Analysis and Prediction of Water Consumption in Nanjing—A Comparative Study Based on Stepwise Regression and Nonparametric Mode

Zhou Xu
* Corresponding author: 178905695@qq.com

College of Finance and Public Economics, Shanxi University of Finance and Economic, Taiyuan, China

Abstract: Accurate prediction of urban water consumption is of great significance to the management and improvement of water supply system. In this thesis, through correlation analysis on 16 factors affecting total water consumption in Nanjing from 2005 to 2019, the factors among them that have more obvious influence are selected, and then they are fitted by stepwise linear regression model and nonparametric regression model respectively. By comparing the fitting results, it is found that nonparametric regression model has better prediction effect than stepwise regression model. The analysis results show that the prediction of urban water consumption is very important for the rational allocation of water resources.

Keyword: urban annual water consumption, stepwise regression, nonparametric regression, prediction

1 Introduction

Water is one of the most important material resources for human survival and development. Abundant available water resources are conducive to accelerating urban development and strengthening its capacity, ensuring the sustainable development of society, economy and ecology, and promoting the construction of a harmonious society. The urban total water consumption can reflect the long-term demand of urban water, thus accurate prediction of the urban total water consumption is not only beneficial to the rational allocation of water resources, but also provides reference basis for ensuring sufficient urban domestic, industrial and agricultural water consumption, which are the vital parts of water resources allocation in urban development. The government has always attached great importance to more accurate prediction of urban water allocation in urban development and plans.[1]

The prediction methods of urban water consumption based on historical observation data mainly include time series analysis, structural analysis and systematic analysis [2]. In recent years, the rapid development of combined analysis is also the main research object of domestic scholars. The time series analysis, mainly based on the periodic change rule of urban water consumption, carries out statistical analysis with a certain time as the change period, and puts forward the mathematical model and calculation method of fitting and prediction, which mainly include moving average method [3] and exponential smoothing method[4], etc. Since this analysis does not show obvious change trend of random disturbance of water consumption, and has stricter requirements on various related influencing variables and higher dependence on historical data,

MSIEID 2022, December 09-11, Chongqing, People's Republic of China

DOI 10.4108/eai.9-12-2022.2327557

it is more suitable for medium- and long-term prediction.

The structural analysis, focusing on the relationship between urban water consumption and various related factors (such as gross regional product, regional water consumption population, and temperature conditions, etc.), finds out the main influencing factors and problems on rational urban water consumption planning, and puts forward reasonable policy suggestions based on the analysis and prediction results. The structural analysis is mainly composed of regression analysis [1,5,6,7] and stepwise regression [8]. However, the factors affecting water consumption are complex, and it is difficult to accurately predict the future values of each influencing factor, so the prediction results will have a small deviation.

The system analysis does not investigate the effect of individual factors, which means it will reflect the comprehensive effect of various factors, weaken the influence of random factors, and thus reflect the inherent regularity of water consumption change. This analysis mainly includes artificial neural network method [9,10], grey prediction method [11], system dynamics [12] and other methods, which can be used to predict medium- and long-term or short-term water consumption. The disadvantages are that the accuracy is low when predicting time series with unstable changes or abrupt structural changes, and the calculation of it is very large. The performance of the algorithm depends on the initial conditions, and the process is easy to fall into local minimum and the convergence is slow.

In recent years, scholars at home and abroad have continuously explored better prediction models that can adapt to the current urban development, and then the combined prediction model [13-15] has gradually developed, which has been studied by most scholars, and is often conducted on a large amount of data combined with neural networks.

By combing the domestic literature on the prediction of urban water consumption, it is found that there are many research results at present, and most of them focus on empirical research. However, there are some differences in the research objects, research indicators and research methods chosen by scholars, and their models have their own advantages and disadvantages. Most of the researches focus on short-term effects. For the long-term sustainable development of urban water consumption and the rational allocation of water consumption, there are few studies on the impact of economic benefits, and there are few literature that use nonparametric models to predict, and most of them are about short-term predictions [16,17]. This thesis selected the actual water consumption data of Nanjing from 2005 to 2019 as the dependent variable, compared the nonparametric model with the stepwise linear model, used the model for prediction, and showed the advantages of nonparametric model in data fitting and prediction under the condition of long-term stable economic development.

