

基于预测误差的用水预测模型选择方法研究

王媛冯骞*

(河海大学环境学院, 南京 210098)

摘要 时间序列法是用水量预测的常用方法,其中预测模型的选择是提高预测精度的关键。针对目前以拟合精度作为模型选择标准的方法,提出了以预测值的置信区间最小为标准选择预测模型的方法。以山西省运城市工业用水量资料为例,分析给出了幂函数、S函数、直线、指数函数、二次和三次抛物线等6个模型95%置信水平下5年和10年两种预测期的预测区间,均以幂函数预测精度最高;对该系列近5年的用水量进行了预测,预测误差也以幂函数最小。表明以预测区间最小选择用水预测模型的方法是合理的、可靠的。

关键词 用水量 模型选择 预测期 预测模型

中图法分类号 TB114; **文献标志码** A

研究用水预测方法、合理预测未来水资源需求,对实现水资源优化配置,促进我国经济社会长期、稳定、快速发展具有重要指导意义^[1]。用水预测有多种方法和模型^[2],如根据对数据处理方式的不同,可以分为时间序列法、结构分析法、系统方法等^[3],在时间序列法中又有灰色预测模型^[4]、自回归模型、组合模型^[5]等。在如此众多的方法和模型中选择一个合适的模型,是用水预测研究的一个重要问题。对于用水预测模型的选择,目前的方法主要是以模型对历史数据的拟合精度为依据。如汤成友等^[6]利用金沙江屏山站1950年~2005年的资料,用前46年的资料率定模型,用后10年的资料检验模型。但是选取合适的预测模型是为了使预测值与实际值的误差最小,而不是模型与现有数据的拟合精度最高。据此,现提出以预测误差最小为标

准,即以一定的置信度,根据预测值置信区间大小来选择预测模型的方法。利用山西省运城市工业用水量资料对此方法进行了分析研究与合理性检验。

1 方法与材料

1.1 拟定待选模型

拟定待选模型为幂函数、直线、指数、逻辑斯蒂函数(S函数)、二次抛物线和三次抛物线模型等,其数学表示式分别为 $y = ax^b$ 、 $y = a + bx$ 、 $y = ae^{bx}$ 、 $y = c/(1 + ae^{bx})$ 、 $y = a_2x^2 + a_1x + a_0$ 和 $y = a_3x^3 + a_2x^2 + a_1x + a_0$ 。式中, y 为用水量,亿 m^3 , x 为时间, a ; a 、 b 、 c 、 a_0 、 a_1 、 a_2 、 a_3 为模型参数。

1.2 模型参数估计与检验

1.2.1 参数估计

采用多元线性回归分析方法建立数学模型,多元线性函数的数学模型为

$$y = \beta_0 + \beta_1x_1 + \cdots + \beta_kx_k + \varepsilon \quad (1)$$

式(1)中, $\beta_0, \beta_1, \beta_2, \cdots, \beta_k$ 为 $k+1$ 个未知参数, ε 是随机变量。

通过 n 次观察,可得到 n 组观察值,用矩阵表示

2010年4月12日收到 国家自然科学基金(50679055)资助
第一作者简介:王媛(1988—),女,山西太原人,研究方向:给排水工程, E-mail: wyrf@163.com。

*通讯作者简介:冯骞(1977—),男,浙江绍兴人,讲师,博士,研究方向:水处理理论与技术、给水排水工程优化。E-mail: xiaofq@hhu.edu.cn。

为 $Y = X\beta + \varepsilon$, 式中 $Y = [y_1, y_2, \dots, y_n]^T$,

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix},$$

$$\beta = [\beta_0, \beta_1, \dots, \beta_k]^T, \varepsilon = [\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n]^T.$$

用最小二乘法求 $\beta_0, \dots, \beta_k$ 的估计量 $\hat{\beta}_0, \dots, \hat{\beta}_k$, $\hat{\beta} = (X^T X)^{-1} (X^T Y)$, $\hat{\beta}_0, \dots, \hat{\beta}_k$ 称为经验回归系数。

