t = X_{t-k}$, 并令 $\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p$, $\theta(B) = 1 -$

收稿日期: 2011-04-26

通信作者: 刘隽 (1978-), 男, 工程师, 主要从事的研究. E-mail: liujun_fjlight@126.com.

基金项目: 福建省福州市科技计划项目 (2008-S-87)

$\theta_1 B - \theta_2 B^2 - \cdots - \theta_p B^p$, 则可将 ARIMA 模型算式表达为

$$\varphi(B)Y_t = \theta(B)\epsilon_t.$$

通过参数估计方法将表达式中的未知参数求解出来, 形成完整的计算式用于具体的预测计算.

2 ARIMA 建模与预测

2.1 平稳性检验与差分处理

将福州市年均雷暴日看成是一个随机性的时间序列进行分析. 由于随机过程的均值和自协方差要求都不依赖于序列时间点的取值, 即随机过程是一个平稳过程. 因此, 在进行 ARIMA 建模前需要对福州市雷暴日的数据平稳性进行检验和处理, 以保证用于建模的数据的稳定性. 这里采用自相关系数 (ACF) 和偏自相关系数 (PACF) 来判断时间序列的平稳性. 自相关系数的计算式为

$$\rho = \frac{\text{cov}(Y_t, Y_{t-1})}{\sqrt{\text{var } Y_t} \cdot \sqrt{\text{var } Y_{t-1}}}.$$

其中: $\text{cov } \epsilon$ 是协方差计算; $\text{var } \epsilon$ 是方差计算.

对于时间序列 Y_t , k 阶偏自相关系数 $\alpha_{k,k}$ 用时间序列的条件密度期望及相应方差的比值来表示, 即偏自相关系数的计算式为

$$\alpha_{k,k} = \frac{E[(Y_t - \hat{Y}_t)(Y_{t-k} - \hat{Y}_{t-k}) | Y_{t-1}, Y_{t-2}, \dots, Y_{t-k+1}]}{\sqrt{\text{var}(Y_t | Y_{t-1}, Y_{t-2}, \dots, Y_{t-k+1})} \cdot \sqrt{\text{var}(Y_{t-k} | Y_{t-1}, Y_{t-2}, \dots, Y_{t-k+1})}}.$$

按上述算式, 对福州市 1961—2006 年的年均雷暴日进行自相关系数和偏自相关系数计算, 如图 1 (a), (b) 所示. 图 1 中: N 为滞后值; ρ, α 分别为自相关系数和偏自相关系数. 从图 1(a), (b) 中可看出, 福州市雷暴日的基本符合稳定性数据的要求, 但整体的截尾和拖尾特征不明显, 不利于进行 ARIMA 建模取值. 于是对样本数据进行差分处理, 即用差分算子 $\nabla Y_t = Y_t - Y_{t-1}$ 进行数据差分, 并引入后移算子 B , 则可推导得 $\nabla^d = (1 - B)^d$. 按照上式对福州市 1961—2006 年的年均雷暴日做二阶差分处理并计算自相关系数和偏自相关系数, 如图 1(c), (d) 所示. 从图 1(c), (d) 可以发现, 二阶差分后的时间序列截尾和拖尾性质很明显, 可以进行 ARIMA 的建模^[6-7].

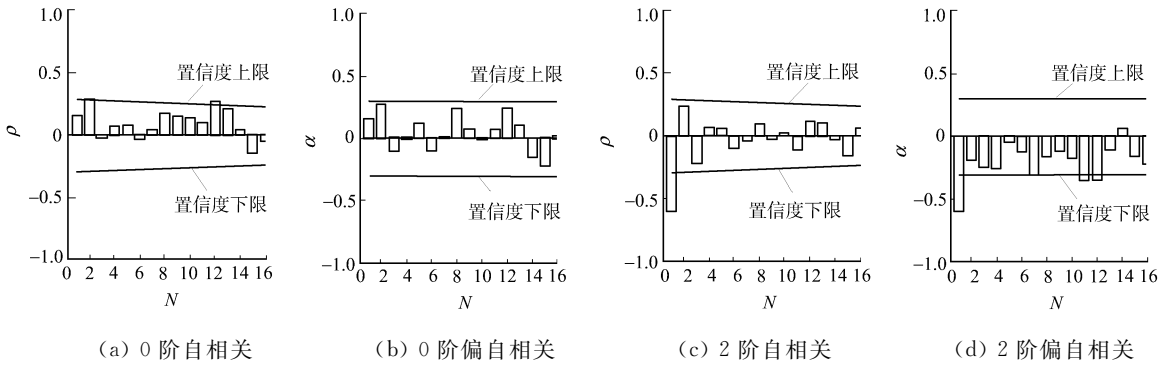


图 1 福州市 1961—2006 年均雷暴日自相关与偏自相关分析图

Fig. 1 ACF and PACF drawing of Fuzhou 1961 to 2006 annual thunderstorm days

2.2 ARIMA 建模

ARIMA 建模的关键在于确定一个最优的自回归 p 、滑动平均指数 q 及相应的差分阶数 d , 在确认了 ARIMA 的 3 个参数后, 采用极大似然函数进行 ARIMA 模型参数的参数估计. 对于平稳序列的 ARIMA 模型, 有

