论文题目：西安市日用水量的非线性时间序列模型

学科专业：应用数学

研 究 生：陈战波　　签 名：陈战波

指导教师：张德生　教授　　签 名：张德生

摘　要

用水量预测在城市建设规划和供水系统调度管理中具有重要的作用。用水量预测主要包括：年用水量预测，季度用水量预测，月用水量预测，日用水量预测和时用水量预测。其中，日用水量预测占据着重要的地位，它不仅能直接指导水厂的生产实践，更能为水厂间的优化调度提供可靠的技术支持。本文在分析西安市供水系统原始观测数据资料的基础上，建立了西安市日用水量多元非参数回归模型、部分线性自回归模型和其它预测模型，具体内容如下：

1.建立了西安市日用水量的多元线性回归模型，对模型进行了 χ^2 检验和残差修正，同时建立了西安市日用水量的综合预测模型，计算结果表明，该模型预测精度不高。

2.利用核估计和局部线性估计方法建立了西安市日用水量的多元非参数回归模型。通过对日用水量的拟合与预测，结果表明多元非参数回归模型拟合与预测误差较小，能满足供水系统的需要。

3.根据 BJ 时间序列建模方法，建立了西安市日用水量的线性自回归模型。结果表明，该模型预测精度不高。

4.建立了西安市日用水量的部分线性自回归模型，其中线性部分考虑日用水量，非线性部分考虑当天的最高温度，该模型综合了非参数回归模型和线性自回归模型的优点，与前三种建模方法比较，该模型在拟合与预测精度上有所提高，证明该方法在西安市日用水量预测这四种模型中效果最优。

关键词：日用水量；非参数回归模型；部分线性自回归模型；核估计；局部线性估计

Title: THE NONLINEAR TIME SERIES MODEL OF XI'AN URBAN DAILY WATER DEMAND

Major: Applied Mathematics

Name: Zhanbo CHEN **Signature:** Zhanbo CHEN

Supervisor: Prof. Desheng ZHANG **Signature:** Desheng ZHANG

Abstract

Water-use forecasting is important for municipal construction planning and the operation of water-supply system. Water-use forecasting is composed of the yearly, seasonal, monthly, daily and hourly municipal water-use forecasting. In the several parts, daily one has a key-position. It may not only directly guide the production of water-supply corporations, but also offer the technical service for optimal scheduling among several water-supply corporations. This paper has established nonparametric regression model, partially linear AR model and other forecasting models of xi'an urban daily water demand based on the xi'an daily water-use data. The four aspects are as follows:

The multi linear regression model has been set up and the prediction error is checked by the χ^2 method. The prediction error is corrected, at last xi'an daily water-use synthesis forecasting model is set up and the prediction error is not good.

Based on the regression function estimated by the kernel estimation and local linear estimation, the nonparametric multi regression model of xi'an urban daily water demand is set up. By the comparison of fact date, it was proved that the nonparametric multi regression model can meet the practical requirement of water supply dispatch system.

According to BJ time series method, the linear AR model of xi'an water daily demand is set up. The computation result shows that the forecasting model for the daily water-consumption is not good.

The partially linear AR model of xi'an water daily demand is set up. The linear aspect takes into account the water demand and the nonlinear aspect takes into account tiptop temperature. This model not only has the merit of regressive model but also has the linear autoregressive model merit, so its forecasting result is better. By the comparison of the former three models, this method is best for the xi'an water daily demand forecasting.

Key words: Water daily demand; Nonparametric regressive model; Partially linear AR model; Kernel estimation; Local linear estimation

独 创 性 声 明

秉承祖国优良道德传统和学校的严谨学风郑重申明：本人所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的成果。尽我所知，除特别加以标注和致谢的地方外，论文中不包含其他人的研究成果。与我一同工作的同志对本文所论述的工作和成果的任何贡献均已在论文中作了明确的说明并已致谢。

本论文及其相关资料若有不实之处，由本人承担一切相关责任

论文作者签名：陈战波 07年 3月27日

学位论文使用授权声明

本人陈战波在导师的指导下创作完成毕业论文。本人已通过论文的答辩，并已经在西安理工大学申请博士/硕士学位。本人作为学位论文著作权拥有者，同意授权西安理工大学拥有学位论文的部分使用权，即：1）已获学位的研究生按学校规定提交印刷版和电子版学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文，可以将学位论文的全部或部分内容编入有关数据库进行检索；2）为教学和科研目的，学校可以将公开的学位论文或解密后的学位论文作为资料在图书馆、资料室等场所或在校园网上供校内师生阅读、浏览。

本人学位论文全部或部分内容的公布（包括刊登）授权西安理工大学研究生部办理。

（保密的学位论文在解密后，适用本授权说明）

论文作者签名：陈战波 导师签名：张付生 07年 3月27日

1 绪论

1.1 课题背景、研究现状及意义

水是人类一切经济活动的命脉，是生命的源泉，是人们赖以生存、生产、生活不可替代的物质资源之一，是现代社会发挥城市中心作用，实现经济繁荣、环境舒适的决定因素，更是城市可持续发展的重要物质基础。随着人类活动的加剧和世界城市化进程的加快，特别是随着工农业生产的不断发展与人口的成倍增长，人类社会的用水量迅速增加，在水资源有限的情况下，水资源供需矛盾日益尖锐，许多城市严重缺水，特别是在工业和人口过度集中的大都市和超大都市区，情况更加严重，水资源问题越来越引起人们的关注。由于水资源不足而出现的供水紧缺和供水危机，已成为全球性的事实，缺水已威胁到人类的生存。现阶段，在洁净水资源短缺的情况下，为了国民经济的持续健康快速发展，为了改善环境，减少污染，使经济与社会、环境协调发展，建立宽泛的和谐社会，除了需要经济有效地使用本地的水资源外，更为重要的是在时空结构上合理调配水资源[1]。

随着我国城市化进程的加快，城市给水系统是合理利用水资源的重要一环，城市供水系统的范围和规模在逐年扩大，与此相应的城市供水量与供水系统复杂性也在逐年提高，传统的经验方法面临前所未有的挑战，城市供水系统的优化调度势在必行。城市供水系统的优化调度研究通常包括用水量预测，工况模拟和供水系统调度决策三个环节的工作。给水系统的第一步工作就是对用水量进行预测。准确的确定城市的用水量是进行给水管网系统计算和分析的基础，是进行给水管网系统模拟的先决条件。因此用水量预测的研究至关重要，是后两个环节工作的基础和前提，它的准确与否直接关系到供水系统工况模拟结果的合理性和调度决策模型的针对性和可靠性[2]。因此长期以来，用水量预测一直是城市供水企业和运行管理部门最为关注的问题之一，与此相关的理论研究一直没有中断过。

服务于供水系统调度运行的用水量预测研究国外起步于七十年代 Yamauchi and Huang、Cassuto and Ryan、HansenR.D 等用时间序列方法或回归分析方法对月用水量预测问题做了研究；Jain D.A(2000)等对周用水量进行了预测讨论；Hartley and Powell[3](1991)、吕谋[4]等(1997)、何文杰[5](2001)、周建华[6](2003)、张雄[7](2005)、侯煜坤[8](2004)等对日用水量预测进行了研究；吕谋[9]等(1998)、张宏伟[10](2001)、刘洪波[11](2002)、柳景青[12](2005)、Vicente J.[13](2004)等从不同的角度，采用不同的方法对调度时用水量预测问题进行了研究。

综上所述，准确地分析、研究城市用水量的变化，建立正确合理的数学模型，进行用水量预测，对于有效利用有限的水资源，避免水资源的浪费和能源的紧缺现象具有非常重要的意义。日用水量预测是根据过去一段时期的用水记录及影响用水量的因素，对未来一天或几天的用水量做出预测，为供水管网系统的优化调度运行提供依据。西安是我国西北重镇，也是水资源匮乏城市之一。因此对西安市日用水量进行建模研究不仅可以为供水部门提供依据，而且能为水资源的合理开发和使用提供参考。

1.2 常用用水量预测方法

用水量预测在城市建设规划和供水系统（优化）调度管理中都具有举足轻重的作用。常用的用水量预测方法可分为两类：一种是解释性预测方法，一种是时间序列分析方法。具体地讲一般采用的模型（或方法）主要有：回归模型（包括多元线性回归和多元非线性回归），线性滑动平均模型，自适应（单）指数平滑模型，季节性指数平滑模型，自回归模型（AR），滑动平均模型（MA），求和滑动平均模型（IMA），自回归滑动平均模型（ARMA），求和自回归滑动平均模型（ARIMA），乘积季节性模型（季节性 ARIMA），灰色预测模型（GM），神经网络模型等，下面分别对这些模型（方法）做以说明：

（1）回归模型

回归模型是一种解释性预测模型，在利用回归模型进行水量预测时，常用的主要有多元线性回归和多元非线性回归两种，这两种回归模型一般都是以温度、节假日和气象状况等因素作为解释变量进行回归分析的。据统计表明，目前在利用回归模型进行用水量预测时，一般大多采用多元线性回归模型，因为它方法简单、模型简洁、易于开发且预测费用相对较低，但不足之处就是预测误差稍大；而多元非线性回归模型相对来讲预测精度较高，不足之处就是模型较复杂，且不易开发。

（2）线性滑动平均模型

线性滑动平均就是把时间序列中连续 n 期的观测值取算术平均作为下一期的预测值，并且每当获取一个新的观测值时就立即将它作为有效数据加入求和平均数中，同时把最老的那个时间数据剔除掉，以此类推进行预测。

线性滑动平均法理论和计算都很简单，但在利用计算机进行预测计算时保存的历史数据量大，若预测目标较多，则需保存极大的数据量。同时，模型参数 n 也难于确定，从原则上讲，一般须选择若干个可能的 n 值建立一个线性滑动平均模型集，并分别计算它们所对应的均方差，从中找出均方差最小的那个 n 值作为模型参数。另外，线性滑动平均模型仅适用于平稳时间序列的预测，若时间序列的数据基础发生震荡（异化），这种方法就不再适用。因此，线性移动平均模型适用于短期平稳时间序列的预测。

（3）自适应单指数平滑模型

自适应单指数平滑模型亦称自动调整平滑参数的单指数平滑模型，它是对单指数平滑模型的改进，能反映时间序列的变化情况，能告诉预测者时间序列是否发生变化、预测是否失去控制（通过追踪信号）。一般来讲，它适用于平稳时间序列的预测。

（4）季节性指数平滑模型

对既有季节性（周期性）因素影响又有趋势性因素影响的时间序列进行预测时，采用一般的指数平滑模型效果不是很明显，而应采用对周期性时间序列预测精度较高的季节性指数平滑模型。季节性指数平滑模型的基本原理是把具有季节性影响因素的时间序列中的趋势性因素、周期性因素和水平因素分离出来，然后再合起来进行预测。

（5）自回归模型（AR）

自回归模型对平稳性或有随机扰动性（趋势性）时间序列的预测精度较高，模型中不存在其它变量，不受模型“相互独立”假设条件的约束，可以消除或改进普通回归预测中由自变量选择、多重共线性及序列相关性等造成的困难。在利用 AR 模型进行用水量预测时，应注意分析用水量时间序列的数据模式（平稳性），这是模型应用的前提。如果在预测过程中用水量时间序列发生变化，则预测的准确性下降。另外值得一提的是 AR 模型还常用于预测残差的修正，即建立残差的自回归模型。

（6）滑动平均模型（MA）

同自回归模型一样，滑动平均模型对平稳性（或弱趋势性）时间序列的预测有较高的精度， 在利用 MA 模型进行用水量预测时，需注意用水量时间序列应是平稳的。否则，应利用差分方法对时间序列进行预处理，使之先变成平稳序列，然后再建立模型进行预测，此时建立起来的模型称之为求和滑动平均模型（IMA）。

（7）自回归滑动平均模型（ARMA）

自回归滑动平均模型（ARMA）是一种针对平稳时间序列的常用预测模型，简单的说它可以认为是自回归模型和滑动平均模型的混合。

一般来说，在利用 ARMA 模型进行用水量预测时，多用于用水量时间序列为平稳序列的情况，模型预测效果较为理想。如果用水量时间序列为非平稳时间序列，且需运用自回归滑动平均方法建模，则可采用差分技术先对原时间序列进行预处理（直到处理后的新序列为平稳时间序列为止），然后再对新序列建立模型进行预测，这种方式建立起来的（用水量）预测模型称为求和自回归滑动平均模型（ARIMA）。再进一步，如果用水量时间序列具有周期性，则可采用一类称为季节性 ARIMA 的模型进行预测。

（8）灰色预测模型（GM）

灰色系统理论是基于关联度收敛原理、生成数、灰导数和灰微分方程等观点和方法建立微分方程模型的一种系统理论，它又简称灰色理论或灰理论，基于灰色系统理论中 GM（1，1）模型的预测称为灰色预测。

灰色预测不要求有很多数据，在利用计算机作为预测计算工具时，不需占过多内存，同时它也不需知道原始数据的数据特征，而只需通过有限次的数据生成，便可将无规则的离散原始序列转化为有规则序列并且序列模型中参数的分布是灰色的，可保持原始序列的特征，能较好地反应系统的实际情况，建模预测精度较高。在利用灰色预测模型进行用水量预测时，特别适用于用水量历史记录较少而用水量影响因素又较多的一类，例如年用水量的预测。

（9） 神经网络模型

（人工）神经网络模型是指模拟人脑神经系统的结构和功能，运用大量的处理部件（如神经元）由人工方式建立起来的网络系统。神经网络的运用很多（如智能控制、模糊控制、模式识别和模拟仿真等），在这之中有相当的一类就是进行（智能）预测。常用于预测的神经网络主要有向前神经网络（如 BP 神经网络）和回归神经网络（如 Jordan 网络和 Elman

网络），一般来说它们都由输入层、隐含层和输出层构成，同层之间各神经元互不相连，相邻神经元之间通过权相连。在利用神经网络模型进行预测时，神经网络应具有很强的接受训练性和自适应性，这其中关键的影响因素就是神经网络学习训练算法的选择，通常采用的算法主要有：Williams 和 Zipser 提出的实时递归学习算法，Narenda 和 Parthasarathy 提出的动态 BP 算法，Rumelhart 等提出的时间反转算法，Baldi 提出的梯度下降学习算法等等。具体在利用神经网络进行水量预测时，输入（矢）量可取历史水量观测值，输出（矢）量可取未来水量预测值，其间的学习、训练过程全由神经网络自动完成，不需人为干预，具有一定的智能性。同时，由于神经网络的适应性较强（值得注意的是神经网络对季节性规律的学习较慢），因此它基本可用于各种水量的预测。

以上方法，在城市用水量预测中，发挥了较好的作用。本文考虑到用水量的非线性特征，建立了西安市日用水量的多元非参数回归模型。另外，由于线性自回归模型没有考虑到温度对用水量过大的影响，因此本文在前人工作的基础上尝试用非线性时间序列中的部分线性自回归模型方法对西安市日用水量进行预测。

1.3 部分线性模型介绍

1.3.1 部分线性回归模型研究现状

Engle et al.(1986)在研究天气和电力销售之间关系时首先引入了部分线性回归模型，Robinson (1988)[14] 则首次将部分线性回归模型引入时间序列。关于模型

$$Y_i = X_i^T \beta + g(t_i) + \varepsilon_i, i=1,2,\cdots,n \tag{1.1}$$

自 Eagle et al (1986)在研究气候条件对电力需求影响这一实际问题时提出上述模型以来，已出现一系列研究成果。Schick 应用 Bickel 中的一些结论研究了上述模型的一类特殊情形中，β 的渐近有效估计的构造；Heckman(1986)[15] 研究了 (X_i,t_i) 是 $i.i.d$ (独立同分布)随机样本，且 $\{X_i\}$ 和 $\{t_i\}$ 是相互独立的，并且 $g(\cdot)$ 的估计取一类样条估计时，β 的加权最小二乘估计 $\hat{\beta}_n$ 的渐近正态性；Rice(1986)[16] 研究了 (X_i,t_i) 是固定设计点列，其 $g(\cdot)$ 的估计取一类样条时，β 的估计的协方差函数的渐近性质；Chen(1988)[17] 研究了当 $h_j(t)=E(X_j|T=t)$ 关于 X 满足 $\alpha(0\le\alpha\le1)$ 阶 Lipschitz 条件，且 $g(\cdot)$ 的估计取一分段多项式估计时，β 的加权最小二乘估计 $\hat{\beta}_n$ 的渐近正态性及其 $g(\cdot)$ 的估计的弱收敛速度；其后,一些学者还研究了当 $g(\cdot)$ 的估计取一些样条估计时，β 的若干估计的性质。关于未知函数 $g(\cdot)$ 取核估计的情形，Speckman(1988)[18] 和 Robinson[19] 分别独立地研究了，当 $h_j(t)=E(\bar{X}_{1j}|T=t)$ 关于 t 满足 $\alpha(0\le\alpha\le1)$ 阶 Lipschitz 条件，且 $g(\cdot)$ 的估计取 Parztn---Roseblatt 核估计时，β 的加权最小二乘估计 $\hat{\beta}_n$ 的渐近正态性及其 $\hat{\beta}_n$ 和 $\hat{g}_n$ (g 的估计)的弱收敛速度，该文去掉了 Speckman(1988)[18] 中对核函数所附加的一些不易验证的条件；而后 Gao[20] 又进一步研究了当 g 的估计取一类核估计序列时，β 的加权最小二乘估计 $\hat{\beta}_n$ 的渐近正态性及其 $g(\cdot)$ 的估计的最优强收敛速度；Hong(1993)[21] 又研究了模型 (1.1)中，当 $g(\cdot)$ 的估计取一类近邻估计时的 $\hat{\beta}_n$ 渐近正态性 $\hat{\beta}_n$ 和 $\hat{g}_n$ (g 的估计)的弱收敛速度,得到了一些深刻的结果。我国学

者，胡舒合(1994)[22]研究了半参数回归模型估计的强相合性，钱伟民(2000)[23]研究了半参数回归模型的误差小波估计。其他如高集体、洪圣岩、赵选民等也在模型估计方面做了一定的工作。柴根象(1995)[24]对半参数回归模型的二阶段估计进行了研究，考虑回归模型（1.1）中 g 为 R^1 上未知函数，β 为 $p\times 1$ 维待估参数向量，并且基于模型的可加性得到了 β 和 g 的估计量 $\hat{\beta}_n, \hat{g}_n$ 并证明了它们具有很好的大样本性质。

