DOI:10.20040/j.cnki.1000-7709.2023.20220550

水效率相关因素的偏最小二乘回归分析

李可柏,卢 慧,陶 军

(南京信息工程大学管理工程学院,江苏 南京 210044)

摘要:为分析用水效率的影响因素及其相关性,纳入所有自变量且克服多重共线性,采用偏最小二乘法建模,研究中国2019年的用水效率相关因素。结果显示,该方法有效描述了变量的相关关系。与万元GDP用水量相关性较高的因素为自然条件、社会发展水平、一产增加值的GDP占比和水行业劳动力比重。整体经济用水效率东高西低。与耕地实际亩均用水量相关性较高的因素为自然条件和一、二产业增加值的GDP占比。农业用水效率为中高西低。与万元工业增加值用水量相关性较高的因素为自然条件、污水处理率、城镇化率和水行业资本比重。工业用水效率东高中低。与人均公共用水量相关性较高的因素为人均水资源量、年降水量、污水处理率、人均GDP和一、二产业增加值的GDP占比。地区因素与公共用水效率的相关性小。研究成果可为水资源优化管理与合理利用提供参考。

关键词:产业用水;用水效率;相关分析;偏最小二乘回归

中图分类号:TV697.1^{+}4 **文献标志码**:A **文章编号**:1000-7709(2023)01-0034-04

1 引言

研究用水效率的影响因素对于水资源开发与利用有重要意义。目前,研究用水效率影响因素的方法有主成分分析、对数平均迪氏指数法、系统动力学法、偏最小二乘法等。为纳入所有自变量且克服多重共线性,拟采用偏最小二乘法建模,研究2019年中国31个省级行政区的用水效率影响因素,发现整体经济用水效率与自然因素、农业增加值占比、社会发展水平和水行业劳动力比重相关性较高,地区表现为东高西低;农业用水效率与自然因素和一、二产增加值占比相关性较高,地区表现为中高西低;工业用水效率与自然因素、社会发展水平和水行业资本比重相关性较高,地区表现为东高中低;人均公共用水量与自然因素、污水处理率、人均GDP和一、二产增加值占比相关性较高,与地区因素的相关性较小。

2 数据与方法

2.1 数据收集与指标处理

数据来自《中国统计年鉴》、《中国水资源公报》和各省级行政区的统计年鉴。因变量为4个,包括万元GDP用水量Y_1(m^3)、耕地实际灌溉亩均用水量Y_2(m^3)、万元工业增加值用水量Y_3(m^3)、人均公共用水量(含第三产业及建筑业等用水)Y_4(L/d),分别代表整体和三次产业用水效率。

自变量为15个,包括人均水资源量X_1(m^3)、常住人口城镇化率X_2(%)、人均GDPX_3(元)、工业增加值的GDP占比X_4(%)、农林牧渔业增加值的GDP占比X_5(%)、第三产业增加值的GDP占比X_6(%)、年平均气温X_7(℃)、年降水量X_8(mm)、污水处理率X_9(%)、水行业劳动力占工业劳动力比重X_{10}(%)、水行业资本占工业资本比重X_{11}(%)、水资源模数X_{12}(m^3/km^2)、东部X_{13}、中部X_{14}、西部X_{15}。其中,X_{13}、X_{14}、X_{15}为虚拟变量,每个模型中需同时使用以表示某地区。东部、中部、西部省份,X_{13}、X_{14}、X_{15}的取值分别为100、010、001。

自变量归为自然条件(X_1、X_7、X_8、X_{12})、社会发展水平(X_2、X_3)、产业增加值的GDP占比(X_4、X_5、X_6)、水行业发展水平(X_9、X_{10}、X_{11})、地区(X_{13}、X_{14}、X_{15})五类。

区域划分参照国家统计局标准,东部包括北京、天津、河北、辽宁、上海、江苏、浙江、福建、山

收稿日期:2022-03-22,**修回日期**:2022-04-25

基金项目:国家社会科学基金项目(17BGL220)

作者简介:李可柏(1974-),男,博士、副教授,研究方向为管理系统工程,E-mail:lzlkb@163.com

东、广东、海南;中部包括山西、吉林、黑龙江、安徽、江西、河南、湖北、湖南;西部包括内蒙古、广西、重庆、四川、贵州、云南、西藏、陕西、甘肃、青海、宁夏、新疆。

2.2 数据处理

2.2.1 数据标准化

将自变量矩阵 $\boldsymbol{X}$ 和因变量矩阵 $\boldsymbol{Y}$ 标准化:

$$\begin{cases} x_{ij}^{*} = (x_{ij} - \mu_j^{(1)})/s_{x_j} & i = 1,2,\cdots,n;j = 1,2,\cdots,p \\ y_{ik}^{*} = (y_{ik} - \mu_k^{(2)})/s_{y_k} & i = 1,2,\cdots,n;k = 1,2,\cdots,q \end{cases} \tag{1}$$

