Journal of Sustainability Science and Management
Volume 19 Number 6, June 2024: 36-53

eISSN: 2672-7226
© UMT Press

MODELLING DOMESTIC WATER DEMAND IN MALAYSIA TO IDENTIFY INFLUENCING FACTORS: A COMPARATIVE ANALYSIS

NATRAH JEFRI AND NORSHAHIDA SHAADAN*

School of Mathematical Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia.

**Corresponding author: shahida@tmsk.uitm.edu.my*

http://doi.org/10.46754/jssm.2024.06.004
Submitted final draft: 15 February 2024 *Accepted: 16 March 2024* *Published: 15 June 2024*

Abstract: Water crises are often experienced by many developing countries worldwide. Predicting future domestic water demand and identifying the influential factors are vital to managing water supply effectively. This study aims to determine the best predictive models among Multiple Linear Regression (MLR), Multi-layer Perceptron (MLP), and Radial Basis Function (RBF) Neural Networks as well as to identify the significant influential factors towards domestic water demand. Based on the yearly records from 2000 to 2018 obtained from the Malaysian Water Association, the Department of Environment, and the Department of Statistics Malaysia the analysis results indicate an increasing pattern of domestic water in Malaysia with the demand for non-domestic water twice lower than domestic water. Based on RMSE and R-squared, Multi-layer Perceptron is the best model for predicting domestic water demand. The MLR model shows that the two most significant influential factors towards domestic water demand are price and design capacity, with negative and positive relationships. The results describe that an increase in price affects a decrease in water demand, while an increase in design capacity will reduce the water demand. The findings suggest that the water utilities in Malaysia should focus more on these identified factors.

Keywords: Water sustainability, water management, water prediction model, water demand.

Introduction

Water is a vital human need. It could be a source of food insecurity that would lead to conflicts such as increased competition over resources, reduction in agricultural production, high food prices, and food/water shortage. Providing sufficient water demand to the public is an important matter to be considered. This aim is tallied with global goals and SDGs to ensure the availability and sustainable management of water and sanitation for everyone. The success of SDG6 requires sustainable management of water resources and access to safe water, which is critical to the survival of people and the planet. Malaysia is rich in water resources. However, an increase in the population, expansion in urbanisation, and rapid socioeconomic development impose high pressure on water resources. Therefore, Malaysia's water supply has changed from relative abundance to relative scarcity (Payus *et al*., 2020). Malaysia's growing demand for water is to sustain its growing population and industrialisation. The term 'demand' refers to the volume that is desired. An accurate water demand estimation helps determine the water quantities to be used. Water demand is divided into domestic and non-domestic (Anang *et al*., 2019). Domestic water demand refers to water allocation from water agencies for household purposes or residential uses such as drinking and washing clothes. Non-domestic demand refers to industrial, commercial, institutional, and public water allocation such as shops, schools, and hospitals. Due to the current phenomenon of high progress in industrial and economic activity and population growth in Malaysia, water demand has been increasing, especially in the urbanised states. The unforeseen contamination in water resources such as rivers, lakes, and aquifers and the impact of climate change alter the availability, quantity, and quality of water supply. As a result, Malaysians suffered from frequent interruptions to the clean water supply.

The most recent incident reported in December 2020 was the water disruption due to pollution in the water resources. Livestock Farming and agriculture are some of the leading causes of water pollution due to chemical dumping from farming and livestock operations.

For example, in coping with water crises in Iran, a group of researchers Mikaiil *et al.* (2023) recommended basic studies are very important in planning water supply for urban populations. Choudhary and Mushtaq (2023) suggest reusing and recycling wastewater as a practical solution. An example of the successful experience of several countries is how China can handle the country's water scarcity by tackling sustainable agricultural water management and "sponge city initiatives" (Qi *et al*., 2020). Kenya implemented "agroforestry". For Chile, by way of better governance, committing to net-zero infrastructure, and implementing a new constitution able to solve the problem. An example of an infrastructure-based solution to water scarcity is the "smart-water management system" utilised in South Korea, an innovative system that helps improve water management's reliability, soundness, and efficiency (Kim & Kang, 2020). Despite the crisis, the water authorities in Malaysia continue to adopt an approach to supply management (Chan, 2004), where water resource management is the main focus. Yet, many areas still need improvement to achieve better governance in water management, including reducing leakage, waste, ineffective privatisation, and demand management for water for both domestic and non-domestic consumers. Several surveys have also been conducted on the demand issues for improvement purposes for Malaysians. Research by Nur Syuhada *et al*. (2020) aimed to ascertain consumers' willingness to pay based on customers' preferences for water quality, water pressure, and reduction of water service disruptions using a series of stated choice experiments with Conditional Logit (CL) model and Mixed Logit (ML) models. The findings contributed to a new viewpoint of the water provider. Earlier research by Yaacob *et al.* (2011) showed that the sample respondents are willing to pay more for drinking water provided that water quality, frequency of water interruption, and trust in tap water have been improved.