1.3.2 部分线性自回归模型研究现状

部分线性自回归模型的一般形式为：

$$Y_t = \beta^T X_t + g(Z_t) + \varepsilon_t \tag{1.2}$$

其中，$X_t = (X_{t1}, \cdots, X_{tp})^T$，$X_{ti}(i=1,\cdots,p)$ 和 Z_t 为 $Y_t \in R$ 的滞后值，$\beta = (\beta_1, \cdots, \beta_p)^T$ 为未知参数向量，p 是正整数，g 为未知可测函数，$\{\varepsilon_t\}$ 为 i.i.d.随机变量序列，均值为 0，方差为 σ^2，且 ε_t 与 $X_{ti}(i=1,\cdots,p)$ 和 Z_t 独立。Gao(1995)[25]考虑模型：

$$Y_t = \beta Y_{t-1} + g(Y_{t-2}) + \varepsilon_t (t \geq 3) \tag{1.3}$$

基于 g 的核光滑研究 β 和 σ^2 的估计量的渐近性质；Gao and Liang (1995)[26]也对 Gao (1995)研究的模型予以考虑，基于 g 的分段多项式估计研究 β 的估计量的渐近正态性，同时还研究了 β 的伪最小二乘估计量和误差方差 σ^2 的估计量的渐近正态性；Gao[27] (1998)在研究半参数自回归模型中提出非参数函数的有限级数近似，研究了级数近似中求和数的适应选择并给出了大样本性质；Schick(1999)[28]在 V 一致历经条件下建构 Gao (1995)[26]研究的模型中参数的有效估计，还考虑了局部渐近正态性和局部渐近最小最大性；Gao and Yee(2000)[29]基于非参数函数的核估计给出部分线性自回归模型中参数估计的渐近正态性；Linton and Mammen(2003)[30]研究一类半参数(∞)模型，基于核光滑和 profiled likelihood 提出一种估计方法，建立参数的分布理论和非参数函数的逐点分布，讨论参数部分和非参数部分的有效性。Härdle, Liang and Gao(2000)[31]对部分线性模型作了详细的介绍。

Gao and Tong (2004)[32]对西澳大利亚(WA)渔业捕鱼量和捕鱼船在海中作业天数建立了部分线性自回归模型。反映上述现象的时间序列模型可表示为：

$$Y_t = \phi(Y_{t-1}, \cdots, Y_{t-d}) + g(Z_t) + \varepsilon_t \tag{1.4}$$

其中，$\phi(\cdot)$ 和 $g(\cdot)$ 分别为 R^d 和 R^q 上的可测函数，d 和 q 皆为正整数，Z_t 为影响因素，随机误差序列 $\{\varepsilon_t\}$ 为白躁声序列，满足条件 $E(\varepsilon_t)=0$，$E(\varepsilon_t^2)=\sigma^2<\infty$，且 ε_t 与 $\{Y_s, s<t\}$ 相互独立。若 $\phi(\cdot)$ 和 $g(\cdot)$ 均为线性函数，则(1.4)为线性 ARMAX 模型 (autoregressive moving-average systems with exogenous variable)，这方面已有大量研究。然而，对于许多实际问题，$\phi(\cdot)$ 或 $g(\cdot)$ 为未知非线性函数，在此情况下研究模型（1.4）颇有意义。Bosq and Shen[33] (1998)针对 $\phi(\cdot)$ 为未知非线性函数，$g(\cdot)$ 为线性和非线性函数两种情形进行了讨论，在 α 混合条件下基于核光滑给出估计的 a.s.和 L_s 收敛性。对于 $\phi(\cdot)$ 为线性函数，$g(\cdot)$ 为未知非参数函数，Liang (1996)[34]考虑了一类特殊情形：

$$X_t = \beta X_{t-1} + g(U_t) + \varepsilon_t \tag{1.5}$$

其中，U_t为服从[0,1]均匀分布的独立同分布随机变量，利用分段多项式估计给出了一定条件下未知参数β的一类二阶渐近有效估计量；Zhu 和 An (1994)[35]考虑了模型(1.4)中$g(\cdot)$为未知非常数光滑函数的情形，在一定条件下给出了θ的强相合估计和g的相合核估计；Teräsvirta, Tjøstheim and Granger (1994)[36]介绍了当模型(1.4)中$g(\cdot)$为标量时，可用向后拟合算，在非常一般的条件下可以得到如 Heckman (1986)[15]和 Robinson (1988)[14]中参数项的$\sqrt{n}$相合性，Powell, Stock and Stoker (1989)进一步提出了理论并给出了经济应用。部分线性自回归模型的优点是它不仅集中了主要部分(即参数分量部分) 的信息并具有较强的解释能力，而且能有效地减少或克服非参数方法信息损失过多的问题，因此部分线性模型是一个在实用上有重大意义的领域[37]，目前所见有关部分线性自回归模型的研究文献还不多。另外，除了上述所述核估计和其它估计方法外，局部多项式估计、样条估计和小波估计的应用还不多见。

1.4 本课题的主要目的及内容

本课题主要目的是利用回归方法、线性时间序列方法、多元非参数方法以及非线性时间序列方法对西安市日用水量进行拟合与预测，并检验其可行性，为西安市供水系统预测方法提供参考，并且也可以为非参数和非线性时间序列方法的应用研究提供一定的参考价值。

具体研究内容包括：

（1）对西安市日用水量、气温以及节假日数据进行分析，建立能服务于供水系统的综合预测模型，并对该模型进行检验。通过检验后的模型用来作预测，与实际数据进行对比判别该模型的优劣。

（2）对多元非参数回归估计方法以及非参数预测方法进行研究，建立相应的模型，结合西安市日用水量数据，用该模型对城市日用水量作出预测。辨别出该模型能否满足供水系统的需要。

（3）根据线性时间序列中 Box-Jenkins 建模方法，对一段西安市日用水量数据进行分析，按照建模的基本步骤，最终确定该时间序列数据适合何种线性时间序列模型，并用最后建立的模型对日用水量作出预测，与实际数据对比，辨别模型的好坏。

（4）综合日用水量的特点，把非线性时间序列方法引入，研究部分线性自回归模型的建模理论，建立城市日用水量的部分线性自回归模型，线性部分考虑城市日用水量，非线性部分考虑当天的最高温度。用建立的模型对日用水量作出预测，辨别出模型的优劣性。

1.5 小结

本章综述了城市用水量的预测方法以及部分线性模型的研究现状，阐述了本课题的背景、研究现状、意义及本课题的研究目的和研究内容。

2 西安市日用水量的综合预测模型

2.1 城市日用水量影响因素相关性分析

一般来说，城市日用水量受到具体工商业分布、居民活动(如节假日与平常日的区别)、气象条件、人口增长及突发事件等因素的影响，并与这些因素间存在某种相关性 ，表现出一定的变化特征，日用水量的回归预测模型正是基于这种相关性和变化特征建立起来的。下面就城市日用水量的影响因素作出分析：

第一，工商业分布：工商业用水在城市用水中占了很大一部分，工商业分布对日用水量的结构影响较大，但对日用水总量变化的影响较小（因为它的用水较为稳定），在日用水量变化中可作为一个常量考虑。

第二，气象因素：气象因素是影响日用水量变化的各种因素中最主要的一个，是影响日用水量变化的“控制”因素。气象因素主要包括日最高温度、日平均温度及天气阴晴状况，它们对日用水量变化都有极大影响。通常随着日最高温度、日平均温度的升高，日用水明显增大。另外，其它气象因素例如日最低温度、空气湿度、气压值等一般可包含或部分包含在上述因素中，对日用水量变化影响不是很显著，可不予考虑。

第三，节假日：节假日从历史经验和原始数据的对比分析中可以得出：节假日用水比同条件下的平常日用水低。故可考虑用节假日进行用水量的修正（一般可采用线性修正）。

第四，其它因素：除上述主要影响因素外，其它因素例如人口增长、突发事件（包括消防、管网维修）等也会对日用水造成影响，但它们对日用水量的影响要么是微弱的，要么是暂时的，都不具有显著性，在考虑影响因素时可忽略不计。

2.2 西安市日用水量的综合预测模型

2.2.1 西安市日用水量的多元线性回归模型

从上述分析可以得出：西安市日用水量的主要影响因素包括日平均温度、日最高温度及节假日等（工商业分布影响可单独考虑），且它们与日用水量的相关关系根据已有的研究结果如张雄[7]吕谋[4]可以用线性关系拟合。故可分别以上述各主要影响因素为解释变量，以日用水量为因变量，建立关于日用水量预测的多元线性回归模型，具体模型形式如下：

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + e_t \qquad (2.1)$$

其中，Y_t表示日实际用水量；

X_{1t}表示日最高温度；

X_{2t}表示日平均温度；

X_{3t}表示节假日，根据影响程度可取X_{3t}=0~3，0：平常日；1：普通节假日，如周末；2：较重要节假日，如五一，国庆，元旦等；3：重要节假日，如春节等；

$\beta_0,\beta_1,\beta_2,\beta_3$表示回归系数；

e_t表示回归残差。

随着季节、气候的变化，回归系数$\beta_0,\beta_1,\beta_2,\beta_3$是动态变化的，特别是在冬季西安市用水结构发生变化，有大量的采暖用水，回归系数发生较大的变化，需进行重新回归。但在不需要采暖时（春、夏、秋）回归系数$\beta_0,\beta_1,\beta_2,\beta_3$变化较小，可视作静态。下面利用西安市2003年6月1号到8月24号的用水资料（数据见附录，数据来源于西安市自来水公司与西安市气象局网站）对它们进行估计。

2.2.2 模型参数的估计[38]

回归系数$\beta_0,\beta_1,\beta_2,\beta_3$的估计方法很多，估计回归参数的最基本的方法是最小二乘法，这个方法不仅仅在统计学中，就是在数学的其它分支，例如运筹学、计算数学、逼近论和控制论等，都是很重要的求解方法。因此本文通过最小二乘法对多元回归模型的系数做出估计。

最小二乘法介绍：假设Y为因变量，$X_1,\cdots,X_{p-1}$为对Y有影响的$p-1$个自变量，并且它们之间具有线性关系：

$$Y=\beta_0+\beta_1X_1+\cdots+\beta_{p-1}X_{p-1}+e \tag{2.2}$$

其中e为误差项，它表示除了$X_1,\cdots,X_{p-1}$之外其它因素对Y的影响以及试验或测量误差。$\beta_0,\beta_1,\cdots,\beta_{p-1}$是待估计的未知参数。假定我们有了因变量$Y$和自变量$X_1,\cdots,X_{p-1}$的$n$组观测值：

$$(x_{i1},\cdots,x_{i,p-1},y_i),i=1,\cdots,n,$$

它们满足：

$$y_i=\beta_0+x_{i1}\beta_1+\cdots+x_{i,p-1}\beta_{p-1}+e_i,i=1,\cdots,n \tag{2.3}$$

误差项$e_i,i=1,\cdots,n$满足$Gauss-Markov$假设。若用矩阵形式表示为

$$\begin{pmatrix}y_1\\y_2\\\vdots\\y_n\end{pmatrix}=\begin{pmatrix}1&x_{11}&\cdots&x_{1,p-1}\\1&x_{21}&\cdots&x_{2,p-1}\\\vdots&\vdots&\vdots&\vdots\\1&x_{n1}&\cdots&x_{n,p-1}\end{pmatrix}\begin{pmatrix}\beta_0\\\beta_1\\\vdots\\\beta_{p-1}\end{pmatrix}+\begin{pmatrix}e_1\\e_2\\\vdots\\e_n\end{pmatrix} \tag{2.4}$$

因此线性回归模型可以表示为：

$$Y=X\beta+e,E(e)=0,Cov(e)=\sigma^2I_n \tag{2.5}$$

获得参数向量β的估计的一个最重要方法是最小二乘法。这个方法是找β的估计，使得偏差向量$e=y-X\beta$的长度之平方$\|y-X\beta\|^2$达到最小，通过求解可以得到β的估计：

$$\hat{\beta}=(X'X)^{-1}X'Y \tag{2.6}$$

根据最小二乘法和用水量、节假日以及气温数据可以得到$\hat{\beta}=(477980,1346,12821,-10924)'$，因此最后可以得到多元线性回归方程为：

$$Y_t = 477980 + 1346X_{1t} + 12821X_{2t} - 10924X_{3t} + e_t \tag{2.7}$$

其中Y_t为回归用水量。回归方程是否具有显著性还需通过显著性检验。

2.2.3 模型显著性检验

多元线性回归方程是否真正成立，是否确实具有线性相关性，在参数估计出来以后还必须通过显著性检验。通常显著性检验包括回归方程的显著性检验（F 检验）和回归系数的显著性检验（t 检验）。回归系数的显著性检验，根据张雄[7]、吕谋[4]的研究结果可以省去。下面只进行回归方程的检验：

假设 H_0：$\beta_0 = \beta_1 = \beta_2 = \beta_3 = 0$，若其被接受，则回归方程不具有显著性。根据《线性统计模型》[38]的理论有：

$$R = \sqrt{\frac{\sum_{t=1}^{t}\left(\hat{Y}_t - \bar{Y}\right)^2}{\sum_{t=1}^{t}\left(Y_t - \bar{Y}\right)^2}} \tag{2.8}$$

其中，R表示复相关系数。取$t=85$，代入各用水相关数据，利用 Matlab 进行计算，最终结果为：$R=0.7110$。在假设成立的条件下，统计量服从F分布，故有：

$$F_c = \frac{\dfrac{R^2}{m}}{\dfrac{1-R^2}{t-m-1}} \tag{2.9}$$

其中，F_c表示F统计量；m表示自变量个数，这里$m=3$；t表示统计的样本总数，这里$t=85$；R表示复相关系数，这里$R=0.7110$。代入以上各数据，经计算得：F_c=40.0537，查“F分布表”，其中第一自由度$df_1 = m = 3$，第二自由度$df_2 = t-m-1 = 81$，置信度为 95%（即显著性水平$\alpha = 0.05$），有：$F_{0.05}(3,81)=2.72$，故F_c=40.0537＞$F_{0.05}(3,81)=2.72$所以拒绝原假设，即在 95%的置信度范围内可认为方程线性回归显著。

从上述分析可知，回归方程具有较好的显著性，但这并不意味着它就能直接用于实际预测，因为回归残差是否是白噪声序列，是否具有自相关性，这些都不是很明白，因此还需要进行残差序列的自相关分析。

2.2.4 回归残差自相关分析

一般来说，残差序列应是一个白噪声序列（既随机序列）。否则，若残差序列具有一定的自相关性，则将给预测带来极大偏差。因此有必要进行残差序列的自相关分析以便确认其是否属白噪声序列。所谓自相关分析就是借助某些数学手段进行的序列自相关性诊断，而自相关性则是指序列中前后观测值之间存在的一定的相关关系属性。常用的自相关分析方法有图示检验法和自相关系数法，图示检验法直观简单，但有时候很难根据图形做出明确的判断，而自相关系数法正好避免了上述缺点，因此下面我们采用自相关系数法进

行残差序列的自相关分析。

根据自相关分析理论[39]有：

$$\rho_k = \frac{\sum_{i=1}^{t-k}(e_i - \bar{e})\ (e_{i+k} - \bar{e})}{\sum_{i=1}^{t}(e_i - \bar{e})^2} \tag{2.10}$$

其中，ρ_k表示时间延迟为k的自相关系数，e_i表示第i期的回归残差值，$\bar{e}$表示回归残差序列的均值，对于回归残差值满足$\bar{e}$=0（由计算引起的误差可忽略不计），t表示观测样本总数，这里取t=85,一般只需要计算t/4 个即可。带入上述数据，公式(2.10)可利用 Matlab 编程进行计算和分析，可得残差自相关系数如图 2-2，图 2-1 为残差变化曲线。

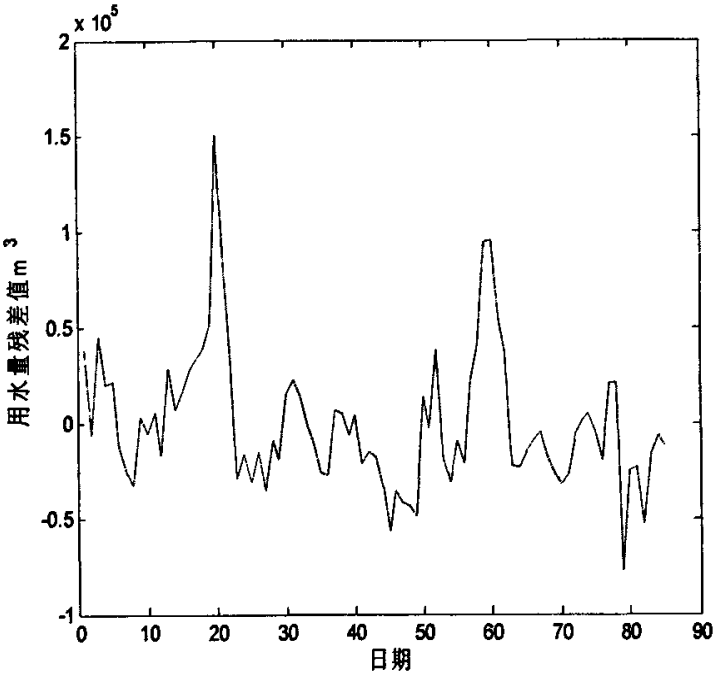

图 2-1　回归预测残差变化曲线

Figure2-1 the prediction residual error curve of regressive model

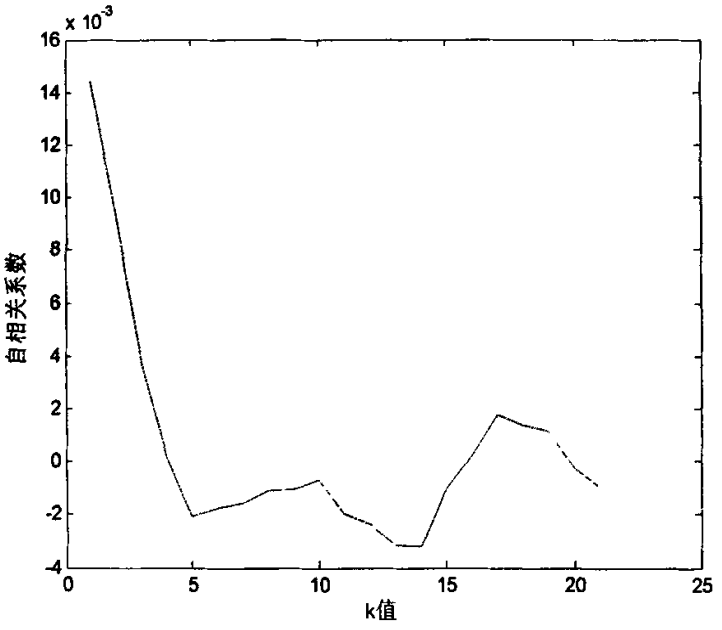

图 2-2 残差序列自相关系数变化曲线

Figure2-2 the autocorrelation coefficient changeable curve of residual error series

从图 2-2 明显可以看出自相关系数变化呈现一定的规律性，这表明残差序列不属于白噪声序列，仍具有自相关性，须对原模型残差进行修正。

2.2.5 模型残差的自回归修正

由于在本文的第 4 章将会详细介绍自回归模型的建模过程，所以在此只给出结果，详细方法可见第 4 章。由于回归模型残差序列仍具有自相关性，因此需对模型进行残差修正。结合残差的变化曲线通过分析，知道残差序列数据不平稳，通过差分平稳后可考虑建立差分后数据的一阶自回归模型 AR(1)，模型具体形式如下：

$$\Delta e_t = -0.0028\Delta e_{t-1} + \varepsilon_t \tag{2.11}$$

转化为原来的残差序列可得：

$$e_t = 0.9972e_{t-1} + 0.0028e_{t-2} + \varepsilon_t \tag{2.12}$$

2.2.6 自回归模型有效性检验[39]

