Principal Component Regression Analysis in Water Demand Forecasting: An Application to the Blue Mountains, NSW, Australia[1]

M. M. Haque[a, *], A. Rahman[a], D. Hagare[a] and G. Kibria[b]
[a]School of Computing, Engineering and Mathematics, University of Western Sydney, Australia
[b]Sydney Catchment Authority, NSW 2750, Australia

Abstract: *Accurate forecast of water demand is very crucial in developing a water resource management strategy to check the balance of future water supply and demand to ensure proper water supplies to the people. In order to forecast water demand, different models have been adopted in the literature. Among these the multiple regression analysis is quite popular for long term water demand forecasting. In spite of their evident success in modelling water demands, it can face difficulties in the case of multicollinearity, which implies highly correlated variables. Since water demand depends on many factors such as population, household size, rainfall, temperature, age of population, education, water price and policy, a multicollinearity problem may arise during the development of a multiple regression model which may lead to the incorrect estimation of future water demand. To avoid multicollinearity problem, principal component regression analysis has been used in several environmental studies which demonstrated its ability to eliminate the multicollinearity problem and to produce better model results. However, application of principal component regression in water demand forecasting is limited. In this study, principal component regression model was developed by combining multiple linear regression and principal component analysis to forecast future water demand in the Blue Mountains Water Supply systems in New South Wales, Australia. In addition, performances of the developed principal component regression model were compared with multiple linear regression model by adopting several model evaluation statistics such as relative error, bias, Nash-Sutcliffe efficiency and accuracy factor. It was found that the developed principal component regression model was able to predict future water demand with a higher degree of accuracy with an average relative error, bias, Nash-Sutcliffe efficiency and accuracy factor values of 3.47%, 2.92%, 0.44 and 1.04, respectively. Moreover, it was found that the principal component regression model performed better than the multiple linear regression model and could be used to eliminate the multicollinearity problem. The method presented in this paper can be adapted to other cities in Australia and the world.*

Keywords: *Principal component regression, multicollinearity, water demand, forecasting, Blue Mountains, principal component analysis*

1 Introduction

Availability of adequate potable water is becoming an increasing concern around the world due to many factors, including population growth, increased water demand, rapid urbanization, water pollution and changing climate. Therefore, integrated water resources management is very much needed which considers both water supply and demand measures to cope up with the limited water resources. Water demand measures include but are not limited to the installation of water efficient appliances, implementation of water use restrictions, development of awareness programs and use of effective water pricing policy (Adamowski and Karapataki, 2010). In order to implement these demand management programs effectively, an accurate estimate of future water demand such as peak demand, daily demand and long-time demand, is very crucial. Future estimates of peak, daily and weekly demand are considered as short term forecasting, which is mainly required for operation of reservoirs and pumping stations, and maintenance of a water supply system (Jain et al., 2001; House-Peters & Chang, 2011). On the other hand, long term forecasting is usually greater than one year which is required for planning and design of system expansion and future resilience analysis (Bougadis et el., 2005; Nasseri et al., 2011).

[1] *Paper JHR008 submitted 12/08/2013; accepted for publication after peer review and subsequent revision on 21/10/2013*
**Corresponding author: E-mail: m.haque@uws.edu.au*

Different types of models such as time series, regression and artificial neural networks have been adopted in the literature to forecast/model residential water demand. Short term water demand is normally forecasted by a time series model (Zhou et al., 2000; Caiado, 2007) and artificial neural networks models (Ghiassi et al., 2008; Firat et al., 2009; Herrera et al., 2010). In long term forecasting, regression based analysis especially multiple regression analysis (e.g. Lahlou and Coyler, 2000; Babel et al., 2007; Polebitski and Palmer, 2009) is one of the most commonly used techniques in the water demand related literature. Though multiple regression techniques have been quite successful in modelling residential water demand, it can face some severe difficulties in the case of multicollinearity (Rajab et al., 2012). A multicollinearity problem arises when the independent variables experience high correlation. This mulicollinearity problem may produce biased standard error of estimate in the regression analysis and may lead to incorrect identification of the most significant variables in the modelling exercise. Importantly, water demand depends on many factors, such as population number, household size, dwelling type, age of population, household income, water price, temperature, rainfall, evaporation and water conservation and restriction programs (Franczyk and Chang, 2009, Harlan et al., 2009, Arbués et al., 2010, Babel and Shinde, 2011). These variables are often correlated with each other and may introduce a multicollinearity problem in water demand modelling exercise.