Predictive modelling is a promising approach to predicting domestic water demand as it uses data mining and probability to forecast or estimate more granular (specific outcomes) with the ability to know the significant contribution of some factors. From a statistical point of view, studies on predicting domestic water demand in Malaysia using predictive modelling are still limited. In common practice, the Multiple Linear Regression (MLR) model is the most popular parametric form of regression analysis and is widely used for prediction. It has a predetermined structure, so the residuals must be normally distributed. Anang *et al*. (2019) have also reported that MLR is one of the predictive models that can be used in the water industry. Besides that, previous studies that used machine learning techniques such as Artificial Neural Networks with Multi-layer Perceptron (MLP) function in predicting domestic water demand in Malaysia were limited, and the results outperformed the statistical methods. A similar study by Hassan (2013) in Malaysia indicated that Artificial Neural Networks with MLP produced more reasonable results than MLR. In addition, the application of Artificial Neural Network with Radial Basis Function (RBF) is lacking, particularly for the Malaysia dataset for water demand prediction. Lee and Derrible (2019) confirmed that Artificial Neural Network models are designed to capture nonlinear and complex relationships between variables using stochastic local optimisation. Thus, it tends to decrease biases during estimation and provide more accurate predictions than linear models. Furthermore, studies to identify the contributing factors towards water demand in Malaysia are still scarce.

A detailed understanding of water usage patterns and the factors that affect water use is essential for the proper management of water supply and the implementation of relevant public policies. Water use patterns are extremely complicated processes that depend on a variety of influential elements, such as seasonal variability and water availability (Machingambi & Manzungu, 2003; Arouna & Dabert, 2010), water supply restrictions (Andey & Kelkar, 2009), tariff structure and pricing (Renwick & Green, 2000), household characteristics (Syme & Shao, 2004) and attitudes and intentions towards water conservation (Corral-Verdugo *et al.*, 2002). These elements directly and indirectly influence water consumption and usage patterns (Jorgensen *et al*., 2009).

Several limited studies have considered several factors associated with water demand in Malaysia. For instance, Anang *et al*. (2019) considered water resources the dependent variable while real income, total consumption per capita, population density, and climate change were the independent variables in their study. The findings show that water demand is positively impacted by total per capita use, agriculture, and population density. The demand for water resources from the agricultural sector is significant. The rise in demand for water during dry spells, which causes water stress, is a correct indication of climate change. This discovery helps enhance climate change forecasting and manage water resources sustainably, especially in the agricultural sector. Another previous study conducted in Malaysia was conducted by Hassan (2013). The analysis based on 247 records revealed that among the independent variables considered in the study, domestic water use, water price, per capita GDP and averaged annual rainfall; GDP and water price were identified to be the most influential factors. In summary, factors such as water consumption, precipitation, temperature, water price, and population are the commonly applied input variables in predicting domestic water demand in Malaysia (Li & Feng, 2019). Several other potential factors such as water treatment plant capacity, water production, water quality, gross domestic product, and other meteorological variables such as relative humidity, evaporation and wind speed may significantly influence water demand and the variation. However, the involvement of these variables in modelling Malaysia's water demand is seen as lacking. Thus, this situation allows this study to fill the gap.

This research aims to investigate the important factors of Malaysian water demand based on a comparative study of MLR, MLP, and RBF Network predictive models. Given the significant factors, the identified best model is proposed for use in predicting water demand. This study implements a comparative analysis to determine the best predictive model among the three considered models. The results would also serve the purpose of more effectively choosing the water consumption forecasting mode parameters for regional water consumption analysis and water resource planning and management.

Materials and Methods

Data and Sources

The data gathered in the yearly record from the year 2000 to 2018 of 14 states in Malaysia including the Federal Territory of Kuala Lumpur, Federal Territory of Labuan, Federal Territory of Putrajaya, Johor, Kedah, Kelantan, Malacca, Negeri Sembilan, Pahang, Perak, Perlis, Penang, Sabah, Sarawak, Selangor, and Terengganu over 19 years period has contributed to 266 of the total observations. Various sources were involved in acquiring the data, such as the Malaysian Water Association, the Department of Environment, and the Department of Statistics.

Study Variables

For this study, other than some selected factors from previous studies, due to data availability, several new factors were considered to be modelled, including water quality index, temperature, relative humidity, evaporation and wind speed. The factors were grouped into three major components: Water, economics, and meteorology. The variables included in this study are shown in Figure 1, which consists of 13 independent variables and one dependent variable. The description of the variables is as follows.

Table 1: Data description

Sources	Variables	Unit
Malaysian Water Association	Population served	Number of people (N)
	Water treatment plant design capacity	Millions of litres per day (MLD)
	Water production	Millions of litres per day (MLD)
	Domestic water	Millions of litres per day (MLD)
	Non-domestic water	Millions of litres per day (MLD)
	Non-revenue water	Millions of litres per day (MLD)
	Water pricing	Ringgit Malaysia per cubic meter (RM/m^3)
Department of Environment	Water quality index	Index
Department of Statistics	Gross domestic product per capita	Ringgit Malaysia (RM)
Department of Irrigation and Drainage	Rainfall	Millimetres (mm)
Meteorological Department	Temperature	Celsius (ºC)
	Relative humidity	Percentage (%)
	Evaporation	Millimetres (mm)
	Wind speed	Kilometre per hour (km/h)

Conceptual Framework

The conceptual framework of the relationship between 13 independent variables was divided into three categories: Water factors, economic factors, meteorological factors, and one dependent variable (Domestic water demand) as shown in Figure 1.