自回归模型是否有效，同样需通过其残差序列$\{\varepsilon_t\}$的自相关分析来获得。同回归残差自相关分析一样，通过计算可得到如图 2-3 残差自相关系数及其变化曲线，根据χ^2分布理论有：

$$\chi_c^2 = t\sum_{i=1}^{m}\rho_i^2 \tag{2.13}$$

其中，χ_c^2 表示 χ^2 统计量；ρ_i 表示自相关系数；m 表示自相关系数的个数，在这里取 m=20；t 表示观测值的总数，在这里取 t=85. 代入上述数据，经计算有：χ_c^2=1.674，查“χ^2 分布表”，其中自由度为 df=m-1=19，置信度为95%（即显著性水平 $\alpha=0.05$），有：$\chi^2_{0.05}(19)=30.144$。由于 $\chi^2_{0.05}(19)=30.144>\chi_c^2=1.674$ 故有 95%的置信度认为这 20 个自相关系数与 0 没有显著性差异，残差序列属白噪声序列。

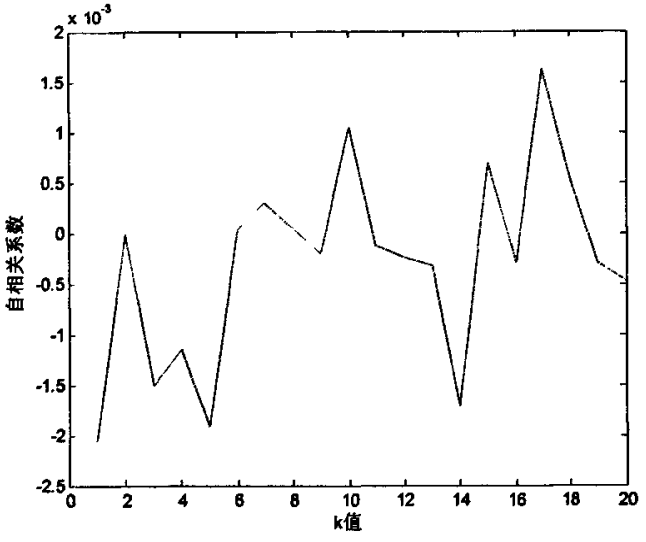

图 2-3 修正后残差序列自相关系数变化曲线

Figure2-3 the autocorrelation coefficient changeable curve of corrected residual error series

2.2.7 西安市日用水量的综合预测模型

结合(2.7)和(2.12)式，西安市日用水量预测的综合模型可表示如下：

$$\begin{aligned} Y_t &= 477980+1346X_{1t}+12821X_{2t}-10924X_{3t} \\ &\quad +0.9972e_{t-1}+0.0028e_{t-2}+\varepsilon_t \end{aligned} \tag{2.14}$$

其中，Y_t 表示第 t 日用水量预测值；

X_{1t} 表示第 t 日最高温度；

X_{2t} 表示第 t 日平均温度；

X_{3t} 表示第 t 日节假日情况，根据节假日的影响程度其取值为 0~3；

e_{t-1}，e_{t-2}，分别表示从 t 日开始（向前）t-1,t-2 期的回归残差量；

ε_t 表示修正后残差项。

日用水量预测综合模型（2.14）是多元线性回归模型和自回归模型的有机结合，其预测残差等于上述进行自回归修正后的残差。前面已论述了自回归模型修正的有效性，故模型（2.14）是有效的。下面利用模型（2.14）对日用水量作出预测，预测结果见表 2-1 与图 2-4。

表 2-1 6.03～8.24 实际值与预测值比较

Table2-1 the comparison of real data and forecasting data from 6.03 to 8.24

日期	实际值 /m^3	预测值 /m^3	绝对误差 /m^3	相对误差 /%	日期	实际值 /m^3	预测值 /m^3	绝对误差 /m^3	相对误差 /%
6.01	870265	-------	---------	---------	7.14	751511	766911	-15400	-2.05
6.02	852089	-------	---------	---------	7.15	734782	758169	-23387	-3.18
6.03	839438	797618	41820	4.98	7.16	737928	716994	20934	2.84
6.04	802901	828369	-25468	-3.17	7.17	803530	810439	-6909	-0.85
6.05	859839	858614	1225	0.14	7.18	847766	848820	-1054	-0.12
6.06	854856	887970	-33114	-3.87	7.19	851843	857358	-5515	-0.65
6.07	843568	856987	-13419	-1.59	7.20	873839	811857	61982	7.09
6.08	851760	859387	-7672	-0.90	7.21	879179	895243	-16064	-1.83
6.09	830944	795107	35837	4.13	7.22	910541	869172	41369	4.54
6.10	784532	792875	-8343	-1.06	7.23	882903	939972	-57069	-6.46
6.11	844146	833042	11104	1.32	7.24	885238	898165	-12927	-1.46
6.12	831512	854371	-22859	-2.75	7.25	946915	925146	21769	2.30
6.13	880478	835018	45460	5.16	7.26	923076	934725	-11649	-1.26
6 14	873326	895054	-21728	-2.49	7 27	923460	881889	41571	4.50
6.15	904991	895381	9610	1.06	7.28	965979	946132	19847	2.05
6.16	915223	904086	11137	1.22	7.29	1014790	960186	54604	5.38
6.17	942777	936980	5797	0.61	7.30	1042459	1041625	834	0.08
6.18	964709	959222	5487	0.57	7.31	977334	1019860	-42526	-4.35
6.19	964373	951334	13039	1.35	8.01	889001	904386	-15385	-1.73
6.20	986222	897246	98976	8.06	8.02	768613	828316	-59703	-7.77
6.21	954107	1023784	-69677	-7.30	8.03	745256	746685	-1429	-0.19
6 22	850633	904179	-53546	-6.29	8.04	797046	788491	8555	1.07
6.23	852424	908790	-56366	-6.61	8.05	843197	837291	5906	0 70
6.24	890963	878761	12202	1.36	8.06	880528	876104	4424	0.50
6.25	831124	845721	-14597	-1.75	8.07	885245	897785	-12540	-1.42
6.26	810729	794805	15924	1.96	8.08	826712	835150	-8438	-1.02
6.27	857331	877063	-19732	-2.30	8.09	789677	796365	-6688	-0.84
6.28	878399	852135	26264	2.99	8.10	766365	760681	5684	0.74
6.29	840524	851273	-10749	-1.27	8.11	773343	752381	20962	2.71
6.30	815263	780848	34415	4.22	8.12	768931	761681	7250	0.94
7.01	792317	784473	7844	0.99	8.13	780359	777071	3288	0.42
7.02	831177	840251	-9074	-1.09	8.14	768013	777580	-9564	-1.25
7.03	821353	836203	-14850	-1.81	8.15	755659	770873	-15214	-2.01
7.04	842671	851811	-9140	-1.08	8.16	781812	740931	40881	5.22
7.05	835624	850798	-15174	-1.82	8.17	803163	823217	-20054	-3.51
7.06	851529	853339	-1810	-0.21	8.18	812807	881563	-78756	-7.05
7.07	857987	824213	33774	3.93	8.19	815311	763031	52280	6.41
7.08	851434	862558	-11124	-1.30	8.20	810067	807407	2660	0.32
7.09	813155	824520	-11365	-1.39	8.21	836303	866862	-30559	-3.65
7.10	846894	836392	10502	1.24	8.22	862114	824785	37329	4.33

续表									
7.11	803895	830069	-26174	-3.26	8.23	877021	837611	39410	4.49
7.12	756685	719941	46744	6.18	8.24	862650	887623	-24973	-2.89
7.13	737839	770479	-42640	-5.77					

注：相对误差$=\frac{\text{实际值}-\text{预测值}}{\text{实际值}}\times 100\%$

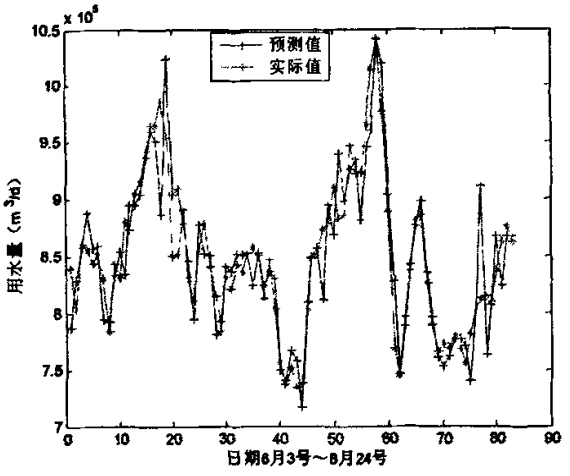

图 2-4 西安市日用水量预测变化曲线

Figure2-4 the prediction changeable curve of xi'an daily water demand

为了便于与其它预测方法做比较，把 8 月 25 号～31 号预测结果单独罗列出来。预测结果见表 2-2

表 2-2 8.25～8.31 实际值与预测值比较

Table2-2 the comparison of real data and forecasting data from 8.25 to 8.31

日期	实际值 / m^3	预测值 / m^3	绝对误差 / m^3	相对误差 / %
8.25	856551	825810	30741	3.59
8.26	811914	802819	9095	1.12
8.27	849784	914200	-64416	-7.58
8.28	791414	812020	-20606	-2.60
8.29	747147	761096	-13949	-1.86
8.30	785090	724088	61002	7.77
8.31	764637	719893	44744	5.85

2.3 本章小结

本章对西安市日用水量、气温以及节假日数据进行分析，首先建立了日用水量的多元线性回归模型，对模型进行了检验和残差修正，最后建立了日用水量的综合预测模型，由预测结果可以看出，该模型预测精度不理想，需要探讨其它预测方法进一步提高精度以满足供水系统的需要。

3 西安市日用水量的多元非参数回归模型

非参数回归模型研究是当前应用统计研究中的一个重要方面，是继协整理论之后，在经济、管理、水文等应用研究中的又一个热点研究方向。由于线性和非线性回归模型都假定变量的关系已知，现实中，事物变量之间的关系未必是线性关系或可线性化的非线性关系，而变量之间的参数非线性关系又很难确定。所以，传统线性或非线性统计模型在实际应用中往往存在模型的设定误差，不能满足经济、管理、水文等应用研究的需要。而非参数回归模型假定变量的关系未知，要对整个回归函数进行估计，因而非参数回归模型是较线性和非线性回归模型更符合现实的模型。非参数回归模型包括完全非参数回归模型（简称非参数回归模型）和半参数回归模型两类。非参数单方程应用研究模型的研究在近 30 年间得到了迅速的发展，非参数联立方程的理论研究也取得了较大发展。近十几年来，非参数回归模型在国际上得到了广泛的应用，已成为应用最多的模型之一[40]。

3.1 多元非参数回归模型介绍[40]

设 Y 为被解释变量，是随机变量，X 为 d 维解释变量向量，是影响 Y 的若干重要因素，它既可以是确定性变量，也可以是随机变量。给定样本观测值 $(X_1, Y_1),(X_2, Y_2),\cdots(X_n, Y_n)$，假定 $\{Y_i\}$ 独立同分布，于是可以建立多元非参数回归模型：

$$Y_i = m(X_i) + \varepsilon_i, \quad i = 1,2,\cdots,n \tag{3.1}$$

其中：$m(\cdot)$ 是未知函数，ε_i 为随机误差项，它反映了除解释变量外，其它影响被解释变量的可观察或不可观察的因素对被解释变量的影响，以及模型的设定误差等。当解释变量为确定性变量时，假定随机误差项的数学期望为零，即 $E\varepsilon_i = 0$。此时，被解释变量的数学期望 $EY_i = m(X_i)$。当解释变量为随机变量时，假定解释变量与随机误差项独立，假定随机误差项的条件数学期望为零即 $E(\varepsilon_i | X_i)=0$。此 时，被解释变量的数学期望 $E(Y_i | X_i)=m(X_i)$。

3.2 多元回归函数 $m(\cdot)$ 的估计[40]

3.2.1 回归函数 $m(\cdot)$ 的核估计

Nadaraya-Watson 提出著名的 N-W 核估计。其思路如下：选定原点对称的概率密度函数 $K(\cdot)$ 为核函数，其中：

$$\int K(u)du = 1 \tag{3.2}$$

窗宽 $h_n > 0$，定义核权函数为：

$$W_{ni}(x) = \frac{K_{h_n}(X_i - x)}{\sum_{j=1}^{n} K_{h_n}(X_j - x)}, \tag{3.3}$$

其中 $K_{h_n}(u) = h_n^{-1} K(uh_n^{-1})$ 也是一个概率密度函数。Nadaraya-Watson 核估计定义为：

$$\hat{m}_n(x) = \sum_{i=1}^{n} W_{ni}(x) Y_i, \tag{3.4}$$

容易推得：

$$\min_{\theta}\sum_{i=1}^{n}W_{ni}(x)(Y_i-\theta)^2=\sum_{i=1}^{n}W_{ni}(x)(Y_i-\hat{m}_n(x))^2, \tag{3.5}$$

所以核估计等价于局部加权最小二乘估计。若$K(\cdot)$是[-1,1]上的均匀概率密度函数，则$m(x)$的Nadaraya-Watson核估计就是落在$[x-h_n,x+h_n]$的X_i对应的Y_i的加权算术平均值。所以，称参数h_n为窗宽，h_n越小，参加平均的Y_i就越少；h_n越大，参加平均的Y_i就越多。若$K(\cdot)$是$(-\infty,\infty)$上原点对称的标准正态密度函数，则$m(x)$的Nadaraya-Watson核估计就是Y_i的加权算术平均值。当X_i落在离x越近时，权数就越大；落在离x越远时，权数就越小；当X_i落在$[x-3h_n,x+3h_n]$之外时，权数基本上为零。

多元非参数模型的不变窗宽核估计为：

$$\hat{m}_n(x,\ h_n)=\frac{\sum_{i=1}^{n}K_{h_n}(X_i-x)Y_i}{\sum_{i=1}^{n}K_{h_n}(X_i-x)}, \tag{3.6}$$

其中h_n为窗宽，$K_{h_n}(u)=h_n^{-d}K(h_n^{-1}u),K(\cdot)$是$d$维对称密度函数，$K(u)\geq 0,\int K(u)du=1$，最常用的核函数为：

$$K(u)=\frac{d(d+2)}{2S_d}(1-u_1^2-,\cdots,-u_d^2)_+ \tag{3.7}$$

其中$S_d=2\pi^{d/2}/\Gamma(d/2)$，当窗宽$h_n=cn^{-1/(d+4)}$时，核估计的收敛速度为$O(n^{-2/(d+4)})$。可以证明，多元密度函数的核估计在内点处具有一致性和渐近正态性。多元非参数回归模型的核估计与一元非参数回归模型的核估计类似，存在边界效应，在边界点处的收敛速度慢于内点处的收敛速度，并且核估计具有渐近正态性。

3.2.2 回归函数$m(\cdot)$的局部线性估计

我们看到，模型（3.1）的核估计具有一致性和渐近正态性，那么，核估计是不是模型（3.1）的最佳估计呢?答案是否定的，核估计存在边界效应，即核估计边界处收敛于实际函数的速度慢于在内点处的收敛速度，而且核估计是局部加权平均，其偏差较大，核估计的偏差还与解释变量的密度函数有关。本节介绍的局部线性估计是模型（3.1）的线性估计类中的最佳估计，它不存在边界效应问题，即其在边界点的收敛速度与内点的一样，且等于核估计在内点处的收敛速度，它的偏差比核估计小，而且其偏差与解释变量的密度函数无关。此外，局部线性估计在估计出回归函数$m(x)$的同时也估计出回归函数的导函数$m'(x)$，这正好符合经济学的乘数分析和弹性分析等需要。

多元非参数模型的不变窗宽局部线性估计为最小化：

$$\sum_{i=1}^{n}\left\{y_i-m(x)-D_m^T(x)(X_i-x)\right\}^2K_{h_n}(X_i-x) \tag{3.8}$$

其中 $D_m(x)=\left(\frac{\partial m(x)}{\partial x_1},\cdots,\frac{\partial m(x)}{\partial x_d}\right)^T$，其矩阵表达式为：

$$\hat{m}_n(x,h_n)=e_1^T(X_x^TW_xX_x)^{-1}X_x^TW_xY \tag{3.9}$$

其中 $e_1=(1,0,\cdots,0)^T$，$X_x=(X_{x,1},\cdots,X_{x,n})^T$，$X_{x,i}=(1,(X_i-x)^T)^T$，$Y=(Y_1,\cdots,Y_n)^T$，$W_x=diag\left\{K_{h_n}(X_1-x),\cdots,K_{h_n}(X_n-x)\right\}$。

存在不变窗宽局部线性估计的先决条件是逆矩阵 $(X_x^TW_xX_x)^{-1}$ 存在，并不是对所有的窗宽 h_n 该逆矩阵都存在，所以在实际应用中为了获得局部线性估计，有时必须调整窗宽 h_n。不变窗宽局部线性估计即适合于解释变量是确定性变量的情形，也适合于解释变量是随机变量的情形。当窗宽 $h_n=cn^{-1/(d+4)}$ 时，局部线性估计的收敛速度为 $O(n^{-2/(d+4)})$。多元非参数回归模型的局部线性估计与一元非参数回归模型的局部线性估计的性质类似，不存在边界效应，在边界点处的收敛速度达到了在内点处的收敛速度。

3.3 窗宽的选择[40]

3.3.1 核估计窗宽的选择

对于核估计而言：当 $h_n\to 0$ 时，有 $\hat{m}_n(X_i)\to K(0)Y_i/K(0)=Y_i$，

$$\hat{m}_n(x)\to 0,(x\neq X_i,i=1,\cdots,n)。 \tag{3.10}$$

可见，太小的窗宽得到了除了数据点外其他点的函数值都为零的函数。所以，太小的窗宽会使得随机误差项产生的噪音没有被排除，是没有意义的估计。当 $h_n\to\infty$ 时，

$$K(\frac{X_i-x}{h_n})\to K(0)\ ,$$

所以：

$$\hat{m}_n(x)\to\frac{n^{-1}\sum_{i=1}^{n}K(0)Y_i}{n^{-1}\sum_{i=1}^{n}K(0)}=n^{-1}\sum_{i=1}^{n}Y_i\ , \tag{3.11}$$

可见，太大的窗宽得到过分光滑的曲线，接近于直线，此时的估计也是没有任何意义的。在核估计的实际应用中，如果回归函数的核估计接近于一条直线，则窗宽肯定过大，参加局部加权的观察点过多，此时可减少窗宽。如果回归函数的核估计很不光滑，则窗宽肯定过小，此时，随机误差项产生的噪音没有被排除，应该加大窗宽，使得在局部参加加权平均的观察点增多，从而更多地消除随机误差项产生的噪音。 由上述可见，窗宽是控制核估计精度的重要参数，最佳的窗宽应当是即不过小也不过大。

