Environmental Engineering Research

Environ. Eng. Res. 2010 September,15(3) : 135-140

Research Paper

DOI:10.4491/eer.2010.15.3.135
pISSN 1225-1025 eISSN 2005-968X

Water Demand Forecasting by Characteristics of City Using Principal Component and Cluster Analyses

Taeho Choi[1], **Oeun Kwon**[2], **Jayong Koo**[1†]

[1]Department of Environmental Engineering, University of Seoul, Seoul 130-743, Korea
[2]Korean Intellectual Property Office, Daejeon 302-701, Korea

Abstract

With the various urban characteristics of each city, the existing water demand prediction, which uses average liter per capita day, cannot be used to achieve an accurate prediction as it fails to consider several variables. Thus, this study considered social and industrial factors of 164 local cities, in addition to population and other directly influential factors, and used main substance and cluster analyses to develop a more efficient water demand prediction model that considers unique localities of each city. After clustering, a multiple regression model was developed that proved that the R^2 value of the inclusive multiple regression model was 0.59; whereas, those of Clusters A and B were 0.62 and 0.74, respectively. Thus, the multiple regression model was considered more reasonable and valid than the inclusive multiple regression model. In summary, the water demand prediction model using principal component and cluster analyses as the standards to classify localities has a better modification coefficient than that of the inclusive multiple regression model, which does not consider localities.

Keywords: Cluster analysis, Multiple regression model, Principal component analysis, Water demand prediction

1. Introduction

The water supply service is used by all people living in a city as it is an important facility for maintaining everyday life and all other activities within urban areas. Today, 100 years after the first introduction of the water supply service, the rate of penetration of this service has almost approached 100% and; thus, is regarded as a basic requisite of urban life and the most fundamental service of the government for the people. In addition, it is expected that quantitative and qualitative demands for water will further increase as the future standard of living and economy are raised [1].

As the "Water Supply Facility Standards" of South Korea clearly state, the purpose of a water supply facility is to supply qualitatively safe and quantitatively sufficient water at a reasonable construction and maintenance expense. An impending task of the current water supply service is to distribute and supply the best quality and quantity of water to meet the customer's needs. Therefore, the current and future water demands should be accurately estimated and predicted to meet these needs. Thus, the amount of water consumption should be accurately predicted as a prerequisite for the stable and efficient operation of a water supply service facility.

Since the on-set of water supply services, the conventional estimation and prediction of the amount of long-term water consumption has principally been carried out based on the average liter per capita day, which was not of major concern in the past when high-density urbanization and regional characteristics were not commonplace [2]. Currently; however, due to many changes, such as many urban redevelopments, diversification of regional characteristics and changes in housing, consumption and life patterns, the previous method for the estimation of water demand, based on the average liter per capita day of entire cities, can not fully account for regional characteristics, which leads to large errors in the actual amount of water consumption as well as the over- or under-design of the capacity of the water supply service facilities across regions, which makes the optimal operation and reasonable maintenance/management of the facility difficult, even after its construction [3].

Currently in Korea, only one single factor, population, is used in analyzing long-term water demands, which causes many problems. To solve these problems, there is a need for research on the various factors that may affect water consumption to assess other affecting factors, other than the population factor, for the purpose of diversification [4, 5].

Therefore, in this study, a comprehensive analysis on the amount of water consumption in nationwide urban areas, with the factors that can characterize the urban areas, was conducted using both principal component and cluster analyses, which are widely applied for the classification of regional characteristics.

Received November 20, 2009 **Accepted** August 03, 2010

[†]Corresponding Author
E-mail: jykoo@uos.ac.kr
Tel: +82-2-2210-2946 Fax: +82-2-2244-2245

In addition, this study proposed a new equation for the prediction of water demand, which may be acceptable for the water consumption patterns for all regional characteristics across all urban area groups. This study also aimed to improve the previous predictive equation for use in the future basic modification plan for the water supply system and in the evaluation index of the water supply system.