Methodology Framework

This study obtained data from the agencies (MWA, DOSM, DOE, DID, MET), followed by the next step, data pre-processing. The pre-processing stage included data quality checking, transformation, and normalisation. Next, data visualisation was used to fulfil the first objective: To identify the pattern of domestic water demand in Malaysia. The model establishment of Multiple Linear Regression and Artificial Neural Network was the step taken to fulfil the next objectives, followed by model training and validation. The model performance evaluation used performance indicators such as the root mean square error (RMSE) and coefficient of determination (R-squared) to compare and identify the best model approach. Finally, the significant factors influencing Malaysia's domestic water demand were identified. The framework of the study is illustrated in the flowchart in Figure 2.

Data Analysis and Method

The study used two statistical software, namely R programming and Statistical Package for the Social Sciences (SPSS). Microsoft Power BI is used for data visualisation since it is the most popular tool to develop a dashboard. The R programming was applied for data cleaning, descriptive statistics, and Multiple Linear Regression. The SPSS software was used to model Multi-layer Perceptron Neural Networks and Radial Basis Function Networks.

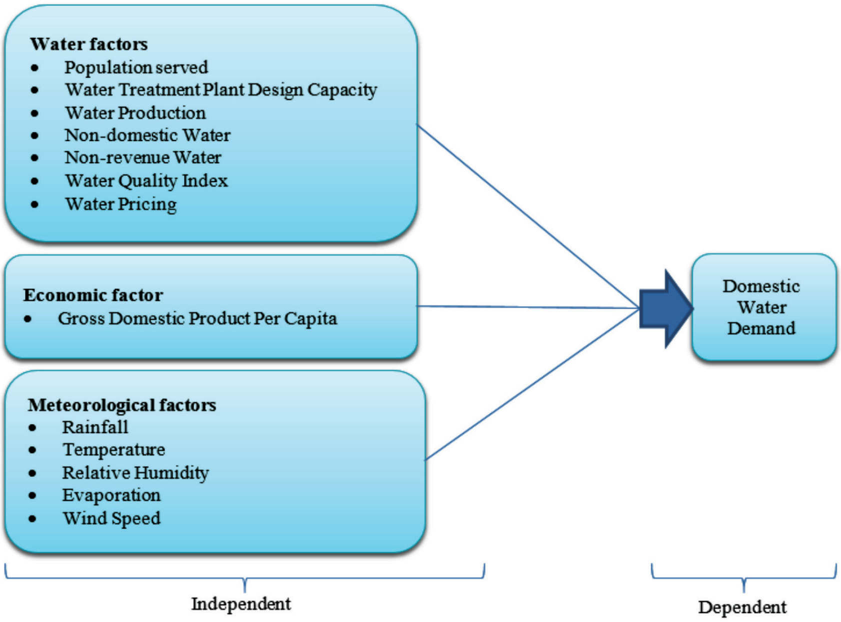


Figure 1: Conceptual framework for predicting domestic water demand

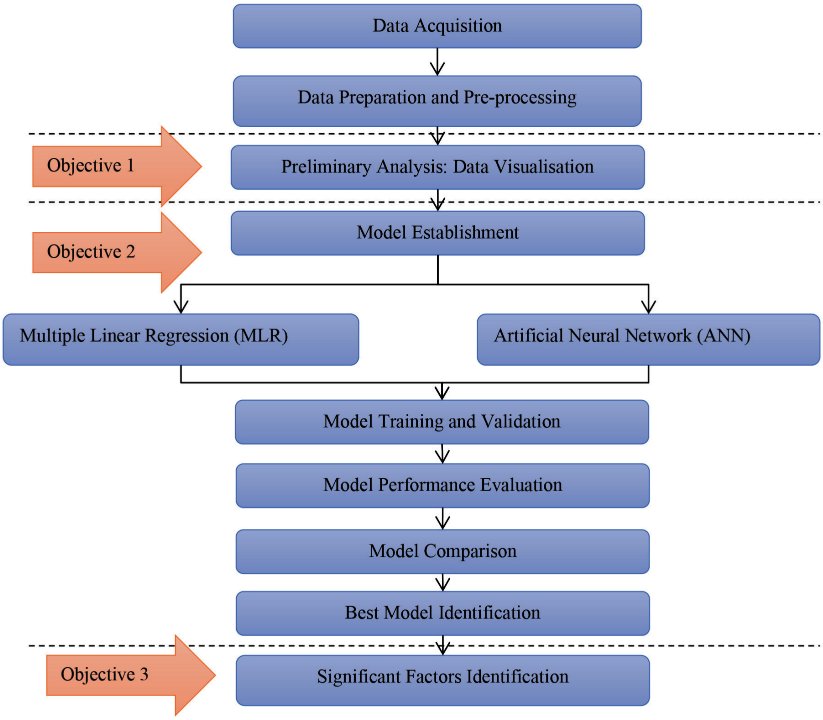

Data Acquisition
Data Preparation and Pre-processing
Objective 1
Preliminary Analysis: Data Visualisation
Objective 2
Model Establishment
Multiple Linear Regression (MLR)
Artificial Neural Network (ANN)
Model Training and Validation
Model Performance Evaluation
Model Comparison
Best Model Identification
Objective 3
Significant Factors Identification

Figure 2: Research methodological framework

The secondary data that had been gathered was analysed using the appropriate data analyses that aligned with the objectives. Generally, three main steps had to be implemented to achieve the goals: Data visualisation, model building, and model comparison. In the modelling stage, the data set is divided into 70% training and 30% testing. The training data set is used to develop the models and the testing data set is used to validate the models.