3.3.2 局部线性估计窗宽的选择

对于局部线性估计而言：窗宽的选择在局部线性估计中是很重要的。理论上每一个选择的规则都面临着在估计的方差和偏差的平方之间做出权衡，而方差和偏差是从均方误差

(MSE)的角度提出来的，因此合适的窗宽应使得均方误差达到最小。达到最小的最佳理论窗宽具有形式$h_n = cn^{-1/(d+4)}$，其中c与n无关，只与回归函数，解释变量的密度函数和核函数有关。应用最佳理论窗宽$h_n = cn^{-1/(d+4)}$，必须先估计c，而对c估计会产生偏差，就会导致h要么选的过大，拟合不好，要么h选的过小，造成过度拟合，不宜预测，所以在实际应用中窗宽的选择就是不断的调整c，使得采用窗宽$h_n = cn^{-1/(d+4)}$的估计达到满意的效果。

3.3.3 样本窗宽的交错鉴定选择方法

交错鉴定方法是选择窗宽h_n的一个常用方法，其基本思路是：在每个局部观察点$x = X_i$，首先，在样本中剔除该观察点（X_i, Y_i）,其次，将将剩下的$n-1$个观察点在$x = X_i$处进行核估计：$\hat{m}_{n,-i} = \sum_{j \neq i} W_{nj}(X_i) Y_j$，最后通过比较平方拟合误差：

$$CV(h_n) = n^{-1} \sum_{i=1}^{n} (Y_i - \hat{m}_{n,-i}(X_i))^2 \omega(X_i) \tag{3.12}$$

选择使平方拟合误差达到最小的的窗宽h_n，其中$\omega(x) \geq 0$为某权函数。该方法的关键是在样本中剔除观察点（X_i, Y_i）。如果不这样的话，由于核权函数$W_{ni}(x)$在观察点$x = X_i$处达到最大值，就会使得$x = X_i$的重要程度过分夸大而其他观察点的重要程度降低。所以采用交错鉴定方法就避免了因没剔除观察点（X_i, Y_i）而将有用的数据排除在外的情形。

3.4 西安市日用水量的多元非参数回归模型

接下来建立城市日用水量多元非参数模型$Y_t = m(X_t) + \varepsilon_t$，其中$X_t = (X_{1t}, X_{2t}, X_{3t})'$，$\varepsilon_t$为随机误差项，利用核估计与局部线性估计对$m(\cdot)$进行估计， 本文采用的核函数为：

$K(u) = \frac{d(d+2)}{2S_d}(1 - u_1^2 - , \cdots, - u_d^2)_+$，其中$S_d = 2\pi^{d/2} / \Gamma(d/2)$，d=3，$\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1} e^{-x} dx$。

通过交错鉴定法确定窗宽见表3-1，对西安市2003.6.01～2003.8.24日用水量数据及其它影响因素（数据见附录）通过Matlab计算得到拟合结果如下表3-1，局部线性估计拟合见图3-1，核估计拟合见图3-2。

表3-1 6.01～8.24实际值与拟合值比较

Table3-1 the comparison of real data and simulation data from 6.01 to 8.24

日期	实际值	多元非参数模型					
	$/m^3$	核估计(h=0.9869)			局部线性估计(h=1.9809)		
		拟合值 $/m^3$	拟合残差 $/m^3$	拟合相对误差/%	拟合值 $/m^3$	拟合残差 $/m^3$	拟合相对误差/%
06.01	870265	860352	9913	1.14	859858	10407	1.20
06.02	852089	852625	-536	-0.06	852451	-362	-0.04
06.03	839438	839769	-331	-0.04	838671	767	0.09

续表

06.04	802901	792398	10503	1.30	788167	14734	1.83
06.05	859839	844520	15319	1.78	843364	16475	1.92
06.06	854856	854320	536	0.06	854592	264	0.03
06.07	843568	860734	-17166	-2.03	868338	-24770	-2.94
06.08	851760	857518	-5758	-0.68	860887	-9127	-1.07
06.09	830944	821803	9141	1.10	829564	1380	0.16
06.10	784532	769336	15196	1.93	757353	27179	3.46
06.11	844146	842235	1911	0.23	850427	-6281	-0.74
06.12	831512	850614	-19102	-2.30	846710	-15198	-1.82
06.13	880478	859694	20784	2.36	850116	30362	3.44
06.14	873326	853326	20000	2.20	869120	4206	0.48
06.15	904991	923752	-18761	-1.97	924910	-20081	-2.53
06.16	915223	858091	57132	6.24	861113	54110	5.91
06.17	942777	940820	1957	0.21	927380	15397	1.63
06.18	964709	962990	1719	0.18	971174	-6465	-0.67
06.19	964373	946366	18007	1.87	938119	26254	2.72
06.20	986222	932158	54064	5.48	910759	75463	7.65
06.21	954107	882046	72061	7.55	870766	83341	8.73
06.22	850633	827064	23569	2.77	830431	20202	2.37
06.23	852424	876725	-24301	-2.85	864798	-12374	-1.45
06.24	890963	884358	6605	0.74	883411	7552	0.84
06.25	831124	837173	-6049	-0.72	850606	-19482	-2.34
06.26	810729	823225	-12496	-1.54	825076	-14347	-1.77
06.27	857331	853777	3554	0.41	865888	-8557	-1.00
06 28	878399	874929	3470	0.40	876224	2175	0.25
06 29	840524	861004	-20480	-2.44	862400	-21876	-2.60
06.30	815263	799928	15335	1.88	797320	17943	2.20
07.01	792317	771814	20503	2.59	767365	24952	3.15
07.02	831177	822510	8667	1.04	820494	10683	1.29
07.03	821353	813880	7473	0.91	826500	-5147	-0.63
07.04	842671	859022	-16351	-1.94	851321	-8650	-1.02
07.05	835624	861321	-25697	-3.08	863634	-28010	-3.35
07.06	851529	862649	-11120	-1.31	866746	-15217	-1.79
07.07	857987	851669	6318	0.74	854120	3867	0.45
07.08	851434	839590	11844	1.39	845394	6040	0.70
07.09	813155	813761	-606	-0.75	813676	-521	-0.06
07.10	846894	838565	8329	0.98	842299	4595	0.54
07.11	803895	810640	-6745	-0.84	819454	-15559	-1.94
07.12	756685	761903	-5218	-0.69	765577	-8892	-1.18
07.13	737839	747797	-9958	-1.35	742494	-4655	-0.63
07.14	751511	763316	-11805	-1.57	774227	-22716	-3.02
07.15	734782	759048	-24266	-3.30	756452	-21670	-2.95
07.16	737928	765720	-27792	-3.77	760208	-22280	-3.02
07.17	803530	825966	-22436	-2.79	837013	-33483	-4.17

续表

07.18	847766	855423	-7657	-0.90	860184	-12418	-1.46
07.19	851843	856260	-4417	-0.52	908076	-56233	-6.60
07.20	873839	862934	10905	1.25	860976	12863	1.47
07.21	879179	882635	-3456	-0.39	868396	10783	1.23
07.22	910541	861377	49164	5.40	863787	46754	5.13
07.23	882903	881052	1851	2.09	881695	1208	0.14
07.24	885238	951393	-66155	-7.47	948636	-63398	-7.16
07.25	946915	915981	30956	3.38	948295	-1380	-0.15
07.26	923076	903076	20000	2.56	964510	-41434	-4.49
07.27	923460	886521	36939	4.00	888324	35136	3.80
07.28	965979	981957	-15978	-1.65	974843	-8864	-0.92
07.29	1014790	957976	56814	5.60	959289	55501	5.47
07.30	1042459	1011046	31413	3.01	994601	47858	4.59
07.31	977334	976628	706	0.07	964160	13174	1.35
08.01	889001	858961	30040	3.38	850727	38274	4.31
08.02	768613	789561	-20958	-2.98	795952	-27339	-3.56
08.03	745256	744666	590	0.07	739407	5849	0.78
08.04	797046	793695	3351	0.42	803746	-6700	-0.84
08.05	843197	847692	-4495	-0.53	851213	-8016	-0.95
08.06	880528	867679	12849	1.46	868068	12460	1.42
08.07	885245	883432	1813	0.20	883297	1948	0.22
08.08	826712	854465	-27753	-3.36	851927	-25215	-3.05
08.09	789677	803741	-14064	-1.96	878468	-88791	-1.12
08.10	766365	796666	-30301	-3.95	790296	-23931	-3.12
08.11	773343	771359	1984	0.26	762210	11133	1.44
08.12	768931	756320	12611	1.64	769591	-660	-0.08
08.13	780359	774243	6116	0.78	768313	12046	1.54
08.14	768013	767986	27	0.00	775739	-7726	-1.01
08.15	755659	767027	-11368	-1.50	774698	-19039	-2.52
08.16	781812	768919	12893	1.65	777313	4499	0.58
08.17	803163	807010	-3847	-0.48	809993	-6830	-0.85
08.18	812807	855787	-42980	-5.29	858870	-46063	-5.67
08.19	815311	836382	-21071	-2.58	835460	-20149	-2.47
08.20	810067	824451	-14384	-1.78	828707	-18640	-2.30
08.21	836303	882714	-46411	-5.55	874223	-37920	-4.53
08.22	862114	861287	827	0.10	863039	-925	-0.10
08.23	877021	866954	10067	1.15	870622	6399	0.73
08.24	862650	866295	-3645	-0.42	865369	-2719	-032

注：$\text{相对误差}=\frac{\text{实际值}-\text{预测值}}{\text{实际值}}\times 100\%$

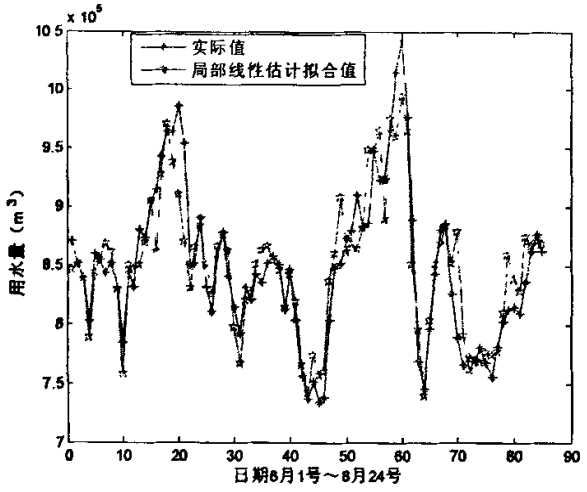

图 3-1 西安市日用水量局部线性估计拟合曲线

Figure3-1 the local linear estimation simulation curve of xi'an daily water demand

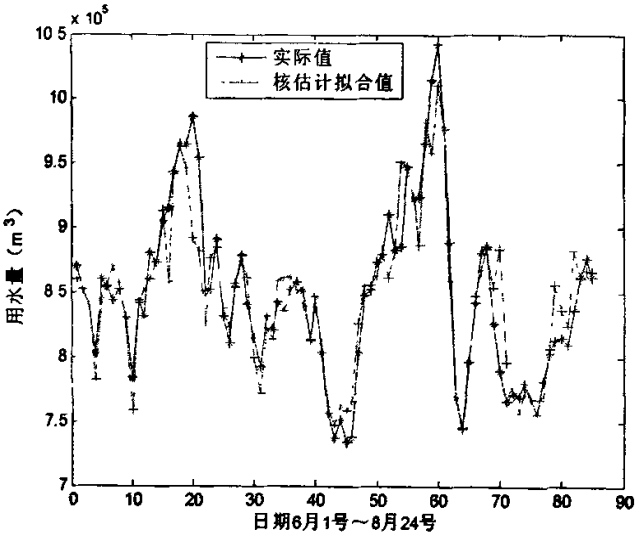

图 3-2 西安市日用水量核估计拟合变化曲线

Figure3-2 the kernel estimation simulation curve of xi'an daily water demand

利用非参数回归模型对 8 月 25~8 月 31 一周的日用水量进行事后预测，采取循环预测法，这样上一时期较大的误差，不会对下一时期造成较大的影响。非参数预测通过 MATLAB 编写程序，采取固定一个点，减少一个循环的思想，预测出一个值，再把预测值放入数据，预测第二个，依次类推，得到 8 月 25 号到 8 月 31 号预测值，实际值与预测值见表 3-2。

表 3-2 8.25～8.31 实际值与预测值比较

Table3-2 the comparison of real data and forecasting data from 8.25 to 8.31

日期	实际值	非参数回归模型预测值			
	/m^3	核估计预测值 (h=0.9869) /m^3	相对误差 /%	局部线性估计预测值 (h=1.9809) /m^3	相对误差 /%
8.25	856551	859986	-0.4%	855029	0.18%
8.26	811914	797505	1.77%	802825	1.12%
8.27	849784	886521	-4.32%	878324	-3.36%
8.28	791414	798887	-0.94%	797251	-0.74%
8.29	747147	746901	0.03%	747602	-0.06%
8.30	785090	770438	1.87%	771654	1.71%
8.31	764637	773289	-1.13%	759569	0.66%

3.5 本章小结

本章首先给出了多元非参数回归模型的定义，以及估计方法，最后通过多元非参数回归模型对城市日用水量进行了拟合与预测，由表3-1，表3-2及图3-1，3-2可知，多元非参数回归模型误差较小。另外，在日用水量问题中，由于非参数估计中局部线性估计可以减少核估计量的偏差和核估计量的方差，因此拟合与预测效果要比核估计稍好；最后通过比较可以得出无论用核估计还是局部线性估计非参数回归预测模型在城市日用水量预测这一实际问题中效果都比较好，基本能满足供水系统的需要。

4 西安市日用水量的线性自回归预测模型

城市日用水量预测常用的方法主要分为两类：①解释性预测方法，即：回归分析方法；②时间序列分析方法。前者是认为系统的输入量与输出量之间存在着某种因果关系，以此来构造预测模型进行预测。该模型对输入变量的精度及可靠性要求较高，特别是在进行离线控制时，需对次日整天的用水量进行预测，这就要求次日的天气、居民活动等情况的预报精度较高，否则误差可能较大。后者是把系统看作一个“暗箱”，可以不管其影响因素，而只关心预测和预测的结果，其预测过程只依赖于历史观测数据。在一个发展较成熟的城市中收集到的一定时间内的城市日用水量数据，根据用水量的变化特性，即用水量的随机性、趋势性及周期性，可知道该数据是一串随时间变化而相互关联的数字序列(动态数据)，序列中不同时刻的随机变量彼此之间有一定的相互关系，因此该数据序列符合时间序列分析方法建模的条件，可以采用时间序列分析方法建模。

本章针对城市日用水量的特点，对西安市某时间段日用水量数据进行分析，最终建立了西安市日用水量的自回归模型。

4.1 线性自回归模型定义[39]

自回归模型的定义为：对于时间序列$\{X_t\}$，如果$\{\varepsilon_t\}$是白噪声，并且实数$a_1,a_2,\cdots,a_p(a_p\neq 0)$使得多项式$A(Z)$的零点都在单位圆外，即：

$$A(Z)=1-\sum_{j=1}^{p}a_jz^j\neq 0,|z|\leq 1, \tag{4.1}$$

就称p阶差分方程：

$$X_t=\sum_{j=1}^{p}a_jX_{t-j}+\varepsilon_t,\ t\in z \tag{4.2}$$

是一个p阶自回归模型，简称为$AR(p)$模型。满足$AR(p)$模型（4.2）的平稳时间序列$\{X_t\}$称为平稳解或$AR(p)$序列。称$a=(a_1,a_2,\cdots,a_p)^T$是$AR(p)$模型的自回归系数，称条件（4.1）是稳定性条件或最小相位条件。

4.2 西安市日用水量线性自回归模型的建模过程分析

人们从大量的实测数据序列中要寻求它的统计规律及及统计特性，就必须先建立模型，其中包括如何识别模型的类型，估计模型参数，以及确定模型阶数等。但是时间序列的建模过程往往是动态进行的，也就是说在建模过程中难免出现判断和估计方面的错误，直接影响到模型的精度与正确性，因此要反复地进行认证和修改，甚至可能推翻重来。以下对自回归模型建立过程步骤作详细分析。

4.2.1 数据的平稳性检验

时间序列数据的平稳性是我们建模的重要前提。平稳数据的主要特点是它的一阶和二阶统计性质不随时间改变，即均值和方差为常数。对时间序列的平稳性检验主要有两种方法，一是根据时序图和自相关图显示的特征做出判断的图检验方法；一种是构造检验统计

量进行假设检验的方法。图检验方法是一种操作简便、运用广泛的平稳性检验方法，它的缺点是判别结论带有很强的主管色彩。所以最好能用统计检验方法加以辅助判断。本文主要用到平稳性的非参数检验法，其它检验方法可以参考文献【39】【41】【47】【53】。

平稳性的非参数检验法又称为游程检验法[41]（或轮次检验法）。该方法只涉及一组实测数据，而不需要假设数据的分布规律，因此本方法具有很好的实用性。在保持随机序列原有顺序的情况下，游程定义为具有相同符号的序列，这种符号可把观测值分成两个互相排斥的类。例如观测序列的值是$x_i(i=1,2,\cdots,N)$，其均值为$\bar{x}$，用符号“+”表示$x_i \geq \bar{x}$，而“－”表示$x_i < \bar{x}$。每个游程的长短在这里并不重要。游程太多或太少都被认为是存在非平稳趋势。游程检验所判断的原假设为：“样本数据出现的顺序没有明显的趋势，就是平稳的”。我们采用的样本统计量有：

N_1= 一种符号出现的总数，

N_2=另一种符号出现的总数，

γ=游程的总数，

其中γ为检验统计量。对于显著水平$\alpha=0.05$的双边检验，由游程检验γ分布表给出概率分布左右两侧为$\alpha/2=0.025$时的上限γ_U和下限γ_L。如果γ在界限以内，则接受原假设；否则拒绝原假设。游程检验γ分布表中N_1和N_2分别表示符号+和－的数目，γ_U和γ_L给出显著水平$\alpha=0.05$时游程总数的下限和上限。由于游程检验属于双边检验，故应将显著水平划分成两边$\alpha/2=0.025$，只要实际的游程总数γ（双边）在表中给出的γ_U和γ_L界限内，则平稳性的假设就可以接受。当N_1或N_2超过 15 时就可以用正态分布来近似，即可用正态分布表来确定检验的接受域和否定域。此时用的统计量为：

$$Z=\frac{\text{游程数}-\text{游程的期望数}}{\text{游程标准差}}=\frac{\gamma-\mu_\gamma}{\sigma_\gamma}, \tag{4.3}$$