Principal component regression (PCR) is a type of regression analysis which considers principal components (PC) as independent variables instead of adopting original variables (Pires et al., 2008). The PCs are the linear combination of the original variables which can be obtained by principal component analysis (PCA). The PCA transforms the original set of inter correlated independent variables to a new set of uncorrelated variables (i.e. PCs). The use of these PCs as independent variables is quite useful in the multiple regression models to avoid the mulicollinearity problem and to identify the variables which are the most significant in making the prediction (Abdul-Wahab et al., 2005; Camdev´yren et al., 2005; Sousa et al., 2007 and Rajab et al., 2013). Abdul-Wahab et al. (2005) used the combination of multiple linear regression and PCA technique to model tropospheric ozone and to identify the significant factors that control ozone levels. Camdev´yren et al. (2005) adopted PCs in multiple linear regression analysis in water quality studies. Sousa et al. (2007) developed the PCR model using PCs as inputs to predict ozone concentrations and compared that model with multiple linear regression and a feed forward artificial neural network model. Rajab et al. (2013) combined a multiple regression model with PCA technique to improve the prediction of ozone levels. All of these studies have found that the incorporation of PCs as independent variables in the regression models improved the model prediction as well as reduced the model complexity by eliminating multicollinearity.

Although PCR has been adopted in many fields of water engineering, there have been limited applications of PCR in water demand forecasting. Few studies have been found in the literature that have adopted PCA to forecast water demand, such as Koo et al. (2005) and Choi et al. (2010). Koo et al. (2005) used the PCA and cluster analysis to divide the region into two groups and then developed a new multiple regression model for each of the groups to overcome the negative coefficient issue (the regression coefficient for the population of the Jung-gu model had a negative value, which was not sensible) of the population variable with the water demand in the existing single multiple regression model for each district in Seoul, Korea. They found that the newly developed model solved the irrationality of the regression model and were able to produce better water demand estimates. Choi et al. (2010) developed three multiple regression models: a comprehensive multiple regression model with six independent variables, a multiple regression model for each cluster and a multiple regression model adopting two principal components (derived from eight number of original variables) for each cluster to estimate the water demand in 164 regions in Korea. They found that the clustering multiple regression model performed best in modelling the observed water demand values and the multiple regression model based on PCs were better than the comprehensive multiple regression model. Due to the potential of PCR in reducing the complexity of multiple regression models by mitigating the multicollinearity problem, it is worth exploring the application of PCR in the water demand forecasting.

In this study, water demand modelling is done by multiple linear regression and PCR techniques to estimate the future water demand in the Blue Mountains Water Supply System in New South Wales, Australia. The objectives of this paper are twofold: (i) to evaluate the effectiveness of PCs as independent variables in multiple linear regression models of water demand forecasting (i.e. checking the applicability of PCR in water demand forecasting); (ii) to evaluate the performance of developed multiple linear regression and PCR models by estimating and comparing several goodness of fit statistics. Here, PCs are obtained by undertaking the PCA of ten water demand variables.

2 Study area and data

The Blue Mountains region (Figure 1) of New South Wales, Australia is selected as the study area. The Blue Mountains Water Supply System provides water to a population of around 48,000 from Faulconbridge to Mount Victoria, which are considered as Upper and Middle Blue Mountains area (Sydney Catchment Authority, 2009). Data on monthly metered residential consumption and number of dwellings were gathered from Sydney Water for the period of 2003-2011. It was found that the single dwelling residential sector (i.e. free standing/semi-detached houses) accounts for about 94% residential water consumption while the multiple dwelling sector (i.e. apartment blocks/units) is for the rest 6%. In this study, analysis was done for the single dwelling residential sector. Collected monthly metered water consumption values were divided by the number of dwellings to get the 'per dwelling monthly water consumption' to be used in the modelling.