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \theta_q \epsilon_{t-q},$$

估计参数 $\boldsymbol{\phi} = [\phi_1, \phi_2, \dots, \phi_p]$, $\boldsymbol{\theta} = [\theta_1, \theta_2, \dots, \theta_q]$ 及 ϵ_t , 需要建立 $p+q+1$ 个方程. 由于 ϵ_t 是一个白噪声, 服从正态分布, 则有似然函数为

$$L(\boldsymbol{\theta}, \boldsymbol{\phi}, \sigma^2) = f(\epsilon_1, \epsilon_2, \dots, \epsilon_n; \boldsymbol{\theta}, \boldsymbol{\phi}, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{\sum_{k=1}^n \epsilon_k^2}{2\sigma^2}\right).$$

对上式取对数可得

$$\ln[L(\boldsymbol{\theta}, \boldsymbol{\varphi}, \sigma^2)] = -\frac{n}{2} \ln 2\pi - n \ln \sigma - \frac{\sum_{k=1}^n \epsilon_k^2}{2\sigma}.$$

对上式进行偏导计算, 即可得相应参数的极大似让估计表达式. 在确定了各个参数的估计量后, 由于在进行计算时认为序列的随机误差 ϵ_t 是一个白噪声, 可以通过对 ϵ_t 构造一定的统计量来进行结论的假设检验, 以保证参数估计结果的准确性. 这里, 借鉴 Box-Pierce 提出的 Q 统计量, 即

$$Q = \sum_{k=1}^m [\sqrt{N} \hat{\rho}_k(N, \epsilon_k)]^2 = N \sum_{k=1}^m \hat{\rho}_k^2(N, \epsilon_k) \simeq \chi^2(m - p - q),$$

进行参数估计结果的假设检验. 为了保证所建立的 ARIMA 模型的最优性, 还需要对所选择的 (p, d, q) 进行最佳准则定阶, 参考 AIC 准则函数, 即

$$AIC(k, j) = n \ln \hat{\rho}(k, j)^2 + 2(k + j),$$

取通过 Q 检验量假设检验的, 且 AIC 值最小的, 那对自回归 p 、滑动平均指数 q 阶数即为最优解.

基于上述方法, 通过对比发现 ARIMA(2, 2, 0) 为最优化结果 (其余参数的计算、对比过程略), 按照 ARIMA(2, 2, 0) 模型进行计算, 如图 2 所示.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
AR(1)	-1.110507	0.140848	-7.884412	0.0000
AR(2)	-0.414125	0.143458	-2.886727	0.0062
R-squared	0.686698	Mean dependent var		0.619048
Adjusted R-squared	0.678866	S.D. dependent var		25.72647
S.E. of regression	14.57886	Akaike info criterion		8.243470
Sum squared resid	8501.728	Schwarz criterion		8.326216
Log likelihood	-171.1129	Hannan-Quinn criter.		8.273800
Durbin-Watson stat	2.249916			
Inverted AR Roots	-.56-.33i	-.56+.33i		

图 2 ARIMA(2, 2, 0) 参数估计与判定结果

Fig. 2 Parameter estimation and hypothesis test result for ARIMA(2, 2, 0)

2.3 ARIMA 预测与结果分析

通过计算得 ARIMA(2, 2, 0) 的两个参数 $AR(1) = -1.11, AR(2) = -0.414$. 将参数代入 ARIMA 一般式 $\varphi(B)(1-B)^d Y_t = \theta(B)\epsilon_t$ 中, 整理可得

$$Y_t = -0.423Y_{t-4} - 0.236Y_{t-3} + 0.777Y_{t-2} + 0.9Y_{t-1}.$$

按照上式, 计算得福州市 1996—2006 年雷暴日的预测值并与真实值对比, 如图 3 所示. 从图 3 可知, 按照 ARIMA(2, 2, 0) 拟合的福州市年均雷暴日分布函数能够比较好的符合福州市近几十年来的雷暴分布趋势. 通过该函数计算的 1996—2006 年福州市的二阶差分雷暴日与实际二阶差分雷暴日的叠合程度较好, 除了 2004 年的差异较大之外, 其他数值基本持平在同一个层次及误差可接受的范围内, 可以认为该预测方程较好地拟合了福州市雷暴日的总体趋势.