(a) Multiple Linear Regression Model (MLR)

Multiple Linear Regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Multiple Linear Regression refers to a statistical technique that is used to predict the outcome of a variable based on the value of two or more variables. The predicted variable is the dependent variable, while the variables used to predict the dependent variable are independent or explanatory (Kutner *et al*., 2004). However, the study used time-series recorded data, hence before the Multiple Linear Regression model was developed, the data needed to be randomised. This step removes the autocorrelation effect on model parameter estimation (Montgomery *et al*., 2012).

The mathematical equation of the model is as follows:

$$Y = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \varepsilon \qquad (1)$$

where Y is the dependent or predicted variable, β_0 is the y-intercept (i.e., the constant term), β_1, β_2, β_k are the regression coefficient for each explanatory variable, X_i are the explanatory variables and ε is the model's random error term. The error terms are assumed to follow a normal distribution with mean zero and constant variance. The MLR's important assumptions include: There is a linear relationship between the dependent and independent variables, the independent variables are not highly correlated with each other (i.e., no multicollinearity) and no outliers and highly influential points in the data.

(b) Multi-layer Perceptron Neural Network

Among the neural networks model, the Multi-layer Perceptron Neural Network is the most popular. This model uses a back propagation (BP) training algorithm. The number of layers and neurons in each hidden layer is optimised by trial-and-error procedure even though the user has determined the number of neurons in the input and output layers. The activation function [*f*(*NET*)] used in the study is the Hyperbolic Tangent function, where it takes real-valued arguments and transforms them to the range (–1, 1). The most commonly used weight optimisation method is the back-propagation algorithm, which iteratively analyses the errors and optimises each weight value based on the errors generated by the next layer (Li *et al*., 2017). The mathematical equation of the model is as follows:

$$NET = \sum_{i,j}^{n} w_{ij} x_i + b \qquad (2)$$

$$f(NET) = tanh\ (NET) \qquad (3)$$

where w_{ij} represents the weight value of a connection, x_i represents an inputted independent variable and b represents a bias.

(c) Radial Basis Function Network

The Radial Basis Function transforms the input signal into another form, which can feed into the network for linear separability (Chandradevan, 2017). It may require more neurons than a multi-layer perceptron. In addition, the Radial Basis Function is strictly limited to have exactly one hidden layer, namely a feature vector. One hidden layer has been proven to approximate any function, and it is also known as a universal approximator (Liu, 2013). Although structurally less complicated than Multi-layer Perceptron, it can achieve better function approximation with only one hidden layer (Markopoulos *et al*., 2016). The basic structure of a Radial Basis Function includes an n dimension input layer, a larger dimension m hidden layer ($m > n$) and the output layer.

Radial Basis Function activates neurons at the hidden layer. The typical Radial Basis Function uses Gaussian and Logistic functions, where the input units distribute the values to the hidden layer units uniformly without multiplying them with weights. Each hidden node contains a centre $\boldsymbol{c}$ vector that is a parameter vector of the same dimension as the input vector $\boldsymbol{x}$; the Euclidean distance between the centre $c_j(t)$ and the network input vector $\boldsymbol{x}$ is defined by $\left\|x(t) - c_j(t)\right\|$.

The study used the Gaussian function which the sensitivity can be tuned by adjusting the spread or variance (δ). A larger spread implies less sensitivity (Chandradevan, 2017).

The mathematical equation of the model is as follows:

$$f(x) = \sum_{j=1}^{m} w_{ij}\varphi_j(x), \quad i = 1, \dots, n \tag{4}$$

$$\varphi_j(x) = exp\left[-\frac{\left\|x(t) - c_j(t)\right\|^2}{2\sigma_j^2}\right], \; j = 1, \dots, m \tag{5}$$

where; w_{ij} represents the weight value of a connection, $\varphi_j(x)$ represents the activation function and $\sigma_j = \sqrt{\sigma_j^2}$ is the spread parameter or the square root of variance σ_j^2.

Performance Indicator

The prediction performance of the Multiple Linear Regression and Artificial Neural Network models are evaluated using root mean square error (RMSE) and the coefficients of determination (R-squared) is used. The models that provided the best prediction values were chosen as the best prediction models.

(a) Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) is the square root of the mean of the square of all of the errors. RMSE is very common and is considered an excellent general-purpose error metric for numerical predictions. The deviation between actual and predicted domestic water demand values:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{6}$$

where n is the number of data points, y_i are the observed values and $\hat{y}_i$ are the predicted values. Using RMSE, the lower the value the better the performance of the estimated model.

(b) Coefficients of Determination (R-squared)

The coefficient of determination (R-squared) is calculated to see how close the predicted values are to the true (observed) value. Statistically, R-squared indicates how much variation in the dependent variable (i.e., water demand) is explained by the independent variables. Other than the error measure (RMSE), in the predictive modelling methodology, this indicator is important as a measure to assess the capability of a predictive model, regardless of whether the model is from the linear or non-linear. Using R-square, the higher the value, the better the performance is the estimated model. R-squared has been popularly used to compare several predictive models' performance (Lee & Derrible, 2020; Shuang & Zhao, 2021) in predicting water demand. Model with a high R-square (> 0.8) shows a good capability to be used for prediction (Mahmud *et al*., 2023).