2. Methods

The current study conducted a multiple regression analysis, where the factors affecting water consumption were used as the independent variables to account for the dependent variables from 164 nationwide local self-governing regions. Each independent variable was evaluated to determine which significantly affected water consumption. A comprehensive multiple regression model was developed and compared with the multiple regression model, which was developed after clustering of their goodness of fit using the observed values [6, 7].

However, the inter-independence of the selected factors could not be guaranteed. The data that were not inter-independent may cause erroneous results in the subsequent analysis, which; in turn, reduces the reliability of the subsequent analysis. Therefore, the data need to be made inter-independent; that is, a principal component analysis was performed [8].

One of the purposes in this study was to account for the regional characteristics in water demand estimations. Thus, a step for classifying the study target areas by their regional characteristics was required, i.e. a clustering analysis. Clustering analysis was conducted to classify the target areas into regional clusters that had similar impacts on the independent variables. The social and industrial factors were also taken into account in this clustering process [5].

After clustering the self-governing regions via this process, multiple regression models for each of the clustered self-governing regions were developed. By comparing these models with the comprehensive model, a model reflecting the social and industrial characteristics of each of the self-governing regions was developed.

2.1. Principal Component Analysis

A principal component analysis is a form of factor analysis, where the correlations between multiple variables are analyzed to account for the common underlying dimension across all factors. The purpose of a principal component analysis is to derive a comprehensive index, which combines multiple indices, and to divide the observation targets into a few groups in order to recognize the value of the data for a multiple regression analysis or discriminant analysis from different perspectives [1, 9].

Although a regression analysis assumes that the variables are independent of each other, the observed variables are not often completely independent, even though they are the variables affecting the dependent variable, since they usually do affect each other. In a principal component analysis, a new variable ensuring inter-independence, named a ‘principal component’, is derived using a linear combination of variables in order of the size of their accountability to the variance of the original variables. Therefore, there are no independent or dependent variables in a principal component analysis, with its primary purpose being to find a common foundation for the variables by analyzing their correlations.

Thus, this study used a principal component analysis to determine the principal components from the social and industrial factors, which were later used in the water demand estimation.

2.2. Clustering Analysis

A clustering analysis refers to a grouping method where the objects of interest are classified into a few clusters in a way that the objects in same cluster are similar, while those in different clusters are dissimilar. A clustering analysis makes no assumption on the number of clusters and their structure, but forms the cluster based only on the similarity or distance between the objects. The major purpose of a clustering analysis is to identify the characteristics of the objects and the relationships among the clusters. In other words, a clustering analysis is a technique that can be used when there are no or unknown classification criteria. When P variables are measured with N of objects, a clustering analysis measures the similarity or dissimilarity between the N variables to connect the object based on their closer distance. Therefore, this study analyzed 164 self-governing regions, using a clustering analysis, to group the self-governing regions with similar water demand characteristics into the same cluster, with the estimation equation then calculated for the water demand [1, 9].

2.3. Data Collection

The present study was conducted using 164 self-governing regions in Korea. Although there have been many studies on the development of models for the estimation of the water demands in specific cities or regions, there have also been many difficulties associated with the studies on all regions in Korea, such as difficulty in choosing the factors affecting the amount of water consumption in an individual self-governing region and the limitation in the practical value using a comprehensive model for the estimation of the water demands, including accuracy problem.

Therefore, the present study extracted the water supply service-related information for 164 self-governing regions from the Water Supply Service Statistics (2006) to determine the influential independent variables and collect the statistics for social/industrial factors via a DB search from Statistics Korea.

2.4. Selection of Factors

The amount of water consumption in urban areas is a direct index representing changes in the demand for water use, which tends to be parallel with the population size, but is largely influenced by the characteristics and location of the city, living standard, life style and socioeconomic conditions. Therefore, this study aimed to examine the relationship between the level of water consumption and the social/industrial factors, which may reflect the regional characteristics as well as the direct affecting factors, such as population and water supply service [10].

The social factors included the water-supply population and the size of the administrative area (1,000 persons/km^2) as the density index for water demand, as well as the number of independent houses/total households, and the number of apartment households/total households as the index of the amount of water consumption across housing types. In addition, this study adopted the ratio of beneficiaries for National Basic Livelihood to investigate the correlation between the standard of living and the level of water consumption.