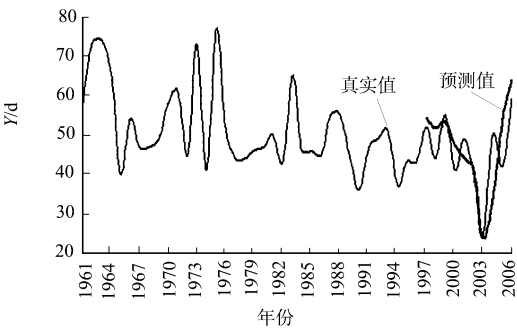


图 3 福州市 1996—2006 年的年均雷暴日预测值与实际值折线对比图

Fig. 3 Predicted compared with actual value of 1996—2006 fuzhou annual thunderstorm days

3 结论

在分析 ARIMA(p, d, q) 预测模型的基础上, 以福州市 1961—2006 年的雷暴日为时间序列基础, 通过对该序列进行平稳性分析、差分处理、自相关、偏自相关系数计算、ARIMA 建模、参数估计、假设检验

及模型预测,并将 ARIMA 模型运用在雷暴日的趋势分析上. 结果显示,ARIMA 能很好的拟合计算出未来短时段内的数据. 为了提高 ARIMA 模型在雷暴日趋势分析中的应用,提出如下的几点建议.

(1) 截止目前为止,可利用的雷暴日观测数据仅有 40 几年,ARIMA 预测是通过过去数据的规律分析及判断来确定参数的,因此,样本量的多少对预测结果有很大的影响.

(2) 可以将年雷暴日预测转换为月雷暴日预测. 由于雷暴活动呈现出很强的季节性规律,如果采用月雷暴日进行数据排序,那么 ARIMA 模型中的季节因子可以得到很好的体现^[8].

(3) 采用一定的科学方法来确定 ARIMA 中的 p, d, q 值. 在目前对 ARIMA 建模过程中, p, d, q 值的确定方法还是主要通过人为的判断自相关、偏相关图进行确定,再通过一定的假设检验进行确认. 这使得所建立的 ARIMA 模型存在很大的主观成分. 当然,也可以通过逐个 p, d, q 取值进行结果对比,但处理方式计算量非常庞大,且某个 p, d, q 值可能适用于后几个数据的预测,对于总体而言,可能有其他的 p, d, q 值更加符合要求. 因此,如何通过一种科学的算法来确定 p, d, q 值,是提高 ARIMA 预测精确度的重要研究内容.

参考文献:

[1] 王锡中,叶玉珍,钟颖颖,等. 江苏省城市雷暴日分布特征[J]. 气象科学,2011,31(1):93-99.
[2] 黄小红,古名岸. 吉安雷暴日统计及其特征分析[J]. 井冈山大学学报:自然科学版,2010,31(6):53-56.
[3] 黄荆鹏. 荆州市雷暴日数的时间分布特点及其预报[J]. 湖北气象,2000(4):20-21.
[4] 汤琴,黄宜坚. 采用 AR 模型双谱估计的概率筛筛分效率[J]. 华侨大学学报:自然科学版,2011,32(3):253-257.
[5] 田霆,陈祥钟,黄春棋,等. 定时截尾缺失数据下指数分布的统计推断[J]. 华侨大学学报:自然科学版,2010,31(1):109-112.
[6] 张旭晖,高苹,许祥,等. 江苏雷暴日发生规律及其大气环流预报模型的建立[J]. 气象科技,2006,34(10):532-537.
[7] 郑小平,高金吉,刘梦婷. 事故预测理论与方法[M]. 北京:清华大学出版社,2009.
[8] 贾俊平. 统计学[M]. 北京:中国人民大学出版社,2003.

Fuzhou Thunderstorm Days Trend Analysis
by the ARIMA Model

LIU Jun, ZHANG Ye-fang, HUANG Yan-bin

(Fujian Meteorological Administration, Fuzhou 350001, China)

Abstract: Paper is based on the analysis of auto regressive integrated moving average (ARIMA) model, leveraging historical thunderstorm day data of Fuzhou which was collected from year 1961 to 2006 to forecast the trend of thunderstorm day by stability analysis, differential treatment, autocorrelation, partial and autocorrelation coefficient calculation and drawing, ARIMA modeling, parameter estimation, hypothesis testing and predictions. Applying ARIMA model in the thunderstorm day trend analysis, the result indicates that ARIMA model has a better short-term prediction and can be applied in the actual forecast of thunderstorm days.

Keywords: thunderstorm day; auto-regressive integrated moving average model; forecast; short term; Fuzhou City

(责任编辑: 陈志贤 英文审校: 吴逢铁)