2 Research Methods and Data Sources

2.1 Research methods

In this thesis, 16 influencing variables are selected from the long-term influencing factors of actual water consumption in Nanjing, and the statistical data from 2005 to 2019 in Nanjing, Jiangsu Province are taken as samples. By using correlation analysis to eliminate the variables with smaller influencing factors, nonparametric regression simulation is carried out, and the simulation results are compared with those of stepwise regression model to analyze the effects of the two models on the prediction of urban annual water consumption.

2.2 Data Source

In view of the availability of data, the data in this thesis comes from Nanjing Statistical Yearbook from 2006 to 2020, and the total water consumption TWC (10,000 cubic meters) is selected as the dependent variable. Industrial water consumption IWC (10,000 cubic meters), domestic water consumption DC (10,000 cubic meters), water consumption population WCP (10,000 people), production water reuse PWR (10,000 cubic meters), annual sewage treatment consumption ASTC (10,000 tons), gross regional product GRP (100 million yuan), per capita GDP GDPpc (yuan), average temperature AT(°C), average maximum temperature AMAXT(°C), average minimum temperature AMINT(°C), extreme maximum temperature EMAXT(°C), extreme minimum temperature EMINT(°C), total precipitation TP (mm), cultivated land area at the end of the year CY (thousand hectares), green coverage rate of built-up area GCR(%) and year-round sunshine YRC (hours) are the preset influencing variables.

3 Empirical Analysis of Annual Water Consumption Prediction in Nanjing

3.1 Multivariate Stepwise Linear Regression

The basic idea of stepwise regression is to introduce variables into the model one by one. After each explanatory variable is introduced, F test should be carried out, and the selected explanatory variables should be tested under t test one by one. When the original explanatory variables become no longer significant due to the introduction of later explanatory variables, they should be deleted to ensure that only significant variables are included in the regression equation before each new variable is introduced. This is an iterative process until neither significant explanatory variables are selected into the regression equation nor insignificant explanatory variables are eliminated from the regression equation to ensure that the final set of explanatory variables is the best and simplest.

Based on the selected dependent variable and 16 independent variable data, the stepwise regression results and analysis are as follows:

Table 1. Model summary[e]

Model	R	R^2	Adjust R^2	Standard estimate error
1	.885[a]	.783	.766	6210.11813
2	.964[b]	.930	.919	3663.62443
3	.982[c]	.965	.955	2725.84894
4	.989[d]	.977	.968	2287.40535

a. predicted variables: (constant), GRP (100 million yuan)
b. predicted variables: (constant), GRP (100 million yuan), IWC (10,000 cubic meters)
c. predicted variables: (constant), GRP (100 million yuan), IWC (10,000 cubic meters) and ASTC (10,000 tons)
d. predicted variables: (constant), GRP (100 million yuan), IWC (10,000 cubic meters), ASTC (10,000 tons) and DC (10,000 cubic meters)

e. dependent variable: TWC (10,000 cubic meters)

Table 1 is the result of each step of model simulation, from which we can find that R, R^2 and adjusted R^2 of model 4 are higher than other models, indicating that the model 4 has the highest fitting degree and the best fitting effect. Therefore, the stepwise regression model selects GRP (100 million yuan), IWC (10,000 cubic meters), ASTC (10,000 tons) and DC (10,000 cubic meters) as the main influencing factors affecting the urban TWC.

Table 2. variance analysis[e]

Model		Sum of square	df	Mean square	F	Sig.
4	return	2.257E9	4	5.641E8	107.820	.000[d]
	Residual	52322232.54	10	5232223.254		
	total	2.309E9	14			

Table 2 is the variance analysis of each model. Since this thesis chooses the best model 4 in the stepwise regression model, the variance test results of model 4 are analyzed. When the F statistic of this model is 107.820, the significance test Sig. = 0.000 < 0.001, which means that the regression equation has passed the significance test (F test), indicating that the established linear regression model has statistical significance.

Table 3 is the regression coefficient test of the model 4. The constant coefficient of this model is 35,541.626, and the independent variable coefficients are GRP=2.587, IWC=0.940, ASTC=0.204, DC=0.163. After t significance test, it can be seen that all Sig. values pass the significance test at 5% significance level. Therefore, this model is the best regression model after screening by stepwise regression model. Then we can establish the following regression model formula:

$$\begin{aligned} T\hat{W}C &= 35541.626 + 2.587GRP + \\ &0.940IWC + 0.204ASTC + 0.163DC \end{aligned} \tag{1}$$

This model has good significance.