式中：

$$\mu_\gamma=\frac{2N_1N_2}{N}+1,$$

$$\sigma_\gamma=\left[\frac{2N_1N_2(2N_1N_2-N)}{N^2(N-1)}\right]^{1/2},$$

$$N=N_1+N_2。$$

对于$\alpha=0.05$的显著水平，如果$|Z|\leq 1.96$(按2σ原则)，则可接受原假设，否则就拒绝[41]。如果时间序列数据不平稳可以做差分或对数变换使其达到平稳。利用游程检验方法对对西安市 2003 年 6 月 1 日到 8 月 24 日日用水量数据x_t（数据见附录二，共 85 个数据，数据来源于西安市自来水公司）进行分析，按照游程检验法通过计算可以得出，未差分前$|Z|=5.77>1.96$所以该时间序列数据是不平稳的，由图 4-1 也可观测出不平稳。对该数据做一次差分以后见图 4-2，可以求得$|Z|=0.0472<1.96$，因此差分后序列达到平稳。

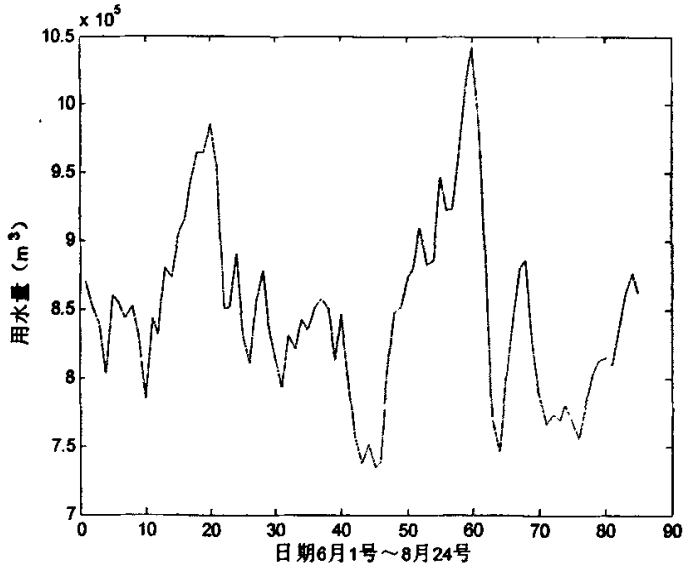

图 4-1 日用水量原始数据图形

Figure4-1 the raw data plot of daily water demand

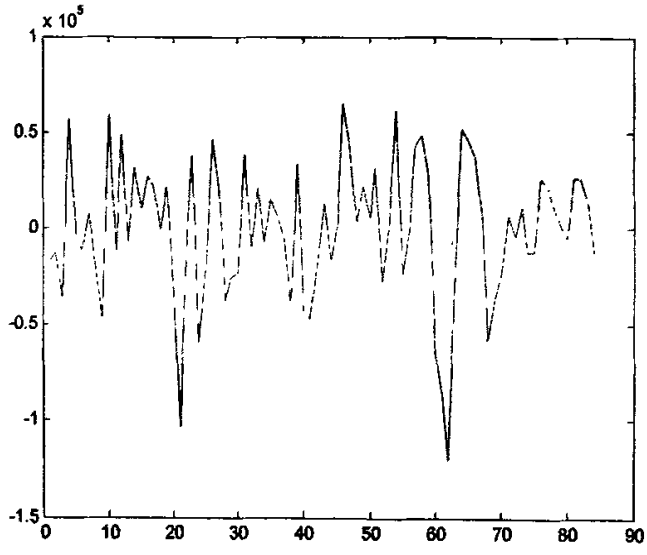

图 4-2 一次差分后日用水量图形

Figure 4-2 the plot of daily water demand data (after the first order difference)

4.2.2 模型识别

对数据进行平稳化处理以后，接下来就是根据处理后的时间序列数据分析何种模型适合该数据序列。建立一个时间序列的数学模型，首先要根据数据信息的先验知识，以及所提供的时间序列的数据概貌，提出一个相适应的模型类别。时间序列模型包括有 *AR(p)*，*MA(q)*，*ARMA(p,q)*及 *ARIMA(p,d,q)*等多种类型。接下来讨论的问题是：假定我们并不知

道构造观测数据序列的模型，如何根据它们的一段样本值$x_1,x_2,\cdots,x_N$，对它们应属的模型类别进行判断，这就是模型识别。

根据文献【41】提出的时间序列模型识别方法，即根据样本自相关系数及样本偏相关系数识别模型。该方法为：如果$\{x_t\}$的样本自相关系数$\hat{\rho}_k$在$k>q$后截尾，则判断$\{x_t\}$是*MA(q)*序列；如果偏相关系数$\hat{\varphi}_{kk}$在$k>p$后截尾，则判断$\{x_t\}$是*AR(p)*序列；如果$\hat{\rho}_k$和$\hat{\varphi}_{kk}$都不截尾，只是按负指数衰减趋于零（即拖尾的），则应判断其为*ARMA*序列，但尚不能判定阶次。这里简单的对$\hat{\rho}_k$和$\hat{\varphi}_{kk}$截尾性的判断做一个说明。理论的自相关系数$\hat{\rho}_k$和偏相关系数$\hat{\varphi}_{kk}$截尾性是指它们从某个p或q值后全为零。但是，由于参数估计的随机性，$\hat{\rho}_k$和$\hat{\varphi}_{kk}$都是随机变量，即使$\{x_t\}$是*MA(q)*或*AR(p)*序列，当$k>q$（或$k>p$）后，$\hat{\rho}_k$（或$\hat{\varphi}_{kk}$）也不会全为零，只是在零附近上下波动。因此对于$\hat{\rho}_k$和$\hat{\varphi}_{kk}$截尾性的判断只能凭借统计手段进行检验和判别。在实际使用时采用检验数据独立性和正态性的方法，即对于*AR(p)*过程，当$k>p$时$\{\hat{\varphi}_{kk}\}$近似为独立分布，且每个$\hat{\varphi}_{kk}$具有零均值和近似为$1/N$的方差（N是建模所用的观察数据个数），而且当N较大时，$\hat{\varphi}_{kk}$的估计值近似为正态分布。因此检验$\hat{\varphi}_{kk}$是否为零，只要看$\hat{\varphi}_{kk}$是否落在$\pm2\sqrt{1/N}$范围内[41]。

日用水量样本数据一次差分后的序列的偏相关系数$\hat{\varphi}_{kk}$见图4-3，从图4-3可以看出$|\hat{\varphi}_{kk}|$虽有波动，但总是向零衰减，所以可以认为该时间序列数据平稳，因为除了$|\hat{\varphi}_{11}|$和$|\hat{\varphi}_{22}|$大于$2\sqrt{1/85}=0.2182$外，其余偏相关系数均在$\pm2\sqrt{1/N}$之内，可以理论上认为$k>2$时，$\hat{\varphi}_{kk}=0$，因此考虑该时间序列为二阶自回归模型*AR*(2)。

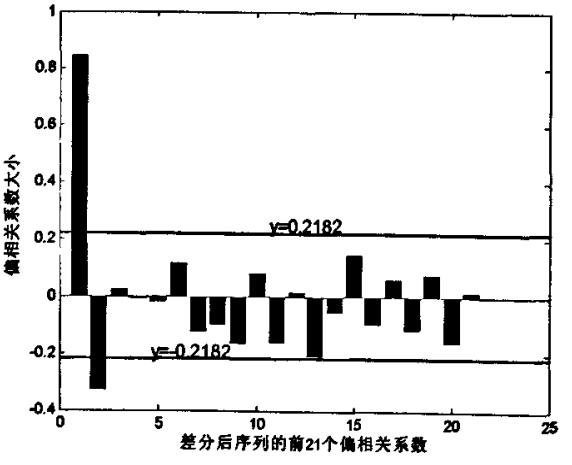

图4-3 一次差分后的序列的偏相关系数图

Figure 4-3 the partial correlation plot of the first order difference series

4.2.3 模型的定阶[39]

识别出模型以后，其次就要根据实际的观测数据具体地确定该类数学模型所包含的阶数及各项系数的数值。前者叫做模型识别，后者叫做模型定阶及参数估计。目前流行的多

种定阶准则，通常可从下述四个方面进行分类：

（1） 利用时间序列的相关特性，即判断模型的自相关系数 $\hat{\rho}_k$ 和偏相关系数 $\hat{\varphi}_{kk}$ 的拖尾和截尾性来确定模型其合适阶次。这是一种初步定阶方法，可在建模开始时加以粗略地估计。

（2） 利用数理统计方法，有：①检验高阶模型新增加的参数是否近似为零？参数的置信区间是否为零来确定模型阶次；②检验残差的相关特性；③F 检验方法。

（3） 利用信息准则，即定义一个与模型阶数信息有关的特征参数，从而选取使它达到最小值的阶数作为模型的阶数，其中包括 AIC，BIC，FPE 及其他准则。

（4） 根据经验提出的定阶方法。

本文主要用到 AIC，BIC 定阶准则， 定义 AIC 准则函数如下：

$$AIC = -2\lg(\text{模型的最大阶数})+2(\text{模型的独立参数个数}) = -2L(\hat{\beta})+2k \tag{4.4}$$

式中：k 为独立参数个数，$\hat{\beta}$ 为参数的最大似然估计值，$L(\cdot)$ 为似然函数。可见，AIC 准则函数由两项构成。第一项体现模型拟合的好坏，它随着阶数的增大而变小；第二项标志了模型参数的多少，随着阶数的增大而变大。取二者的最小值意味着对上述两个量的一种权衡。所谓 AIC 准则就是当欲从一组可供选择的模型中选择一个最佳模型时，选取 AIC 为最小模型是适宜的。若事先给定阶数上界 n_h，则

$$AIC(k_0) = \min_{0<k<n_h} AIC(k) \tag{4.5}$$

其中 k_0 为最佳参数的个数。下面给出 *AR(p)* 时序模型的 AIC 准则的具体形式：

$$AIC(p) = N\lg\hat{\sigma}_a^2 + 2(p+1), \tag{4.6}$$

AIC 准则还可引申到其他模型的定阶，它是迄今应用最广泛的一种定阶方法。这个准则的优点在于它是借助信息论而提出的一个完全客观的定阶准则，而 F 检验则要求建模人员主观地选择置信度。然而对于 AIC 准则的使用尚需注意以下几个问题。

（1）运用 AIC 准则时，仍然需要人为地预先框定阶数的最大范围 n_h。从经验来看，阶数的上限取 $\sqrt{N}$，$N/10$，$\lg N$ 均可。在比较 AIC 值大小的过程中，如果已接近阶数上限仍不能确定 AIC 的极小点，则应加大上限，继续进行比较。

（2）AIC 准则要求参数由最大似然估计，但当序列不服从正态分布时计算表明此准则对于最小二乘法估计也仍然适用。

（3）AIC 准则是模型优化的一种宏观度量，但不宜机械地以绝对最小值来选择模型阶数，而是要在所对应的模型进行多方比较，确定理想的模型阶数及相应参数。

AIC 准则虽然为时序模型的定阶带来很多方便，然而 AIC 准则也有不足之处，有人从理论上证明了 AIC 方法不能给出相容估计。也就是说，当样本长度 $N \to \infty$ 时，用 AIC 定出的模型阶数估计值，并不能依概率收敛到真值。为了改进这一方法，Akaike(1976 年)和 E.J.Haman(1979)年等人又提出了 BIC，$\varphi(K)$ 等准则。BIC 准则函数的定义如下：

$$BIC(p) = N\lg\hat{\sigma}_a^2 + p(\lg N)/N \quad , \tag{4.7}$$

若某一阶数 p_0 满足：

$$BIC(p_0) = \min_{0\le p\le p_n} BIC(p), \tag{4.8}$$

其中 p_n 是阶数上限，则取 p_0 为最佳阶数。

本章利用偏相关确定模型时已经初步估计出模型阶数为 2 阶，现在利用 AIC 与 BIC 准则对日用水量模型做进一步确定。

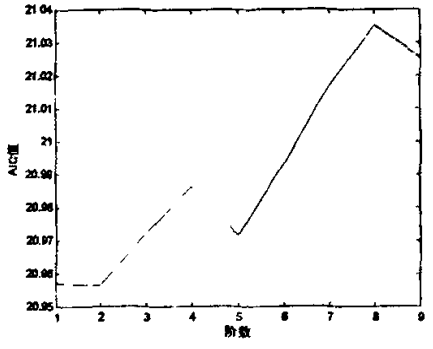

图 4-4 AIC 图
Figure 4-4 AIC plot

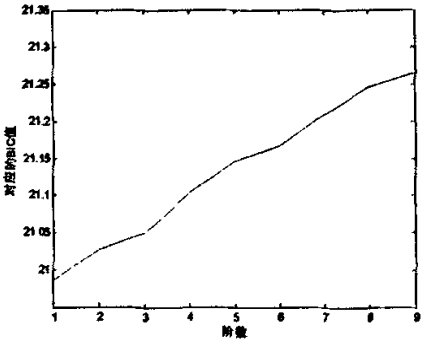

图 4-5 BIC 图
Figure 4-5 BIC plot

一般说来 AIC 达到极小时所对应的阶数往往比 BIC 准则相应定出的阶数高，但是从 *AR(p)*模型的 *Yule－Wal*ker 方程看，略有高估并不引起严重的后果，而低估了阶数会带来很大的模型误差。另一方面，实际数据中阶数 p 并不存在，把阶数 p 估计的略高一点在预测问题中还有利于多用历史数据，所以在应用工作中，当样本量不是很大时，人们还是乐于使用 AIC 定阶。因此确定模型阶数为 2 阶。

4.2.4 模型参数的估计及原始预测[39]

选择好了拟合模型之后，下一步就是要利用序列的观察值确定该模型的参数，即估计模型中未知参数的值。本文主要讨论 *AR(p)*模型经常使用的参数估计方法，*AR(p)*模型经常使用的参数估计方法主要有最小二乘估计，矩估计，最大似然估计等。下面我们主要介绍介绍一下矩估计方法。从 AR 模型的基本知识知道，其自回归系数 α 由 *AR(p)*序列的自协方差函数 $\gamma_0,\gamma_1,\cdots,\gamma_p$ 通过 Yule-Walker 方程：

$$\begin{bmatrix}\gamma_1\\ \gamma_2\\ \vdots\\ \gamma_p\end{bmatrix} = \begin{bmatrix}\gamma_0 & \gamma_1 & \cdots & \gamma_{p-1}\\ \gamma_1 & \gamma_0 & \cdots & \gamma_{p-2}\\ \vdots & \vdots & \cdots & \vdots\\ \gamma_{p-1} & \gamma_{p-2} & \cdots & \gamma_0\end{bmatrix}\begin{bmatrix}\alpha_1\\ \alpha_2\\ \vdots\\ \alpha_p\end{bmatrix} \tag{4.9}$$

唯一决定。白噪声 σ^2 的方差由

$$\sigma^2=\gamma_0-(\alpha_1\gamma_1+\alpha_2\gamma_2\cdots+\alpha_p\gamma_p) \tag{4.10}$$

决定。现在从观测样本 $y_1,y_2,\cdots,y_N$ 可以构造出自协方差函数的估计：

$$\hat{\gamma}_k=\frac{1}{N}\sum_{j=1}^{N-k}y_jy_{j+k}\quad,\quad k=0,1,\cdots,p \tag{4.11}$$

所以 AR（p）的自回归系数和白噪声方差的矩估计：

$$(\hat{\alpha}_1,\hat{\alpha}_2,\cdots,\hat{\alpha}_p)^T\quad,\quad\hat{\sigma}^2$$

就由样本 Yule-Walker 方程

$$\begin{bmatrix}\hat{\gamma}_1\\ \hat{\gamma}_2\\ \vdots\\ \hat{\gamma}_p\end{bmatrix}=\begin{bmatrix}\hat{\gamma}_0 & \hat{\gamma}_1 & \cdots & \hat{\gamma}_{p-1}\\ \hat{\gamma}_1 & \hat{\gamma}_0 & \cdots & \hat{\gamma}_{p-2}\\ \vdots & \vdots & \cdots & \vdots\\ \hat{\gamma}_{p-1} & \hat{\gamma}_{p-2} & \cdots & \hat{\gamma}_0\end{bmatrix}\begin{bmatrix}\hat{\alpha}_1\\ \hat{\alpha}_2\\ \vdots\\ \hat{\alpha}_p\end{bmatrix} \tag{4.12}$$

和 $\hat{\sigma}^2=\hat{\gamma}_0-(\hat{\alpha}_1\hat{\gamma}_1+\hat{\alpha}_2\hat{\gamma}_2\cdots+\hat{\alpha}_p\hat{\gamma}_p)$ 决定。在实际工作中，对于较大的 p 为了加快计算速度还可以采用如下的 Levinson 递推方法：

$$\begin{cases}\hat{\sigma}_0^{\,2}=\hat{\gamma}_0,\\ \hat{\alpha}_{1,1}=\hat{\gamma}_1/\hat{\sigma}_0^{\,2},\\ \hat{\sigma}_k^{\,2}=\hat{\sigma}_{k-1}^{\,2}(1-\hat{\alpha}_{k,k}^{\,2}),\\ \hat{\alpha}_{k+1,k+1}=\dfrac{\hat{\gamma}_{k+1}-\hat{\gamma}_k\hat{\alpha}_{k,1}-\hat{\gamma}_{k-1}\hat{\alpha}_{k,2}-\cdots-\hat{\gamma}_1\hat{\alpha}_{k,k}}{\hat{\gamma}_0-\hat{\gamma}_1\hat{\alpha}_{k,1}-\hat{\gamma}_2\hat{\alpha}_{k,2}-\cdots-\hat{\gamma}_k\hat{\alpha}_{k,k}},\\ \hat{\alpha}_{k+1,j}=\hat{\alpha}_{k,j}-\hat{\alpha}_{k+1,k+1}\hat{\alpha}_{k,k+1-j},1\le j\le k,k\le p.\end{cases} \tag{4.13}$$

最后得到矩估计

$$(\hat{\alpha}_1,\hat{\alpha}_2,\cdots,\hat{\alpha}_p)=(\hat{\alpha}_{p,1},\hat{\alpha}_{p,2},\cdots,\hat{\alpha}_{p,p}),\quad\hat{\sigma}^2=\hat{\sigma}_p^2。 \tag{4.14}$$

由于上述的矩估计由 Yule-Walker 方程得到，所以又被称为 Yule-Walker 估计。

对西安市 2003 年 6 月 1 号到 8 月 24 号日用水量数据，求出自相关系数后，根据(4.12)估计出模型的参数，其结果为 $\hat{\phi}_1=1.0674$，$\hat{\phi}_2=-0.3114$，相应的二阶自回归模型为：

$$X_t-X_{t-1}=0.2623\times(X_{t-1}-X_{t-2})-0.1491\times(X_{t-2}-X_{t-3})+e_t \tag{4.15}$$

其中，e_t 为随机误差项，最后可以得到

$$X_t=1.2623X_{t-1}-0.4114X_{t-2}+0.1491X_{t-3}+e_t \tag{4.16}$$

利用(4.16)式对 2003 年 6 月 4 号到 8 月 24 号日用水量作预测，以便检验模型的有效性，采取直接预测方法，预测结果见表 4-1 与图 4-6。