The mathematical equation of R-squared value:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{7}$$

where n is the number of observations, y_i are the observed values, $\hat{y}_i$ are the predicted values and are the average values.

Results and Discussion

Descriptive Statistics

Based on Table 2, the dependent variable, which was domestic water demand, had the minimum and maximum values of 12.54 and 824.25, respectively. While non-domestic water demand had a minimum and maximum value of twice as low as domestic water demand. The average water production was approximately 741 million litres per day (MLD), indicating that the water was sufficient for Malaysians. However, the average unbilled or lost water was higher (276.42 MLD) compared to non-domestic used (159.95 MLD). The average rainfall in Malaysia was about 2,300 mm with an average evaporation of 4 mm and an average temperature of 27ºC. It should be noted that all variables were approximately normally distributed as the skewness values were within the range of ±2 standard deviation, even though the variables of relative humidity, temperature, and wind speed were recorded negatively.

This study used a line chart to visualise the pattern of the three main water types: Domestic water demand, non-domestic water demand, and non-revenue water in Malaysia over 19 years.

Figure 3 depicts domestic water having the highest demand compared to non-domestic water over 19 years. This indicates that the demand for household use was greater than that of the consumers of the agricultural, industrial, and institutional sectors. However, the water loss in Malaysia due to several issues, such as pipe leakages, was greater, as the line plot of non-revenue water (NRW) was higher than non-domestic water. Overall, the three main types of water revealed an increasing pattern over time. The study used a horizontal bar chart to display the proportions of the three main types of water (domestic, non-domestic, and non-revenue).

Referring to Figure 4 above, domestic and non-domestic demand proportions were approximately 40% and 23%, respectively. It was almost a 20% difference between both demands. However, a mere 3% difference between domestic and non-revenue water (NRW) indicated that the unbilled water or water loss in Malaysia was high and almost similar to the domestic water demand.

Table 2: Descriptive statistics

Variables	Statistics				
	Min	Max	Mean	Median	Skewness
Population served (N)	76,100.00	3,741,901.00	1,536,436.00	1,506,211.00	0.32
Design capacity (MLD)	57.43	2,109.45	917.85	929.00	0.05
Water production (MLD)	32.14	1,839.60	740.66	755.50	0.04
Non-revenue water (MLD)	6.61	712.95	276.42	265.80	0.34
Price (RM/m^3)	0.52	0.91	0.68	0.67	0.59
Gross domestic product (RM)	2,146.00	286,297.00	31,678.63	265.80	0.31
Rainfall (mm)	1,299.60	3,358.80	2,256.17	2,285.4	0.13
Evaporation (mm)	3.10	5.20	4.13	4.10	0.04
Relative humidity (%)	75.10	87.50	81.73	81.50	–0.02
Temperature (ºC)	25.70	28.50	27.28	27.30	–0.29
Wind speed (km/h)	1.00	2.70	1.81	1.90	–0.46
Non-domestic water (MLD)	7.37	406.35	159.95	167.50	0.14
Domestic water (MLD)	12.54	824.25	299.67	281.50	0.50