3.2 Correlation Test

In order to analyze the influence degree of the important factors that affect the annual TWC in Nanjing more comprehensively, this thesis referred to relevant data, analyzed papers and literature, consulted experts from all sides, and finally selected 16 related independent

Table 3. Regression coefficient test

model		Non-standardized coefficient		Standard coefficient	t	Sig.	Correlation		
		B	Standard error	trial version			Zero order	Partial	section
4	(constant)	35541.626	7152.186		4.969	.001			

	GRP (100 million yuan)	2.587	.207	.774	12.486	.000	.885	.969	.594
	IWC (10,000 cubic meters)	.940	.127	.356	7.388	.000	.322	.919	.352
	ASTC (10,000 tons)	.204	.056	.225	3.645	.004	.730	.755	.174
	DC (10,000 cubic meters)	.163	.069	.114	2.371	.039	.120	.600	.113

Table 4. Correlation coefficients between independent variables and dependent variables

Independent variable	Pearson correlation	Kendall correlation	Spearman correlation
IWC (10,000 cubic meters)	.322 .121	.371* .027	.514* .025
DC (10,000 cubic meters)	.120 .335	-.029 .441	-.118 .338
WCP (10,000 people)	.853** .000	.790** .000	.904** .000
PWR (10,000 cubic meters)	.468* .039	.524** .003	.686** .002
ASTC (10,000 tons)	.730** .001	.619** .001	.771** .000
GRP (100 million yuan)	.885** .000	.752** .000	.875** .000
GDPpc (yuan)	.872** .000	.752** .000	.875** .000
AT(℃)	.419 .060	.398* .022	.390 .075
AMAXT(℃)	.286 .150	.206 .147	.235 .199
AMINT(℃)	.510* .026	.343* .040	.422 .058
EMAXT(℃)	.332 .114	.144 .228	.292 .146
EMINT(℃)	.304 .136	.153 .214	.189 .249
TP (mm)	-.032 .455	-.048 .402	-.004 .495
CY (thousand hectares)	-.779** .000	-.752** .000	-.875** .000
GCR(%)	-.166 .277	-.010 .480	-.125 .329
YRC (hours)	.264 .171	.219 .128	.311 .130

variables that have influence on the dependent variables. For the accuracy of correlation analysis results, Pearson, Kendall and Spearman correlation coefficients are used to analyze each variable.

The results obtained by the above three correlation analysis methods are basically the same, which shows that each factor has different influence degree on the dependent variable, and the influence effect is also different. Therefore, in order to reduce the problem of inaccurate prediction results when the nonparametric research dimension is large, this thesis eliminates the variables with low and insignificant influence, and uses the variables with significant residual influence to study.

3.3 Nonparametric Regression

Since the factors affecting urban water consumption are complex and varied, and the influence degree, method and effect of each variable are different, it is easy for ordinary parametric models to have larger "specification error" when fitting and even predicting later. For example, if the real population is not normal, or even far away from normal, there may be a big deviation if statistical inference is made by using the model with normal hypothesis premise. That is to say, the parametric model is highly dependent on the model setting, so it may not be robust enough. Therefore, this thesis uses nonparametric regression method to simulate the research object, and compares the results with those of stepwise regression method.

Providing y is a to-be-explained variable, a random variable, and x is a k dimensional explanatory vector, representing a number of important factors. It can be either a deterministic variable or a random one. Now consider the following nonparametric multivariate regression model:

$$y_i = m(x_i) + \varepsilon_i = m(x_{1i}, x_{2i}, \ldots, x_{ki}) + \varepsilon_i \tag{2}$$

where $\varepsilon_i (i = 1,2, \ldots, n)$ is a random error term, $m(\cdot)$ is an unknown multivariate function, which reflects the influence of other observable or unobservable factors on the explanatory variables besides the to-be-explained variables, and the specification error of the model.

The difficulty (and advantage) in estimating the model (2) is that $m(\cdot)$ is an unknown function (even the form of the function is unknown). Because of this, this model can better reflect the real regression relationship. The idea of nonparametric regression is to estimate $m(x_i)$ respectively for each $i(i = 1,2\ldots, n)$, so as to get the estimation of regression function $m(x)$. In a sense, we are not looking for the analytical solution to $m(x)$, but for its numerical solution.