表 4-1 6.04～8.24 实际值与预测值比较

Table4-1 the comparison of real data and forecasting data from 6.04 to 8.24

日期	实际值 /m^3	预测值 /m^3	绝对误差 /m^3	相对误差 /%	日期	实际值 /m^3	预测值 /m^3	绝对误差 /m^3	相对误差 /%
6.01	870265	------	------	------	7.14	751511	803468	-51957	-6.91
6.02	852089	------	------	------	7.15	734782	763667	-28885	-3.93
6.03	839438	------	------	------	7.16	737928	731759	6169	0.83
6.04	802901	843146	-40245	-5.01	7.17	803530	776368	27162	3.38
6.05	859839	849960	9879	1.15	7.18	847766	843738	4028	0.48
6.06	854856	857551	-2695	-0.31	7.19	851843	787316	64527	7.57
6.07	843568	835695	7873	0.93	7.20	873839	892535	-18696	-2.14
6.08	851760	859463	-7703	-0.90	7.21	879179	849976	29203	3.32
6.09	830944	859038	-28094	-3.38	7.22	910541	896870	13671	1.50
6.10	784532	800300	-15768	-2.00	7.23	882903	877114	5789	0.66
6.11	844146	850308	-6162	-0.73	7.24	885238	866832	18406	2.07
6.12	831512	852066	-20554	-2.46	7.25	946915	918139	28776	3.04
6.13	880478	867012	13466	1.53	7.26	923076	891487	31589	3.42
6.14	873326	854761	18565	2.13	7.27	923460	895506	27954	3.03
6.15	904991	877602	27389	3.03	7.28	965979	953225	12754	1.32
6.16	915223	887075	28148	3.08	7.29	1014790	989315	25475	2.51
6.17	942777	896546	46231	4.90	7.30	1042459	985585	56874	5.46
6.18	964709	986415	-21706	-2.25	7.31	977334	957301	20033	2.05
6.19	964373	951266	13107	1.36	8.01	889001	897822	-8821	-1.00
6.20	986222	996974	-10752	-1.09	8.02	768613	746371	22242	2.89
6.21	954107	968055	-13948	-1.46	8.03	745256	782554	-37298	-5.00
6.22	850633	903854	-53221	-6.25	8.04	797046	807096	-10050	-1.26
6.23	852424	879218	-26794	-3.14	8.05	843197	792461	50736	6.02
6.24	890963	881515	9448	1.06	8.06	880528	838556	41972	4.77
6.25	831124	855909	-24785	-2.98	8.07	885245	890507	-5262	-0.59
6.26	810729	833393	-22664	-2.80	8.08	826712	834108	-7396	-0.89
6.27	857331	903618	-46287	-5.40	8.09	789677	770563	19114	2.42
6.28	878399	843422	34977	3.98	8.10	766365	795076	-28711	-3.75
6.29	840524	801647	38877	4.63	8.11	773343	732950	40393	5.22
6.30	815263	846158	-30895	-3.78	8.12	768931	796832	-27901	-3.63
7.01	792317	784567	7750	0.98	8.13	780359	763877	16482	2.11
7.02	831177	843729	-12552	-1.51	8.14	768013	776204	-8191	-1.07
7.03	821353	827076	-5723	-0.70	8.15	755659	764093	-8434	-1.11
7.04	842671	790659	52012	6.17	8.16	781812	781755	57	0.00
7.05	835624	836932	-1308	-0.16	8.17	803163	795153	8010	1.00
7.06	851529	814711	36818	4.32	8.18	812807	791983	20824	2.56
7.07	857987	846891	11096	1.29	8.19	815311	797650	17661	2.17
7.08	851434	832415	19019	2.23	8.20	810067	801067	9000	1.11
7.09	813155	848858	-35703	-4.39	8.21	836303	811368	24935	2.98

续表									
7.10	846894	853998	-7104	-0.83	8.22	862114	840598	21516	2.50
7.11	803895	866505	-62610	-7.78	8.23	877021	857034	19987	2.28
7.12	756685	797894	-41209	-5.45	8.24	862650	841755	20895	2.42
7.13	737839	751134	-13295	-1.80					

注：相对误差$=\frac{\text{实际值}-\text{预测值}}{\text{实际值}}\times 100\%$

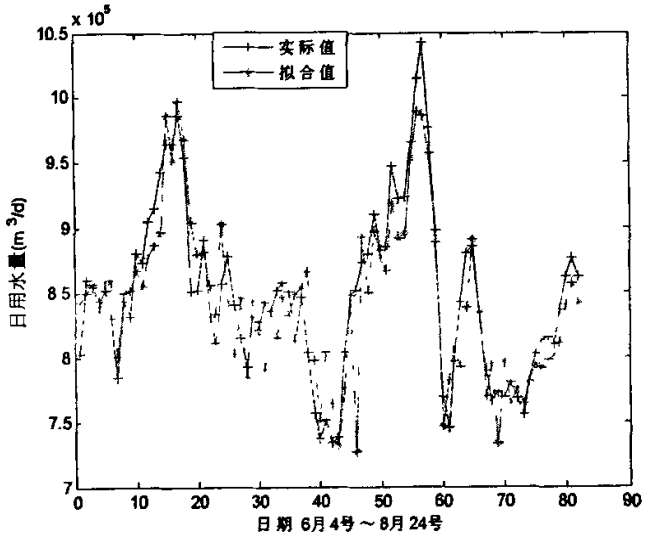

图 4-6 AR 模型预测值与实际值比较图

Figure4-6 the comparison of forecasting data and real data of AR model

4.2.5 模型的检验及预测[39]

模型建立以后，还要对其初始预测残差的独立性进行检验。检验方法主要包括：散点图法，估计相关系数法，F 检验法，χ^2 检验法。这里主要介绍 χ^2 检验方法。对于由残差项所构成的时间序列可用自相关分析的方法判别是否具有随机性。这里采用 χ^2 检验法，设其前 m 个自相关系数为 $\hat{\rho}_1,\hat{\rho}_2,\cdots,\hat{\rho}_m$，则统计量 $Q=n\sum_{k=1}^{m}\hat{\rho}_k^2\sim\chi^2(m-1)$。置信度为 95%时，所对应的 χ^2 数值如果比 Q 大，就有 95%的置信度认为原假设成立，否则就否定 e_t 为白噪声序列。计算出该$\{e_t\}$时间序列数据中的前 22 个自相关系数，Q=4.947<11.591，所以有 95%的置信度认为残差序列具有随机性，因此该自回归模型是有效的。

模型通过检验以后，就可以用模型用进行预测。预测方法主要包括直接预测，和循环预测，一般采取循环预测（动态预测），这样上一时期较大的误差，不会对下一时期造成较大的影响。预测结果见表 4-2。

表 4-2 8.25～8.31 实际值与预测值比较

Table4-2 the comparison of real data and forecasting data from 8.25 to 8.31

日期	实际值 / m^3	预测值 / m^3	绝对误差 / m^3	相对误差 / %
8.25	856551	835297	21254	2.48
8.26	811914	835881	-23967	-2.95
8.27	849784	836997	12787	1.50
8.28	791414	837203	-45789	-5.79
8.29	747147	804360	-57213	-7.66
8.30	785090	818120	-33030	-4.21
8.31	764637	794250	-29613	-3.87

4.2.6 线性自回归模型流程图

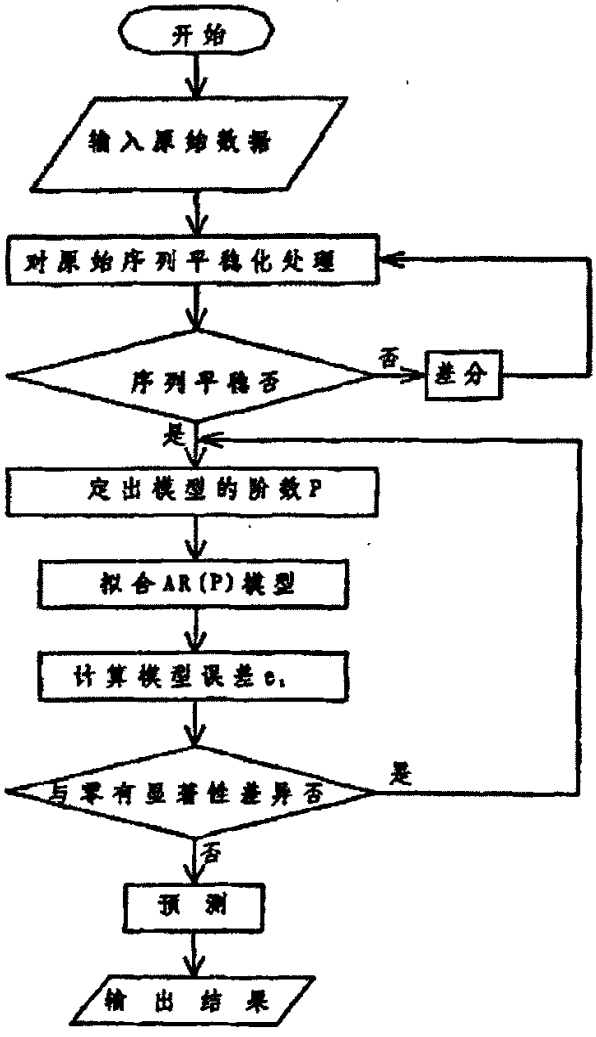

图 4-7 线性 AR 模型流程图

Figure4-7 the flow plot of linear AR model

4.3 本章小结

本章根据 BJ 时间序列建模方法，对一段西安市日用水量数据进行分析，按照建模的基本步骤，最终确定该时间序列数据适合自回归模型，建立了能服务于供水系统的自回归模型，根据预测结果表明，该种方法预测精度不高，因此可以把对日用水量影响较大的温度因素考虑进去，建立部分线性自回归模型。

5 西安市日用水量预测的部分线性自回归模型

对于城市日用水量预测问题来说，多元线性回归模型和非参数模型由于考虑了日用水量的众多影响因素，取得了较为理想的预测效果，但是由于日用水量的影响因素过多以及数据难以采集等缺点，因此还具有一定的局限性；自回归预测模型把系统看作一个“暗箱”，可以不管其它影响因素，而只关心预测和预测的结果，其预测过程只依赖于历史观测数据，因此省去了采集其它影响因素数据的麻烦。但是对于严重地受到各种外在因素（例如天气、节假日等）影响的日用水量来说，采用仅依赖于历史观测值的时间序列分析法进行预测，误差较大，特别是在天气变化频繁的时段内，利用自回归分析法进行的用水量预测更是如此。因此本章综合日用水量的特点，把非线性时间序列方法引入，建立城市日用水量的部分线性自回归模型，线性部分考虑日用水量，非线性部分考虑当天的最高温度。该模型综合了非参数回归模型和线性自回归模型的优点，因此在拟合与预测精度上有所提高。

5.1 部分线性时间序列模型介绍[42]

部分线性模型的一般形式为：

$$Y_t = U_t^T\beta + \phi(X_t) + e_t, \quad 1 \le t \le T, \tag{5.1}$$

其中 $U_t = (U_{t1},\cdots,U_{tp})^T$ 与 $X_t = (X_{t1},\cdots,X_{tq})^T$ 为不同的两组时间序列数据，U_t 和 X_t 可以是时间序列数据 Y_t 的滞后变量。$\beta = (\beta_1,\cdots,\beta_p)^T$ 为未知参数向量，$\phi(\cdot)$ 为 R^q 维上的非线性函数，$e_t = Y_t - E[Y_t \mid U_t, X_t]$，$e_t$ 为误差项。在模型（5.1）里 $U_t^T\beta$ 为线性时间序列部分，$\phi(X_t)$ 非线性时间序列部分。下面再介绍一下 β 和 $\phi(\cdot)$ 的估计问题。令 $A_p = \{1,2,\cdots,p\}$，$D_q = \{1,2,\cdots,q\}$，AA 表示所有非空子集 A_p，DD 表示所有非空子集 D_q，对于任何子集 $A \in AA$，U_A 定义为由 $\{U_{it}, i \in A\}$ 组成的所有元素构成的向量，β_A 定义为由 $\{\beta_i, i \in A\}$ 组成的所有元素构成的向量。对任何子集 $D \in DD$，X_{tD} 为由 $\{X_{it}, i \in D\}$ 组成的所有元素构成的向量。$d_E = |E|$ 表示集合 $|E|$ 的势。假设有唯一的 (β_*, ϕ_*) 满足（5.1）式，则（5.1）变为：

$$Y_t = U_{t*}^T\beta_* + \phi_*(X_{tD_*}) + e_t, \tag{5.2}$$

对每一个给定的 $A \in AA$ 和 $D \in DD$ 考虑一个部分线性模型：

$$Y_t = U_{tA}^T\beta_A + \phi_D(X_{tD}) + e_t(A,D), \tag{5.3}$$

先估计 β_A 和 $\phi_D(\cdot)$，作以下定义：

$$\hat{\phi}_{it}(D) = \sum_{s=1}^{T} W_D(t,s)Y_s, \tag{5.4}$$

$$\hat{\phi}_{2t}(A,D)=\sum_{s=1}^{T}W_D(t,s)U_{sA}, \tag{5.5}$$

$$Z_t(D)=Y_t-\hat{\phi}_{1t}(D), \tag{5.6}$$

$$Z(D)=(Z_1(D),\cdots,Z_T(D))^T, \tag{5.7}$$

$$V_t(A,D)=U_{tA}-\hat{\phi}_{2t}(A,D), \tag{5.8}$$

$$V(A,D)=(V_1(A,D),\cdots,V_T(A,D))^T, \tag{5.9}$$

$$\phi_1(x)=E[Y_t\mid X_t=x], \tag{5.10}$$

$$\phi_2(x)=E[U_t\mid X_t=x], \tag{5.11}$$

$$V_t=U_t-hat\phi_2(X_t), \tag{5.12}$$

$$V=(V_1(A,D),\cdots,V_T(A,D))^T, \tag{5.13}$$

其中 $\hat{\phi}_2(x)$ 定义为：

$\hat{\phi}_{2t}(A,D)=\sum_{s=1}^{T}W_D(t,s)U_{sA}$，$A=A_p$，$D=D_q$ $W_D(t,s)=\dfrac{K_D((X_{tD}-X_{sD})/h)}{\sum_{l=1}^{T}K_D((X_{tD}-X_{lD})/h)}$，$T$ 为数据量，K_D 为多元核函数，h 为带宽参数，h 满足：

$$h\in H_{TD}=[a_DT^{-\frac{1}{4+|D|}-C_D},b_DT^{-\frac{1}{4+|D|}+C_D}], \tag{5.14}$$

其中 a_D,b_D,c_D 满足 $0<a_D<b_D<\infty$，和 $0<c_D<\dfrac{1}{2(4+|D|)}$，显而易见，对 (A,D) 存在 $(2^p-1)\times(2^q-1)$ 可能，通过数据 $\{(Y_t,U_t,X_T):t=1,2,\cdots,T\}$ 选择 (A,D)，其中 $\{(Y_t,U_t,X_T):t=1,2,\cdots,T\}$ 满足：

$$Y_t=U_t^T\beta+\phi(X_t)+e_t \tag{5.15}$$

通过公式（5.3），（5.13），β_A 的最小二乘估计为：

$$\hat{\beta}(A,D)=(V(A,D)^TV(A,D))^{-1}V(A,D)^TZ(D) \tag{5.16}$$

现在根据交错鉴定（*cross-validation*）方法选择 $A\in AA$ 和 $D\in DD$，我们把数据分为两部分：$\{(Y_t,U_t,X_t):t\in S\}$ 和 $\{(Y_t,U_t,X_t):t\in S^C\}$，其中 S 为 $\{1,2,\cdots,T\}$ 的子集，$\{1,2,\cdots,T\}$ 包

含T_v个整数，S^c为S的补集，S^c包含T_c个整数。$T_v+T_c=T$。部分线性模型（5.3）通过数据$\{(Y_t,U_t,X_t):t\in S^C\}$来拟合，预测误差通过$\{(Y_t,U_t,X_t):t\in S\}$来评价，与（5.16）式一样：

$$\hat{\beta}_c(A,D)=(V_c(A,D)^TV_c(A,D))^{-1}V_c(A,D)^TZ_c(D), \tag{5.17}$$

其中：

$$V_c(A,D)=(V_{1,c}(A,D),\cdots V_{lT_c}(A,D))^T, \tag{5.18}$$

$$Z_c(D)=(Z_{1,c}(D),\cdots,Z_{lT_c,c}(D))^T, \tag{5.19}$$

$t\in S^c$或者$t\in S$，

$$V_{t,c}(A,D)=U_{tA}-\hat{\phi}_{2t}^c(A,D), \tag{5.20}$$

$$\hat{\phi}_{2t}^c=\sum_{s\in S^c}W_D(t,s)U_{sA}, \tag{5.21}$$

$$Z_{t,c}(D)=Y_t-\hat{\phi}_{1t}^c(D), \tag{5.22}$$

$$\hat{\phi}_{1t}^c(D)=\sum_{s\in S^c}W_D(t,s)Y_s, \tag{5.23}$$

对于$t\in S$，我们用$V_{t,c}(A,D)^T\hat{\beta}_c(A,D)$来预测$Z_{t,c}(D)$，用$t\in S^c$中的数据来估计（5.16）式。因此均方预测误差可以表示为：

$$CV(A,D;h)=CV(A,D;h,T_v)=CV_s(A,D)=\frac{1}{T_v}\sum_{t\in S}(Z_{t,c}(D)-\hat{Z}_t^c(A,D))^2w(X_t), \tag{5.24}$$

这里$w(X_t)$是去除X_t的异常值的权函数。我们采用*Monte Carlo* CVT_v方法选择(A,D)。随机抽取$\{1,2,...,T\}$中的n个子集中的一个集合R，该集合大小为T_v，选择(A,D)使

$$MCCV(A,D;h)=\frac{1}{n}\sum_{S\in R}CV_S(A,D;h,T_v)=\frac{1}{nT_v}\sum_{S\in R}\sum_{t\in S}(Z_{t,c}(D)-\hat{Z}_t^c(A,D))^2\omega(X_t) \tag{5.25}$$

达到最小。其中：

$$(\hat{A},\hat{D},\hat{h})=\arg\min_{\{A\in AA,D\in DD,h\in H_{TD}^c\}}MCCV(A,D;h), \tag{5.26}$$

H_{TD}^c在H_{TD}里定义，用T_c代替T。$(\hat{A},\hat{D})$即为所求，$\hat{h}$为最佳带宽。部分线性自回归模型程序计算框图见下图5-1。

