

主成分回归在需水预测中的应用

王春娟¹, 冯利华¹, 罗伟², 庞小笑¹

(1. 浙江师范大学 地理与环境科学学院, 浙江 金华 321004;

2. 江西省庐山自然保护区管理处, 江西 庐山 332900)

摘要: 城市需水量的准确预测对区域的发展具有十分重要的意义。城市需水受多重因素的影响,且这些因素大多存在较强的相关性。通过主成分分析法的计算分析,以金华市为例,采用2000-2010年的城市用水资料建立回归模型对需水量进行预测。结果表明:该模型应用于城市用水预测,其结果与当地实际情况较为吻合,模型的拟合程度和预测准确度较好。

关键词: 主成分分析; 回归模型; 需水预测

中图分类号: TU997.31

文献标识码: A

文章编号: 1672-643X(2013)01-0050-04

Application of principal component regression model to water demand forecast

WANG Chunjuan¹, FENG Lihua¹, LUO Wei², PANG Xiaoxiao¹

(College of Geography and Environmental Sciences, Zhejiang Normal University, Jinhua 321004, China;

2. Nature Reserve Management Office of Lushan of Jiangxi Province, Lushan 332900, China)

Abstract: The accurate prediction of water demand has great significance for the development of city. The water demand is affected by multiple factors and most of these factors have strong relevance. Based on the calculation steps of the principal component analysis, and introduced an instance (water consumption data of Jinhua city from 2000 to 2010), the regression model of predicting water demand was established. The results showed that the model was applied to forecast water demand, and the conclusion agreed quite well with the local actual conditions, the model fitting degree and prediction accuracy were ideal.

Key words: principal component analysis; regression model; water demand forecasting

由于城市的高速发展,对水资源的需求急剧增长,同时,由于供水工程建设的滞后与水环境的污染,我国许多城市都出现了严重的缺水现象。水的短缺已经成为这些城市经济、社会发展的重要制约因素。因此,为了满足城市对水资源的需求,科学准确的需水预测是一项重要的研究课题,它可以为供水决策、水利投资、水资源规划和管理提供必要的依据^[1]。需水预测是以经济发展特征和用水现状的研究为基础,重点研究工业发展趋势与产业结构的调整、人民生活水平提高的速度与程度等。目前,需水预测方法有人均综合用水量推算法、神经网络法、用水定额法、时间序列预测、回归分析预测、指数法预测及灰色预测等。预测水资源需求量,不仅要考虑经济结构、工农业生产布局、城乡居民生活水平等

诸多因素,同时还要考虑水资源开发条件约束和节约用水等因素。单一的方法很难保证预测的科学性和精确度要求,因此,本文采用组合预测方法。需水预测总量采用主成分分析法与回归方法的结合,并与多元回归分析法做对比分析。

1 金华市水资源现状

金华市位于浙江西山区金衢盆地东段,地势南北高、中部低,属于中亚热带季风气候区,四季分明,年温适中,热量丰富,雨量较多,有明显的干、湿两季。其中多年平均降水量为1 512.9 mm,多年平均水资源利用率为26.59%。

市域内江河分属钱塘江、瓯江、曹娥江、椒江3大水系,流域面积分别为9 332.73、949.71、341.62

收稿日期:2012-09-19; 修回日期:2012-10-14

基金项目:国家自然科学基金项目(41171430、40771044)

作者简介:王春娟(1984-),女,陕西宝鸡人,硕士研究生,研究方向为水文学与水资源、资源地理。

通讯作者:冯利华(1955-),男,浙江建德人,教授,主要从事水文学与水资源的教学与研究工作。

和 293.96 km², 分别占全市江河总面积的 85.49%、8.69%、3.13% 和 2.69%。集雨面积在 100 km² 以上的江溪有 40 多条。

由于经济快速发展, 金华市区生活及工业用水, 从 2000 年开始, 其市辖区义乌市出资购买东阳市横锦水库的淡水资源永久性使用权, 这也成为了即“水权与水市场”理论提出之后的首个水权交易^[2]。