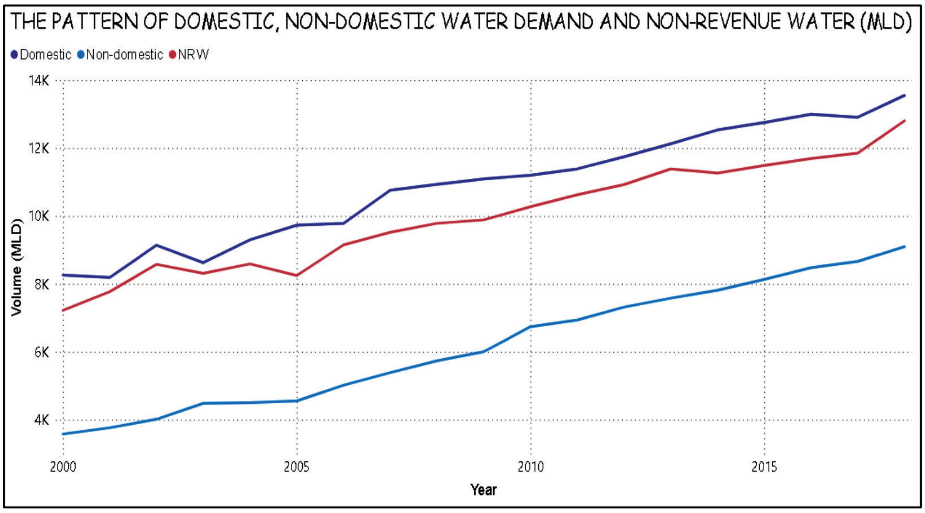


Figure 3: The pattern of the water demands and loss in Malaysia

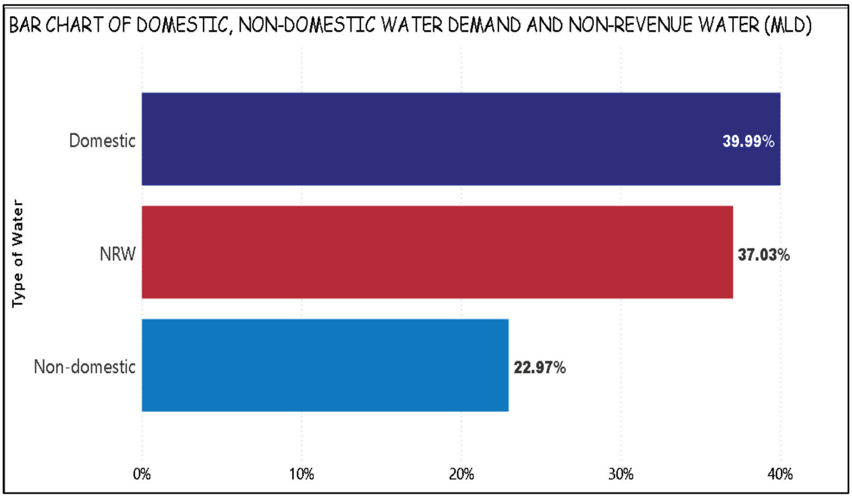


Figure 4: The proportion of water demands and loss in Malaysia

Results of the MLR Model

Several prior analyses have been conducted to ensure the MLR assumptions are met. The dependent variable, domestic water has skewed distribution, indicating the data are not normally distributed, thus the dependent variable requires log transformation. The result of correlation analysis describes the relationship between the independent and dependent variables.

Based on Table 3, at a 5% significance level, as shown by p-value < 0.05, all independent variables have a significant relationship with water demand except rainfall, temperature and relative humidity. As demonstrated by the high correlation coefficient (r > 0.75), the important variables with a strong positive linear relationship with domestic water demand are population served, design capacity, water production and non-domestic water. Meanwhile, non-revenue water and gross domestic product have a moderate relationship (0.5 < r < 0.75).

Table 3: Correlation between the dependent (Log Domestic Water Demand) and the independent variables

Independent variables	**Correlation value (r)**	**P-value**
Population served	0.79	< 0.001
Design capacity	0.88	< 0.001
Water production	0.90	< 0.001
Non-domestic water	0.81	< 0.001
Non-revenue water	0.69	< 0.001
Water price	–0.17	0.005
Gross domestic product	0.73	< 0.001
Rainfall	0.13	0.033
Temperature	–0.12	0.048
Relative humidity	0.02	0.782
Evaporation	–0.29	< 0.001
Wind speed	0.16	0.011

Additionally, water price, rainfall, evaporation and wind speed had a weak relationship ($r < 0.5$) towards log domestic water demand. A negative sign shows an indirect relationship, while a positive sign indicates a direct relationship.

Although scatter plots and correlation matrices can also detect multicollinearity, their results only reveal the bivariate relationship between the independent variables and are just an indication of insights; meanwhile, VIF is the most common and confirmed method. The method can demonstrate the relationship between a variable and several other variables. Referring to Montgomery *et al*. (2012), multicollinearity checking was conducted using the Variance Inflation Factor (VIF) for this study. Based on the VIF value in Table 4, the multicollinearity problem existed since the two variables, water production (WP), and design capacity (DC), had a VIF value of more than 10. After removing water production (WP) variables, the multicollinearity problem was

Table 4: Table of Variance Inflation Factor (VIF)

Variables	**VIF (All Variables)**	**VIF (Remove Water Production)**
Population served	8.11	5.95
Design capacity	15.10	8.34
Water Production	39.27	-
Non-revenue water	6.31	2.84
Price	1.16	1.16
Gross domestic product	5.35	4.92
Rainfall	1.43	1.34
Evaporation	1.85	1.83
Relative humidity	1.95	1.96
Temperature	1.84	1.79
Wind speed	1.32	1.29
Non-domestic water	8.04	5.70

solved with all VIFs less than 10. Hence, a total of 11 independent variables were employed in this study.

MLR model in this study was developed using enter method, where all the independent variables were entered in a single step. In the first trial full model, three variables were insignificant: Ggross domestic product (GDP), evaporation, and wind speed. Next, the development of the final model was obtained by excluding the three non-significant variables from the model. The result of the final estimated model is given in Table 5 as follows.

The equation of the final model is as follows:

$$\begin{aligned}&\text{Log (Domestic water demand)}\\ &= 18.65 + (2.833E-7)\ \text{(Population served)}\\ &\quad + (8.883E-4)\ \text{(Design capacity)} - 1.599\\ &\quad \text{(Water price)} + (6.070E-4)\ \text{(Non-revenue}\\ &\quad \text{water)} + (1.500E-3)\ \text{(Non-domestic)}\\ &\quad - (2.015E-4)\ \text{(Rainfall)} - (8.010-2)\\ &\quad \text{(Relative humidity)} - 2.507\ \text{(Temperature)}\end{aligned}$$

Based on the p-value (< 0.000), the two most significant influential factors of water demand are water price and design capacity. Water prices have a negative association which indicates an indirect relationship, a one-unit increase in price will impact a 1.599 (MLD) reduction in water demand. Meanwhile, design capacity indicated a direct relationship; a one-unit increase in design capacity will impact water demand to increase by (8.883E – 4) (MLD).

(a) MLR Model Diagnostic Checking

Results in Table 6 and Figure 5 provide evidence of the accuracy and validity of the final model.