1.2.2 参数检验

常用的检验方法有 F 检验法、 t 检验法、 r 检验法, 这里采用 F 检验法。

$$F = \frac{\frac{U}{k}}{\frac{Q_e}{(n-k-1)}} \tag{2}$$

式(2)中 $U = \sum_{i=1}^n (y_i - \bar{y})^2$, 称为回归平方和, $Q_e = \sum_{i=1}^n (y_i - \hat{y}_i)^2$, 称为剩余平方和。 y_i 为实测值, $\hat{y}_i$ 为预测值, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_k x_{ik}$, $i = 1, 2, \dots, n$ 。 F 值反映了拟合精度的高低, F 值越大, 表示拟合精度越高。

当显著水平 α 给定后, 如果 $F > F_{1-\alpha}(k, n-k-1)$, 则认为 y 与 $x_1, \dots, x_k$ 之间显著地有线性关系; 否则就认为 y 与 $x_1, \dots, x_k$ 之间线性关系不显著。

1.3 预测值置信区间的计算

模型预测值的置信区间在一定程度上表示了模型预测精度, 置信区间越小表示模型的预测精度越高。在置信水平为 α 条件下, 模型预测值 $\hat{y}$ 的置信区间为 $(\hat{y} - K, \hat{y} + K)$, 其中^[7],

$$K = \hat{\sigma}_e t_{1-\frac{\alpha}{2}}(n-k-1) \sqrt{1 + GCG^T} \tag{3}$$

式(3)中, $\hat{\sigma}_e = \sqrt{\frac{Q_e}{n-k-1}}$, 称为剩余标准差; $t_{1-\frac{\alpha}{2}}(n-k-1)$ 为置信度为 $1 - \alpha$ 条件下的 t 分布值, 可查表求得; $G = [1, x_1, \dots, x_k]$ 为预测点, 系已知值; $C = (X^T X)^{-1}$ 。

1.4 预测精度的评价

以预测值的相对误差作为预测精度, 按照式(4)计算。若 $MAPE \leq 15\%$, 认为预测是成功的, 若

$MAPE \leq 10\%$, 认为是高精度的预测^[8]。

$$MAPE = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{y_i} \times 100 \tag{4}$$

式中 $MAPE$ 为平均相对误差, %; 其余符号意义同前。

1.5 资料

山西省运城市是一个水资源极为短缺的地区, 水资源短缺已成为制约该地区经济建设和工业发展的重要因素。为了实现该地区水资源的可持续利用, 以及经济社会的可持续发展, 现利用 1985 年~2007 年用水系列资料, 对两种模型选择方法进行了分析与比较, 并对 5 年和 10 年两个预测期的工业用水量进行了预测分析。

2 两种模型选择方法的比较

2.1 以拟合精度为标准选取模型

剩余标准差、复相关系数 R 和 F 值表示了模型拟合精度的大小, 剩余标准差越小(复相关系数越大)表明拟合精度越高。由表 1 可见, 三次抛物线的拟合精度最高, 次之为幂函数, 以指数函数拟合精度最低, 相应的复相关系数分别为 0.963 9、0.923 3 和 0.838 2。因此, 以拟合精度为标准选取的模型应为三次抛物线。

2.2 以预测精度为标准选取模型

直线模型可采用一元线性回归分析确定模型参数, 即令式(1)中 $k = 1$, 利用上述方法求得参数的估计值; 幂函数、指数函数、二次抛物线及三次抛物线模型通过变量代换可化为直线模型, 然后按照上述方法求取参数; S 函数须采用非线性拟合方法求其参数。并相应地求得各模型剩余标准差 $\hat{\sigma}_e$ 、复相关系数 R 和 F 值, 以及置信度 $1 - \alpha = 0.95$ 时两种预测期(5 年和 10 年)的置信区间 K 值, 结果见表 1。由表 1 可见所选择的模型均达到了极显著水平, 均可用于运城市工业用水量的预测, 但是各个模型预测结果的精度(K 值)是不相同的, 因此, 合理地选择预测模型, 可以提高预测精度。

