Analysis of Annual Water Use Time-series and Influencing Factors in China

Jinxin Li, Wandi Jiang and Youdong Wen*

School of Statistics, Jiangxi University of Finance and economics Jiangxi Nanchang 330013

*Corresponding author e-mail: jxufewyd@163. com

Abstract. Water resources are one of the most precious resources on the earth. Faced with the increasingly severe shortage of water, human awareness of water conservation is constantly increasing. This paper analyzes the time series of annual average water consumption in China from 2004 to 2016 and the influencing factors for 13 years. It is concluded that China's per capital water consumption is mainly affected by total water supply, total surface water supply, total ground water supply, total agricultural water supply and total industrial water supply, and the influence of these five factors are all positive, however, it is worth noting that the real factor affecting China's per capital water consumption is the total amount of water used for agriculture and industry, which has nothing to do with the total amount of domestic water.

1. Introduction

Although the water resources on Earth are very rich, with a total of 1.45 billion cubic kilometers, there are very few that can be directly used by people for living and production. The main reason is that the sea water, which accounts for an absolute proportion of water, is salty and bitter, and cannot be used for drinking water or for irrigating land or for industrial use. The secondary reason is that freshwater resources account for only 2.5% of the total amount of water, and more than 70% of them are frozen in the ice sheets of the Antarctic and Arctic, in addition to difficult-to-use mountain glaciers and permafrost snow, a total of 87% of freshwater resources are difficult to use. The global freshwater resources are not only short of resources but also extremely unevenly distributed. By region, freshwater resources in 9 countries, including Brazil, Russia, Canada, China, the United States, Indonesia, India, Colombia, and the Congo, account for 60% of the world's freshwater resources. About 1.5 billion people in 80 countries and regions that account for about 40 percent of the world's population have insufficient fresh water, of which about 300 million people in 26 countries are extremely deprived of water [1-2]. What's more, by 2025, three billion people are expected to face water shortages, with 40 countries and regions facing severe shortages.

According to the latest data, the total amount of water resources in China is 2.8 trillion cubic meters, including 2.7 trillion cubic meters of surface water and 0.83 trillion cubic meters of groundwater, since surface water and groundwater are mutually converted and mutually replenished, the amount of groundwater resources that does not overlap with the river runoff is about 0.1 trillion cubic meters after deducting the duplicate calculation amount of 0.73 trillion cubic meters. According to internationally recognized standards, less than 3,000 cubic meters of water per person is a mild
shortage, less than 2000 cubic meters of water per person is moderately deficient, less than 1000 cubic meters of water per person is a serious shortage, less than 500 cubic meters of water per person is an extreme shortage. China currently has 16 provinces (regions, cities) which per capital water resources (excluding transit water) below the severe water shortage line, and 6 provinces, regions (Zhejiang, Shandong, Henan, Shanxi, Jiangsu) which per capital water resources below 500 cubic meters [3-5]. The total amount of water resources in China is not very large, ranking sixth in the world, and the per capital occupancy is less than 2,240 cubic meters, ranking 88th among the 153 countries that are counted by the World Bank. The distribution of China's water resources is also uneven. The Yangtze River basin and its southern areas only account for 36.5% of the country's total land area and 81% of country's water resources. To the north of the Yangtze River, the land area accounts for 63.5 percent of the country's total land area and only 19 percent of country's water resources.

This article analyzes the changes in annual average water use of Chinese residents from 2004 to 2016, studies its law of change and analyze its influencing factors.

