Characteristics and Influence Factors of Water Consumption in China Pro-vincial Capital Cities by Means of Multivariate Regression Algorithm

Jinjun Zhou[1,2], Jiahong Liu[2,a], Hao Wang[1,2], Zhongjing Wang[1], and Weiwei Shao[2]

[1] Department of Hydraulic Engineering, Tsinghua University, Beijing 100084, China
[2] China Institute of Water Resources and Hydropower Research State Key Laboratory of Simulation and Regulation of Water Cycle in River Basin, Beijing 100038, China

Abstract. The city is a typical natural and social dual water circulation area. The water use characteristics of the city have obvious dual attributes, and there are many factors (both on nature and social side) affecting the urban water consumption (UWC). This article aimed to research the structure and characteristics of the UWC. Taking the provincial capital cities of China as the research objects, 24 index factors and data of the year 2015 were selected to construct a multivariate regression model between urban UWC and index factors. The results showed that the combination of correlation analysis and full subset regression could effectively screen the prediction variables of UWC. Principal component analysis could effectively reduce variable dimensions of UWC while preserving raw dataset information as much as possible. The main factors affecting UWC on the social side include the built-up area, the urban population, road cleaning area, residential electricity consumption, and per capital water consumption, and the main factors of the natural side include per capital green land and precipitation.

1 Introduction

The urban hydrological cycle process becomes more complicated under the influence of climate change and urbanization [1]. Urban water consumption (UWC) is an important part of urban hydrological process, which determines the process change and flux of urban hydrological cycle [2]. In addition, UWC is an important reference for urban water resources management and allocation, as well as an important index for urban planning, design and construction [3]. In recent years, the shortage of urban water resources has become a bottleneck restricting urban development [4]. Therefore, it is necessary to understand the influencing factors of UWC and analyze their characteristics.

However, due to the interference of strong human activities, the process of UWC becomes more complicated [5]. It is not only disturbed by human activities, but also affected by natural factors such as hydrology and meteorology [6]. The growth of urban population promotes the increase of the total amount and intensity of UWC, while the water price controls the amount and intensity of UWC [7]. Water consumption per capital per day is highly influenced by meteorological factors, socioeconomic status, water supply and conservation factors [8]. From a technical point of view, water supply facilities, water use efficiency and reuse rate also have a great impact on urban water use [9].

Multiple linear regression, time series methods and artificial neural networks were often used to forcast the UWC in many existing research [10-11]. But these methods or methods are basically a black box model, so it is difficult to analyze the influence characteristics of influencing factors on UWC [12], which is the premise and basis for improving the situation of urban water use, the scientific basis for realizing the fine management of urban water resources [13]. Data mining is a process of extracting hidden and potentially useful information and knowledge from a large, incomplete, noisy, fuzzy, random real data [14], which is applicable to the study of complex relationships that are not of a non-mechanical rationality [15].

This paper selected the methods of correlation analysis, regression analysis, and principal component analysis (PCA) in data mining methods to study the influencing factors and their influencing characteristics of UWC. It was expected that the natural and social duality of UWC can be analyzed, and the influence of natural and social factors on UWC could be considered as comprehensively as possible.

2 Materials and methods

The objects of this study were the 31 provincial capitals and municipalities cities directly under the central government of China in year of 2015, which did not include Taipei because of insufficient information. Data

[a] Corresponding author: liujh@iwhr.com

sources were “China City Statistics Yearbook”, “China City Construction Statistical Yearbook”, “Statistical Yearbook” of each city and so on. According to statistics data, there were six main categories of water consumption in cities, namely, residential water, water for public administration and services, water for production, free supply water, leakage and other water supply. Figure 1 showed the structure of sample cities in 2015. The residential water was the largest type of water for most cities[16]. In 2015, more than 50% of the total consumption of water for residents and public services was more than 80% in these cities.