表 1 模型参数拟合结果

模型	参数	$\hat{\sigma}_e$	R^2	F	$K/\text{亿 m}^3$	
					5 年	10 年
1. 幂函数	$a=2.436\ 7, b=-0.274\ 8$	0.105\ 3	0.923\ 3	252.7 * *	0.256\ 0	0.277\ 0
2. 直线	$a=0.673\ 3, b=0.054\ 4$	0.1118	0.919\ 4	239.4 * *	0.264\ 7	0.282\ 8
3. S 函数	$a=2.266\ 7, b=0.116\ 9, c=2.122\ 8$	0.109\ 1	0.914\ 1	223.4 * *	0.265\ 2	0.287\ 0
4. 指数	$b=-0.134\ 3, a=1.034\ 7$	0.140\ 3	0.838\ 2	108.8 * *	0.341\ 2	0.369\ 3
5. 二次抛物线	$a_0=5.918\ 9 \times 10^{-1}, a_1=7.392\ 6 \times 10^{-2}, a_2=-8.146\ 0 \times 10^{-4}$	0.1093	0.926\ 6	126.2 * *	0.365\ 2	0.555\ 2
6. 三次抛物线	$a_0=3.420\ 7 \times 10^{-1}, a_1=1.869\ 9 \times 10^{-1},$ $a_2=-1.234\ 4 \times 10^{-2}, a_3=3.202\ 7 \times 10^{-4}$	0.076\ 7	0.963\ 9	266.9 * *	0.488\ 7	1.183\ 0

* * 表示 F 检验达到极显著水平 ($F > F_{0.01}$)。

由表 1 可见,幂函数的预测精度最高(K 值最小),次之为直线,再次为 S 函数,以三次抛物线的预测精度最低。相应 5 年预测期的 K 值分别为 0.256 0 亿 m^3 、0.264 7 亿 m^3 、0.265 2 亿 m^3 、0.488 7 亿 m^3 ,10 年预测期的 K 值分别为 0.277 0 亿 m^3 、0.282 8 亿 m^3 、0.287 0 亿 m^3 、1.183 0 亿 m^3 。由此可见,应选择幂函数作为山西运城市工业用水量预测模型。

2.3 模型选择方法的合理性检验

统计理论要求预测点(G)必须在观测值(X)范围内^[9]。用水预测是对未来某年的用水量做预测,其预测点必然在观测值范围之外,因而利用 K 值选择模型的方法须加以检验。参照“未知数据法”的检验方法^[5],选取早期数据作为建模数据,选取最新的数据作为“未知数据”,由此建立模型、做预测,将预测值与“未知数据”做比较,求预测误差。若各模型的预测误差变化顺序与相应的 K 值变化顺序一致,表明以 K 值最小选择用水预测模型的方法是合理的。

这里仍采用运城市工业用水量系列资料,求取 5 年和 10 年两个预测期的预测误差,为了消除偶然性,对于每种预测期连续求出 5 年的预测误差及其平均值(MAPE)。该平均值越小,表明模型的预测精度越高。现以 5 年预测期为例说明其计算过程。

首先选取 1985 年~2002 年系列资料建立模型,可求得上述六种模型的参数,相应地可求出 2007 年用水量预测值和预测误差,如幂函数模型,

其预测值为 $\hat{y} = ax^b = 0.549\ 8 \times 23^{0.363\ 2} = 1.7171$,预测误差的绝对值为 $|y - \hat{y}| = |2.0161 - 1.717\ 1| = 0.299\ 0$,相对误差为 $0.299\ 0/2.016\ 1 \times 100 = 14.83\%$ 。同理,选取 1985 年~2001 年系列资料,建立模型,可求得 2006 年用水量预测值和预测误差,以此类推,可求得 2005 年、2004 年和 2003 年用水量预测值和预测误差,由此求得 5 年预测误差的平均值,见表 2。