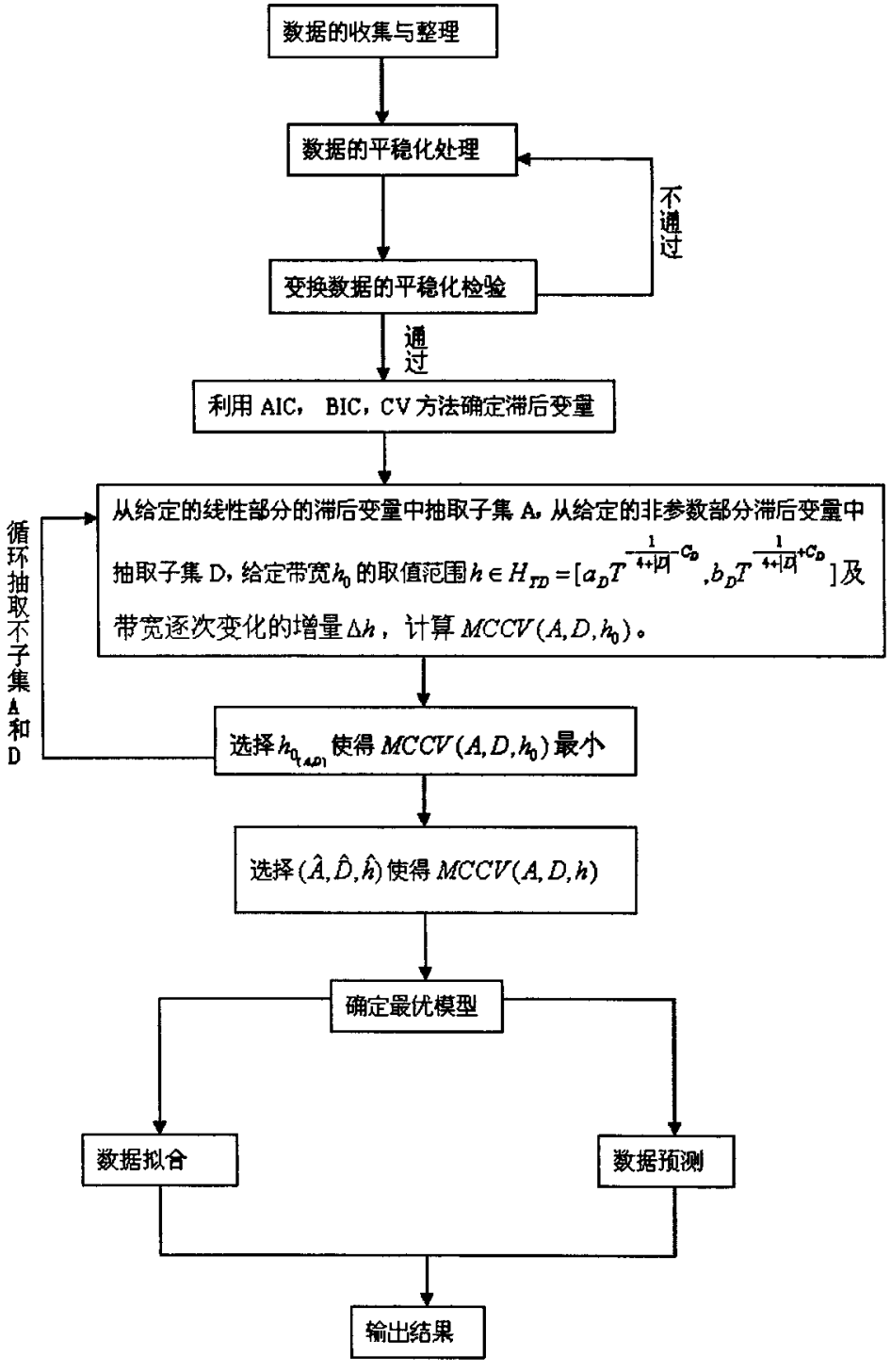

图 5-1 部分线性自回归模型流程图

Figure5-1 the flow plot of partially linear AR model

5.2 西安市日用水量部分线性自回归模型的建立

Gao and Tong (2004)[42]在研究西澳大利亚(WA)渔业捕鱼量和捕鱼船在海中作业天数之间关系时，根据已有的研究表明捕鱼量和海中作业天数之间呈非线性关系，而现在的捕鱼量与过去的捕鱼量呈线性关系，最终建立了部分线性自回归模型。反映上述现象的时间序列模型可表示为：

$$C_t = \beta_1 C_{t-1} + \cdots \beta_p C_{t-p} + \phi(E_t, E_{t-1}, E_{t-2}, \cdots, E_{t-q}) + e_t, \quad t \geq r, \tag{5.27}$$

其中 $r = \max(p,q)$, $\{e_t\}$ 是随机误差，其中 C_t 和 E_t 分别代表捕鱼量和捕鱼船在海中作业的天数。本文根据部分线性自回归模型的理论建立西安市日用水量的部分线性自回归模型。因变量为日用水量，根据现有的研究结果线性部分考虑日用水量的滞后变量，非线性部分考虑对日用水量影响较大的日最高气温。

5.2.1 数据的平稳性检验与差分变换处理

由于本文欲建立部分线性自回归模型，因此先对数据进行平稳性检验和预处理，以满足时间序列建模的需要。在上一章中已经详细讲述了数据平稳性检验的非参数检验方法，这里只给出结果，对日用水量数据和气温做对数变换和差分处理。对日用水量数据处理后可得到 $|z| = 0.0473 < 1.96$，满足平稳性条件；对日最高温度数据进行处理后可得到 $|z| = 1.1036 < 1.96$，同样满足平稳行条件。假设原模型为（5.27）式，变换数据后模型可变为：$Y_{t+r} = \beta_1 Y_{t+r-1} + \cdots + \beta_p Y_{t+r-p} + \phi(X_{t+r}, X_{t+r-1}, \cdots, X_{t+r-q}) + e_t, t \geq 1,$ （5.28）

其中 $\{e_t\}$ 是严平稳时间序列，具有零均值与有限方差，$Y_t = \log_{10}(C_t)$，$X_t = \log_{10}(E_t)$。

5.2.2 部分线性自回归模型滞后变量个数的确定

在线性时间序列模型中，滞后变量与顺序的确定通常用到 AIC，BIC 或者 FPE 准则，在非参数时间序列分析中，Auestad and Tjoshteim(1990)[43] 建议用 FPE，

$$\hat{FPE}(i) = \frac{1}{n}\sum_t [Y_t - \hat{f}\{X_t(i)\}]^2 \omega\{X_t(i)\} \frac{1 + (nh^p)^{-1} J^p B_p}{1 - (nh^p)^{-1}\{2K^p(0) - J^p\} B_p} \tag{5.29}$$

其中 $J = \int K^2(x)dx$，$B_p = n^{-1}\sum_t \frac{\omega^2\{X_t(i)\}}{\hat{p}\{X_t(i)\}}$，$\hat{f}(X_t(i))$ 为核均值估计根据 $N-W$ 核估计，

$\hat{f}(y_1, y_2, \cdots y_p) = \frac{\sum_{t=p+1}^{n} \prod_{i=1}^{p} K\{(y_t - X_{t-i})/h_i\} X_t}{\sum_{t=p+1}^{n} \prod_{i=1}^{p} K\{(y_t - X_{t-i})/h_i\}}$，通过（5.29）式确定模型的最大阶数。

Cheng和tong[43] 建议用cross validation准则确定滞后变量个数，定义

$X_t(d) = (Y_{t-1}, Y_{t-2}, \cdots, Y_{t-d})$，则：

$$CV(d) = \frac{1}{N-r+1}\sum_t \{Y_t - \hat{f}_{-t}(X_t(d))\}^2 \omega\{X_t(d)\}, \tag{5.30}$$

其中 $\hat{f}_{-t}$ 去掉第 t 个数据后函数的核估计。可以应用线性自回归模型中的 AIC，BIC 准则

进行自回归模型阶数的初步确定。然后应用*leave－one－out*交错鉴定法进行修正与确定。

表 5-1 AIC 和 BIC 确定日用水量数据滞后阶数

Table 5-1 utilize AIC and BIC methods to ascertain daily water demand data lag orders

k	1	2	3	4	5	6	7	8	9
AIC	-8.0235	-8.0256	-8.0060	-7.9917	-8.0090	-7.9888	-7.9650	-7.9458	-7.9569
BIC	-7.9945	-7.9628	-7.9192	-7.8760	-7.8644	-7.8151	-7.7624	-7.7143	-7.6965

表5-2 AIC和BIC确定日最高温度滞后阶数

Table 5-2 utilize AIC and BIC methods to ascertain tiptop temperature lag orders

k	1	2	3	4	5	6	7	8	9
AIC	-5.6836	-5.6843	-5.6704	-5.6579	-5.6078	-5.6826	-5.6608	-5.6409	-5.6186
BIC	-5.6547	-5.6244	-5.5836	-5.5422	-5.5631	-5.5109	-5.4582	-5.4094	-5.3582

初步确定q=2, p=2, 用*leave－one－out* cross validation方法可得：

表5-3 *leave－one－out* cross validation方法确定日用水量数据滞后阶数

Table 5-3 utilize *leave－one－out* cross validation method to ascertain daily water demand data lag orders

k	1	2	3	4	5	6	7	8	9
cv	0.3224	0.3698	0.5981	1.5984	0.8450	0.4692	1.5843	2.3958	1.6549

表 5-4 *leave－one－out* cross validation 方法确定日最高温度滞后阶数

Table 5-4 utilize *leave－one－out* cross validation method to ascertain tiptop temperature lag orders

k	1	2	3	4	5	6	7	8	9
cv	0.0033	0.0029	0.0098	0.1564	0.3697	0.2549	0.3654	0.3987	0.4261

鉴于滞后变量阶数过少会影响到拟合与预测效果，所以最终确定 q=2, p=2。

5.2.3 最优部分线性自回归模型的选取

在开始使用模型之前，我们需要选择出模型（5.28）的适宜与简洁的模型。我们用 6 月 1 号～8 月 24 号 84 个数据 6 月 1 号～8 月 24 号经过差分后，我们选择$n=T=84$，$T_c=[T^{3/4}]=27$，$T_v=T-T_c=57$，我们考虑用核函数$k(u_1,\cdots,u_j)=\prod_{i=1}^{j}k(u_i)$，其中$1\le j\le 4$，$k(\cdot)=\frac{1}{\sqrt{2\pi}}\exp(-u^2/2)$，$H_{TD}^{c}=[0.1\cdot T_c^{-\frac{2}{9}},3\cdot T_c^{-\frac{1}{9}}]$，$|D_v|=2$，根据 MATLAB 编程，$MCCV(T_v)$选择最后确定最优模型为：

$$Y_{t+2}=\hat{\beta}_1Y_{t+1}+\hat{\beta}_2Y_t+\hat{\phi}(X_{t+2},X_{t+1}),1\le t\le 82 \qquad (5.31)$$

其中$\hat{\beta}_1=0.9843$，$\hat{\beta}_2=-0.0408$，$\hat{\phi}(\cdot)$为非参数估计，模型（5.12）的中的最适宜带宽为

$\hat{h}=0.1482$，拟合结果见下表 5-5 与图 5-3。

表 5-5 部分线性自回归模型拟合结果

Table5-5 the simulation result of partially linear AR model

日期	实际值 $/m^3$	预测值 $/m^3$	绝对误差 $/m^3$	相对误差 /%	日期	实际值 $/m^3$	预测值 $/m^3$	绝对误差 $/m^3$	相对误差 /%
6.01	870265	-------	--------	---------	7.14	751511	768943	-17432	-2.32
6.02	852089	-------	--------	---------	7.15	734782	748956	-14174	-1.93
6.03	839438	--------	---------	---------	7.16	737928	737895	33	0.00
6.04	802901	815987	-13086	-1.62	7.17	803530	791567	11963	1.49
6.05	859839	849031	10808	1.25	7.18	847766	847894	-128	-0.02
6.06	854856	856934	-2078	-0.24	7.19	851843	845698	6145	0.72
6.07	843568	835987	7581	0.90	7.20	873839	887865	-14026	-1.60
6.08	851760	859567	-7807	-0.92	7.21	879179	874569	4610	0.52
6.09	830944	845698	-14754	-1.78	7.22	910541	906543	3998	0.44
6.10	784532	794502	-9970	-1.27	7.23	882903	889564	-6661	-0.75
6.11	844146	841365	2781	0.33	7.24	885238	885698	-460	-0.05
6.12	831512	829846	1666	0.20	7.25	946915	938962	7953	0.84
6.13	880478	866541	13937	1.58	7.26	923076	925689	-2613	-0.28
6.14	873326	859850	13476	1.54	7.27	923460	914568	8892	0.96
6.15	904991	895067	9924	1.09	7.28	965979	968954	-2975	-0.31
6.16	915223	906548	8675	0.95	7.29	1014790	989315	25475	2.51
6.17	942777	938905	3872	0.41	7.30	1042459	998465	43994	4.22
6.18	964709	970326	-5617	-0.58	7.31	977334	968954	8380	0.86
6.19	964373	957236	7137	0.74	8.01	889001	885694	3307	0.37
6.20	986222	991032	-4810	-0.49	8.02	768613	759862	8751	1.14
6.21	954107	980030	-25923	-2.72	8.03	745256	754569	-9313	-1.25
6.22	850633	840974	9659	1.13	8.04	797046	807454	-10408	-1.31
6.23	852424	860132	-7708	-0.90	8.05	843197	859845	-16648	-1.97
6.24	890963	889841	1122	0.12	8.06	880528	858956	21572	2.45
6.25	831124	841369	-10245	-1.23	8.07	885245	889562	-4317	-0.49
6.26	810729	809865	864	0.10	8.08	826712	831564	-4852	-0.59
6.27	857331	861239	-3908	-0.46	8.09	789677	782152	7525	0.95
6.28	878399	871234	7165	0.82	8.10	766365	774840	-8475	-1.11
6.29	840524	838945	1579	0.19	8.11	773343	764845	8498	1.10
6.30	815263	828494	-13231	-1.62	8.12	768931	769165	-234	-0.03
7.01	792317	784861	7456	0.94	8.13	780359	774552	5807	0.74
7.02	831177	834594	-3417	-0.41	8.14	768013	778552	-10539	-1.37
7.03	821353	828946	-7593	-0.92	8.15	755659	755515	144	0.02
7.04	842671	834206	8465	1.00	8.16	781812	784561	-2749	-0.35
7.05	835624	834568	1056	0.13	8.17	803163	844165	-41002	-5.11
7.06	851529	848953	2576	0.30	8.18	812807	808123	4684	0.58
7.07	857987	851233	6754	0.79	8.19	815311	807452	7859	0.96
7.08	851434	841235	10199	1.20	8.20	810067	815953	-5886	-0.73

续表									
7.09	813155	828465	-15310	-1.88	8.21	836303	820841	15462	1.85
7.10	846894	854642	-7748	-0.91	8.22	862114	835616	26498	3.07
7.11	803895	813207	-9312	-1.16	8.23	877021	868489	8532	0.97
7.12	756685	767898	-11213	-1.48	8.24	862650	851654	10996	1.27
7.13	737839	749867	-12028	-1.63					

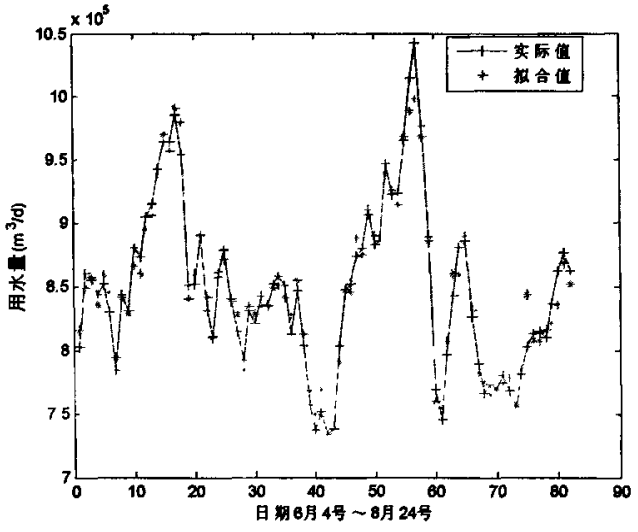

图 5-3 部分线性自回归模型拟合图

Figure5-3 the simulation result of partially linear AR model

5.2.4 部分线性自回归模型的检验及预测结果分析

模型建立以后，还要对其拟合残差的有效性进行检验。这里采用 χ^2 检验法，设其前 m 个自相关系数为 $\hat{\rho}_1,\hat{\rho}_2,\cdots,\hat{\rho}_m$，则统计量 $Q=n\sum_{k=1}^{m}\hat{\rho}_k^2\sim\chi^2(m-1)$。置信度为 95%时，所对应的 χ^2 数值如果比 Q 大，就有 95%的置信度认为原假设成立，否则就否定 e_t 为白噪声序列。计算出该 $\{e_t\}$ 时间序列数据中的前 22 个自相关系数， Q=4.860<11.591，所以有 95%的置信度认为残差序列具有随机性，因此该部分线性自回归模型是有效的。

模型通过检验以后，就可以根据模型（5.31）做预测，采取动态预测方法进行预测。先对第 85 个数据进行预测，再把第 85 个预测数据补充到用水量数据中，去预测第 86 个数据，依次类推。这样上一时期较大的误差，不会对下一时期造成较大的影响。8 月 25 号到 8 月 31 号实际值与预测值的比较见表 5-6。

表 5-6 8.25～8.31 实际值与预测值比较

Table5-6 the comparison of real data and forecasting data from 8.25 to 8.31

日期	实际值 / m^3	预测值 / m^3	绝对误差 / m^3	相对误差 / %
8.25	856551	861560	-5009	-0.58%
8.26	811914	804764	7150	0.88%
8.27	849784	852367	-2583	-0.30%
8.28	791414	780694	10720	1.35%
8.29	747147	760103	-12956	-1.73%
8.30	785090	783260	1830	0.23%
8.31	764637	786347	-21710	-2.84%

5.3 四种模型的拟合与预测效果比较

通过拟合与预测效果（用均方误差 MSE、平均绝对误差 MAE 及预测相对误差）对本文所采用的城市日用水量预测方法优劣性作出评价，结果可见表 5-7 与表 5-8

$$MSE=\sqrt{\frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{n}} \qquad MAE=\frac{1}{n}\sum_{i=1}^{n}|Y_i-\hat{Y}_i|$$

表 5-7 四种模型的拟合的平均绝对误差与均方误差比较表

Table5-7 the comparison of MSE and MAE of four models

	线性综合模型 / m^3	多元非参数回归模型		线性自回归模型 / m^3	部分线性自回归模型 / m^3
		核估计 / m^3	局部线性估计 / m^3		
MAE	21121.47	16007.48	14786.69	22856.93	10864.41
MSE	28630.19	24973.50	22653.92	29415.15	16956.20