The industrial factors included the number of businesses/the unit administrative area (No./km^2), the number of factories/the unit administrative area (No./km^2), the number of employees in the industrial sector/the residential population of the area per total population (person/person) as the index for the impact of the number of businesses and factories on the level of water consumption, as well as the degree of financial independence (%) as the index of financial ability related to urban development.

3. Results and Discussion

3.1. Development of Comprehensive Multiple Regression Model

To determine which factors will affect water consumption among the social and industrial factors anticipated to do so, this study used a multiple regression model to investigate the correlation between the level of water consumption and the affecting factors.

The R^2 value (coefficient of determination) from a multiple regression equation is used to explain the correlation between independent and dependent variables, which increases with the addition of new independent variables. Generally speaking, the accountability of a regression equation increases with increasing R^2 value. However, it can be inefficient to add more independent variables only to slightly increase the R^2 value. To address this problem, an adjusted R^2 value is used, where the value of the adjusted-R^2 decreases, but the R^2 value increases slightly when a certain variable to be added has very little additional accountability. In this type of case, it is inappropriate to include the variable as an independent variable.

Therefore, this study also adopted adjusted R^2 values in developing an appropriate multiple regression model for the independent variables.

In the multiple regression analysis, the dependent variable was the level of daily water consumption of each self-governing region, while the independent variables were the 6 factors, with the exceptions of the proportion of apartment households and number of businesses in a unit area, which were excluded to solve the multicollinearity problem.

Table 1. Coefficient value (B) and standardized coefficient (β) in the comprehensive multiple regression model

Name of factor	B	(β)
(Constant)	-706043.7	
Water supplying population density (X_1)	91850.7	0.769
Ratio, independent housings (X_2)	443772.8	0.406
Ratio, households in apartment (X_3)		
Ratio, beneficiaries of National Basic Livelihood (X_4)	1556112.6	0.131
No of firms per unit area (X_5)		
No of factories per unit area (X_6)	-19199.4	-0.426
Rate of employees in industrial sector (X_7)	412349.3	0.091
Degree, financial independence (X_8)	11382.9	0.704

The results from the multiple regression analysis gave an adjusted R^2 of 0.59, which was slightly lowered than the R^2 value of 0.605. However, the significance level was 0.000, indicating that this model was significant. The coefficients (B) and standardized coefficients (β) of the comprehensive multiple regression model are presented in Table 1.

Conversely, the R^{2}calculated from the multiple regression model for the classification into administrative territory showed that the R^2 values for special metropolitan cities and Jeju Special Self-Governing Province was 0.935, indicating that the observed and predicted values were a good fit. However, the R^2 values for other cities and 'Gun' (similar to county) were 0.397 and 0.04, respectively, indicating that the comprehensive multiple regression model was not sufficient to represent the observed value.

To solve this problem, a new model for water consumption needs to be developed to reflect the social and industrial characteristics of the self-governing regions after their grouping together, rather than using the comprehensive multiple regression model for the water consumption for all 164 regions.

3.2. Analysis of the Affecting Factors and Classification of Self-Governing Regions through Statistical

3.2.1. Analysis of Affecting Factors through Principal Component Analysis

The principal component analysis in this study used the 4 social factors and 4 industrial factors from the 164 self-governing

Table 2. Total variance explained by the principal components against the social-industrial factors

Component	Initial Eigen value		
	Total	% Variance	% Accumulation
1	4.978	62.219	62.219
2	1.574	19.677	81.896
3	0.712	8.899	90.795
4	0.324	4.054	94.849
5	0.236	2.956	97.805
6	0.151	1.893	99.698
7	1.458E-02	0.182	99.880
8	9.580E-03	0.120	100.000

Table 3. Accumulative factor by each principal component (social/industrial factors)

Name of factor	Component	
	1	2
Ratio, households in apartment (X_3)	0.878	0.311
Ratio, independent housings (X_2)	-0.868	-0.405
Ratio, beneficiaries of National Basic Livelihood (X_4)	-0.851	-0.275
Degree, financial independence (X_8)	0.762	0.532
Rate of employees in industrial sector (X_7)	0.684	-0.125
Water supplying population density (X_1)	0.197	0.960
No of firms per unit area (X_5)	0.186	0.948
No of factories per unit area (X_6)	0.184	0.834