2. Data sources and time series analysis

2.1. Data sources

The data source for this study is the official publication of the Ministry of Water Resources of the National Bureau of Statistics of China. The data details are shown in Table 1, Y represents the per capital water supply (m$^3$ per person), X1 represents the total water supply (100 million m$^3$), X2 represents the total surface water supply (100 million m$^3$), and X3 represents the total groundwater supply (100 million m$^3$), X4 indicates total other water supply (100 million cubic meters), X5 indicates total water use (100 million cubic meters), X6 indicates total agricultural water use (100 million cubic meters), X7 indicates total industrial water use (100 million cubic meters), X8 represents total domestic water use (100 million cubic meters), X9 represents total ecological water use (100 million cubic meters).

|   | Y    | X1  | X2   | X3    | X4    | X5    | X6    | X7   | X8    | X9    |
|---|------|-----|------|-------|-------|-------|-------|------|-------|-------|
| 2016 | 438.12 | 6040.16 | 4912.4 | 1057 | 70.85 | 6040.2 | 3768 | 1308 | 821.6 | 142.6 |
| 2015 | 445.09 | 6103.2  | 4971.5 | 1069.2 | 62.5  | 6103.2 | 3851.5 | 1334.8 | 794.2 | 122.7 |
| 2014 | 446.75 | 6094.88 | 4920.46 | 1116.94 | 57.46 | 6094.86 | 3868.98 | 1356.1 | 766.58 | 103.2 |
| 2013 | 455.54 | 6183.45 | 5007.29 | 1126.22 | 49.94 | 6183.45 | 3921.52 | 1406.4 | 750.1 | 105.38 |
| 2012 | 454.71 | 6141.8  | 4963.02 | 1134.22 | 44.55 | 6141.8 | 3880.3 | 1423.88 | 728.82 | 108.77 |
| 2011 | 454.4  | 6107.2  | 4953.3 | 1109.1 | 44.8  | 6107.2 | 3743.6 | 1461.8 | 789.9 | 111.9 |
| 2010 | 450.17 | 6021.99 | 4881.57 | 1107.31 | 33.12 | 6021.99 | 3689.14 | 1447.3 | 765.83 | 119.77 |
| 2009 | 448.04 | 5965.15 | 4839.47 | 1094.52 | 31.16 | 5965.15 | 3723.11 | 1390.9 | 748.17 | 102.96 |
| 2008 | 446.15 | 5909.95 | 4796.42 | 1084.79 | 28.74 | 5909.95 | 3663.46 | 1397.08 | 729.25 | 120.16 |
| 2007 | 441.52 | 5818.67 | 4723.9 | 1069.06 | 25.7  | 5818.67 | 3599.51 | 1403.04 | 710.39 | 105.73 |
| 2006 | 442.02 | 5794.97 | 4706.8 | 1065.52 | 22.7  | 5794.97 | 3664.45 | 1343.76 | 693.76 | 93.9 |
| 2005 | 432.07 | 5632.98 | 4572.19 | 1038.83 | 21.96 | 5632.98 | 3580 | 1285.2 | 675.1 | 92.68 |
| 2004 | 428    | 5547.8  | 4504.2 | 1026.4 | 17.2  | 5547.8 | 3585.7 | 1228.9 | 651.2 | 82   |
2.2. Time Series Analysis

Figure 1. Time Series Analysis.

From Figure 1, it can be seen that each data maintained a relatively stable trend in the past thirteen years and there was no large-scale fluctuations. And only the total amount of water shows a clear growth trend, while the total surface water supply shows a slight increase.

3. Analysis of influencing factors

3.1. Method selection

Because there are different degrees of correlation between the factors explored by the Institute, it is necessary to adopt the statistical method of Ridge Regression to overcome the problem of multicollinearity. Ridge regression is a biased estimation regression method dedicated to the analysis of collinear data [6-7]. It is essentially a kind of improved least squares estimation method, by abandoning the unbiasedness of the least squares method, the regression coefficient is more realistic and more reliable at the cost of losing some information and reducing precision, the fitting of pathological data is stronger than the least square method.

For some matrices, a small change in one of the elements of the matrix can cause large errors in the final calculation. This type of matrix is called a "morbid matrix". Sometimes the incorrect calculation method will also make a normal matrix appear sick in the operation. For the Gaussian elimination method, if the elements on the pivot element (the diagonal element) are small, they will show morbid characteristics in the calculation.