表 5-8 预测相对误差比较表

Table5-8 the comparison of prediction comparatively error

日期	线性综合模型相对误差/%	多元非参数回归模型		线性自回归模型相对误差/%	部分线性自回归模型相对误差/%
		核估计相对误差/%	局部线性估计相对误差/%		
8.25	3.59%	-0.4%	0.18%	2.48%	-0.58%
8.26	1.12%	1.77%	1.12%	-2.95%	0.88%
8.27	-7.58%	-4.32%	-3.36%	1.50%	-0.30%
8.28	-2.60%	-0.94%	-0.74%	-5.79%	1.35%
8.29	-1.86%	0.03%	-0.06%	-7.66%	-1.73%
8.30	7.77%	1.87%	1.71%	-4.21%	0.23%
8.31	5.85%	-1.13%	0.66%	-3.87%	-2.84%

通过比较可以得出，在日用水量预测这一非线性问题中，部分线性自回归模型与多元非参数回归模型效果好于线性回归模型与线性自回归模型，线性自回归模型效果最差，这是因为最高温度对日用水量影响过大造成的。部分线性自回归模型在四种建模方法中效果最优。

5.4 本章小结

本章综合日用水量的特点，把非线性时间序列方法引入，建立城市日用水量的部分线性自回归模型，首先介绍了部分线性自回归模型建立的理论推导，其中线性部分考虑城市日用水量，非线性部分考虑当天的最高温度。该模型综合了多元非参数回归模型和线性自回归模型的优点，因此在拟合与预测精度上有所提高，最后通过与前三种方法比较，精度有所提高，说明该方法在西安市日用水量预测中是有效的。

6 结论

用水量预测在城市建设规划和供水系统调度管理中具有重要的作用。用水量预测主要包括：年用水量预测，季度用水量预测，月用水量预测，日用水量预测和时用水量预测。其中，日用水量预测占据着重要的地位，它不仅能直接指导水厂的生产实践，更能为水厂间的优化调度提供可靠的技术支持。本文在分析西安市供水系统原始观测数据资料的基础上，建立了西安市日用水量多元非参数回归模型和部分线性自回归模型以及其它预测模型。

6.1 主要研究成果

1.建立了西安市日用水量的多元线性回归模型，对模型进行了 χ^2 检验和残差修正，同时建立了西安市日用水量的综合预测模型，计算结果表明，该模型预测精度不高。

2.利用核估计和局部线性估计方法建立了西安市日用水量的多元非参数回归模型。通过对日用水量的拟合与预测，结果表明多元非参数回归模型拟合与预测误差较小，能满足供水系统的需要。

3.根据 BJ 时间序列建模方法，建立了西安市日用水量的线性自回归模型。结果表明，该模型预测精度不高。

4.建立了西安市日用水量的部分线性自回归模型，其中线性部分考虑日用水量，非线性部分考虑当天的最高温度，该模型综合了非参数回归模型和线性自回归模型的优点，与前三种建模方法比较，该模型在拟合与预测精度上有所提高，证明该方法在西安市日用水量预测这四种模型中效果最优。

6.2 尚待研究的问题

本文所建立的模型均有其特殊的意义，在实际中表示了一定相关关系。这些说明所建模型是科学合理的，具有一定的实用价值。但所建模型也还有不完善的地方，有待进一步改进：

1.综合线性预测模型回归系数随着季节的变化，应是动态变化的，特别在是冬季用水结构发生变化时（冬季有取暖用水）更是如此。上述模型中的各系数均是恒定的，这在温度发生突变（低温）的时候会引起较大的误差。

2.多元非参数回归模型可以尝试考虑更多的影响因素，估计方法可以用小波估计，或者样条和稳健估计来完成。

3. 部分线性自回归模型中，非线性部分可以考虑更多的影响因素，对非线性时间理论进行深入分析，以减小预测误差。

致 谢

本论文是在我的导师张德生教授的严格要求和悉心指导下完成的，在此我要深深地感谢张老师对我的谆谆教诲和殷殷关切。在我攻读硕士学位期间，张老师的对待知识一丝不苟的态度使我在学业上受益匪浅，并且在树立正确的人生观上也给了我很大的启发。张老师渊博的知识、严谨的学风、敏锐的思维、和蔼的态度以及对科学的敬业精神使我受益很大。在此，对张老师表示衷心的感谢！

需要特别感谢的是西北工业大学博士武新乾师兄，在本论文的开题和进行过程中得到了师兄的大力支持和帮助，在论文思路的形成和估计方法理论的推导上他给了我很多宝贵的意见和建议。师兄的博学、勤奋也深深地影响了我，给我树立了一面旗帜，激励我上进。在这里，再次对师兄表示衷心的感谢！

在完成论文过程中，感谢西安市自来水公司提供的数据，感谢水利水电学院张雄同学给予的热情帮助。感谢我的同学巩永丽、姜爱平、张小静的互相讨论和勉励，使大家共同按时完成学业。

在硕士学习期间，理学院的老师们为我的学习提供了许多帮助和便利条件，老师们渊博的知识对我学业的帮助很大，在跟随老师们学习过程中，学到了许多有用的知识，在此，向老师们表示衷心的感谢！

最后，衷心感谢我的家人对我的全力支持和照顾，使我能全身心地投入到学习当中，按时完成学业！

陈战波

2007 年 3 月

参 考 文 献

【1】李红艳. 城市给水系统优化调度模型研究[D] . 太原：太原理工大学，2003：26.

【2】柳景青. 调度时用水量预测的系统理论方法及应用研究[D]. 杭州：浙江大学，2005：3.

【3】Hartly J.A,powell R.S.The development of a combined demand prediction system[J].Civil Engineering Systems,1991,8(4):231-236.

【4】吕谋，赵洪宾.城市日用水量预测的组合动态建模方法[J]. 给水排水，1997，23(11):25-27.

【5】何文杰. 供水管网动态模拟技术的研究[D]. 哈尔滨：哈尔滨建筑大学，2001：14-19.

【6】周建华. 大规模城市给水管网系统优化运行模型研究[D]. 哈尔滨：哈尔滨建筑大学，2003：1-83.

【7】张雄，党志良，张贤洪，马丁. 城市用水量预测模型综合研究[J] . 水资源研究，2005，26(1)：21-24.

【8】侯煜坤，刘遂庆，陶涛. 城市日用水量的自回归模型（AR）预测方法[J]. 河南科学，2004，22(4)：502-504.

【9】吕谋，赵洪宾. 时用水量预测的自适应组合动态建模方法[J]. 系统工程理论与实践，1998，18（8）:101-107，112.

【10】张宏伟. 城市供水系统运行决策支持系统[D]. 天津：天津大学，2001：1-86.

【11】Liu Hongbo,Zhang Hongwei,et al.Comparision of the City water consumption Short-term Forecasting Methods[J].Transactions of Tianjin University,2002,8(3):211-215.

【12】柳景青，张土乔. 时用水量的分时段混沌建模方法[J]. 浙江大学工学版，2005，39(1)：11-15.

【13】Vicente J.Rafael J.B.,et al.Stochastic model to evaluate residential water demands[J].Journals of Water resources Planning and management,2004,130(5):386-394.

【14】Robinson P.M. Root -consistent semiparametric regression[J]. Econometrica,1988,56(9):931-954.

【15】Heckman,N.Spline smoothing in a partly linear model[J].Roy.Stat.Soc.Ser,1986,B48:244-248.

【16】Rice,J.Convergence rates for partial spline models[J].Stat.Prob.Lett. 1986,4(4):203-208.

【17】Chen.H.Convergence rate for parametric components in a partly linear model[J]. Ann.Stat 1988,16(3):136-146.

【18】Speckman,P.Kernal smoothing in partial linear models[J].Roy.Statist.Soc.Ser ,1988,B50:413-436.

【19】Robinson,P.M.Asymptotically efficiency estimation in the presence of heteroscedasticity of unknown form[J].Econometrika,1987,55(7):875-891.

【20】Gao,J.T and Zhao,L.C. Adaptive estimation in partly linear models[J].Science in China,1988,A: 791-803.

【21】Hong, S. Y and Zhao, Z. B. Asymptotic of kernel-least square methods in a partly linear model[J]. Ann. Math, 1993, 14A:717-731.

【22】胡舒合. 固定设计半参数回归模型估计的强相合性[J]. 数学学报，1994，37（2）:393-401.

【23】钱伟民，柴根象. 半参数回归模型的误差方差的小波估计[J]. 数学年刊，2000，21A(3):341-350.

【24】柴根象，孙平. 半参数回归模型的二阶段估计[J]. 应用数学学报，1995，18(3):353-363.

【25】Gao J. Asymptotic properties of some estimators for partly linear stationary autoregressive models[J].

Commun. Statist.-Theory and Methods,1995,24(3):211-226.

【26】Gao J. and Liang H. Asymptotic normality of pseudo-LS estimator for partly linear autoregression models[J]. Statist. Probab. Lett,1995,23(1):27-34.

【27】Gao J. Semiparametric regression smoothing for non-linear time series[J]. Scandinavian Journal of Statistics,1998,25 (8): 521-539.

【28】Schick A. Efficient estimation in a semiparametric autoregressive model[J]. Statistical Inference for Stochastic Processes,1999,2(1):69-98.

【29】Gao J. and Yee T. Adaptive estimation in partially linear autoregressive models[J]. The Canadian Journal of Statistics,2000,28 (9): 571-586.

【30】Linton D. and Mammen E. Estimating semiparametric ARCH(∞) models by kernel smoothing methods (with discussion) [J]. Econometrica,2003,73(8):771-886.

【31】Härdle W., Liang H. and Gao J. Partially Linear Models[M]. New York.Springle Series in Contributions to Statistics. Physica-Verlag, 2000:12-96.

【32】Gao J. and Tong H. Semiparametric non-linear time series model selection[J]. J. R. Statist. Soc. B, 2004,66 (4):321-336.

【33】Bosq D. and Shen J. Estimation of an autoregressive semiparametric model with exogenous variables[J]. J. Statist. Plann. Inference,1998,68(2):105-127.

【34】Liang H. Second order asymptotic efficiency in a partly autoregressive model[J]. Systems Science and Mathematical Sciences, 1996,9 (2):164-170.

【35】Zhu L.X. and An H.Z. A note on the strong consistency of estimates in partially linear models[J]. Acta Mathematica Scientia, 1994,14 (2): 146-152.

【36】Teräsvirta T., Tjøstheim D. and Granger C.W.J.Aspects of modelling nonlinear time series[J]. Handbook of Econometrics,1994, 4(3):219-257.

【37】柴根象，洪圣岩.半参数回归模型[M]. 安徽合肥：安徽教育出版社,1999：22-35.

【38】王松桂. 线性统计模型[M]. 北京：高等教育出版社，1999：1-54.

【39】何书元. 应用时间序列分析[M]. 北京：北京大学出版社,2003：1-79.

【40】叶阿忠. 非参数计量经济学[M]. 天津：南开大学出版社，2003：1-86.

【41】张树京等. 时间序列分析简明教程[M] . 北京：清华大学出版社,2003：1-49.

【42】JiTI Gao. And Howell Tong. Model selection in nonparametric and semiparametric time series regression[J].Statist Infer Stochastic Processes,1999,2(1):69-98.

【43】Wolfgang Hardle and Rong Chen .Nonparametric time analysis, a selective review with examples[J]. Journal of Nonparametric Statistics,1995,5(2):157-184.

【44】陈家鼎等. 数理统计学讲义[M]. 北京：高等教育出版社，1993：10-78.

【45】安鸿志. 时间序列的分析与应用[M]. 北京：科学技术出版社，1983：1-89.

【46】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，1988：1-92.

【47】田铮翻译. 时间序列的理论与方法第二版[M]. 北京：高等教育出版社，2001：2-76.

【48】钱伟民，柴根象. 半参数回归模型小波估计的强逼近[J]. 中国科学（A 辑），1999,29(3):233-240.

【49】薛留根，韩建国. 半参数回归模型中二阶段估计的渐近性质[J] . 应用数学学报（A 辑），2001，16(1)：87-94.

【50】任哲，胡舒合. 部分线性模型中参数估计的 Bootstrap 逼近[J]. 数学杂志，2002，22(3)：23-27.

【51】程正兴. 小波分析算法与应用[M]. 西安：西安交通大学出版社，1998：2-65.

【52】王振龙. 时间序列分析[M]. 北京：中国统计出版社，2000：1-74.

【53】王燕. 应用时间序列分析[M]. 北京：中国人民大学出版社，2005：1-44.

附 录

本人在攻读硕士学位期间发表的论文：

1. 陈战波，张德生.城市日用水量预测的非参数模型，青岛科技大学学报，2007，28（1）：65-68.
2. 陈战波，张德生，武新乾.自回归模型在西安市日用水量预测中的应用，江汉大学学报，2006，34（4）：10-12.
3. 韩有旺，李九红，陈战波.ANSYS 在纤维布加固钢筋混凝土梁的非线性分析的应用，西北水利发电，2006，22（5）：53-55.
4. 姜爱平，张德生，武新乾，陈战波.中国城镇居民消费函数的非参数模型，海南师范学院学报，2006，19（3）：214-218.
5. 陈战波，张德生，武新乾.城市日用水量预测的部分线性自回归模型研究，已投《数理统计与管理》。

西安市自来水公司日用水资料

2003.06.01~30

日期	用水量	日最高温度	日最低温度	温差	日平均温度	阴晴状况	降水	节假日
1	870265	32.0	27.0	5.0	25.1		0.0	Sun.
2	852089	35.0	22.7	12.3	26.0		0.7	
3	839438	28.0	18.0	10.0	21.7			
4	802901	25.6	17.0	8.6	21.1			
5	859839	33.6	18.8	14.8	24.6			
6	854856	34.5	19.3	15.2	26.7			
7	843568	32.7	22.4	10.3	27.9			Sat.
8	851760	34.9	24.3	10.6	28.9		0.0	Sun.
9	830944	30.4	20.7	9.7	24.1		0.0	
10	784532	24.1	19.5	4.6	21.8			
11	844146	32.5	15.6	16.9	24.7			
12	831512	32.5	17.2	15.3	25.5			
13	880478	32.2	20.5	11.7	25.8			
14	873326	36.0	19.5	16.5	27.4			Sat.
15	904991	37.1	21.6	15.5	29.0			Sun.
16	915223	35.0	22.4	12.6	28.3			
17	942777	36.9	22.3	14.6	29.8			
18	964709	37.7	23.5	14.2	31.0			
19	964373	37.3	24.5	12.8	30.0			
20	986222	26.7	23.7	3.0	30.1			
21	954107	33.4	24.2	9.2	28.2		32.0	Sat.
22	850633	29.5	20.8	8.7	24.7		0.3	Sun.
23	852424	34.0	21.8	12.2	27.9			
24	890963	34.4	24.2	10.2	29.9			
25	831124	33.2	20.9	12.3	26.5		14.0	
26	810729	30.0	19.7	10.3	24.0			
27	857331	35.5	22.0	13.5	28.6			
28	878399	34.0	23.6	10.4	29.2			Sat.
29	840524	32.0	26.0	6.0	27.3		0.0	Sun.
30	815263	26.2	20.6	5.6	22.4			

西安市自来水公司日用水资料

2003.07.01~31

日期	用水量	日最高温度	日最低温度	温差	日平均温度	阴晴状况	降水	节假日
1	792317	23.4	19.2	4.2	20.3		19.8	Tue.
2	831177	29.5	19.3	10.2	23.4			
3	821353	27.5	21.7	5.8	24.0		0.9	
4	842671	32.0	19.4	12.6	25.9			
5	835624	32.8	21.5	11.3	27.3			Sat.
6	851529	34.5	23.2	11.3	28.5		0.0	Sun.
7	857987	23.9	20.7	3.2	26.0		47.0	
8	851434	30.7	21.2	9.5	25.5			
9	813155	28.8	21.6	7.2	23.6			
10	846894	30.8	19.4	11.4	25.2			
11	803895	28.8	21.0	7.8	24.1		13.7	
12	756685	24.4	19.7	4.7	21.2		13.5	Sat.
13	737839	21.9	19.3	2.6	20.2		23.7	Sun.
14	751511	24.9	17.9	7.0	21.3		0.0	
15	734782	24.1	20.4	3.7	21.9		17.0	
16	737928	23.0	19.2	3.8	20.6		0.5	
17	803530	33.1	19.9	13.2	25.2			
18	847766	35.3	22.3	13	28.5			
19	851843	35.3	25.3	10	30.1			Sat.
20	873839	31.3	24.2	7.1	27.4			Sun.
21	879179	33.5	23.9	9.6	28.0			
22	910541	32.7	22.3	10.4	27.3		0.1	
23	882903	33.7	24.8	8.9	29.5		0.3	
24	885238	37.3	24.3	13	30.3		0.4	
25	946915	39.3	27.4	11.9	33.2			
26	923076	36.5	28.9	7.6	33.4			Sat.
27	923460	33.5	27.5	6	30.5			Sun.
28	965979	36.6	27.2	9.4	31.1			
29	1014790	37.1	25.8	11.3	30.6			
30	1042459	38.1	27.1	11.2	32.6			
31	977334	35.6	27.9	7.7	31.1			

西安市自来水公司日用水资料

2003.08.01~31

日期	用水量	日最高温度	日最低温度	温差	日平均温度	阴晴状况	降水	节假日
1	889001	31.8	22.5	9.3	25.8		4.4	
2	768613	32.9	20.9	12	21.8		12.9	Sat.
3	745256	22.2	19.9	2.3	21.2		9.4	Sun.
4	797046	26.2	20.6	5.6	23.3			
5	843197	30.4	23.2	7.2	26.0			
6	880528	32.0	24.9	7.1	28.4			
7	885245	33.4	25.9	7.5	29.6		14.1	
8	826712	31.4	23.1	8.3	25.9		16.8	
9	789677	26.5	22.8	3.7	24.9		1.1	Sat.
10	766365	25.9	19.1	6.8	22.7		7.0	Sun.
11	773343	23.6	18.7	4.9	21.0		0.5	
12	768931	22.6	18.3	4.3	20.2		13.6	
13	780359	23.9	17.7	6.2	20.7		0.2	
14	768013	24.7	19.5	5.2	20.4		19.1	
15	755659	24.9	18.0	6.9	20.6		1.3	
16	781812	24.9	15.2	9.7	20.3			Sat.
17	803163	26.6	16.9	9.7	21.8			Sun.
18	812807	35.2	22.3	12.9	28.5			
19	815311	30.4	20.5	10	25.1			
20	810067	29.4	20.9	8.5	24.6			
21	836303	33.5	23.9	9.6	28.6			
22	862114	32.5	23.2	9.3	27.8			
23	877021	34.3	24.6	9.7	28.9			Sat.
24	862650	32.1	25.8	6.3	28.4		0.0	Sun.
25	856551	28.8	23.7	5.1	25.6		0.0	
26	811914	26.0	21.3	4.7	22.6		0.0	
27	849784	33.5	27.5	6.0	30.5			
28	791414	27.4	21.8	5.6	23.2		20.6	
29	747147	21.9	17.7	4.2	19.8		34.8	
30	785090	19.7	15.1	4.6	17.2			Sat.
31	764637	19.2	15.2	4.0	17.3		38.9	Sun.