

文章编号:1005-3085(2008)04-0611-05

基于ARIMA模型的用电量时间序列建模和预报*

杨建萍

(浙江理工大学数学科学系, 杭州 310018)

摘 要: 本文主要研究电力系统的窃电问题。采用ARIMA (p, d, q)(自回归求和移动平均)动态模型, 对客户用电规模的时间序列数据进行了拟合和预测。若客户的实际用电量低于置信度为95%预测下限, 就认为该客户有窃电嫌疑。最后采用滁州电力局的客户历史用电数据进行实证分析, 结果显示该方法比较适合国内电力MIS系统的现状, 能有效的提高工作效率, 缩减堪察成本。

关键词: MIS系统; 用电规模; 窃电; 时间序列; ARIMA

分类号: AMS(2000) 91B70

中图分类号: 0212

文献标识码: A

1 引言

随着我国经济的不断发展和城乡人民生活水平的不断提高, 电力已成为人们赖以生存的基本能源。同时, 窃电问题也变得越来越突出, 并且近年来这种违法行为的主体已变得多元化, 窃电手段也日趋隐蔽并向高科技发展, 国家为此蒙受了巨大的损失。由此, 反窃电的工作亦显得越来越重要。在这种情况下, 仅仅依靠群众的举报或派几名工作人员去检查来开展反窃电工作, 已经不合适了。针对这一情况, 参考了大量的文献资料及电力行业中其它一些问题的解决方法, 如: 电力负荷问题等^[1-3], 并结合电力MIS系统的实际情况, 提出了利用Box-Jenkins在1970年提出的ARIMA(自回归求和移动平均)动态模型^[4,5], 对电力MIS系统中客户用电规模数据进行拟合与预测, 并根据客户用电规模变化判断该客户是否有窃电嫌疑的方法。最后, 选取滁州市电力局提供的客户用电量的时间序列数据对此方法的可行性进行了实证分析, 结果表明此方法是可行、有效的。在实证分析中, 所有数据采取了如下处理: 1) 除去了最近半年内没有用电量记录的客户的数据记录, 认为这是一部分已停户的客户; 2) 客户用电量记录的异常值处理^[6,7]。

2 用电规模时间序列的ARIMA模型

由Box-Jenkins提出的ARIMA(自回归求和移动平均)模型是近几年国际上最流行、最基本的一种时间序列模型, 它可根据动态数据自身的相关特征来解释其本身结构与规律, 从而进行预测。目前已广泛应用于各社会学科领域中。首先, 我们简单的介绍一下ARMA(自回归滑动平均)过程和ARIMA(自回归求和移动平均)过程的定义。

定义2.1 (ARMA(p,q)过程) 称过程 $\{X_t, t = 0, \pm 1, \dots\}$ 是ARMA(p,q)过程, 如果 $\{X_t\}$ 是平稳过程, 且对任意的 t , 有

$$X_t - \Phi_1 X_{t-1} - \dots - \Phi_p X_{t-p} = Z_t + \Theta_1 Z_{t-1} + \dots + \Theta_q Z_{t-q},$$

收稿日期: 2006-07-18. 作者简介: 杨建萍(1979年1月生), 女, 硕士, 讲师. 研究方向: 应用统计方向.

*基金项目: 滁州电力公司、倍多科技有限公司、卡耐信息科技有限公司资助.

其中 $Z_t \sim WN(0, \sigma^2)$, $\Phi_p(B) = 0$ 和 $\Theta_q(B) = 0$ 的根全在单位圆内, 则称 $\{X_t, t = 0, \pm 1, \dots\}$ 是均值为 μ 的平稳可逆的 ARMA(p, q) 过程, 如果 $\{X_t - \mu\}$ 是平稳可逆的 ARMA(p, q) 过程。

定义 2.2 (ARIMA(p, d, q) 过程) 设 d 是非负整数, 令 $Y_t = (1 - B)^d X_t$, 如果 $\{Y_t\}$ 是因果的 ARMA(p, q) 过程, 称 $\{X_t\}$ 是 ARIMA(p, d, q) 过程。