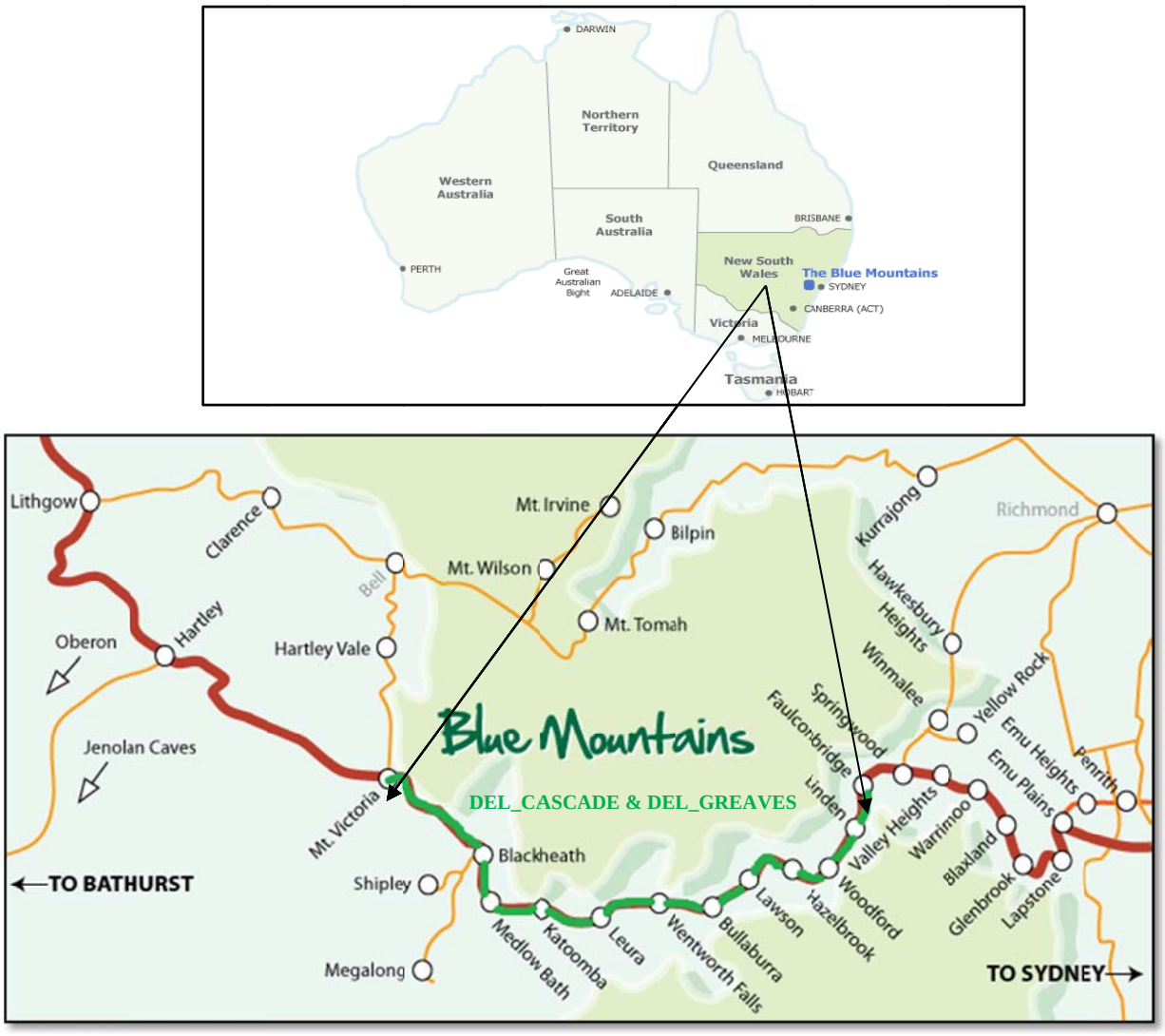

Figure 1 Blue Mountains region in Australia and Cascade and Greaves creeks water supply area (Bluemountainsaustralia, 2013)

Data on water usage price and water conservation savings (WCS) were also gathered from Sydney Water for the same period. In this study, WCS refers to the number of dwellings that have participated in the water demand management programs in the Blue Mountains, such as rain water tank, WaterFix (installation of new showerheads, flow restrictions and minor leak repairs undertaken by a licensed plumber), DIY (Do-It-Yourself) kits (self-installed flow restrictors), water efficient washing machines and toilets (Sydney Water, 2010). New South Wales Government imposed three different levels of water restriction based on the severity during the drought periods (2003-2009) to manage the

short water supply. Level 1 and Level 3 were the most liberal and the most severe level of restrictions, respectively. Level 1, Level 2 and Level 3 water restriction were imposed to three different occasions during the 2003-2009 in the Sydney region. In this study, three dummy variables were considered for these three levels of water restriction. The value of a dummy variable was considered as 1 (one) when the water restrictions were in place; otherwise its value was considered to be zero in the data matrices. Meteorological data, such as rainfall, number of rain days, temperature, evaporation and solar exposure were collected from Sydney Catchment Authority.

3 Methods

In this study, multiple linear regression (MLR) equation and PCA were combined together to perform PCR analysis. This PCR model was adopted to predict the future water demand. Brief description of PCA, MLR and PCR are given in the following sections.

3.1 Principal component analysis

Principal component analysis transforms the original data set of *n* variables which are correlated among themselves to various degrees to a new data set containing *n* number of uncorrelated principal components (PCs). The PCs are linear functions of the original variables in a way that the sums of the variances are equal for both the original and new variables. The PCs are sequenced from the highest variance to the lowest one. The first PC explains the highest amount of variance in the data. The next highest variance is explained by the second PC and so on for all *n* PCs. The values of all the PCs can be obtained by the same equation as Equations 1 and 2. These two equations are for PC 1 and PC 2. Although the number of PCs and original variables are equal, normally most of the variance in the data set can be explained by the first few PCs that can be used to represent the original observations (Abdul-Waheb et al., 2005; Olsen et al., 2012). This helps in reducing the dimensionality of the original data set.

$$PC1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = \sum_{j=1}^{n} a_{1j}x_j \quad (1)$$

$$PC2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = \sum_{j=1}^{n} a_{2j}x_j \quad (2)$$

Where $x_1, x_2, \ldots, x_n$ are the original variables in the data set and a_{jj} are the eigenvectors.

The eigenvalues are the variances of the PCs and the coefficients a_{jj} are the eigenvectors extracted from the covariance or correlation matrix of the data set. The eigenvalues of the data matrix can be calculated by Equation 3 as shown below:

$$\left|C - \lambda I\right| = 0 \quad (3)$$

Where C is the correlation/covariance matrix, λ is the eigenvalue and I is the identity matrix.