由表 2 可见,模型预测误差绝对值的平均值和相对误差平均值(MAPE)依模型顺序变化是完全一致的,均以幂函数最小,直线次之,再次为 S 函数,以三次抛物线最大,分别为 0.134 6 亿 m^3 、0.162 5 亿 m^3 、0.307 6 亿 m^3 和 0.542 7 亿 m^3 ,相应的 MAPE 为 6.9%、9.3%、16.4% 和 29.6%。与表 1 中置信区间 K 值变化顺序基本一致。由此说明,以 K 值最小为标准选择用水预测模型的方法是合理的。

对 10 年预测期也进行了同样的分析,也以幂函数模型预测误差最小,次之为 S 函数,再次为直线,其预测误差绝对值的平均值分别为 0.109 0 亿 m^3 、0.297 9 亿 m^3 和 0.419 1 亿 m^3 ,相应的 MAPE 为 5.6%、15.8% 和 23.9%。预测误差最大为三次抛物线,其预测误差绝对值的平均值为 3.276 1 亿 m^3 ,相应的 MAPE 为 183.5%,且出现了预测值为负值的不合理情况。这一结果表明,幂函数用于中长期预测是较为合理的,次之为直线或 S 函数,与以 K 值最小为标准选择的预测模型基本一致。

表 2 模型预测误差分析(5 年预测期)

预测年份	建模系列	实测值/亿 m ³	模型、预测误差(实测值 - 预测值)及其绝对值的平均值/亿 m ³					
			幂函数	直线	S 函数	指数	二次抛物线	三次抛物线
2003	1985 ~ 1998	1.651 1	-0.012 7	-0.282 4	0.095 3	-0.723 0	0.073 6	-0.138 5
2004	1985 ~ 1999	1.716 7	0.043 4	-0.198 8	0.205 4	-0.607 0	0.300 1	0.463 5
2005	1985 ~ 2000	1.722 0	0.049 2	-0.163 3	0.245 2	-0.535 6	0.455 2	0.789 9
2006	1985 ~ 2001	1.955 2	0.268 5	0.070 5	0.488 0	-0.287 1	0.698 7	0.853 0
2007	1985 ~ 2002	2.016 1	0.299 0	0.097 3	0.504 3	-0.255 2	0.657 9	0.468 8
平均值			0.134 6	0.162 5	0.307 6	0.481 6	0.437 1	0.542 7
MAPE/%			6.9	9.3	16.4	27.5	23.3	29.6

由上述比较可知,拟合精度的高低与预测精度的高低是不一致的,预测的目的是希望预测值与实际值的差异最小,而不是模型与现有数据的拟合精度最高。因此,相对于用拟合精度高低来选取模型,以预测误差最小选取模型更为合理。据此,对于山西省运城市工业用水量的预测模型应选取幂函数,而不是以拟合精度最高为标准选择三次抛物线。

3 运城市工业用水量的预测及其预测精度

根据上述分析,应选用幂函数作为运城市工业用水量的预测模型,相应的模型参数为 $a = 0.531\ 1$, $b = 0.386\ 8$, $R = 0.974\ 8$ 。该模型符合国内外工业用水量变化趋势^[10],随着时间的增加,用水量增加幅度在减小。由此求得预测期为 5 年和 10 年的预测值分别是 1.927 3 亿 m³ 和 2.053 8 亿 m³,相应的置信区间 K 值(置信度为 95%)分别为 0.256 0 亿 m³ 和 0.277 0 亿 m³。若以置信区间 K 值与相应预测值的比值作为相对误差,可得出两个预测期的相对误差分别为 13.3% 和 13.5%,均小于 15%,预测是成功的。

4 结论

(1) 以预测值的置信区间最小为依据选取预测模型,要好于以拟合误差最小选取预测模型的常规

方法。该方法具有较严密的理论基础,据此选出的预测模型与用实测值分析结果一致;