Table 4. Status of local bodies by cluster according to the results of the clustering analysis

Cluster	Local bodies
A	Dongducheon, Gwacheon, Guri, Osan, Uiwang, Hanam, Icheon, Anseong, Gimpo, Gwangju, Yangju, Pocheon, Yeoju, Yeoncheon, Gapyeong, Gangneung, Donghae, Taebaek, Sokcho, Samcheok, Hongcheon, Hoengseong, Yeongwol, Pyeongchang, Jeongseon, Cheorwon, Whacheon, Yanggu, Inje, Goseong, Yangyang, Chungju, Jecheon, Cheongwon, Boeun, Okchun, Youngdong, Jincheon, Goesan, Eumseong, Danyang, Jeungpyeong, Gongju, Boryeong, Asan, Seosan, Nonsan, Gyeryong, Geumsan, Yeongi, Buyeo, Seocheon, Cheongyang, Hongseong, Yesan, Taean, Dangjin, Jeongeup, Namwon, Gimje, Wanju, Jinan, Muju, Jangsu, Imsil, Sunchang, Gochang, Buan, Naju, Gwangyang, Damyang, Gokseong, Gurye, Goheung, Boseong, Hwasun, Jangheung, Gangjin, Haenam, Yeongam, Muan, Hampyeong, Yeonggwang, Jangseong, Wando, Jindo, Shinan, Gimcheon, Andong, Yeongju, Yeongcheon, Sangju, Mungyeong, Gunwi, Uiseong, Cheongsong, Yeongyang, Youngduk, Cheongdo, Goryeong, Seongju, Chilgok, Yecheon, Bonghwa, Uljin, Ulleung, Jinhae, Tongyeong, Sacheon, Miryang, Geoje, Uiryeong, Haman, Changnyong, Goseong, Namhae, Hadong, Sancheong, Hamyang, Geochang, Hapcheon (121)
B	Gwangju (Jeonnam), Daejeon, Ulsan, Suwon, Seongnam, Uijeongbu, Anyang, Bucheon, Gwangmyeong, Pyeongtaek, Ansan, Namyangju, Siheung, Gunpo, Yongin, Paju, Hwaseong, Yangpyeong, Chuncheon, Wonju, Cheongju, Cheonan, Jeonju, Gunsan, Iksan, Mokpo, Yeosu, Suncheon, Pohang, Gyeongju, Gumi, Gyeongsan, Changwon, Masan, Jinju, Gimhae, Yangsan, Jejudo (39)
C	Busan, Daegu, Incheon (3)
D	Seoul (1)

Table 5. Coefficient value (B) and standardized coefficient (β) in multiple regression model by cluster

Cluster	Name of factor	B	(β)
A	(Constant)	56701.8	
	Water supplying population density (X_1)	-2459.6	-0.099
	Ratio, independent housings (X_2)	-52681.5	-0.697
	Ratio, households in apartment (X_3)		
	Ratio, beneficiaries of National Basic Livelihood (X_4)	46089.9	0.065
	No of firms per unit area (X_5)		
	No of factories per unit area (X_6)	905.2	0.079
	Rate of employees in industrial sector (X_7)	-34910.9	-0.127
	Degree, financial independence (X_8)	363.0	0.241
B	(Constant)	-389111.6	
	Water supplying population density (X_1)	-6926.6	-0.232
	Ratio, independent housings (X_2)	306618.9	0.433
	Ratio, households in apartment (X_3)	87689.7	0.113
	Ratio, beneficiaries of National Basic Livelihood (X_4)	4407723.9	0.588
	No of firms per unit area (X_5)		
	No of factories per unit area (X_6)	839.3	0.094
	Rate of employees in industrial sector(X_7)	-687462.1	-0.491
	Degree, financial independence(X_8)	11127.9	1.572
C	(Constant)	568883.7	
	Water supplying population density (X_1)		
	Ratio, independent housings (X_2)		
	Ratio, households in apartment (X_3)		
	Ratio, beneficiaries of National Basic Livelihood (X_4)	-4769292.5	-0.207
	No of firms per unit area (X_5)		
	No of factories per unit area (X_6)	71547.6	0.985
	Rate of employees in industrial sector (X_7)		
	Degree, financial independence (X_8)		

regions for 2006. As shown in Table 2, the 2 principal components of the socio-industrial factors from the principal component analysis accounted for 82% of the total variances.