式中,x_{ij}、x_{ij}^{*} 分别为标准化前后自变量的值;y_{ik}、y_{ik}^{*} 分别为标准化前后因变量的值;$\mu_j^{(1)}$、$\mu_k^{(2)}$ 分别为 x_j、y_k 的均值;s_{x_j}、s_{y_k} 分别为 x_j、y_k 的标准差,得到标准差矩阵 $\boldsymbol{E}_0$、$\boldsymbol{F}_0$。

2.2.2 共线性检验

自变量的共线性检验见表 1。由表 1 可知,最小特征值接近 0,最大条件指标远大于 30,说明自变量存在严重共线性,不能用普通最小二乘法,偏最小二乘法则是较好选择。

表 1　自变量共线性检验

Tab. 1　Collinearity test of independent variables

模型维数	特征值	条件指标	模型维数	特征值	条件指标
1	10.565	1.000	9	0.061	13.150
2	1.307	2.843	10	0.031	18.409
3	0.960	3.317	11	0.025	20.528
4	0.729	3.807	12	0.011	30.432
5	0.585	4.249	13	0.002	68.750
6	0.356	5.449	14	0.001	132.072
7	0.242	6.607	15	0	267.305
8	0.124	9.229			

2.3 方法选择

偏最小二乘法是一种多因变量对多自变量的回归方法,从因、自变量集中提取成分 $\boldsymbol{t}_h$($h = 1, 2,\cdots;h \leqslant \boldsymbol{X}$ 的秩)且成分相互独立,提取的成分最大程度包含自变量信息,也能较好解释因变量;再用提取的成分对自变量、因变量作回归方程,通过交叉有效性,使结果满意为止[1]。

2.4 模型的构建

2.4.1 确定成分 $\boldsymbol{t}_h$ 个数

$\boldsymbol{t}_h$ 的交叉有效性定义为:

$$Q_h^2 = 1 - S_{\mathrm{PRESS},h}/S_{\mathrm{ss},h-1} \tag{2}$$

式中,Q_h^2 为交叉有效性系数,是评价拟合方程的预测能力的评价指标;$S_{\mathrm{PRESS},h}$ 为 $\boldsymbol{Y}$ 的预测误差平方和;$S_{\mathrm{ss},h-1}$ 为用全部样本点拟合的具有 $h-1$ 个成分的方差拟合误差。

$S_{\mathrm{PRESS},h}/S_{\mathrm{ss},h-1}$ 的比值越小越好,增加成分 $\boldsymbol{t}_h$ 有益[2]。结果显示第 4 个成分的 Q_h^2 为 0.016 734 4,第 5 个成分 Q_h^2 为负值,且 4 个主成分对自变量和因变量信息的累计贡献率分别为 72.4%、54.6%,因此提取 4 个成分并认为其能较好解释该模型。

2.4.2 成分 $\boldsymbol{t}_h$ 的提取

$\boldsymbol{t}_h$ 的表达式为:

$$\boldsymbol{t}_h = \boldsymbol{E}_0 \boldsymbol{w}_h^{*} \tag{3}$$

其中　$\boldsymbol{w}_h^{*} = \prod_{j=1}^{h-1} (\boldsymbol{I} - \boldsymbol{w}_j \boldsymbol{p}_j^{\mathrm{T}})\boldsymbol{w}_h$;$\boldsymbol{p}_j = \dfrac{\boldsymbol{E}_{j-1}^{\mathrm{T}} \boldsymbol{t}_j}{\| \boldsymbol{t}_j \|^2}$

式中,$\boldsymbol{w}_h$ 为矩阵 $\boldsymbol{E}_{h-1}^{\mathrm{T}} \boldsymbol{F}_0 \boldsymbol{F}_0^{\mathrm{T}} \boldsymbol{E}_{h-1}$ 最大特征值对应的特征向量;$\boldsymbol{E}_i$ 为与 $\boldsymbol{t}_i$ 同阶的残差向量,且对任意 i,$\boldsymbol{t}_i^{\mathrm{T}} \boldsymbol{E}_i = 0$;$\boldsymbol{I}$ 为单位矩阵。