The PC coefficients or the weights of the variables in the PC are then calculated by Equation 4:

$$\left|C - \lambda I\right| a_{jj} = 0 \quad (4)$$

Due to differences in the units of the water demand variables used in this study, a correlation matrix of the variables was used to obtain eigenvalues and eigenvectors. The eigenvectors multiplied by the square root of the eigenvalues produce a $n \times n$ matrix of coefficients, which are called variable loadings. Importance of each original variable to a particular PC is represented by these loadings. Furthermore, the sum of the products of the variable loadings and the values of original variables produce a new set of data values which are called component scores. These scores can be used in the multiple linear equations as new variables to predict the future water demand.

3.2 Multiple regression analysis

Multiple linear regression attempts to model the relationship between two or more independent variables with a dependent variable by fitting a linear equation to the observed data. The general equation of a MLR model can be expressed as below (Montgomery et al. 2001):

$$Y = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_n x_n \quad (5)$$

Where Y is the dependent variable, $a_i (i = 0,\ldots,n)$ are the parameters generally estimated by least squares method and $x_i (i = 0,\ldots,n)$ are the independent variables.

3.3 Principal component regression (PCR)

In the PCR analysis, MLR and PCA are combined together to establish a relationship between the dependent variable and the selected PCs of the input variables (Pires, et al., 2008). Mainly principal component scores obtained from the PCA are taken as the independent variable in the multiple linear regression equation to perform the PCR analysis. The general equation of PCR model is as follows:

$$Y = a_1 \times PC1 + a_2 \times PC2 + \ldots + a_n \times PCn \quad (6)$$

4 Model evaluation criteria

The performances of the developed MLR and PCR models were compared by adopting four statistical performance indices: (i) relative error (RE); (ii) percent bias (PBIAS); (iii) Nash-Sutcliffe efficiency (NSE); and (iv) accuracy factor (A_f). RE measures the relative size of error in the modelled values in terms of percentage with respect to observed values. An ideal value of RE is zero, which indicates that the developed model is perfect, which is not possible in practice. However, the lesser the RE value, the better the performance of the model would be. PBIAS measures the percentage of the residuals with respect to observed values which indicate whether the developed model overestimates or underestimates the observed values overall (Gupta et al. 1999). The perfect value of PBIAS is zero. Low values of PBIAS indicate better simulation results by the model where positive and negative values represent underestimation and overestimation bias, respectively in the modelled results. NSE is a dimensionless model performance indicator which measures the relative magnitude of the residual variance with respect to the observed data variance (Nash and Sutcliffe, 1970). The optimal value of NSE is one; however, any value between 0 to 1 is generally considered as an acceptable range of performance (Moriasi et al., 2007). A_f demonstrates the spread of the model results with respect to observed values. The best value of A_f is one which indicates a perfect agreement between the modeled and observed values. The smaller the value of A_f, the more accurate the model results is (Basant et al., 2010). The values of these evaluation statistics are computed by the equations presented in Table 1 from the modeled and observed values of the dependent variable. In this study, these were computed for both the modelling period (2003-2008) and the forecasting period (2009-2011).

Table 1 Equations used to calculate the model performance indices

Performance Indices	Equation
RE	$\frac{\sum_{i=1}^{n} \lvert (O - P) \rvert}{N} \times 100$
PBIAS	$\frac{\sum_{i=1}^{n} (O - P)}{\sum_{i=1}^{n} O} \times 100$
NSE	$1 - \left[\frac{\sum (O_i - P_i)^2}{\sum (O_i - O_{mean})^2} \right]$
A_f	$10^{\left(\frac{\sum_{i=1}^{n} \left\lvert \log(\frac{P}{O}) \right\rvert}{N} \right)}$

O: Observed water demand, *P*: Model estimated water demand, *N*: Number of observations.

5. Results

Pearson correlation matrices of the water demand variables are presented in Table 2. Statistically significant correlation coefficients ($p < 0.05$) are highlighted in bold. The linear relationship between two variables and the existence of the collinearity between the independent variables can be identified from these coefficients.

Table 2 Pearson correlation matrix of different variables

		PDRC	RF	NRD	MMT	EVP	SE	WP	WCS	RL1	RL2	RL3
PDRC	Pearson Correlation	1	**-.150***	**-.174***	**.284****	**.568****	**.283****	**-.611****	**-.728****	.020	-.130	**-.452****
	Sig. (2-tailed)		.046	.020	.000	.000	.000	.000	.000	.790	.085	.000
RF	Pearson Correlation		1	**.648****	**.227****	.014	.086	.094	.074	-.038	-.012	.063
	Sig. (2-tailed)			.000	.002	.858	.254	.214	.327	.613	.872	.408
NRD	Pearson Correlation			1	**.239****	.039	**.167***	**.256****	**.189***	-.102	-.104	.061
	Sig. (2-tailed)				.001	.609	.026	.001	.012	.179	.169	.424
MMT	Pearson Correlation				1	**.814****	**.852****	-.014	-.008	.130	.008	.019
	Sig. (2-tailed)					.000	.000	.849	.917	.085	.920	.801
EVP	Pearson Correlation					1	**.849****	**-.197****	**-.223****	**.207****	.015	**-.201****
	Sig. (2-tailed)						.000	.009	.003	.006	.842	.007
SE	Pearson Correlation						1	.126	.128	.082	.016	.033
	Sig. (2-tailed)							.094	.090	.276	.829	.663
WP	Pearson Correlation							1	**.906****	-.102	-.108	**.159***
	Sig. (2-tailed)								.000	.176	.154	.034
WCS	Pearson Correlation								1	-.023	.000	**.474****
	Sig. (2-tailed)									.762	.996	.000
RL1	Pearson Correlation									1	-.059	-.135
	Sig. (2-tailed)										.438	.074
RL2	Pearson Correlation										1	**-.167***
	Sig. (2-tailed)											.026
RL3	Pearson Correlation											1
	Sig. (2-tailed)											

