Determinants of domestic water demand for the Beijing region

D. Karimanzira, M. Jacobi & C. Ament
Institute of Information Technology and Automation,
Department of System Analysis, Ilmenau, Germany

Abstract

The analysis of demand for water, including realistically forecasting future levels of demand, is an important and critical step in the economic analysis of water supply projects. The results of demand analysis will enable to determine the service levels to be provided, determine the size and timing of investments, estimate the financial and economic benefits of projects, and assess the ability and willingness to pay of the project beneficiaries. Furthermore, the surveys carried out during the demand assessment will provide data on cost savings, willingness to pay, income and other data needed for economic analysis. In this paper, methods of statistical analysis (correlation, regression, etc) will be used to determine the factors, which influence water demand. Each model region may have its own set determinants for domestic water demand and the importance of a given factor may vary from one region to another. Therefore, this paper focuses on the major determinants of domestic water demand for the Beijing region. Several models analyzing the determinants were compared. The models based on a feasible generalized least squares (FGLS) analysis.
Keywords: decision support system, water demand models, statistical analysis, correlation and significance tests, regression analysis.

1 Background

Increasing population growth and the associated process of urbanization in the semi arid city of Beijing, China requires a reliable source of water. Although the city currently has an inexpensive and abundant supply of water, it is imperative that the city faces the challenge associated with providing safe drinking water. This work is part of the Chinese – German joint project "Towards Water-

WIT Transactions on Ecology and the Environment, Vol 103,
www.witpress.com, ISSN 1743-3541 (on-line)
doi:10.2495/WRM070111

Scarcity Megalopolis' Sustainable Water Management System"[6]. This project takes the challenge of water shortage, the outstanding conflict between water supply and demand. It aims at a decision support system (DSS) for the sustainable development of economics and community in Beijing. An essential requirement for such a DSS is a simulation model of the water resources/supply system. Part of the simulation model is shown in Figure 1.

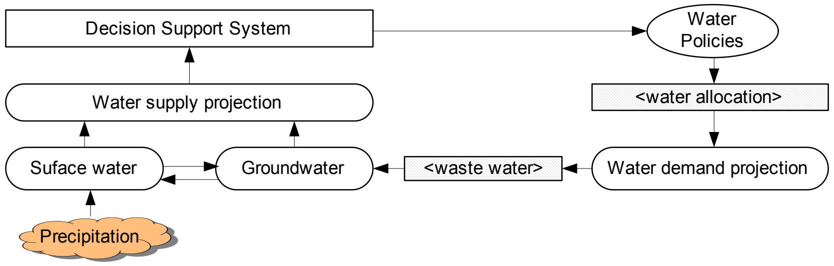

Figure 1: The graphical user interface of the simulation model.

The simulation model comprises the water supply, optimization and the water demand systems. On focus in this paper is part of the water demand system, namely the domestic water demand. Compared to the agricultural and the industrial water demand, the domestic water demand is very difficult to model, because it is determined by several subjective factors.

2 Introduction

Recently, several models for domestic water forecasting have been developed and published in literature. The model from Archibald is a simple component model for calculating the domestic water demand per household. One of the components in the model is "bathing and showering" [2]. The problem with this type of model is that every component has several factors, which should be determined and in most applications there is not enough information to every component and if one component cannot be modelled the whole model would be insufficient. Eqns (1) and (2) describe some of the most used models for domestic water demand forecasting.

$$\Phi_{dwd} = \Phi_{pop} + \eta\Phi_{gdpc} \left[m^3 / Jahr\right] \tag{1}$$

$$DSWI = DSWI_{\min} + DSWI_{\max}\left(1 - e^{-\gamma_d GDPC^2}\right) \tag{2}$$

Eqn (1) describes the IMPACT-Model [1], where ϕ_{dwd} is the growth rate of the domestic water demand, ϕ_{pop} is the growth rate of the population and ϕ_{dwd} is the growth rate of the gross domestic product per capita. Eqn (2) expresses the WATERGAP-Model [4], where $DSWI$ is the domestic structural water intensity and γ_d is the curve retardation.