我们知道如果数据满足: (a) 与平稳序列无显著差异; (b) 有迅速下降的自相关系数, 则可寻求一个合适的 ARMA 模型来拟合, 否则可预先对数据进行处理, 使经过处理后的数据具有性质 (a) 和 (b)。在本文中我们用 $\{X_{t,ghi}\}$ 表示第 i 个客户第 t 个月用电量的离散值, 对于这样一个离散的时间序列先进行如下的处理。

1) 画出数据的曲线。若曲线图大体上呈这样的表现(数据较大时, 对应曲线的变化较大; 而数据较小时, 对应曲线的变化较小)时, 就对此序列采用 Box-Cox 变换

$$Y_{t,ghi} = \begin{cases} X_{t,ghi}^\lambda - 1 & \lambda \neq 1, \\ \log(X_{t,ghi}) & \lambda = 0. \end{cases}$$

通过选取一个合适的 λ , 使 $\{Y_{t,ghi}\}$ 在一个常值附近变化。

2) 对数据 $\{Y_{t,ghi}\}$ 进行平稳性检验, 这里采用序列的 Daniel 检验法。若数据 $\{Y_{t,ghi}\}$ 非平稳, 由 Box-Jenkins(1970) 提出的建模方法, 通过对数据 $\{Y_{t,ghi}\}$ 反复作用差分算子直到差分后的数据序列与平稳过程 $\{Z_{t,ghi}\}$ 相似, 也就是减去均值后就得到一个均值为零的平稳时间序列 $\{W_{t,ghi}\}$ 。

3) 对平稳时间序列 $\{W_{t,ghi}\}$ 进行自相关函数(ACF)和偏相关函数(PACF)识别, 给出 (p, q) 的暂定值。

4) 用极大似然估计及 AIC 准则给出暂定模型的参数估计值。

5) 进行拟合优度检验: 模拟序列是否更好的反映了观测到的时间序列, 及对白噪声进行诊断。

3 实证分析

以客户数据 $\{X_{t,gh5}\}$ 为例, 把数据分为两部分, 1999.02 ~ 2002.01 作为建模用, 2002.02 ~ 2002.11 的数据用来选择最优模型。

首先, 对第一部分的数据, 根据曲线的图形和数据分布的特点(峰度 = 0.621, 偏度 = 0.448), 先作 $\lambda = 0$ 的 Box-Cox 变化, 接着对平稳性 $\{Y_{t,gh5}\}$ 做了显著水平 $\alpha = 0.05$ 的 Daniel 检验, 计算得到 $p = 0.002 < 0.05$, 即该序列为非平稳序列。为了得到一个稳定的序列 $\{Z_{t,ghi}\}$, 对序列 $\{Y_{t,gh5}\}$ 作一阶差分并减去均值。最后, 画出平稳时间序列 $\{Z_{t,ghi}\}$ 样本自相关函数和样本偏相关函数图如下图 1, 由 gh5 样本偏相关函数图知, 用 AR(3) 拟合数据序列是合适的。表 1 是系数的 Yule-Walker 估计值 $\hat{\phi}_{mj}$, $j = 1, \dots, m$, $m = 1, \dots, 4$ 及相应的比值 $r_{mj} = \hat{\phi}_{mj} / 1.96\hat{\sigma}_{mj}$, 其中 $\hat{\sigma}_{mj}$ 是矩阵 $\sigma^2 \Gamma^{-1}(\hat{\phi}_m) / n$ 对角线上第 j 个元素; $\sigma^2 \Gamma^{-1}(\hat{\phi}_m) / n$ 是 ϕ_m 的渐进协方差的估计式。另外, 如果 $|r_{mj}| > 1$, 则以近似于 0.05 的概率拒绝假设: $\phi_{mj} = 0$ 。