*Correlation is significant at the 0.05 level (2-tailed).

**Correlation is significant at the 0.01 level (2-tailed).

In Table 2, PDRC is per dwelling residential consumption (dependent variable) in kL/dwelling/month; RF is total monthly rainfall in mm; NRD is number of rain days in a month; MMT is monthly mean maximum temperature in ^{0}C; EVP is monthly total evaporation in mm; SE is monthly mean daily global solar exposure (MJ/m^2); WP is water price in AUD/kL; WCS is water conservation savings in dwelling numbers; RL1 is dummy variable for Level 1 restriction; RL2 is dummy variable for Level 2 restriction; and RL3 is dummy variable for Level 3 restriction. As can be seen in Table 2, per dwelling residential consumption were negatively correlated with RF, NRD, WP, WCS, RL1, RL2 and RL3. This result was expected as the rainfall and number of rainy days increase; the water requirement for gardening would be less. Hence, total water consumption would be reduced. Furthermore, it is obvious that with the enhancement of water conservation programs and restriction levels, the total water consumption goes down. Water consumption was found to be positively correlated with MMT, EVP and SE (Table 2). As these variables are mostly related with temperature, water consumption will be more if the temperature is relatively high in a day. High correlation coefficients were found between the independent variables, such as MMT and EVP (0.814), MMT and SE (0.852), WP and WCS (0.906), which demonstrate the existence of multicollinearity between the variables.

The PCA was done on the ten independent variables to explain per dwelling water consumption level in the Blue Mountains Water Supply systems. Table 3 and 4 summarise the results of the PCA on the ten independent variables with the amount of variance explained by each PC. From Table 3, it can be seen that the first five PCs had eigenvalues higher than 1. Moreover, these first five PCs explained around 94% of the total variation of variables in PCA. These five PCs were selected for PCR. Contribution of a particular variable within a PC is normally judged by its variable loadings value. The higher the loading of a variable, the more contribution is reflected by that variable within a particular PC. The bold marked loads in Table 4 indicate the high existing correlation between the variables and corresponding PC.

Table 3 Variance explained by the PCs

Value	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9	PC 10
Eigenvalue	2.97	2.68	1.39	1.29	1.04	0.32	0.15	0.08	0.06	0.02
% of Var.	29.68	26.81	13.87	12.92	10.44	3.23	1.46	0.77	0.60	0.21
Cum. %	29.68	56.49	70.36	83.29	93.73	96.96	98.42	99.19	99.79	100.00

Table 4 Component loadings (correlations between original variables and the firs five PCs)

Variable	PC 1	PC 2	PC 3	PC 4	PC 5
RF	0.08	**0.52**	0.37	**-0.61**	-0.26
NRD	-0.01	**0.70**	0.25	-0.48	-0.21
MMT	**0.82**	0.43	0.09	0.11	0.13
EVP	**0.92**	0.24	-0.11	0.16	0.07
SE	**0.78**	0.52	-0.03	0.17	0.23
WP	-0.48	**0.73**	-0.44	0.10	0.02
WCS	-0.53	**0.76**	-0.22	0.22	0.10
RL1	0.48	-0.29	-0.22	0.13	**-0.77**
RL2	0.27	-0.46	-0.25	**-0.64**	0.47
RL3	-0.15	-0.05	**0.90**	0.35	0.14

All ten water demand variables were included in the five selected PCs. However, only certain variables showed high loadings within each PC, such as the first PC was heavily loaded on MMT, EVP and SE, and the second PC was heavily loaded with RF, NRD, WP and WCS. Similarly as can be seen in Table 3, PC 3, PC 4 and PC 5 were loaded heavily with Level 3, Level 2 and Level 1 restrictions, respectively.

Component score coefficients (eigenvectors) and the values of the original variables were then multiplied to obtain PC score values. These score values were used as independent variables in the stepwise multiple linear regression analysis to determine the most significant PCs for water demand prediction. Data from October 2003 to December 2008 were used to develop the PCR model. Then the model was used to

forecast water demand for the period of January 2009 to September 2011. In PCR analysis, PC 1, PC 2 and PC 5 were found to have significant ($p < 0.05$) linear relationship with per dwelling water consumption (PDRC) (Table 5).