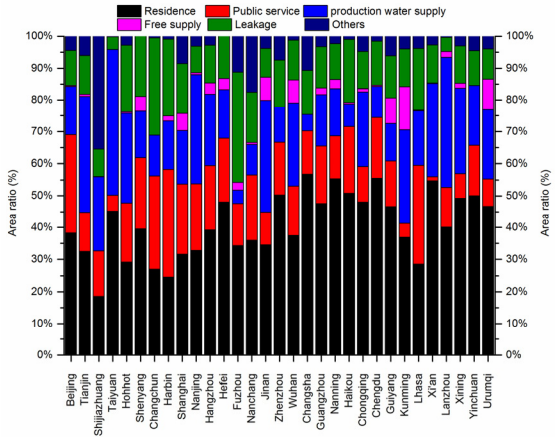

Figure 1. The UWC structure of provincial capitals cities in China.

The main influencing factors included urban construction, social economy, climatic characteristics and so on. Specific indicators included the area of built-up areas, urban population, per capita road area, per capita income per capita, per capita water consumption, per capita road area, per capita green area, tertiary industry ratio, average temperature, rainfall days, sunshine hours and so on. The detailed names and signs of 24 index variables were shown in the table 1.

Table 1. Predictive variables and their signs.

Names	Sign	Names	Sign
Water use population	wsp	Precipitation days	pd
Urban area	ua	Hours of sunshine	s
Urban population	pn	Reclaimed water	rw
Built-up area	ba	Road cleaning area	rca
per capita income	pi	Density of population	dp
Daily water consumption per capita	pw	Water popularity rate	wpr
Per capita road area	pr	Gas popularity rate	gpr
Per capita green area	pg	Water pipe density	pnd
Green coverage rate of built area	gr	Secondary industry ratio	sr
Tertiary industry proportion	tr	Industrial electricity consumption	ie
Precipitation	p	Residential electricity consumption	re
Average temperature	t	Industrial wastewater discharge	iww

R language was the main research tool in this study, which was a powerful tool for data analysis, summary, exploration and mining. The correlation analysis in R language was used to test the index variables and reduce the variables with good correlation. Then, the all-subsets regression was used to select the best predictor variables by adjusting the value of R^2. On the other hand, the principal component analysis (PCA) method was used to analyze the prediction variables, and the principle was to select the latent relationship structure of the predicted variables, and to reduce the dimension of the variables. Then the multivariate linear regression analysis was carried out by ordinary least squares (OLS) method, and data mining research route was shown in figure 2.

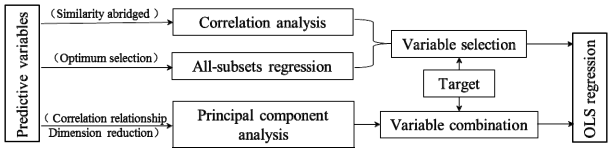

Figure 2. Technical roadmap of data mining.

3 Results and discussions

3.1 Regression fitting of UWC based on optimal variables

3.1.1 Correlation analysis

The corrgram () function in the R language was used to graphically show the correlation between variables, as shown in figure 3. The diagonal line was 24 variables, and the triangle area at the lower left of the diagonal line was a rectangle with slashes of different colors. The blue rectangle pattern indicated that the two variables in the cell were positively correlated, and the direction of the slash was also from the bottom left to the upper right. The deeper the blue was, the greater the correlation was. Similarly, the red rectangle represented a negative correlation between the two variables, and the slash slanting from the top left to the bottom right, the deeper the red was, the greater negative correlation was. The diagonal upper right showed the correlation between the two variables in the form of a pie chart. In the same way, the blue represents the positive correlation, and the red represented the negative correlation. The positive correlation was to fill the pie chart clockwise from 12 points, and the negative correlation was filled in the counter clockwise direction. And the larger the filling range was, the deeper the color was, and the greater the correlation was.

The positive correlation between the variables such as *ua*, *ie*, *pi*, *iww*, *pnd*, *rca*, *re*, *pn*, *ba*, *wsp* was good, and the negative correlation between *s* and *t*, *p*, *pd* was better. It needed to be selectively deleted from these variables to improve the unrelevance and effectiveness of the premeasured variables. Because the variable *s* had good correlation with the other three variables, the variable *s*

was deleted. For other variables with good correlation, half of the variables could be selectively deleted.