The kernel regression estimator of multivariate nonparametric model at x_0 can be written as:

$$\hat{m}(x_0) = \frac{\sum_{i=1}^{n} K[(x_i - x_0)/h] y_i}{\sum_{i=1}^{n} K[(x_i - x_0)/h]} \tag{3}$$

where $K(\cdot)$is $k$dimensional kernel function. The properties of multivariate kernel regression estimators are similar to those of univariate kernel regression. However, the optimal bandwidth is $h^* = O/(n^{-1/(k+4)})$ (greater than the optimal bandwidth in univariate case), and the speed of $\hat{m}(x_0)$ convergence to the true value is slower. The more explanatory variables of multiple regression are, the slower the convergence speed and the greater the demand for sample size are.

This is also the reason for using multiple correlation analysis to eliminate variables with little influence before doing multivariate nonparametric regression.

Table 5. Non-parametric regression results

Local-linear regression	Number of obs =15
Kernel: gaussian	E (Kernel obs) = 15
Bandwidth: improved AIC	R-squared= 0.9974
TWC	Estimate
Mean TWC	122375.7
Effect IWC	1.028123
WCP	21.86301
PWR	.0188933
ASTC	.3668104
GRP	12.74128
GDPpc	-.9639916
AT	-3305.056
AMINT	5568.38
CY	-2551.766

In order to improve the results of nonparametric regression, this thesis uses the minimized improved AIC criterion to calculate the bandwidth.

It can be seen from table 5 that the value of R² is 0.9974, which shows that the fitting effect of the model is better.

3.4 Fitting Effect Comparison

The fitting results of the data are as follows:

Table 6. Comparison of fitting results

years	actual value	Multiple stepwise regression	Stepwise regression deviation	Nonpara metric fit	Non-parametric deviation
2005	118875	120913.6	1.71%	118872.4	0.00%
2006	117604	116415.2	-1.01%	117576.8	-0.02%
2007	104760	105102.9	0.33%	104792.8	0.03%
2008	105129.64	106510.7	1.31%	106468.9	1.27%
2009	108735	106339.1	-2.20%	107052.2	-1.55%
2010	112326	111285	-0.93%	111788.6	-0.48%

2011	118862	118128.7	-0.62%	119031.5	0.14%
2012	121401	123276.5	1.54%	120834.5	-0.47%
2013	126656.16	125148.3	-1.19%	126464.1	-0.15%
2014	122404.08	121719.3	-0.56%	122833.1	0.35%
2015	125255.12	126784.5	1.22%	124816	-0.35%
2016	132652.26	131879.1	-0.58%	132561.2	-0.07%
2017	135182.94	136379.3	0.88%	135685.8	0.37%
2018	133978	137614.8	2.71%	133998	0.01%
2019	152953	149277.2	-2.40%	152860.3	-0.06%

It can be seen from the fitting data in Table 5 that the fitting value of the nonparametric regression model is closer to the actual value than that of the stepwise regression model, which indicates that the nonparametric model is better than the multivariate stepwise regression model in dealing with the complicated and changeable problems of urban water consumption prediction. For further explanation, the mean absolute error E_{MAE} is compared with the mean square error E_{MSE}, and the results are as follows. The calculation formulas of E_{MAE} and E_{MSE} are as follows:

$$E_{MAE} = \frac{1}{n}\sum_{i=1}^{n} | TWC_i - T\hat{W}C_i | \tag{4}$$

$$E_{MSE} = \sqrt{\frac{\sum_{i=1}^{n} (TWC_i - T\hat{W}C_i)^2}{n}} \tag{5}$$

Table 7. Comparison of mean absolute error and mean square error

Statistical indicators	**Stepwise polynomial regression**	**Nonparametric regression**
E_{MAE}	1600.08	408.3253333
E_{MSE}	1867.64971	629.7953032

It can be seen from Table 6 that the mean absolute error and mean square error of nonparametric estimation are much smaller than those of stepwise regression, and the fitting degree of nonparametric estimation is higher than that of stepwise polynomial regression.

The stepwise regression model and nonparametric regression model are used to do out-of-sample prediction on the actual water consumption in 2020, and the actual and predicted values are shown in the table below. The results show that the predicted value of nonparametric regression model is better than that of stepwise regression model.