After analysing several models in literature [1–4] a general model is obtained, which can be expressed as follows:

$$W = f\left(W_{prev}, p, y, d, g, k, v, gdpc, prec, E, Eva, Inc, fam\right) \qquad (3)$$

where W is the domestic water demand, W_{prev} is the previous domestic water demand, p is the water price elasticity, y is the income elasticity, d is the residence density (population), g is the individual preferences (e.g. bathing habit), k is the number of individuals per household, $prec$ is the precipitation, Eva is the evapotranspiration, E is the employment rate and v is the weather.

In this work the influence of the different factors on the domestic water demand is analysed using methods of statistical analysis i.e., correlation, regression, multicollinearity and significance tests. In this way we obtain the most important factors, which should be included in every domestic water demand model for reliable results.

It is known that each model region may have its own set of domestic water demand determinants and the importance of a given factor may differ from one project to another. Therefore to prove generality and robustness of the resulting determinants, the same tests were applied to different data sets of different regions of different development status (Canada, Germany).

3 Test for determinants

3.1 Correlation and significance tests

To test the determinants, correlation tests are used in this paper. A correlation describes the strength of an association between variables. An association between variables means that the value of one variable can be predicted, to some extent, by the value of the other. A correlation is a special kind of association: there is a linear relation between the values of the variables. A non-linear relation can be transformed into a linear one before the correlation is calculated. For a set of variable pairs, the correlation coefficient gives the strength of the association. The square of the size of the correlation coefficient is the fraction of the variance of the one variable that can be explained from the variance of the other variable. The relation between the variables is called the regression line. The regression line is defined as the best fitting straight line through all value pairs, i.e., the one explaining the largest part of the variance. The correlation coefficient is calculated with the assumption that both variables are stochastic (i.e., bivariate Gaussian). If one of the variables is deterministic, e.g., a time series or a series of doses, this is called regression analysis. In regression analysis, the interpretation of the correlation coefficient is different from that of correlation analysis. In regression analysis, tests on statistical significance can only be used when the conditional probability distribution of the other variable is known or can be guessed. However, the regression line can still be used.

If the aim is only to prove a monotonic relation, i.e., if one variable increases the other either always increases or decreases, like in most of our cases, then the rank correlation test is a better test.

The normal procedure for performing correlation and significance tests is as follows: First the hypothesis is H_0 made: “The values of the members of the pairs are uncorrelated, i.e., there are no linear dependencies”. It is also assumed that the values of both members of the pairs are normal (bivariate) distributed.

Procedure:

The correlation coefficient r_{xy} of the pairs (x,y) is calculated as:

$$r_{xy} = \frac{\operatorname{cov}(x, y)}{\operatorname{var}(x) \cdot \operatorname{var}(y)} \tag{4}$$

$$r_{xy} = \frac{\frac{1}{N}\sum_k (x(k) - \bar{x}) \cdot (y(k) - \bar{y})}{\sqrt{\frac{1}{N}\sum_k (x(k) - \bar{x})^2} \cdot \sqrt{\frac{1}{N}\sum_k (y(k) - \bar{y})^2}} \tag{5}$$

The regression line y = a * x + b is calculated as:

$$a = \frac{N\left(\sum xy\right) - \left(\sum x\right) \cdot \left(\sum y\right)}{N\left(\sum x^2\right) - \left(\sum x\right)^2} \tag{6}$$

$$b = \frac{\sum y - a\sum x}{N} \tag{7}$$

Level of Significance:

The value of $t = r_{xy} \cdot \mathrm{sqrt}((N-2)/(1-r_{xy}^2))$ has a Student-t distribution with degrees of freedom N-2. If the degrees of freedom are greater than 30, the distribution of t can be approximated by a standard normal distribution.