Table 5 Results of regression analysis

Included Independent Variables	Regression Coefficients	Standard Error	Standardized Beta Coefficients	t	Sig.	R^2 (%)
Constant	12.463	.096		129.568	.000	66.4
PC 1	.463	.050	.720	9.325	.000	
PC 2	-.285	.060	-.361	-4.728	.000	
PC 5	-.192	.068	-.215	-2.823	.006	

As can be seen in Table 5, the three PCs (i.e. PC 1, PC 2 and PC 5) could explain 66% of the variation in water consumption. PC 1 and PC 2 were found to be the most significant independent variables in the regression analysis as the standardized beta coefficients values are the highest and the second highest for these two PCs, respectively. PC 1 had positive impact on water consumption while PC 2 had a negative impact (Table 5) as the sign of the regression coefficients were found to be positive and negative for PC 1 and PC 2, respectively. This implies that if the value of PC 1 increases, water consumption would be expected to increase and water consumption would decrease as the value of PC 2 increases. Therefore, a total increase in significant variables of PC 1, namely MMT, EVP and SE would lead to an increase in the water consumption level. On the other hand, an increase in the significant variables of PC 2 (RF, NRD, WP and WCS) would lead to a decrease in water consumption level, as expected. Dummy variables for Level 3 and Level 2 water restrictions had significant loadings in PC 3 and PC 4, respectively which were excluded from the PCR model as its *p* values were not statistically significant. However, linear effects of these variables were partially incorporated in the model as RL 3 and RL 2 were also included in PC 1, PC 2 and PC 5. PC 5 is negatively correlated with water consumption. As level of water restriction goes up, the level of water consumption would be expected to decrease.

The developed PCR model can be written as:

$$PDRC = 12.463 + 0.463 \times PC1 - 0.285 \times PC2 - 0.192 \times PC5 \quad (7)$$

The comparison of observed and predicted water consumption values by the PCR model is presented in Figure 2. The forecasted monthly water consumption values were found to be close to the observed values. Average relative error values for all of the predicting months were found to be 3.47%, which indicates that the model is capable of forecasting monthly water demand with a high level of accuracy.

Three forms of multiple linear regression techniques, linear, semi-log and log-log were adopted to develop the multiple linear regression models. In the linear model, the relationship between the dependent variable, per dwelling water consumption and the independent variables (e.g. rainfall, temperature and water price) were assumed to be linear. In semi-log model, only the dependent variable was in logarithmic form, whereas in the log-log models all the independent variables and dependent variable were entered in logarithmic form in the regression equation. After checking the model performances of these three models, it was found that the semi-log model performed best to model the water demand. Therefore, the semi-log model was taken as the final model to report in this study. The variables were retained in the regression equation for which the regression coefficients were significant at 5% significance level. It was found that only five variables were found to be significant in the equations. Five other variables were found not to be statistically significant, including the number of rain days in a month, monthly total evaporation, monthly mean daily global solar exposure, water conservation savings and a dummy variable for Level 1 restriction due to the high correlation with other variables.

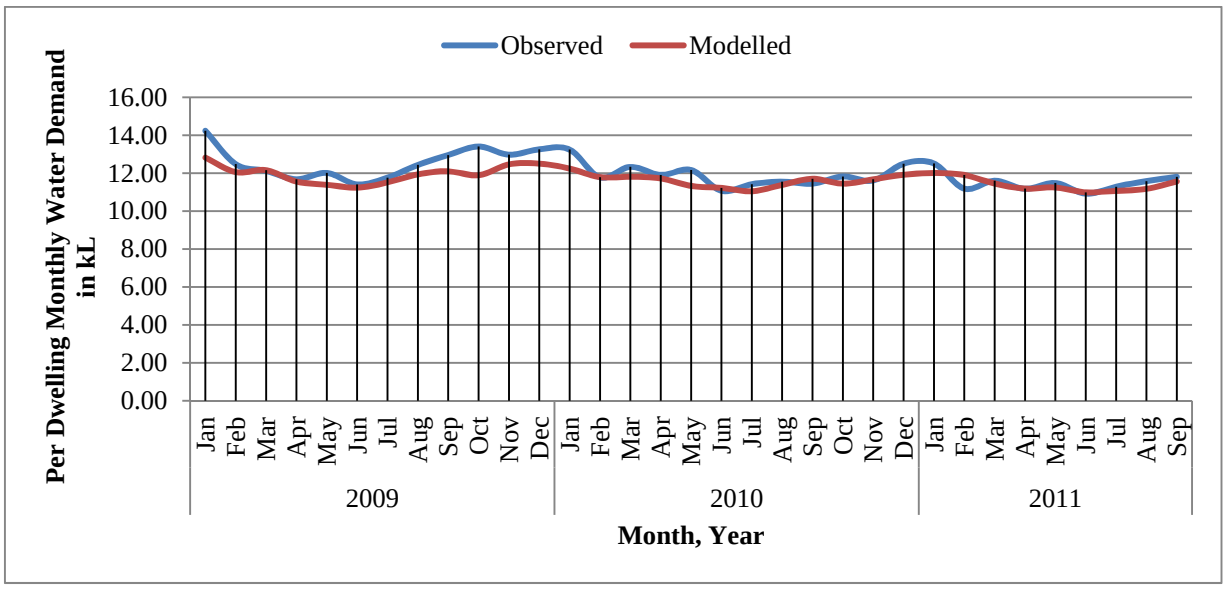

Figure 2 Comparison of predicted and observed water demand

The equation of the developed semi-log model is given below:

$$\log(PDRC) = 1.24 - 0.000079 \times RF + 0.00227 \times MMT - 0.121 \times WP - 0.0376 \times RL2 - 0.0331 \times RL3 \quad (8)$$

where PDRC =per dwelling residential consumption (dependent variable) in kL/dwelling/month; RF = total monthly rainfall in mm, MMT = monthly mean maximum temperature in ^{0}C; WP = water price in AUD/kL; RL2 = Dummy variable for Level 2 restriction; RL3 = Dummy variable for Level 3 restriction.

The comparison of the performance of the developed MLR and the PCR model is presented in Table 6 for both the modelling and forecasting period. It was found that the performances of the models were nearly the same during the modelling period. However, the PCR model outperformed the MLR model during the forecasting period. All of the performance statistics were in favour of the PCR model. The PCR model considered the PCs as independent variables which accounted for the contribution of all the original variables without having any multicollinearity problem. On the other hand, in the developed MLR model, half of the original variables had to be discarded due to the multicollinearity problem, which might be the reason for the underperformance with respect to the PCR model.

Table 6 Results of the performance indices for both the MLR and PCR model

Performance Indices	MLR	PCR
	Modelling period (2003-08)	
RE	3.54	3.57
PBIAS	1.07	0.00
NSE	0.65	0.67
A_f	1.04	1.03
	Forecasting Period (2009-11)	
RE	7.91	3.47
PBIAS	8.04	2.92
NSE	0.37	0.44
A_f	1.09	1.04

6 Conclusions

In this study, the principal component regression (PCR) model was developed by combining multiple linear regression (MLR) and principal component analysis to identify the most important variables for water demand modelling and to forecast future water demand. It was found that PC 1 and PC 2 were the most significant independent variables in the PCR model. Therefore, the variables which had significant loadings within these PCs could be considered as important predictor variables for water demand forecasting. Hence, monthly mean maximum temperature, monthly total evaporation, monthly mean daily global solar exposure, monthly total rainfall, number of rainy days in a month, water price and water conservation savings variables were found to be the most important predictor variables as these variables were within PC 1 and PC 2. However, some of these variables were found to be highly correlated with each other. Inclusion of all these variables in the multiple linear regression model might lead to inconsistent estimation of future water demand. Therefore, the developed PCR model was used to forecast the future water demand which showed a high degree of prediction accuracy with an average relative error value of 3.47%. Moreover, the developed PCR model with three PCs as independent variables was able to explain 66% variation in water consumption level. The performances of the developed PCR model were compared to the MLR for both the modelling and forecasting period. Though both models performed similarly during the modelling period, the PCR model outperformed the MLR during the forecasting period. All the performance statistics, relative error, percent bias, Nash-Sutcliffe efficiency and accuracy factor value were found to be in favour of the PCR model. Moreover, it was found that half of the original variables were discarded in the MLR model due to the multicollienarity problem. To avoid these problems, the PCR model could be used to get a better prediction in water demand forecasting. The method presented in this paper can be applied to other water supply systems to develop water demand forecast models.

Acknowledgements

Water consumption data used in this study was obtained from Sydney Water on 4 May 2012. The best available data at the time of study has been used, which may be updated in the near future. The authors express their sincere thanks to Pei Tillman and Frank Spaninks of Sydney Water for their assistance in collating and providing the data. Further, the authors are grateful to Lucinda Maunsell and Peter Cox of Sydney Water and Mahesh Maheswaran of Sydney Catchment Authority for their cooperation and assistance during data collation and analysis.

References

Abdul-Wahab SA, Bakheit, CS, Al-Alawi SM (2005). Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations, *Environmental Modelling & Software*, 20(10), 1263-1271.

Adamowski J, Karapataki C (2010). Comparison of multivariate regression and artificial neural networks for peak urban water-demand forecasting: Evaluation of different ANN learning algorithms, *Journal of Hydrologic Engineering,* 15(10), 729-743.

Arbués F, Villanúa I, Barberán R (2010). Household size and residential water demand: an empirical approach, *Australian Journal of Agricultural and Resource Economics,* 54(1), 61-80.

Babel M, Gupta AD, Pradhan P (2007). A multivariate econometric approach for domestic water demand modeling: An application to Kathmandu, Nepal, *Water Resources Management,* 21(3), 573-589.

Babel MS, Shinde VR (2011). Identifying Prominent Explanatory Variables for Water Demand Prediction Using Artificial Neural Networks: A Case Study of Bangkok, *Water Resources Management,* 25(6), 1653-1676.

Basant N, Gupta S, Malik A, Singh KP (2010). Linear and nonlinear modeling for simultaneous prediction of dissolved oxygen and biochemical oxygen demand of the surface water-a case study, *Chemometrics and Intelligent Laboratory Systems*, 104(2), 172-180.