2.4.3 构建回归方程

成分提取后,建立回归方程:

$$\boldsymbol{F}_{h-1} = \boldsymbol{t}_h \boldsymbol{r}_h^{\mathrm{T}} + \boldsymbol{F}_h \tag{4}$$

其中　$\boldsymbol{r}_h = \boldsymbol{F}_{h-1}^{\mathrm{T}} \boldsymbol{t}_h / \| \boldsymbol{t}_h \|^2$

式中,$\boldsymbol{r}_h$ 为回归系数。

四个成分 t_1、t_2、t_3、t_4 建立的偏最小二乘回归模型为:

$$\begin{aligned} y_k^{*} = {} & r_{1k}(w_{11}^{*}x_1^{*} + w_{12}^{*}x_2^{*} + \cdots + w_{1,15}^{*}x_{15}^{*}) + \\ & r_{2k}(w_{21}^{*}x_1^{*} + w_{22}^{*}x_2^{*} + \cdots + w_{2,15}^{*}x_{15}^{*}) + \\ & r_{3k}(w_{31}^{*}x_1^{*} + w_{32}^{*}x_2^{*} + \cdots + w_{3,15}^{*}x_{15}^{*}) + \\ & r_{4k}(w_{41}^{*}x_1^{*} + w_{42}^{*}x_2^{*} + \cdots + w_{4,15}^{*}x_{15}^{*}) \end{aligned} \tag{5}$$

标准化指标变量系数见表 2。

表 2　标准化指标变量系数

Tab. 2　Standardized index variable coefficient

自变量	万元 GDP 用水量 Y_1/ m^3	耕地实际灌溉亩均用水量 Y_2/m^3	万元工业增加值用水量 Y_3/m^3	人均公共用水量 Y_4 /m^3
人均水资源量 X_1/m^3	0.021	−0.001	0.250	0.381
常住人口城镇化率 X_2/%	−0.094	−0.050	−0.140	−0.093
人均 GDP X_3/元	−0.115	−0.044	0.016	0.128
工业增加值的 GDP 占比 X_4/%	−0.042	−0.105	−0.095	−0.176
农林牧渔业增加值的 GDP 占比 X_5/%	0.113	0.218	0.022	−0.122
第三产业增加值的 GDP 占比 X_6/%	−0.048	−0.043	−0.057	0.065
年平均气温 X_7/℃	−0.113	0.231	0.125	0.099
年降水量 X_8/mm	−0.107	0.319	0.212	0.147
污水处理率 X_9/%	−0.025	−0.091	−0.210	−0.337
水行业劳动力占工业劳动力比重 X_{10}/%	0.076	0.074	0.008	0.054
水行业资本占工业资本比重 X_{11}/%	−0.018	−0.002	−0.131	−0.083
水资源模数 X_{12}/($m^3 \cdot km^{-2}$)	−0.045	0.057	0.077	0.066
东部 X_{13}	−0.085	0.025	−0.116	−0.007
中部 X_{14}	−0.005	−0.130	0.143	0.045
西部 X_{15}	0.088	0.092	−0.014	−0.033

3　结果与分析

从建立的回归模型来看,各用水效率的相关指标相同,均为 15 个。这是由偏最小二乘法的特点决定,该法能在克服多重共线性的情况下将所有自变量纳入模型。但显然 Y_1、Y_2、Y_3、Y_4 应有

各自的影响因素。对此通过模型的标准化指标变量系数来体现。同一自变量在不同因变量中的此系数值不同。如某自变量在一个因变量中的此系数绝对值较大，则该自变量是此因变量的重要影响因素。如同一个自变量在另一因变量中的此系数绝对值较小，该自变量则是此因变量的轻微影响因素。

3.1 万元 GDP 用水量 Y_1

由表 2 可知，系数绝对值大于 0.07 的指标与 Y_1 相关性较高，包括人均 GDPX_3、农林牧渔业增加值占比 X_5、年平均气温 X_7、年降水量 X_8、城镇化率 X_2 和水行业劳动力占比 X_{10}。其中，X_3 为负向指标，表明随着经济和社会发展水平上升，整体经济用水效率会提高。相比中国北方，南方年平均气温和年降水量更高，降水量是北方的近 4 倍，地表水资源更丰富。X_7、X_8 均为负向指标，显示中国南方整体经济用水效率高于北方整体经济用水效率。X_5 为正向指标，表明农业占比高的地区，整体经济用水效率相对较低。X_{10} 为正向指标，用水效率较高地区的水行业劳动力的生产效率也较高。如黑龙江、海南和西藏的水行业劳动力占比较高，均在 2%以上，这些地区的万元 GDP 用水量均超过 60.8 m^3 的全国水平；而上海、福建和广东的水行业劳动力占比均在 0.5%以下，万元 GDP 用水量则都低于 42 m^3，远低于全国水平。