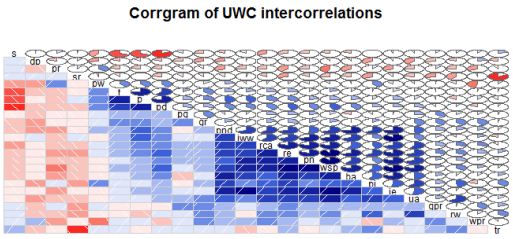

Figure 3. Correlation coefficient diagram of prediction variables of UWC.

3.1.2 ALL-subsets regression

The all-subsets regression is an analysis of all possible models, with the number of variables increasing from 2 to all until R^2 reaches its maximum and becomes stable. We used the regsubsets () function in the leaps package in R language to achieve all-subsets regression. The values of "nbest" was set to 4, which mean that showed four best predictive variable models for each number of variable in the regression process. Figure 4 showed the best model for a all-subsets regression of UWC. In the graph, the horizontal axis was the intercept and 24 predictive variables, and the vertical axis was the R^2 value under the combination of different numbers of variables. In this study, the optimal R^2 reached 0.96, and the fitting effect was better. The corresponding variable number was 8.

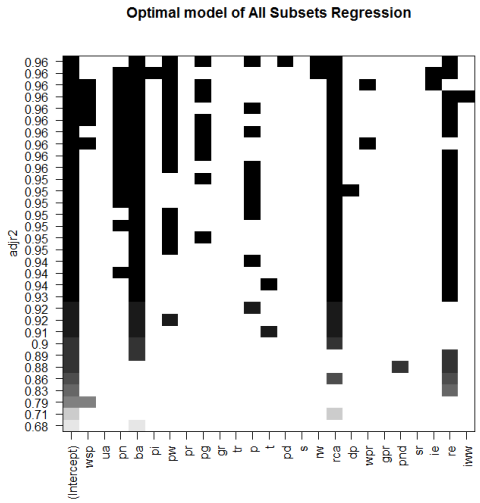

Figure 4. Four optimal models with different subset sizes

In order to quantitatively analyze the frequency of occurrence of different variables in the full subset regression and combine the correlation analysis to select the best variables, table 2 gave the proportion of 24 predictive variables in the all-subsets regression.

Table 2. Frequency distribution of prediction variables in the all-subsets regression process.

Variables	Frequency	Variables	Frequency
wsp	18.75%	pd	3.13%
ua	0.00%	s	0.00%
pn	43.75%	rw	6.25%
ba	84.38%	rca	84.38%
pi	3.13%	dp	3.13%
pw	46.88%	wpr	6.25%
pr	0.00%	gpr	0.00%
pg	34.38%	pnd	3.13%
gr	0.00%	sr	0.00%
tr	0.00%	ie	6.25%
p	34.38%	re	68.75%
t	6.25%	iww	3.13%

Among them, the probability of the occurrence of the *ua*, *t*, *s*, *gr*, *sr*, *tr*, *pr* was 0, so the seven variables were deleted. Because the frequency of *iww* was lower and the rate of occurrence of *ie*, the *iww* was deleted. The occurrence rate of *pi*, *pd*, *pnd*, *dp* was also low, so these variables were also deleted.

By means of correlation analysis and all-subsets regression analysis, 12 variables were deleted from 24 prediction variables. The OLS multivariate linear regression was used to fit the UWC, and the results of 12 variables and 24 variables were compared. In the case of 24 variables, the multiple R-squared was 0.988, the adjusted R-squared was 0.934， and the p-value was 0.00057. In the case of the 12 variables filtered, the multiple R-squared was 0.977, the adjusted R-squared was 0.962，and the p-value was 3.018e-12.

3.2 Regression fitting of UWC based on principal component analysis

Using principal component analysis (PCA), the related variables were converted into unrelated combinatorial variables to reduce the dimension of variables and retain the original information as much as possible.

First, the number of principal components should be determined. In this section, the fa.parallel () function was used to check the 24 variables by parallel analysis. Figure 5 showed the lithotripsy test based on observational eigenvalues, the eigenvalues derived from 100 random data matrices and the feature criteria greater than 1. The three criterions indicated that selecting 5 or 6 principal components can retain most of the information of the dataset.