Bluemountainsaustralia.com (nd), Location and maps, viewed 10 February 2013, http://www.bluemts.com.au/info/about/maps/

Bougadis J, Adamowski K, Diduch R (2005). Short-term municipal water demand forecasting, *Hydrological Processes*, 19(1), 137-148.

Caiado J (2009). Performance of combined double seasonal univariate time series models for forecasting water demand, *Journal of Hydrologic Engineering*, 15(3), 215-222.

Çamdevýren H, Demýr N, Kanik A, Keskýn S (2005). Use of principal component scores in multiple linear regression models for prediction of Chlorophyll-a in reservoirs, *Ecological Modelling*, 181(4), 581-589.

Choi TH, Koo JY (2010). Water Demand Forecasting by Characteristics of City Using Principal Component and Cluster Analyses, *Environmental Engineering Research,* 15(3), 135-140.

Firat M, Yurdusev MA, Turan ME (2009). Evaluation of artificial neural network techniques for municipal water consumption modeling, *Water Resources Management,* 23(4), 617-632.

Franczyk J, Chang H (2009). Spatial analysis of water use in Oregon, USA, 1985–2005, *Water Resources Management,* 23(4), 755-774.

Ghiassi M, Zimbra DK, Saidane H (2008). Urban water demand forecasting with a dynamic artificial neural network model, *Journal of Water Resources Planning and Management,* 134(2), 138-146.

Gupta HV, Sorooshian S, Yapo PO (1999). Status of automatic calibration for hydrologic models: Comparison with multilevel expert calibration, *Journal of Hydrologic Engineering,* 4(2), 135-143.

Harlan SL, Yabiku ST, Larsen L, Brazel AJ (2009). Household water consumption in an arid city: affluence, affordance, and attitudes, *Society and Natural Resources,* 22(8), 691-709.

Herrera M, Torgo L, Izquierdo J, Pérez-García R (2010). Predictive models for forecasting hourly urban water demand, *Journal of hydrology,* 387(1), 141-150.

House-Peters LA, Chang H (2011). Urban water demand modeling: Review of concepts, methods, and organizing principles, *Water Resources Research,* 47(5),W054.

Jain A, Kumar Varshney A, Chandra Joshi U (2001). Short-term water demand forecast modelling at IIT Kanpur using artificial neural networks, *Water Resources Management,* 15(5), 299-321.

Koo J, Yu M, Kim S, Shim M, Koizumi A (2005). Estimating regional water demand in Seoul, South Korea, using principal component and cluster analysis, *Water Science & Technology:* Water Supply, 5(1), 1-7.

Lahlou M, Colyer D (2000). Water conservation in Casablanca, Morocco. JAWRA *Journal of the American Water Resources Association,* 36(5), 1003-1012.

Montogomery, D. C., Peck, E. A., and Vining, G. G. (2001). "Introduction to linear regression analysis." Third edition, John Wiley & Sons, New York, USA.

Moriasi D, Arnold J, Van Liew M, Bingner R, Harmel R, Veith T (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations, *Transactions of the ASABE*, 50(3), 885-900.

Nash JE, Sutcliffe J (1970). River flow forecasting through conceptual models part I—A discussion of principles, *Journal of Hydrology,* 10(3), 282-290.

Nasseri M, Moeini A, Tabesh M (2011). Forecasting monthly urban water demand using Extended Kalman Filter and Genetic Programming, *Expert Systems with Applications*, 38(6), 7387-7395.

Olsen RL, Chappell RW, Loftis JC (2012). Water quality sample collection, data treatment and results presentation for principal components analysis-literature review and Illinois River watershed case study, *Water Research,* 46(9), 3110-3122.

Pires J, Martins F, Sousa S, Alvim-Ferraz M, Pereira M (2008). Selection and validation of parameters in multiple linear and principal component regressions. *Environmental Modelling & Software,* 23(1), 50-55.

Polebitski AS, Palmer RN (2009). Seasonal residential water demand forecasting for census tracts, *Journal of Water Resources Planning and Management,* 136(1), 27-

Rajab JM, Jafri MZM, San Lim H, Abdullah K (2012). Regression analysis in modeling of air surface temperature and factors affecting its value in Peninsular Malaysia, *Optical Engineering*, 51(10).

Rajab JM, MatJafri M, Lim H (2013). Combining multiple regression and principal component analysis for accurate predictions for column ozone in Peninsular Malaysia, *Atmospheric Environment*, 71, 36-43.

Sousa S, Martins F, Alvim-Ferraz M, Pereira M (2007). Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations, *Environmental Modelling & Software,* 22(1), 97-103.

Sydney Catchment Authority (2009). Blue Mountains water supply system: Strategic review. Sydney Catchment Authority, Penrith, Australia.

Sydney Water (2010). Water conservation and recycling implementation report, 2009-10. Sydney Water, New South Wales, Australia.

Zhou SL, McMahon TA, Walton A, Lewis J (2000). Forecasting daily urban water demand: a case study of Melbourne, *Journal of Hydrology,* 236(3), 153-164.