地区因素方面，用水总量中部最大，东部次之，西部最小。Y_1 与东部负相关，与西部正相关。整体经济用水效率表现为东部高而西部低。东部是中国的发达地区，生产技术更先进，水资源利用效率相对较高。而西部的经济和产业发展水平处于相对劣势，水利基础设施和节水技术也相对滞后，导致总用水量和整体经济用水效率较低。

3.2 耕地实际灌溉亩均用水量 Y_2

表 2 中系数绝对值大于 0.1 的指标与 Y_2 相关性较高，包括农林牧渔业增加值占比 X_5、年平均气温 X_7、年降水量 X_8 和工业增加值占比 X_4。其中，X_5 为正向指标，表明农业占比高的地区农业用水效率相对较低。部分可能原因在于中国农业有很大一部分还未实现规模化，个体种植农户较多，农田水利建设并不完善，粗放式的农田漫溉方式导致农业水资源利用效率较低。X_7、X_8 为正向指标。一方面，气温升高导致农作物的水分蒸腾作用增强，需补充更多的农田灌溉用水；另一方面，降水量更高的地区，可用于灌溉的农业用水随之增加，可能导致这些地区不注重农业节水技术的推广使用，从而提高了亩均用水量。X_4 为负向指标，说明工业化程度增加有助于提高农业用水效率。如山西、浙江、陕西的工业增加值占比达到 36%以上，均高于 31.9%的全国平均水平，但其亩均用水量均低于全国平均水平。

地区因素方面，农业用水量中部最大，西部次之，东部最小。中部为我国重要商品粮生产基地，主要包括江汉、松嫩、洞庭湖和鄱阳湖平原，所在省份农业用水量较大。Y_2 与中部负相关，与西部正相关。农业用水效率表现为中部高西部低。相比中部，西部中的西北地区较为干旱，农作物的水分蒸腾作用强；而西南地区虽然水资源相对丰富，但农业节水技术应用不足，农民节水意识较为薄弱，导致西部农业水资源利用效率较低。

3.3 万元工业增加值用水量 Y_3

认为表 2 中系数绝对值大于 0.1 的指标与 Y_3 相关性较高，包括人均水资源量 X_1、年降水量 X_8、污水处理率 X_9、城镇化率 X_2、年平均气温 X_7 和水行业资本占比 X_{11}。其中，X_1、X_8、X_7 均为正向指标，表明用水条件较差的地区会更重视提高工业用水的使用效率，因此丰水地区应增强工业节水意识的培养和节水技术的开发使用。X_9 为负向指标，说明污水处理对于提升工业用水效率效果明显。我国污水处理率普遍在 90%以上，特别是河北和山东两省高于 97%，其万元工业增加值用水量仅约为全国平均水平的 1/3；而西藏和海南的污水处理率低于 75%，其万元工业增加值用水量则均高于全国平均水平。目前，中国工业污水排放量超过印度、美国和欧盟总和[3]。因此，若尽可能将工业污水净化处理循环利用，可大为减少工业新水用水量。X_2 为负向指标，表明随着社会发展程度的提高，往往工业化水平也越高，工业用水效率也随之提高。X_{11} 为负向指标，事实上，水行业是一个典型的资本密集型行业[4]，对于该行业的运营发展非常重要。

地区因素方面，工业用水量东部最大，中部次之，西部最小，西部工业用水量仅为东部的 1/3 左右。Y_3 与东部负相关，与中部正相关。工业用水效率表现为东部高而中部低。东部具备技术优势和人才储备，拥有良好的产业结构。相对而言，中部的技术和人才优势不如东部，且中部崛起战略的实施，使得众多企业向中部转移，形成火电、钢铁、石油石化等高耗水行业集聚群，工业用水结构得不到优化。另外，中部的人均教育支出低于东部地区，这可能也是影响中部地区工业用水效率较低的原因之一[5,6]。