第二步以 Yule-Walker 估计做初值, 用极大似然估计拟合阶数依次为 1, 2, \dots 的自回归模型, 表 2 给出了零均化数据序列的极大似然估计及相应的 AIC 值、BIC 值和 FPE 值。由表 2 可知, 零均化的数据序列 $\{Z_{t,gh5}\}$ 用 AIC 准则所拟合的自回归模型是

$$Z_{t,gh5} + 0.6668Z_{t-1,gh5} + 0.5254Z_{t-2,gh5} + 0.4402Z_{t-3,gh5} = Z_t.$$

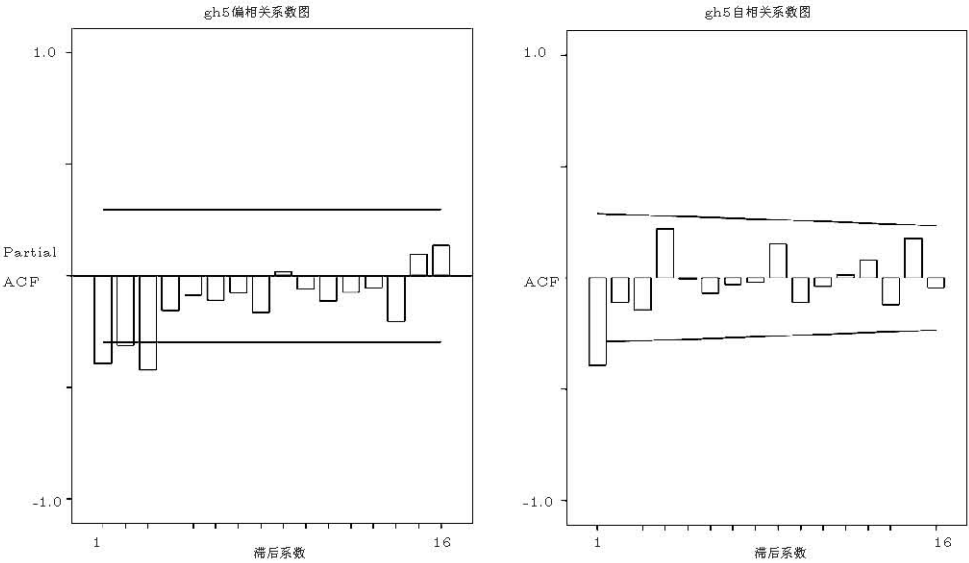


图 1: gh5 样本自相关函数与偏相关函数图

表 1: AR(m) 拟合系数向量与比值 r_{mj}

$m \backslash j$	1	2	3	4
1	0.3982 (1.310)			
2	0.5207 (1.652)	0.3077 (0.976)		
3	0.6332 (2.053)	0.4981 (1.489)	0.3656 (1.186)	
4	0.6969 (2.136)	0.5850 (1.593)	0.476 (1.296)	0.1743 (0.534)

表 2: 数据的极大似然估计 $\hat{\phi}_{mj}$ 及标准误差 $\hat{\sigma}^2$, 相应的AIC、BIC和FPE值

$m \backslash j$	1	2	3	4	$\hat{\sigma}_m^2$	AIC	BIC	FPE
1	0.4428 (0.160)				0.3758	69.664	68.109	0.398
2	0.5533 (0.163)	0.3625 (0.185)			0.3372	68.518	66.917	0.378
3	0.6668 (0.156)	0.5254 (0.182)	0.4402 (0.178)		0.2845	65.685	64.116	0.338
4	0.7199 (0.168)	0.642 (0.207)	0.5181 (0.201)	0.1552 (0.194)	0.2792	67.791	65.366	0.3513

同时由样本ACF知, 用MA(1)模型拟合也是合适的。对滑动平均过程参数以新息估计 $\hat{\theta}$ 作为参数的初始值, 再用极大似然估计, 可得阶数为1的滑动自回归模型为

$$Z_{t,gh5} = Z_t - 0.9724Z_{t-1},$$

白噪声估计、AIC、BIC分别为: $\hat{\sigma}=0.2436$, AIC= 57.046380, BIC= 54.322404。

第三步对上述两个模型进行拟合优度检验, 结果如下。

1) 作出AR(3)模型残差和MA(1)模型残差的样本自相关函数和偏相关函数, 得到AR(3)模型残差样本自相关函数都位于置信界 $\pm 1.96(1 - q_{ii}(\hat{\Phi}_1, \hat{\Phi}_2, \hat{\Phi}_3))^{\frac{1}{2}} 36^{-\frac{1}{2}}$ 之内, MA(1)模型残差的样本偏相关系数全位于置信界 $\pm 1.96(1 - q_{ii}(\hat{\theta}))^{\frac{1}{2}} 36^{-\frac{1}{2}}$ 之内。