Remarks:

This could be called the most misused statistical procedure. It is able to show whether two variables are connected. It is not able to show that the variables are not connected. If one variable depends on another, i.e., there is a causal relation, then it is always possible to find some kind of correlation between the two variables. However, if both variables depend on a third, they can show a sizable correlation without any causal dependency between them. A famous example is the fact that the position of the hands of all clocks is correlated, without one clock being the cause of the position of the others. Another example is the significant correlation between human birth rates and stork population sizes. To overcome this problem the spearman rank correlation is used in combination with regression analysis is applied in this paper.

3.1.1 Spearman’s rank correlation test

This is a test for correlation between a sequence of pairs of values [7, 8]. Using ranks eliminates the sensitivity of the correlation test to the function linking the

www.witpress.com, ISSN 1743-3541 (on-line)

pairs of values. In particular, the standard correlation test is used to find linear relations between test pairs, but the rank correlation test is not restricted in this way. Given N pairs of observations (x_i,y_i), the x_i values are assigned a rank value and, separately, the y_i values are assigned a rank. For each pair (x_i,y_i), the corresponding difference, d_i between the x_i and y_i ranks is found. r_{xy} is:

$$r_{xy} = \sum_{i=1}^{N} d_i^{\,2} \qquad (8)$$

For large samples the test statistic is then:

$$z = \frac{6r_{xy} - N\left(N^2 - 1\right)}{N\left(N+1\right)\cdot\sqrt{N-1}}, \qquad (9)$$

which is approximately normally distributed. A further technique is now required to test the significance of the relationship. The z value must be looked up on the Spearman rank significance curves (see Figure 3).

4 Determinants for Beijing domestic water demand

Possible influencing factors for the domestic water demand for the Beijing region were selected and are listed in Figure 2 for the period from 1996-2003 [5]. They are factors from the weather, population and economy, which are thought to be obviously linked to the domestic water demand. The main objective of this study is to find out, which factor has the greatest influence and which ones are negligible in the models, and therefore simplify them.

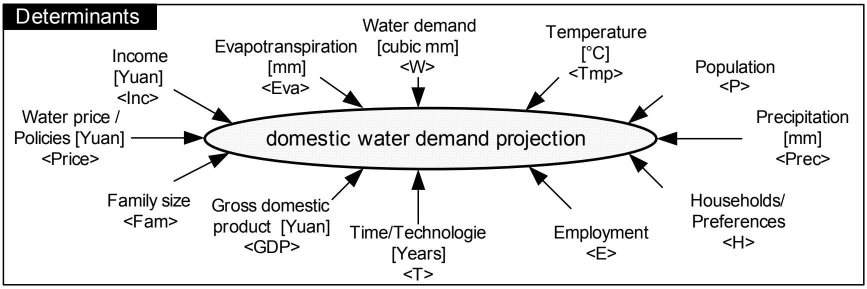

Figure 2: Possible determinants for the domestic water demand for Beijing.

Correlation, regression and significance analysis described in the previous section were performed for the domestic water demand with respect to all input variables including the previous domestic water demand W_{prev}.

5 Results

All estimated cross-correlation coefficients r_{xy} which were significant at the 5% level according to the three-step procedure described above are summarized in Table 1 (bold) and Figure 3. Surprisingly, the magnitude of the total cross correlation coefficient of the domestic water demand and the number of employment is quite large (-0.777). A decreasing domestic water demand by increasing employment cannot be explained logically. The partial correlation coefficient (a measure for the dependence of two variables after switching of the linear influences of other variables) between employment and the domestic water demand confirms this. After switching of the linear influences of other variables the remaining partial correlation coefficient is only -0.065, which practically shows no linear dependence between the two variables. Also unexpected is the minimal correlation of the domestic water demand and the population (correlation coefficient of -0.016). Normally, one would think that the domestic water demand increases with increasing population growth.