The R-squared adjusted was 0.8577, the value is high, showing the strong capability of the predictive model. The model was significant as the p-value was less than 0.05. It can be concluded that all important variables in the model explained 86% of the total variation in log domestic water demand.

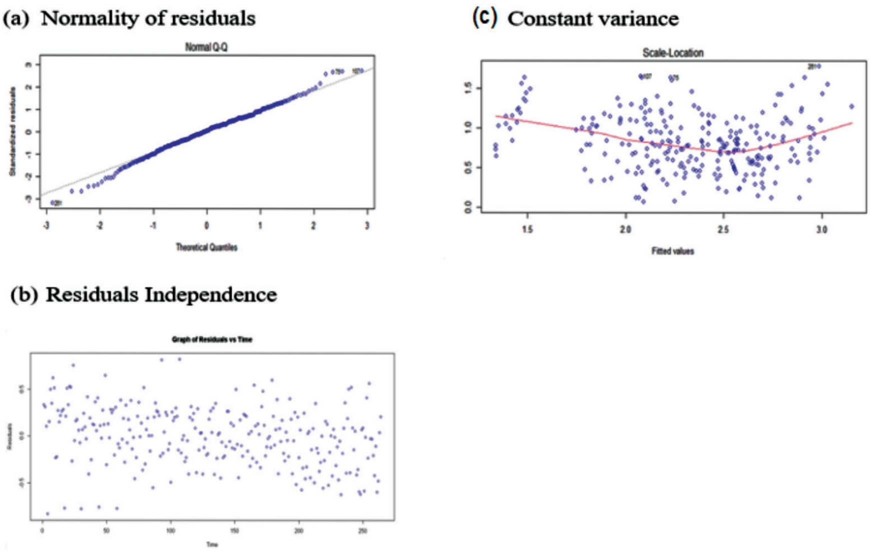
Figure 5 (a) portrays that most points are along a straight line, forming a converged pattern at the distribution's tails. Additionally, the residuals independence plot also exhibits a random pattern [Figure 5 (b)], as does the constant variance plot [Figure 5 (c)]. The results conclude that the residual normality and independent assumptions of MLR are met. The

Table 5: Standardised Coefficient estimates of the reduced model

	Estimate	**Standard Error**	**t-value**	**p-value**
Intercept	18.65	2.696	6.918	0.000
Population served	2.833e-7	5.953e-8	4.759	0.000
Design capacity	8.883e-4	1.410e-4	6.300	0.000
Water price	–1.599	2.380e-1	–6.717	0.000
Non-revenue water	6.070e-4	2.340e-4	2.594	0.010
Non-domestic water	1.500e-3	5.619e-4	2.670	0.008
Rainfall	–2.015e-4	6.593e-5	–3.056	0.003
Relative humidity	–8.010e-2	1.488e-2	–5.383	0.000
Temperature	–0.2507	6.459e-2	–3.881	0.000

Table 6: Summary of the final model obtained

	R-squared	**Adjusted R-squared**	**ANOVA Goodness of Fit (p-value)**
Final Model	0.8639	0.8577	0.000



(a) Normality of residuals

(c) Constant variance

(b) Residuals Independence

Figure 5: MLR model diagnostic checking

results provide evidence that the MLR model obtained is valid.

Results of Neural Network Models: Multi-layer Perceptron and Radial Basis Function

The total of 264 observations was partitioned randomly to 70% for model building and 30% for model validation. However, the exact percentage and the number of observations that were partitioned for the training and testing set using Multi-layer Perceptron and Radial Basis and are shown in Table 7.

The model was validated using 32.6% of the testing dataset. In order to validate the accuracy of the model obtained, the performance indicator of Root Mean Square Error (RMSE) was used. The Relative Error (RE) indicates the percentage of an incorrect prediction. The values of the model performance for both model training and testing are tabulated in Table 8.

The prediction error of the training model was lower than that of the testing model, with the RMSE values of 0.107 and 0.133, respectively. Additionally, there was very little difference in RE value (0.01) between the training and testing models. A smaller error was observed for Multi-layer Perceptron indicating that it is superior to Radial basis Function in the prediction performance.

The variables important to the models can be accessed using the normalised percentage chart, depicted in Figure 6. The percentage was obtained from the value of importance. It measured how much the predicted value changed for different independent variable values.

Table 7: Exact percentage split

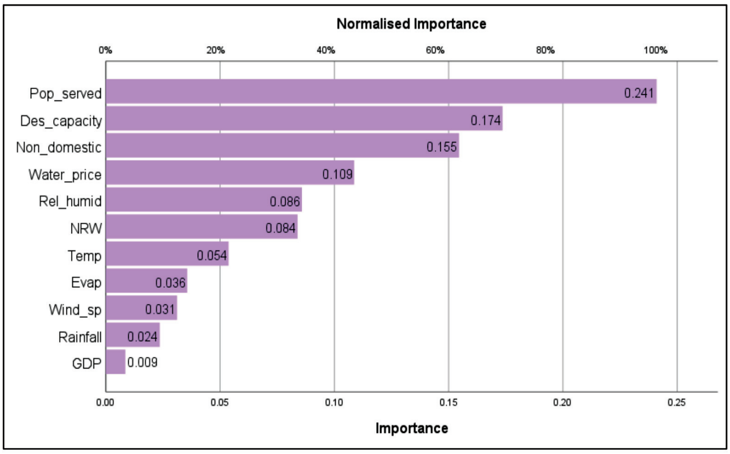
Model	Type of Dataset	Percentage Split	Number of Observations
Multi-layer Perceptron	Training	67.4	178
	Testing	32.6	86
Radial Basis Function	Training	71.2	188
	Testing	28.8	76