The least square method commonly used in regression analysis is an unbiased estimate. For a well-posed problem, $X$ is usually full rank, $X\theta = Y$, Using the least squares method, define the loss function as the square of the residual, minimizing the loss function $\|X\theta - y\|^2$. The above optimization problem can be solved by gradient descent method, or it can be solved by the following formula: $\theta = (X^TX)^{-1}X^T y$, When the rank $X$ is not full, or when the linear correlation between some columns is large, the determinant of $X^TX$ is close to 0, that is, $X^TX$ close to singularity, the above problem becomes an ill-posed problem, at this time, the calculation $(X^TX)^{-1}$ error will be very large, the traditional least squares method lacks stability and reliability.

In order to solve the above problems, we need to convert ill-posed problems into well-posed problems: we add a regularization term to the above loss function, to become $\|X\theta - y\|^2 + \|\Gamma\theta\|^2$, Among them, we define $\Gamma = \alpha I$, then $\theta(\alpha) = (X^TX + \alpha I)^{-1}X^Ty$, in the above formula, $I$ is a
unit matrix. With the increase of $\alpha$, the absolute value of each element in $\Theta(\alpha)$ tends to become smaller, and their deviation from the correct value $\theta_i$ becomes more and more larger. When $\alpha$ tends to infinity, $\Theta(\alpha)$ tends to 0. Among them, the trajectory $\Theta(\alpha)$ that changes with the change of $\alpha$ is called a ridge track. The actual calculation can choose large numbers of values $\alpha$, make a ridge trace, from the figure can be seen at which time the value becomes stable, then determine this value $\alpha$.

3.2. Empirical Research

By constructing an ordinary linear regression equation for annual average water consumption, it was found that although the regression coefficient ($R$) was very high, reaching 0.9983, but all the constant term were not significant, then look through the R language package car( ) for each variable's VIF (variance inflation factor) value, the VIF values of the 9 variables X1 to X9 were found to be very high, both well above 10, therefore, it can be concluded that there is multicollinearity among variables. Next, using the function lm.ridge( ) in the MASS package to achieve the ridge return. The following calculation tests 151 lambda values, and finally selects the one that makes the generalized cross-validation GCV the smallest. It can be known from the operation that the lambda GCV at the minimum GCV is 1 and the corresponding coefficient at the minimum GCV is 2265.6.

![Figure 2. Relationship between lambda and GCV.](image)

From Figure 2, we can see that the choice of lambda is not so important, as long as you are not too close to lambda=0, there is not much difference. Finally, use the linear Ridge( ) function in the ridge package to select the ridge regression parameter automatically.

It can be seen from Table 2 that the coefficients of the respective variables change from insignificant to significant (X1, X2, X3, X4, X5, X6, X7), and the ridge regression parameter value is 0.2538. Finally, the Lasso regression was used to solve the colinearity problem [9].

4
Table 2. Automatic Ridge Regression Parameter Results.

|       | Estimate Scaled | estimate Std. | Error (scaled) | t value (scaled) | Pr(>|t|) |
|-------|-----------------|----------------|----------------|-----------------|----------|
| (Intercept) | 1.878e+02 | NA | NA | NA | NA |
| X1     | 6.64E-03 | 4.60E+00 | 2.93E-01 | 15.692 | < 2e-16 ** |
| X2     | 8.67E-03 | 4.77E+00 | 4.64E-01 | 10.274 | < 2e-16 ** |
| X3     | 5.73E-02 | 6.67E+00 | 9.85E-01 | 6.768 | 1.30E-11 ** |
| X4     | -6.145e-02 | -3.64E+00 | 8.95E-01 | 4.063 | 4.85E-05 ** |
| X5     | 6.64E-03 | 4.60E+00 | 2.93E-01 | 15.692 | < 2e-16 ** |
| X6     | 6.89E-03 | 2.80E+00 | 8.64E-01 | 3.245 | 0.00117 * |
| X7     | 3.96E-02 | 9.20E+00 | 9.82E-01 | 9.37 | < 2e-16 ** |
| X8     | -3.687e-04 | -6.28E-02 | 1.07E+00 | 0.059 | 0.95324 |
| X9     | -2.930e-02 | -1.58E+00 | 1.22E+00 | 1.295 | 0.19529 |