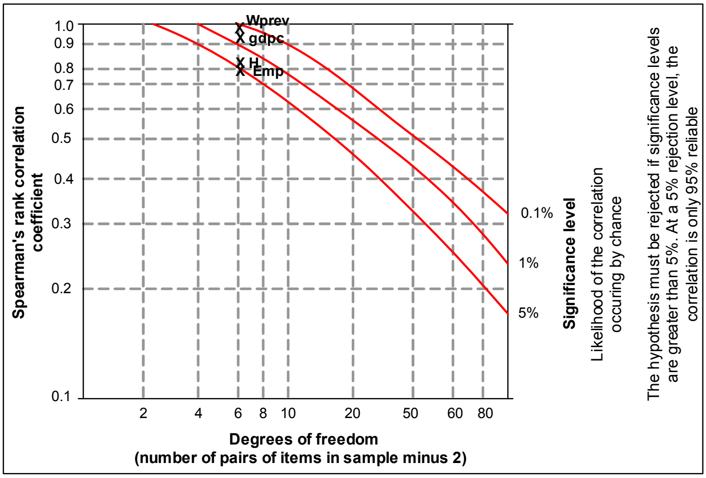

Figure 3: Significance of the Spearman's rank correlation coefficients.

Table 1: Correlation coefficients of the variables.

	T	P	GDP	H	Tmp	prec	E	W	gdpc
T	1.000	**0.934**	**0.991**	**0.993**	0.047	*-0.546*	*0.406*	0.183	-0.294
P	**0.934**	1.000	**0.928**	**0.947**	-0.199	*-0.413*	*0.548*	-0.016	-0.537
GDP	**0.991**	**0.928**	1.000	**0.996**	0.079	*-0.489*	*0.501*	0.078	-0.373
H	**0.993**	**0.947**	**0.996**	1.000	0.002	*-0.486*	*0.499*	**0.676**	-0.384
Tmp	0.047	-0.199	0.079	0.002	1.000	-0.243	-0.133	0.332	0.382
prec	*-0.546*	*-0.413*	*-0.489*	*-0.486*	-0.243	1.000	0.069	-0.500	-0.199
E	*0.406*	*0.548*	*0.501*	*0.499*	-0.133	0.069	1.000	**-0.777**	**-0.899**
W	0.183	-0.016	0.078	**0.676**	0.332	-0.500	**-0.777**	1.000	**0.824**
gdpc	-0.294	-0.537	-0.373	-0.384	0.382	-0.199	**-0.899**	**0.824**	1.000
Eva	—	—	—	—	—	—	—	**-0.726**	—

Only the previous water demand, gdpc, employment, time and number of households (H) were significant at the 5% level (see Figure 2 and Table 1). Therefore, it is recommended to include these variables as inputs in reduced forecasting models. Regression results also show that several combinations of these variables are possible to obtain a reliable model. Most of the coefficients of the explanatory variables have expected signs. The positive value of temperature suggests domestic consumers use more water when the weather is relatively warm. Precipitation contributes negatively to water consumption, meaning that households tend to use less water when there is enough rainfall. Family size and water price (not shown) are not significant at any level, which may be due to the fact that both variables vary little with time.

Important was also to find the robustness and the generality of the influencing parameters. Therefore, the correlation coefficient for the previous water demand, price, employment, time and number of households to domestic water demand were calculated for other regions of different nature where data could be obtained. Data for Germany and Canada was present and the calculated correlation coefficients were almost similar to that of Beijing, *T, E* and *H* had the highest correlation with the domestic water demand, which suggests that models for forecasting domestic water demand that include these variables are quite reliable. To avoid multicollinearity in the models, the population density *P* is excluded to be an explanatory variable as it is highly correlated with *GDP* and *H*.