Table 8. Out-of-sample predicted values

years	Stepwise regression prediction	Nonparametric reprediction
2020	153641.38	158463.27

4 Conclusions and Policy Suggestions

Based on the statistical data of Nanjing city in Jiangsu province from 2005 to 2019, this thesis analyzes the correlation between urban annual water consumption data and its long-term influencing factors, selects the variable data which has obvious influence on total water consumption, uses the method of stepwise regression and nonparametric regression to show the advantages and disadvantages of the model in data fitting, predicts the data respectively, and draws the following conclusions:

Empirical results show that the fitting effect of nonparametric model is better than that of traditional multivariate stepwise linear regression model in annual water consumption fitting. In the prediction problem, the effect of kernel estimation is also obvious. Therefore, in the long-term prediction, we can use the more accurate kernel estimation method to predict the urban annual water consumption, which can provides supports for the long-term stable urban development, national water conservancy projects and water resources scheduling problems, so as to achieve the purpose of saving economy and rationally allocating and utilizing water resources.

References

[1] Zhu Ping. Application of Multivariate Linear Regression Model in Beijing Water Consumption Prediction[J]. Science and Education Guide (First Ten-day Periodical), 2015(**34**):167-168+190.

[2] Niu Ruiwen. Urban Hourly Water Consumption Prediction Based on Grey Elman Neural Network [J]. Fujian Building Materials, 2018(**02**): 1-3.

[3] Chang Shuling, You Xueyi. Study on Tianjin Water Demand Prediction[J]. Resources and Environment in Arid Areas, 2008,**22** (2): 14-19.

[4] Wang Yongling. *Prediction Calculation Method*[M]. Beijing: Science Press, 1986.

[5] Zhou Pengfei, Lu Zeyu. Prediction of Urban Water Consumption Based on SPSS Multivariate Linear Regression Model[J]. Water Conservancy Science and Economics, 2018,**24** (05): 6-10.

[6] Sun Aimin. Multivariate Linear Regression Prediction of Water Demand in Xi'an City[J]. Value Engineering, 2020,**39**(13):42-45.

[7] Li Shangying, Dong Ye. Research and Prediction of the Relationship among Population, Water Consumption and Economy in Xinjiang Province[J]. Water Resources Development and Management, 2020(03):35-43.

[8] Sun Liqin, Chang Anding, Wei Longhu, Wei Weiping. Based on GM ——(1,1) Prediction of Water Consumption by Stepwise Regression Model [J]. Statistics and Decision, 2016(20):95 -97.

[9] Pan Zhongjian. Prediction of Agricultural Water Demand Based on BP-AGA Method[J]. Modern Computer (Professional Edition), 2016(29): 33-35 +40.

[10] Yan Xu, Li Siyuan, Zhang Zheng. Application of BP Neural Network Based on Genetic

Algorithm in Cities [J]. Computer Science, 2016,**43**(S2):547-550.

[11] Shi Lizhong, Xi Lu. Analysis and Prediction of Water Consumption Per 10,000 Yuan GDP in Liaoning Province Based on Grey System Theory[J]. Journal of Shenyang University (Natural Science Edition), 2017,**29**(05):380-383.

[12] Isadila Wangsufu, Anwar Mohammad Ming. Prediction of Water Consumption in the Process of Urbanization in Kashgar, Xinjiang Province[J]. Journal of Glaciology and Permafrost, 2017, **39**(03):688-695.

[13] Zhang Xuan. Annual Water Consumption Prediction of Zhengzhou Based on PCA-BP Neural Network[J]. Science and Technology Innovation, 2020(28):107-111.

[14] Wu Fengbo, Zhao Pan, Lu Qiantong. Urban Water Consumption Prediction Based on GA-BP Neural Network[J]. Modern Electronic Technology, 2020,**43**(08):147-150.

[15] Chen Jiatong, Wen Lishu, Tan Yaxin. National Water Consumption Prediction Based on Grey Prediction and Elman Neural Network[J]. Jiangxi science, 2018,**36**(06):961-967.

[16] Ren Aihong, Chen Zhanbo. Research on Nonparametric Model of Urban Daily Water Consumption Prediction[J]. Journal of Southwest University for Nationalities (Natural Science Edition), 2007(04):767-771.

[17] Chen Zhanbo, Zhang Desheng, Han Youwang, Zhang Jun. Non-parametric Model for Prediction of Urban Daily Water Consumption[J]. Journal of Qingdao University of Science and Technology (Natural Science Edition), 2007(01):65-68.