As shown in Table 3, the first component showed an accumulation frequency in the order; highest to lowest: apartment ratio, independent housing ratio, ratio of beneficiaries for National Basic Livelihood and the number of industrial employees, while the second component showed that an accumulation frequency in the order: ratio of water-supply population density, the number of businesses in a unit area and the number of factories in a unit area.

3.2.2. Segmentation of Local Bodies through Clustering Analysis

Although the above comprehensive multiple regression model, based on the 6 independent variables, may predict the comprehensive water consumption pattern of the 164 self-governing regions, it was adjudged to have limitation in predicting the water consumption of the self-governing regions that have exceptional regional characteristics (e.g. rural areas or newly emerging industrial complex areas). Therefore, this study aimed to appropriately classify the 164 self-governing regions by representing their social and industrial status via the clustering analysis, and develop a multiple regression model for each cluster. The clustering was conducted around the 2 components derived from the prior principal component analysis, the results of which are shown in Table 4.

3.3. Development of Multiple Regression Model for Each Cluster Using Independent Variables

A multiple regression model for each cluster derived from the clustering analysis was developed to predict the level of water consumption after solving the multicollinearity problem. Unlike the previous comprehensive multiple regression model with the adjusted R^2 0.59, the adjusted R^2 of the multiple regression model for cluster A was increased to 0.62 after clustering, while that for cluster B was 0.74. For cluster C, which included only 3 self-governing regions; however, there might be a limitation because the multiple regression model was developed using the data from 2006 only. In addition, the model was developed with the exclusion of variables from the multiple regression analysis other than the ratio of beneficiaries for National Basic Livelihood and the number of factories in a unit area. Furthermore, Seoul was classified into D cluster, which appeared markedly differentiated from other clusters in terms of the level of water consumption and social/industrial factors. Thus, a separate multiple regression model needs to be developed for cluster D only, using multi-year data to consider their time-serial characteristics.

Table 6. Coefficient value (B) and standardized coefficient (β) in multiple regression model based on the principal components and clustering

Cluster	Name of factor	B	(β)
A	(Constant)	30944.7	
	Principal component 1 (X_{10})	14788.5	0.792
	Principal component 2 (X_{11})	18022.5	0.293
B	(Constant)	111336.1	
	Principal component 1 (X_{10})	47687.9	0.329
	Principal component 2 (X_{11})	32009.7	0.475
C	(Constant)	-1204655.8	
	Principal component 1 (X_{10})	1486611.1	1.788
	Principal component 2 (X_{11})	757570.2	2.369

Table 5 shows the coefficient value (B) and standardized coefficient (β) of the multiple regression model for each cluster.

3.4. Development of Multiple Regression Model for Each Cluster Using Principal Components

A multiple regression model was developed with the 2 variables derived from the principal component analysis. Table 6 shows the coefficient value (B) and standardized coefficient (β) of the multiple regression model for the principal component of each cluster. The adjusted R^2 value for clusters A and B were 0.56 and 0.12, respectively.

3.5. Determination of Optimal Model for Predicting Water Demand

As described above, 3 multiple regression models were developed to predict the water demands in 164 self-governing regions; they were a comprehensive multiple regression model, a multiple regression model for each cluster and a multiple regression model based on principal components and clustering. To solve the multicollinearity problem, the models were constructed after eliminating the independent variables with high correlations with others.

Consequently, the clustering multiple regression model best predicted the observed value, and the multiple regression model based on principal components and clustering had a limitation in efficiently expressing the water consumption patterns for the 164 self-governing regions with only the 2 principal components.