据《2010 年金华市水资源公报》的统计数据显示, 全市水资源总量 152.2162 亿 m³ (其中地下水资源量 24.0160 亿 m³), 比常年值偏多 65.9%。全市平均降水量 2 053.3 mm, 折合降水总量 224.1757 亿 m³, 比常年值 (多年平均值) 偏多了 35.7%。全市总供水量 19.4407 亿 m³, 其中地表水源占 91.8%, 地下水源占 8.0%, 其他水源供水量占 0.2%。金兰水库向金华市区供水 0.6174 亿 m³。全市总用水量 19.4407 亿 m³, 全市用水消耗量 11.3954 亿 m³, 占总用水量的 58.6%。全市大中型水库蓄水总量 6.9992 亿 m³, 比上年末增加 1.2981 亿 m³。全市 660.46 km 河流的水质评价结果为 II 类水占 21.8%, III 类水占 32.8%, IV 类水占 12.3%, V 类水占 11.4%, 劣于 V 类水占 21.7%。

若从水资源总量与水资源开发利用率来看, 金华市并不缺水, 但由于降水时空分布不均、水资源调蓄能力不足, 存在水质性和区域性缺水, 并且《金华市国民经济和社会发展第十二个五年规划纲要 (2011-2015 年)》中指出, 到 2015 年, 全市 GDP 将达 3370 亿元, 年均增长率为 10%, 人均 GDP 将达 70560 元, 年均增长率为 9.5%, 人口自然增长率将高达 5.78‰, 如此发展趋势, 为水资源规划、管理部门敲响警钟, 水资源问题将成为制约金华市可持续发展的新“瓶颈”。

2 分析方法

2.1 多重共线性

目前, 在诊断自变量系统中是否存在多重共线性时, 常采用方差膨胀因子 (VIF) 诊断法^[3]。其表达式为: $F_{VIF_i} = (1 - R_i^2)^{-1}$, 其中 R_i 为自变量对其余自变量 X_i 进行回归分析的复相关系数, 所有 X 变量中最大的 F_{VIF_i} 通常用来作为多重共线性严重程度的指标, 并且还度量回归系数的估计方差与自变量线性无关时相比增加了多少。若 $F_{VIF_i} > 10$, 表示多重相关性将严重影响最小二乘估计值。

2.2 主成分分析法

主成分分析是把多个变量化为少数几个综合变

量的一种多元统计分析方法, 其主要功能是简化统计分析的数据结构。在多元变量研究中, 变量之间往往存在着一定的相关性, 使得观测数据的信息有所重叠。主成分分析采取降维方法, 通过确定少数几个综合因子即主成分来代表原来众多的变量, 使这些主成分能尽可能地反映原来的变量的信息, 而且彼此之间互不相关^[4]。设原始变量为 $X_1, X_2, \dots, X_p$, 主成分分析后得到的新变量 (综合变量) 为 $Z_1, Z_2, \dots, Z_m$, 它们是 $X_1, X_2, \dots, X_p$ 的线性组合 ($m \leq p$)。新变量 $Z_1, Z_2, \dots, Z_m$ 构成的坐标系是在原坐标系经平移和正交旋转后得到的, 称 $Z_1, Z_2, \dots, Z_p$ 空间为 m 维主超平面。在主超平面上, 第一主成分 Z_1 对应于数据变异 (贡献率 e_1) 最大的方向。对于 $Z_2, \dots, Z_m$ 依次有 $e_2 \geq e_3 \geq e_4 \geq e_m$ 。因此, Z_1 是携带原始数据信息最多的一维变量, 而 m 维主超平面是保留原始数据信息量最大的 m 维子空间。

主成分分析法的主要步骤如下^[5-6]:

(1) 为了排除数量级和量纲带来的误差, 对数据的标准化处理。公式为:

$$Y_{ij} = \frac{X_{ij} - \bar{X}_j}{S_j} \quad (i = 1, 2, \dots, I; j = 1, 2, \dots, J)$$