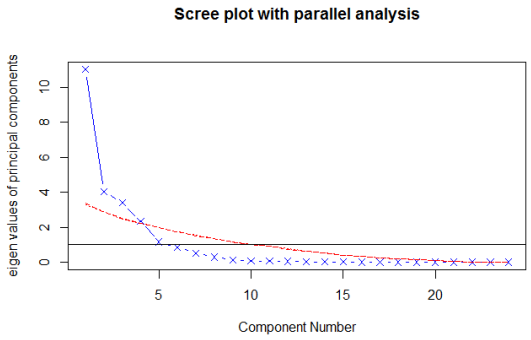

Figure 5. Evaluation of the number of principal components to be retained in UWC prediction variables. Scree Plot (x + full line); Parallel analysis of 100 simulations (dotted line); Criterion of characteristic value greater than 1(y=1).

Then the principal component analysis was carried out using principal component (PC) 1-6, and the main component PC1-6 was set to be 0.34, 0.14, 0.11, 0.09, 0.07, 0.06. When six main components were retained, the corresponding degree of interpretation was 81%. In general , the interpretation of the principal component to the data set was 80%. Therefore, six principal components were retained, and the scores of each main component were as follows:

PC1 = 0.94wsp + 0.74ua + 0.83pn + 0.78ba + 0.61pi + 0.10pw - 0.40pr + 0.22pg + 0.35gr + 0.11tr + 0.48p + 0.35t + 0.57pd - 0.35s + 0.35rw + 0.82rca - 0.20dp + 0.20wpr + 0.28gpr + 0.80pnd - 0.28sr + 0.84ie + 0.94re + 0.75iww.

PC2 = -0.22wsp - 0.29ua - 0.17pn - 0.17ba - 0.15pi + 0.59pw + 0.06pr + 0.17pg + 0.20gr - 0.27tr + 0.76p + 0.74t + 0.68pd - 0.80s - 0.37rw - 0.06rca - 0.06dp - 0.22wpr - 0.20gpr + 0.00pnd + 0.20sr - 0.23ie - 0.12re + 0.01iww.

PC3 = -0.12wsp + 0.25ua - 0.17pn - 0.18ba + 0.39pi + 0.43pw + 0.30pr + 0.28pg + 0.34gr + 0.70tr - 0.09p + 0.07t + 0.12pd + 0.02s + 0.50rw - 0.11rca - 0.56dp - 0.25wpr + 0.32gpr - 0.05pnd - 0.63sr - 0.32ie - 0.02re - 0.44iww.

PC4 = -0.06wsp + 0.09ua + 0.00pn - 0.32ba - 0.43pi - 0.43pw - 0.32pr + 0.56pg + 0.62gr - 0.13tr + 0.10p + 0.15t - 0.06pd - 0.05s + 0.20rw + 0.09rca + 0.08dp + 0.72wp + 0.36gpr - 0.21pnd + 0.01sr - 0.18ie + 0.02re - 0.10iww.

PC5= -0.07wsp + 0.13ua + 0.13pn - 0.11ba + 0.10pi + 0.25pw + 0.57pr + 0.26pg + 0.20gr - 0.47tr - 0.07p - 0.19t - 0.21pd + 0.17s - 0.25rw + 0.13rca - 0.45dp - 0.06wp + 0.41gpr + 0.08pnd + 0.49sr + 0.10ie - 0.03re + 0.12iww.

PC6 = -0.16wsp - 0.39ua -0.28pn + 0.20ba + 0.05pi + 0.08pw + 0.09pr - 0.13pg - 0.03gr + 0.31tr + 0.16p + 0.11t - 0.13pd + 0.14s - 0.48rw - 0.09rca + 0.09dp + 0.38wp + 0.55gpr + 0.25pnd - 0.26sr + 0.13ie + 0.02re + 0.17iww.

Through the score expression of principal component, the value of six principal components were calculated by using the data set of predictive variable, and then the multivariate linear regression fitting was carried out by using lm () function. The six principal components were represented by $z1$, $z2$, $z3$, $z4$, $z5$, and $z6$, and the water consumption of city was represented by y. The relation was obtained as follows:

$$y_d = -35150 + 1.14 \times z1 + 6.81 \times z2 - 2.50 \times z3 - 6.14 \times z4 - 2.06 \times z5 - 8.40 \times z6$$

The regression results showed that the multiple R-squared was 0.879, the adjusted R-squared was 0.848, and the p-value was 7.27e-10. The regression coefficient of principal component PC1 was 3.43, which mean that when the other principal components remain unchanged, the UWC increased by 3.43% with the principal component PC1 increased by 1%.