(2) 对于运城市工业用水量预测以幂函数最好,其预测期为 5 年和 10 年的预测值相对误差较为接近,均不超过 15%。

参 考 文 献

1 王海锋,贺 骥,庞靖鹏.需水预测方法及存在问题研究.水利发展研究,2009;(3):19—22,24

2 李红艳,崔建国,张星全.城市用水量预测模型的优选研究.中国给水排水,2004;20(2):41—43

3 张雅君,刘全胜.需水量预测方法的评析与择优.中国给水排水,2001;17(7):27—29

4 舒诗湖,向 高,何文杰.灰色模型在城市中长期用水量预测中的应用.哈尔滨工业大学学报,2009;41(2):85—87

5 王自勇,王 圃.组合模型在城市用水量预测中的应用.中国给水排水,2008;24(12):37—39

6 汤成友,官学文,张世明.现代中长期水文预报方法及其应用.北京:中国水利水电出版社,2008

7 赵 静,但 琦.数学建模与数学实验.北京:高等教育出版社,2008

8 郑爽英.城市供水量的预测模型研究.成都科技大学学报,1995;(6):19—26

9 袁志发,宋世德.多元统计分析.北京:科学出版社,2009

10 鲁 欣,秦大庸,胡晓寒.国内外工业用水状况比较分析.水利水电技术,2009;40(1):102—105

Research on Selection of Water Consumption Forecast Model

WANG Yuan, FENG Qian *

(College of Environmental Science and Engineering, Hohai University, Nanjing 210098, P. R. China)

[Abstract] The time series method is one of common methods for forecasting water consumption. The prediction accuracy on water consumption can be guaranteed by the selection of forecast models. View of the current selection of forecasting models in order to minimize the fitting error, a method of selecting forecast model was put forward based on minimum error of forecast. Taking the industrial water consumption data of Yuncheng city for example, the prediction interval of forecast period of 5 and 10 years were given in the 95% confidence level for six models such as power function, S function, linear, exponential functions, quadratic and cubic parabola. The results show that power function has the highest prediction accuracy. The series amount of water data in the nearly 5 years was predicted. Power function has the minimum prediction error. The results show that the method of choosing models according to the minimize prediction range is reasonable and reliable.

[Key words] water consumption model selection prediction period prediction model.

(上接第 4754 页)

- 7 Vautz W, Sielemann S. Determination of terpenes in humid ambient air using ultraviolet ion mobility spectrometry, *Aanalytica Chimica Acta*, 2004; 513(2):393—399
- 8 Siems W F, Wu Ching, Tarver E E, *et al.* Measuring the resolving power of Ion mobility spectrometers, *Anal Chem*, 1994; 66: 4195—4201
- 9 Baim M A, Hill Jr H H. Resolution measurement for ion mobility spectrometry. *Anal Chem*, 1985; 57:1902—1907
- 10 Wu Ching, Siems W F, Asbury G R, *et al.* Electrospray ionization high - resolution ion mobility spectrometry - mass spectrometry. *Anal Chem*, 1998; 70 : 4929—4938
- 11 Vlasak P R, Beussman D J. An interleaved comb ion deflection gate for m/z selection in time-of-flight mass spectrometry. *Rev Sci Instrum*, 1996; 67(1):68—72

Research of the Uniformity of the Electric Field in Ion Mobility Spectrometer Drift Tube

SHI Ying-guo, DU Ming-juan

(Liaoning Shihua University, Fushun 113001, P. R. China)

[Abstract] Based on theoretical analysis and computer simulation, we made a ionization ion mobility spectrometer which uses a vacuum ultraviolet light lamp to ionizing source. Detailed studied the space uniformity of electric field in self made in ion mobility spectrometer drift tube. When the inner diameter of the drift tube is about 30 mm, the outer diameter of the drift tube is about 50 mm, the thickness of the metal ring is 1 mm, and the thickness of the PTFE ring is 1 mm could guarantee the space uniformity of electric field in the ion mobility spectrometer drift tube.

[Key words] ion mobility spectrometry drift tube VUV—ionization acetone