2) Pormanteau检验: 采用Mcleod-Li检验统计量: $\tilde{Q}_w = 1368 \sum_{j=1}^h \hat{\rho}_w^2(j)/(36-j)$, 其中

$$\hat{\rho}_w(h) = \frac{\sum_{i=1}^{36-h} (\widehat{W}_t^2 - \overline{W}^2)(\widehat{W}_{t+h}^2 - \overline{W}^2)}{\sum_{t=1}^{36} (\widehat{W}_t^2 - \overline{W}^2)},$$

\widehat{W}_t 为残差, $\overline{W}^2 = \frac{1}{36} \sum_{t=1}^{36} \widehat{W}_t^2$, Mcleod与Li证明了在模型合适、采用kend假设成立的条件下, \tilde{Q}_w 分布接近 χ^2 分布, 对AR(3)模型Pormanteau检验的显著性概率 P 满足 $P(\chi^2(23) > 7.3355) = 0.999 > 0.05$, 故接受原假设; 同理对MA(1)模型也有显著性概率 P , $P(\chi^2(21) > 10.507) = 0.971 > 0.05$, 也接受原假设。由此这两个残差都通过了模型的拟合优度检验。

到这一步为止, 我们已经为编号为gh5的客户用电时间序列数据进行了ARIMA建模, 最后通过ARIMA(1,3,0)模型和ARIMA(1,0,1)模型的8步预报, 且用预报均值作为预报值与实际值作比较, 为数据序列挑了一个最优模型, 即ARIMA(1,3,0), 并用此模型对未来16个月的用电量进行预报, 求出其95%置信度的置信上下界, 结果发现该客户在2000年5月用电量明显偏低, 就认为该客户有窃电嫌疑。这一结论与实际情况相一致。

4 模型的优缺点

从目前的电力MIS系统状况来看, 不可能提供更多的有关用电客户的信息。该方法的优点在于用到的用电客户的信息比较少, 仅仅通过用电规模的ARIMA模型计算出客户在某一时刻用电量预期下限值, 若此值大于真实值, 则我们认为该用户有窃电嫌疑。这样做的好处在于, 通过事先划定的窃电嫌疑的范围, 就缩小了供电企业实地检查的范围, 提高了工作效率, 缩减了堪察成本。所以此种方法较适合于电力MIS系统的现状。缺点在于对于那些“狡猾”的窃电户, 我们需要配合使用其他方法, 如通过和同性质、同规模企业进行比较等。

参考文献:

- [1] 范习辉, 张勇传. 短期电力负荷预测的 $GM(1, 1)$ 模型群法及应用[J]. 水电能源科学, 2002, (3): 77-80
- [2] 方兆本等. 电力企业客户信用风险管理实证研究[J]. 电力系统自动, 2005, 29(1): 61-64
- [3] 史德明, 李林川. 宋建文基于灰色预测和神经网络的电力系统负荷预测[J]. 电网技术, 2001, 25(12): 14-17
- [4] Groschwitz N, Plozoz G. A time series model of long-term NSFNET backbone traffic[C]// Proc IEEE Icc 94. Necooreans, USA., 1994: 1400-1404
- [5] Brockwell P, Davis R. Time Series: Theory and Methods[M]. New York: Springer Verlag, 1991
- [6] 范金城等. 数据分析[M]. 北京: 科学出版社, 2005
- [7] 杨位钦. 时间序列分析与动态数据建模[M]. 北京: 北京理工大学出版社, 1988

ARIMA Time Series Modeling and Forecasting of Electricity Consumption

YANG Jian-ping

(Department of Mathematics, Zhejiang Sci-Tech University, Hangzhou 310018)

Abstract: This paper mainly describes a new method for detecting the electricity stealer in China. According to the Box-Jenkins theory and ARIMA (P, d, q) dynamic model, the time series of the electricity consumption is identified, simulated and predicted. The low 95% confidence limit in prediction is used as the threshold value of the electricity stealing. To verify the method, some power supply bureau's data is adopted and the results show that our method is feasible, effective and cost saving.

Keywords: MIS system; electricity consumption; electricity stealer; time series; ARIMA