To prove our results models, which included combinations of the different variables were implemented and tested. Results of some selected models, described by equations 10 and 14 are shown in Figure 4. All the models are compared with the simple linear model, where the entire yearly water requirement of the households is the product of the estimated domestic water demand per capita and the estimated population.

$$\text{Model 1}: W = \alpha + \beta T + \gamma E + \delta W_{prev} + \Delta Temp + \nu P \tag{10}$$

$$\text{Model 2}: W = \alpha + \beta T + \gamma E + \delta W_{prev} + \Delta Temp \tag{11}$$

$$\text{Model 3}: W = \alpha + \beta T + \gamma E + \delta W_{prev} \tag{12}$$

$$\text{Model 4}: W = \alpha + \beta T + \gamma E + \mu GDP \tag{13}$$

$$\text{Model 5}: W = \alpha + \beta T + \gamma E + \varphi H \tag{14}$$

The parameters in Table 2 were estimated for the different models. Their standard deviation and coefficient of variation in % were also calculated. In Table 3 the results of the correlation tests of the covariance between the calculated parameters are listed. The models could be reduced accordingly.

The results of all the models show a very good adjustment of the model values up to the real water requirement, and also the future development follows a smooth, realistic process. Due to the fact that there could be some effects that are correlated to some explanatory variables the OLS is biased and inconsistent.

Therefore, the parameter estimation was done using feasible generalized least squares analysis.

Table 2: Estimated parameters and their standard deviation.

	Best Estimate (Model)					Standard Deviation (Model)				
	1	2	3	4	5	1	2	3	4	5
α	265.7	-697	-77.947	-1063	-3480	1294	525	532	162	1948
β	-0.0139	0.0363	0.0411	0.055	0.1793	0.0674	0.0271	0.0274	0.0815	0.0998
γ	0.0012	-0.0039	-0.0042	-0.0031	-1.9E-3	0.0064	0.0015	0.0015	9.5E-6	0.0010
δ	0.7722	-0.2478	-0.333	—	—	13.709	0.5554	0.5623	—	—
Δ	-0.013	0.0924	—	—	—	0.1564	0.0849	—	—	—
ν	0.0039	—	—	—	—	0.0047	—	—	—	—
μ	—	—	—	-0.0186	—	—	—	—	0.0512	—
φ	—	—	—	—	-0.0202	—	—	—	—	0.013

Table 3: Correlation coefficient of the parameters.

	α	β	γ	δ	Δ	ν
α	1.0	-0.9999	0.9746	0.9938	-0.7033	0.9029
β	-0.9999	1.0	-0.9757	-0.994	0.6997	-0.9045
γ	0.9746	-0.9757	1.0	0.9744	-0.7655	0.9688
δ	0.9938	-0.994	0.9744	1.0	-0.7044	0.9034
Δ	-0.7033	0.6997	-0.7655	-0.7044	1.0	-0.8183
ν	0.9029	-0.9045	0.9688	0.9034	-0.8183	1.0

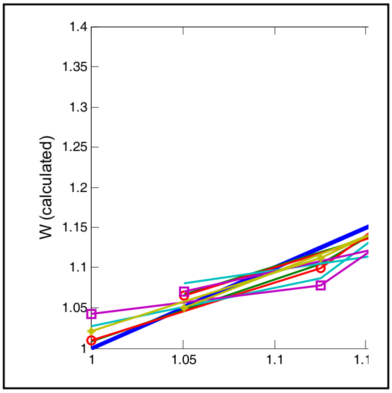

Figure 4: W using different models (calculated vs measured).

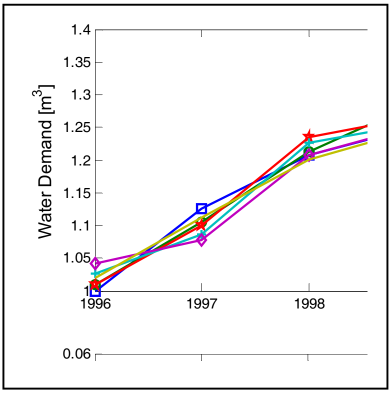

Figure 5: Domestic water demand forecasting using different models.