3.4 人均公共用水量 Y_4

表 2 中系数绝对值大于 0.1 的指标与 Y_4 相关性较高,包括人均水资源量 X_1、污水处理率 X_9、人均 GDPX_3、工业增加值的 GDP 占比 X_4、农林牧渔业增加值的 GDP 占比 X_5 和年降水量 X_8。其中,X_1、X_8 均为正向指标,表明用水条件较好地区的人均公共用水量也更多。X_9 为负向指标,说明提高污水处理率,有助于提升公共用水效率。X_3 为正向指标,表明随着经济和社会的整体发展,一、二产的比重下降,三产的比重上升,造成人均公共用水量随之上升。近十年来,我国一、二产业比重分别下降 3.2%、7.3%,第三产业比重则上升 10.5%,人均公共用水量也从 2009 年的 52.6 L/d 上升至 2019 年的 86.0 L/d。反之,X_4、X_5 均为负向指标,说明一、二产的比重上升,三产的比重下降,导致人均公共用水量减少。尽管第三产业用水量上升,但其对于水资源的压力还较小,污染较少,且产值较高[7]。因此,在环境承载范围内,应促进三产的可持续发展。但在发展三产的同时,也要注重加大污水处理效率,提升用水效率。

地区因素方面,生活用水量东部最大,中部次之,西部最小,西部生活用水量约为东部的 1/2。三个地区的标准化指标变量系数的绝对值均较小,解释能力均较低,即地区因素与 Y_4 相关性不大,表明三个地区的人均公共用水需求具有一致性趋势。

以上结果显示,对于用水量与用水效率的关系,东、中、西部的整体经济和各产业均不是一致的正向或反向关系。

4 结论

a. 自然因素与整体经济用水效率、农业用水效率、工业用水效率和人均公共用水量均相关。

b. 社会因素中整体用水与农业增加值占比和社会、水行业水平相关;农业用水与一、二产增加值占比相关;工业用水与社会、水行业水平相关;公共用水与社会、水行业水平和一、二产增加值占比相关。

c. 整体用水效率东高西低;农业用水效率中高西低;工业用水效率东高中低;公共用水效率与地区相关性小。

参考文献:

[1] 王文圣,丁晶,赵玉龙,等. 基于偏最小二乘回归的年用电量预测研究 [J]. 中国电机工程学报,2003,23(10):17-21.

[2] 王惠文. 偏最小二乘回归方法及其应用 [M]. 北京:国防工业出版社,1999.

[3] 雷英. 工业污水治理中常见问题 [J]. 中国科技信息,2021(13):63-64.

[4] MURPHY M. 2018 state of the water industry: The challenge of building resilience [J]. Journal of the American water works association, 2018, 110(8): 61-71.

[5] 李珊,张玲玲,丁雪丽,等. 中国各省区工业用水效率影响因素的空间分异 [J]. 长江流域资源与环境,2019,28(11):2539-2552.

[6] 雷玉桃,黄丽萍. 中国工业用水效率及其影响因素的区域差异研究——基于 SFA 的省际面板数据 [J]. 中国软科学,2015(4):155-164.

[7] 刘慧敏,周戎星,于艳青,等. 我国区域用水结构与产业结构的协调评价 [J]. 水电能源科学,2013,31(9):159-163.

Partial Least Squares Regression Analysis of Correlation Factors of Water Use Efficiency

LI Ke-bai, LU Hui, TAO Jun

(School of Management Science and Engineering, Nanjing University of Information Science & Technology, Nanjing 210044, China)

Abstract: In order to analyze the influencing factors and their correlations of water use efficiency, incorporate all independent variables and overcome multicollinearity, the partial least squares method was used to study the correlation factors of water use efficiency in 2019 in China. The results show that the partial least square method can describe the correlation between variables effectively. Factors that are highly correlated with water consumption per ten thousand Yuan of GDP are natural conditions, social development level, GDP proportion of added value of primary industry, and the proportion of labor force in the water industry. The overall economic water use efficiency is high in the eastern region and low in the western region. Factors that are highly correlated with actual irrigation water consumption per mu of cultivated land are natural conditions and GDP proportion of added value of primary and secondary industries. Agricultural water use efficiency is high in the central region and low in the western region. Factors highly correlated with the water consumption per ten thousand Yuan of industrial added value are natural conditions, sewage treatment rate, urbanization rate and the capital proportion of water industry. Industrial water efficiency is high in the eastern region and low in the central region. Factors highly correlated with per capita public water consumption are per capita water resources, annual precipitation, sewage treatment rate, per capita GDP and GDP proportion of added value of primary and secondary industries. There is little correlation between regional factors and public water use efficiency. The research results can provide reference for the optimal management and rational utilization of water resources.

Key words: industrial water; water use efficiency; correlation analysis; partial least squares regression