式中: X_{ij} 为第 i 个分区中第 j 个指标的值; $\bar{X}_j, S_j$ 分别为第 j 个指标的样本均值和样本标准差。

(2) 计算 $(Y_{ij}) I \times J$ 相关矩阵 R 。

(3) 计算 R 的 J 个特征值 $\lambda_1 \geq \lambda_2 \geq \dots, \lambda_J$, 以及对应的特征向量 $U_1, U_2, \dots, U_J$, 它们是标准正交, $U_1, U_2, \dots, U_J$ 称为主轴。

(4) 计算贡献率 $e_k = \lambda_k / \sum_{j=1}^J \lambda_j$ 和累计贡献率

$$E = \sum_{k=1}^m \lambda_k / \sum_{j=1}^J \lambda_j。$$

(5) 求主成分: $Z_k = \sum_{j=1}^J U_{kj} X_{kj} \quad (k = 1, 2, \dots, m; k = 1, 2, \dots, J)$ 。式中: U_{kj} 为特征向量 U_k 的第 j 个分量。

(6) 综合分析: 一个 m 维主超平面究竟以多大的精度来近似代替原变量系统以确保尽可能多的数据信息? 可以通过求累计贡献率 E 来判断 $E = \sum_{k=1}^m \lambda_k / \sum_{j=1}^J \lambda_j$, 如取 $E > 80\%$ 时最小 m , 则可得主超平面的维数 m , 从而对 m 个主成分进行综合分析。

3 实例分析

3.1 模型的建立

本文以金华市为例, 采用主成分回归分析法, 来实现对需水量的预测。首先以金华市 2000-2010

年的序列资料(表 1),从中选取 8 个因子, X_1 : GDP,万元; X_2 : 人均 GDP, 10^3 元; X_3 : 固定资产投资,万元; X_4 : 工业个数,个; X_5 : 人口,万人; X_6 : 日供水能力,万 m^3 ; X_7 : 居民生活用水量,亿 m^3 ; X_8 : 污水年排放量, 10^6 t; Y : 需水量,亿 m^3 。从中挑选

出的这些因子相互独立性较差,存在有较多的交叉重叠信息,因此可采用主成分分析法来提取主要影响因素,最终得到较少的综合指标,尽可能的反映原来的影响因素的信息,在通过利用回归分析法,实现对需水量的预测。

表 1 金华市 2000 - 2010 年用水量及其影响因子标准化后的数据

年份	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	Y
2000	-1.2025	-1.2138	-1.7954	-1.4023	-1.2819	1.2206	0.2350	0.7187	1.1819
2001	-1.1053	-1.1109	-1.5130	-1.3575	-1.0596	1.2690	0.6719	0.7714	1.1663
2002	-0.9464	-0.9394	-0.7564	-0.8779	-0.8767	2.0116	1.8440	1.8653	2.0954
2003	-0.7145	-0.6874	-0.1975	-0.6742	-0.7714	-0.4612	-0.0295	0.4182	-0.5897
2004	-0.3763	-0.3208	0.3065	0.1780	-0.5096	-0.8347	-1.3818	-0.0942	-0.8873
2005	-0.2134	-0.1524	0.2572	-0.1658	-0.1541	-0.9001	-1.4296	-0.4260	-0.8760
2006	0.1140	0.1909	0.2547	0.2714	0.2364	-0.6184	-0.7720	-0.7406	-0.5815
2007	0.5560	0.3760	0.4170	0.7149	0.5860	-0.5713	-0.4406	0.1342	-0.5324
2008	0.9695	0.7310	0.6365	0.7407	0.9108	-0.3795	-0.2893	0.2863	-0.3318
2009	1.1304	1.2391	0.8667	1.1791	1.2428	-0.2962	0.7342	-1.2508	-0.2505
2010	1.7887	1.8878	1.5236	1.5538	1.6773	-0.4396	0.8557	-1.6826	-0.3945