The result of PCA was slightly smaller than that of direct regression for 24 predictive variables (initial variable), which the difference was -0.109 and -0.086 respectively for multiple R-squared and adjusted R-squared, but the p-value and the value of Pr(>|t|) decreased significantly. Compared with the results of selected predictive variables in 3.1, the difference of PCA between multiple R-squared and adjusted R-squared were -0.098 and -0.115, respectively. The value of p-value was reduced obviously and the value of Pr(>|t|) also changed from obviously not close to 0 to close to 0.

4 Conclusions

This paper took 31 provincial capitals cities as the research objects, selected the statistical data of 2015, and carried on the analysis to the forecast variables by the means of the correlation analysis, the all-subsets regression analysis, and the PCA. Then, the OLS regression simulation analysis was carried out, and the following main conclusions were drawn.

Correlation analysis can screen the indicators with high correlation among UWC variables, all-subsets regression can select UWC variables under the optimal model, and the combination of two methods can effectively select variables, reduce the dimension of predictive variables, and promote multiple regression fitting effect.

PCA can analyze the internal relationship between variables of UWC, recombine them into principal components, and keep the information of the initial variable dataset as much as possible while reducing the dimension of variables. The confidence interval of predictive variables selected by PCA does not contain 0, which indicates that the relationship between predictive variables and UWC was obvious. The confidence interval of predictive variables in correlation analysis and full subset regression analysis contains 0. This was an obvious advantage over correlation analysis and all-subsets regression analysis, but the R^2 value was slightly lower than other methods.

Based on the analysis results, it could be concluded that the influence on the UWC of urban construction and meteorological characteristics was obvious, and the impact of economic level was slightly weak. The

specific indicators mainly included the built-up area, the urban population, road cleaning area, residential electricity consumption, per capital water consumption, per capital green land, precipitation, and so on, which covered factors on both the social and natural sides.

References

1. H. Wang, Y.W. Jia, G.Y. Yang, Z.H. Zhou, Y.Q. Qiu, C.W. Niu, H. Peng. Chin Sci Bull, **58,** 27 (2013)
2. Y.X. Yuan, J. Zhang, H.F. Xu, S.L. Qu. Water & Wastewater Engineering, **30,** 6 (2004)
3. M.A.A. Khadam. Water International, **13,** (1988)
4. P. Jing. Science & Technology Management Research, (2015)
5. G. Chen, T. Long, J. Xiong, Y. Bai. Water Resources Management, **31,** 15 (2017)
6. L. Ramulongo, N.S. Nethengwe, A. Musyoki, Procedia Environmental Sciences, **37,** (2017)
7. W.Y. Han, X.P. Chen, Z.L. Zhang. Chinese Journal of Ecology, (2018)
8. L. Fan, L. Gai, Y. Tong, R. Li. Journal of Cleaner Production, 166 (2017)
9. D.G.M. Silva, J.G. Erazo, A.M.O. Cruz, Rev.ing.univ.medellín, **11,**21(2012)
10. H.B. Liu, H.W. Zhang. Journal of Tianjin Institute of Textile Science & Technology. (2004)
11. X.U. Shi-Rong, X.K. Yin, L.I. Li-Wu. Journal of Hunan Urban Constructin College, (2002)
12. K.L. Spencer, I.G. Droppo, C. He, L. Grapentine, K. Exall, Water Research, 45,8 (2011)
13. X.R. Wang. Global Science Technology & Economy Outlook, **28,** 11 (2013)
14. A.A. Freitas. Advances in Evolutionary Computing. (2003)
15. Y. Saygin, A. Reisman, Y.T. Wang, Social Science Electronic Publishing, **51,** 4 (2004)
16. G. Romano,N. Salvati, A. Guerrini. Water Resources Management,28,15(2014)