4. Conclusions

This study aimed to diversify the factors affecting water consumption by taking into account the regional characteristics using a statistical approach employing principal component and clustering analyses for predicting water demands. A further aim was to develop a more effective water demand estimation model by supplementing with statistical objectivity. The study findings were as follows:

Although the correlations between the independent variables selected from the social and industrial factors for predicting the level of water consumption of the 164 self-governing regions were high, the correlations between the independent variable and the level of water consumption were low. Thus, due to the confounding regional characteristics, there was a need to group the regions to reflect the characteristics of the independent variables.

In addition, the results from the principal component analysis showed that a social factor, i.e. water supply population density, accounted for 76% of the total variance. As for the industrial factors, two factors, i.e. the number of businesses per unit area and the number of factories per unit area, accounted for 85% of the total variance, which allowed this study to test the importance of the determined independent variables via the standardized coefficients of the comprehensive multiple regression model.

The results from the clustering of the 164 self-governing regions via the clustering analysis showed that the clustering based on the level of water consumption better reflected the regional characteristics than the clustering based on the 8 independent variables, which enabled this study to derive significant clustering results. In addition, the correlation analysis between the level of water consumption and independent variables, which was conducted after clustering of the self-governing regions into 4, showed that the correlation coefficients increased more than those before clustering, which suggests that the clustering in this study increase the correlation between the level of water consumption and the independent variables.

This study generated 2 principal components from the principal component analysis of the social and industrial factors, and developed 3 multiple regression models after clustering of the 164 self-governing regions into 4 clusters via a clustering analysis; the comprehensive multiple regression mode, a clustering multiple regression model and a principal component/clustering multiple regression model. Of the 3 developed models, the clustering multiple regression model clustered around regional characteristics was the found to be the most appropriate. For regions C (Busan, Daegu and Incheon) and D (Seoul); however, there was a limitation in accurately identifying the extent of errors associated with the model because the multiple regression models for those 2 regions were developed based on the data for a single year (2006).

In conclusion, the developed models were adjudged to have sufficient validity in terms of their logical aspects. The developed models may be used as basic data for the planning of water supply facilities, as well as their maintenance, by estimating the level of water consumption. The models are also expected to have economical efficacy due to the saving of water resource and reduced construction and maintenance costs when used for the long-term planning and optimal operation of water supply systems, such as filtration plants, water reservoirs and water pipes/drainpipes/water service pipe. However, it is suggested that more accurate and extensive data should be acquired and further details of the factors affecting water demands taken into account to develop a more scientific and reasonable model for the estimation of water demands.

Acknowledgements

This research was supported by a grant (4-2-3) form the Sustainable Water Resources Research Center of the 21st Century Frontier Research Program.

References

1. Shim MH. Estimating long-term water demand by principal component and cluster analysis in central Seoul [dissertation]. Seoul: University of Seoul; 2005.
2. Arbues F, Garcia-Valinas MA, Martinez-Espineira R. Estimation of residential water demand: a state-of-the-art review. *J. Soc. Econ.* 2003;32:81-102.
3. Kim YS. A study on estimation of domestic water demand by actual survey data [dissertation]. Seoul: University of Seoul; 2007.
4. Reynaud A. An econometric estimation of industrial water demand in France. *Environ. Resource Econ.* 2003;25:213-232.
5. Pyon SS. Long-term water demand forecasting using the system dynamics model [dissertation]. Seoul: University of Seoul; 2003.
6. Billings RB, Jones CV. Forecasting urban water demand. Denver: American Water Works Association; 1996.
7. Forrester JW. System dynamics as an organizing framework for pre-coiiege education. *Syst. Dynam. Rev.* 1993;9:183-194.
8. Morell I, Gimenez E, Esteller MV. Application of principal components analysis to the study of salinization on the Castellon Plain (Spain). *Sci. Total Environ.* 1996;177:161-171.
9. Grove DM, Roberts CA. Principal component and cluster analysis of 185 large towns in England and Wales. *Urban Stud.* 1980;17:77-82.
10. Cochran R, Cotton AW. Municipal water demand study, Oklahoma City and Tulsa, Oklahoma. *Water Resour. Res.* 1985;21:941-943.