注:本文基础数据来源于《金华市统计年鉴》(2001 - 2011 年),《金华市水资源公报》(2000 - 2010 年)。

3.2 多重共线性诊断

根据表 1 的数据,通过统计分析软件 SPSS 进行回归分析的多重共线性诊断,各个影响因素的方差膨胀因子如表 2 所示。因此可以得出, X_1 、 X_3 、 X_4 、 X_5 、 X_6 间均存在严重的多重共线性,只有 X_7 、 X_8 的共线性较小,另外, X_2 与其他因子存在很严重的多重共线性,在 SPSS 软件里操作时,系统已自动排除,为了保证下一步计算的准确性,故将 X_2 、 X_7 、 X_8 删除掉。

表 2 方差膨胀因子

变量	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
F_{VIF}	133.666	10328.6	24.609	67.407	141.720	20.312	8.3314	1.130

3.3 检验是否适合提取主成分

KMO(Kaiser - Meyer - Olkin) (表 3) 检验统计量是用于比较变量间简单相关系数和偏相关系数的指标。KMO 统计量是取值在 0 ~ 1 之间,当所有变量间的简单相关系数平方和远远大于偏相关系数平方和时,KMO 值接近 1。KMO 值越接近于 1,变量间的相关性就越强原有变量就越适合作因子分析;当所有变量间的简单相关系数平方和接近 0 时,KMO 值就接近 0,变量间的相关性就越弱,原有变量就越不适合作因子分析。

通过 SPSS 软件操作可得,KMO 数值为 0.822,所以适合做主成分分析。

同时 Bartlett 球形检验得出的相伴概率为

0.00001,远小于显著性水平 0.05,因此。原变量之间存在相关性,适合进行主成分分析。

3.4 主成分个数确定

通过 SPSS 分析软件得到的表 4(相关系数矩阵)、表 5(主成分特征统计信息)、表 6(主成分载荷矩阵),最终确定 2 个主成分。

表 3 Kaiser 给出的常用的 KMO 度量标准

KMO 值	0.9	0.8	0.7	0.6	0.5
变量是否适合做主成分分析	非常 适合	适合	一般	不太 适合	极不 适合

表 4 相关系数矩阵

影响因子	X_1	X_3	X_4	X_5	X_6
X_1	1.000	0	0	0	0.030
X_3	0.910	1.000	0	0	0.003
X_4	0.979	0.956	1.000	0	0.013
X_5	0.996	0.904	0.977	1.000	0.034
X_6	-0.582	-0.759	-0.664	-0.569	1.000

注:左上角为相关系数,右上角为显著性水平。

表 5 特征值和主成分贡献率及累计贡献率 %

主成分	特征值	贡献率	累计贡献率
Z_1	4.355	87.098	87.098
Z_2	0.563	11.252	98.350
Z_3	0.067	1.333	99.683
Z_4	0.012	0.238	99.921
Z_5	0.004	0.079	100.000

表 6 主成分的载荷矩阵

影响因子	Z_1	Z_2	影响因子	Z_1	Z_2
X_1	0.969	0.227	X_5	0.965	0.243
X_3	0.974	-0.073	X_6	-0.748	0.659
X_4	0.989	0.111			

第一主成分反映的是金华市国民经济发展状况,包含了总信息量的 87.098%;第二主成分反映的是水资源供应量状况,包含了 12.902%。这 2 个主成分包含了原来 5 个影响因素的信息 98.350%,所以可代替原来的 5 个影响因素。