Signif.codes: 0 ‘***’ 0.01 ‘**’ 0.05 ‘.’ 0.1 ‘ ’ 1

Combining Tables 3, 4 and Figure 3, it can be seen that when the value is taken to step 6, the Cp value is the smallest. Therefore, it is considered that the independent variables X3, X7, X1, X2, X6, and X9 have an influence on the dependent variable Y, of which only X9 is insignificant.

Table 3. LASSO variable selection order table.

|       | X3 | X7 | X1 | X2 | X6 | X9 | X5 | X4 | X8 |
|-------|----|----|----|----|----|----|----|----|----|
| Var   | 3  | 7  | 1  | 2  | 6  | 9  | 5  | 4  | 8  |
| Step  | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  |

Figure 3. LAR dynamic change chart.
Table 4 Cp value change table.

| Df | Rss      | Cp        |
|----|----------|-----------|
| 0  | 858.16   | 884.5796  |
| 1  | 463.76   | 474.9858  |
| 2  | 263.79   | 268.2966  |
| 3  | 168.93   | 171.2933  |
| 4  | 8.28     | 5.6455    |
| 5  | 5.53     | 4.7659    |
| 6  | 5.14     | 6.3675    |
| 7  | 3.87     | 7.0413    |
| 8  | 3.76     | 8.9223    |
| 9  | 2.87     | 10.00     |

4. Conclusion
In this paper, through automatic ridge regression and Lass regression methods, it is determined that China's per capital water consumption is mainly affected by the total water supply, total surface water supply, total ground water supply, total agricultural water use, and total industrial water use. And the influence of these five factors is positive, but it is worth noting that the real factor that affects China's per capital water consumption is the total amount of water used for agriculture and industry, which has nothing to do with the total amount of domestic water.

Acknowledgements
This article was supported by the “Research on the Relationship between R&D Input and Middle-income Trap from the Perspective of 'Dual Creation'” (TJ161002).

References
[1] Hanasaki, Naota, et al. "A global hydrological simulation to specify the sources of water used by humans". Hydrology & Earth System Sciences 22(2018):1-53.
[2] Ma, L. , et al. "Impacts of irrigation water sources and geochemical conditions on vertical distribution of pharmaceutical and personal care products (PPCPs) in the vadose zone soils". Science of the Total Environment 626(2018):1148-1156.
[3] Wang, Xiaofeng, et al. "Phthalate esters in main source water and drinking water of Zhejiang Province (China): Distribution and health risks". Environmental Toxicology & Chemistry 34.10 (2015):2205-2212.
[4] Ding, W. X., et al. "Genotoxicity of microcystic cyanobacteria extract of a water source in China". Mutation Research 442.2(1999):69-77.
[5] Ye, Xiao Yan, et al. "Real-Time PCR Detection of Enteric Viruses in Source Water and Treated Drinking Water in Wuhan, China". Current Microbiology 65.3(2012):244-253.
[6] Mcdonald, Gary C. "Ridge regression". Wiley Interdisciplinary Reviews Computational Statistics 1.1(2010):93-100.
[7] Zhang, Yuchen, J. C. Duchi, and M. J. Wainwright. "Divide and Conquer Kernel Ridge Regression: A Distributed Algorithm with Minimax Optimal Rates". Journal of Machine Learning Research 30.1(2013):592-617.
[8] Park, Trevor, and G. Casella. "The Bayesian Lasso". Publications of the American Statistical Association 103.482(2008):681-686.