Figures 3, 4 and Table 4 show the qualitative and quantitative results of the models, respectively. The sum of the squares of the residuals, the multiple correlation coefficients for I/O data and the linear correlation coefficient of measured output and the calculated output show the best results in the following order: 1, 5, 2, 3 and 4. Due to the high correlation of the domestic water demand and the number of households the reduced model (Model 5) that include the number of households shows better results compared to Model 2.

Table 4: Statistical results of the models.

	Model 1	Model 2	Model 3	Model 4	Model 5
Sum of squares of residuals	0.0037	0.0049	0.0069	0.0072	0.0047
Correlation: x - y data					
Multiple Correlation Coefficient (MCC)	0.9756	0.9672	0.9539	0.9514	0.9688
MCC F-test ratio	147.908	192.969	252.363	238.411	382.102
MCC F-test probability	0.0257	0.0077	0.0024	0.0028	0.00093453
Correlation: y(measure) - y(calculated)					
Linear Correlation Coefficient (LCC)	0.9756	0.9672	0.9539	0.9514	0.9688
LCC Probability	3.6625E-6	1.248E-5	3.3178E-5	3.9827E-5	8.5582E-6
Degrees of freedom	2	3	4	4	4
Number of data points	8	8	8	8	8
Number of estimated paramaters	6	5	4	4	4

6 Conclusions

The applied study is an important starting point for the development of simple and robust models. For the residential sector the variables of the economic water use models include income, household size, housing density, air temperature, rainfall, marginal price, and fixed charges for water and wastewater. In most regions there is no data available for most of these parameters. It is therefore very important to know which parameter, can be used at the minimum to produce some reliable results. In this study several variables have been tested for their influence on the domestic water demand. It has been shown that to predict domestic water reliably at least the gdpc, the previous water demand, employment rate, the time and the number of households must be included. The estimation can be improved by using panel data covering a longer time period or more disaggregated sub-regional level analyses. It would also be useful to extend the study with more adequate data especially regarding time series water prices for the domestic sector. Well-designed household surveys would provide richer information and greater insights into the factors influencing domestic water demand. The results of the comparison between Canada, Beijing and Germany shows that in water abundant areas more water will be used and also the increased positive correlation coefficient of income (Beijing 0.345; Canada 0.578; Germany 0.623) implies that consumers who have a high income tend to consume more water.

www.witpress.com, ISSN 1743-3541 (on-line)

References

[1] Rosegrant, M., Ringler, C., Msangai, S., Cline, S. *International Model for Policy Analysis of Agricultural Commodities and Trade (IMPACT-WATER): Model Description*. Int. Food Policy Research Inst, Washington, D.C., 2005.

[2] Archibald G.G. *Forecasting Water Demand – A Disaggregated Approach*. Journal of Forecasting, 1983.

[3] Zena Cook, Scott Urban. *Domestic, Commercial, municipal and Industrial water Demand; Assessment and Forecast in Ada and Canyon Counties, Idaho*. Idaho department of Water resources IDWR, 2001.

[4] Petra Döll, Joseph Alcamo, Thomas Henrichs, Frank Kaspar, Bernhard Lehner, Thomas Rösch and Stefan Siebert. *The Global Integrated Water Model WaterGap 2.*1, 2001.

[5] *Beijing Municipal Bureau of Statistics. Beijing Statistical Information net*, Mai 2005. http://www.bjstats.gov.cn/english/index.html.

[6] Rauschenbach T. The Introduction of Water Resources in Beijing and the Demand Analysis, July 2005.

[7] Hartung J. *Statistik. Lehr- und Handbuch der angewandten Statistik*. 13. Auflage, Oldenbourg Verlag München Wien, 2002

[8] Wernstedt J. Experimentelle Prozessanalyse. 1. Auflage. VEB Verlag Technik Berlin, 1989

[9] Brockwell, P.J. and Davis R. A. Time Series: Theory and Methods. Springer Verlag, New York. 1991