3.5 主成分回归分析

在多重共线性诊断、主成分分析后,将 2 个主成分代替原来的自变量进行多重回归分析,得到标准化自变量与应变量间的回归模型为: $Y = 20.643 + 2.004Z_1 - 4.838Z_2$ ($n = 11, r = 0.998$) 此回归模型是较为准确的,因此可以用第一主成分、第二主成分的因子得分来预测金华市需水量,如表 7 所示,建立的主成分回归方程,各个回归系数均通过了 t 检验,达到了显著水平。而从表 8 可以看出,使用所有的变量建立的城市需水量多元线性回归方程为: $Y = -41.519 - 0.002X_1 + 0.0001X_3 + 0.0001X_4 + 0.093X_5 + 0.037X_6$ ($n = 11, r = 0.999$),其中拟和效果较主成分回归分析较为良好。但是回归系数均为通过 t 检验,与主成分回归方程的回归系数相比较,显著性水平较差。

为了能够更直观的看出金华市需水预测情况,现将主成分回归与多元线性回归模型预测结果作成对比趋势图(图 1)。

表 7 主成分回归方程系数值

变量	回归系数	标准误差	t 值	显著水平
b_0	20.643	0.151	136.919	0.000
b_1	2.004	0.242	8.267	0.000
b_2	-4.838	0.366	-13.222	0.000

表 8 多元线性回归方程系数值

变量	回归系数	标准误差	t 值	显著水平
b_0	-41.5190	41.4550	-1.002	0.363
b_1	-0.0020	0.0010	-1.252	0.266
b_2	0.0001	0.0010	0.264	0.802
b_3	0.0001	0.0001	0.728	0.499
b_4	0.0930	0.0940	0.984	0.370
b_5	0.0370	0.0010	33.832	0.000

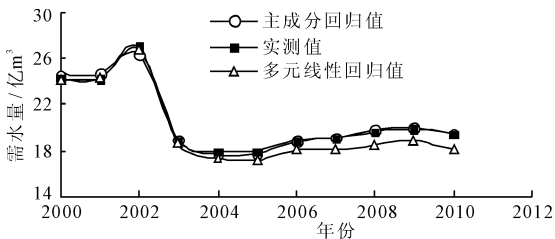


图 1 模型预测结果对比趋势图

4 结 语

本文通过主成分回归分析法预测了金华市需水量,得出以下结论:

- (1)通过主成分回归分析法建立了需水预测模型,并对金华市需求量进行了预测,结果显示,主成分回归值与实际值间的误差较小,回归系数达到了极显著水平,使所建立的模型具有较高的预测精度,从而弥补了最小二乘法不能有效识别及消除变量间多重相关性影响的不足,体现出所建模型的准确度。
- (2)通过主成分分析法消除多重共线性问题,所得结果合理、可靠。以金华市为例进行分析得到,因子自变量间的相关性有时对模型回归拟合效果影响不太明显,但对于模型的回归系数、分量分解的影响较为突出,所以,在需水量资料分析中,不能仅仅只注重模型的拟合效果,更应该考虑对回归系数估计的合理性。
- (3)利用主成分的互不相关性建立应变量与主成分的回归,从理论上而言,消除了多重共线性,实现了变量的降维。在数据无误差的情况下,主成分回归分析法较为准确的预测需水量,此外,通过所建模型的各影响因素系数,根据其大小来直观的了解各影响因子对需水量的影响程度,以便为水资源规划、管理部门提供可靠、准确的依据。

参考文献:

[1] 史经海. 城市需水预测模型研究[J]. 黑龙江水利科技, 2011,39(3):56-58.

[2] 冯利华,王天阳,张建珍,等. 金华市的水资源核算与绿色利用新思路[J]. 热带地理,2006,26(2):162-166.

[3] 刘国旗. 多重相关性的产生原因及其诊断处理[J]. 合肥工业大学学报(自然科学版),2001,24(4):607-610.

[4] 张旭明. 产业集群持续成长因素分析与实证研究[D]. 吉林:吉林大学,2008.

[5] 傅湘,纪昌明. 区域水资源承载能力综合评价:主成分分析法的应用[J]. 长江流域资源与环境,1999,8(2):168-172.

[6] 冯利华,赵浩兴,瞿有甜. 灾害等级的综合评价[J]. 灾害学,2002,17(4):16-20.