Table 8: Performance indicators for training and testing models

Model	Performance Indicators	Training	Testing
Multi-layer Perceptron	Root Mean Square Error (RMSE)	0.107	0.133
	Relative Error (RE)	0.023	0.033
Radial Basis Function	Root Mean Square Error (RMSE)	0.201	0.168
	Relative Error (RE)	0.081	0.040

Multi-layer Perceptron model shows that the highest importance value was 0.241, which was the population served, followed by the design capacity and non-domestic water with 0.174 and 0.155, respectively. The least important variable in the model was gross domestic product (0.009). Meanwhile, the radial basis model found that the variables that were significant to the model could be accessed using the normalised percentage chart. The highest importance value was 0.140, which was the non-revenue water. Followed by population served

(a) Multi-layer Perceptron



(b) Radial Basis Function

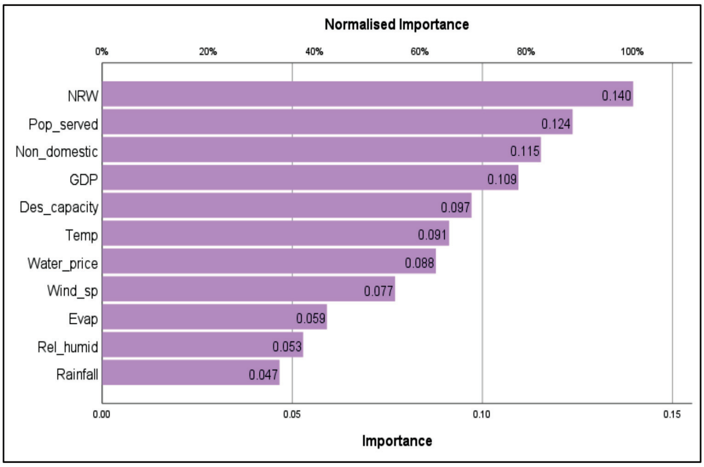


Figure 6: Variables importance of the models

and non-domestic water with 0.124 and 0.115, respectively. The least important variable in the model was rainfall (0.047).

Models Comparison Analysis

The three models' performances were compared to get the best predictive model. The performance indicators, such as the root mean square (RMSE) and coefficient of determination (R-squared), were used to evaluate the accuracy and prediction capability. Table 9 summarises the performance of the three models, Multiple Linear Regression, Multi-layer Perceptron Neural Network, and Radial Basis Function Network, as well as the listing of identified important variables of each model.

Based on the result of the model performances obtained, the Multi-layer Perceptron Neural Network model was chosen as the best predictive model for predicting domestic water demand in Malaysia due to the lowest RMSE and the highest R-squared value. The lowest RMSE value of 0.133 indicated that the prediction error was small thus it was accurate. The highest R-squared value of 0.974 produced by Multi-layer Perceptron Neural Network model confirmed that the model is highly capable of predicting water demand.

Table 9: Comparison of the models' performance

No.	Modelling Technique	Performance Indicator		Important Variables (Rank Order-descending)
		RMSE	R-squared	
1	Multiple Linear Regression	0.281	0.929	Price Design Capacity Population served Relative Humidity Temperature Rainfall Non-domestic water Non-revenue water
2	Artificial Neural Network with Multi-layer Perceptron	0.133	0.974	Population served Design capacity Non-domestic Price Relative humidity Non-revenue water Temperature Evaporation Wind speed Rainfall Gross domestic Product
3	Artificial Neural Network with Radial Basis Function	0.168	0.940	Non-revenue water Population served Non-domestic Gross domestic Product Design capacity Temperature Price Wind speed Evaporation Relative humidity Rainfall

Using the best predictive model obtained, the important factors were determined. The most important factors for predicting domestic water demand in Malaysia were population served (100%), design capacity (72%), non-domestic water (64.1%), price (45.1%), relative humidity (35.6%), non-revenue water (34.8%), temperature (22.3%), evaporation (14.8%), wind speed (12.9%), and rainfall (9.8%). The lowest contributing factor was gross domestic product (3.5%).

Conclusion

The novelty of the paper is to consider the application of Multiple Linear Regression (MLR) and Artificial Neural Networks (ANN) to highlight the important factors in water needs in Malaysia and to determine which one is the best as a prediction tool. Several new considered factors such as water quality, relative humidity and evaporation, were involved in the model. The analysis results indicated that the patterns of water demand for both domestic and non-domestic, as well as non-revenue water, has increased. The proportion of domestic water was almost 40%, and 23% for non-domestic water. The remaining percentage (37%) was unbilled water or water loss. It is imperative for Malaysia better to manage water resources and production in the years ahead.

As mentioned, predictive modelling using Multiple Linear Regression (MLR) and Artificial Neural Networks (ANN) were employed in this study to achieve the best prediction model and to identify the significant contributing factors. MLR modelling began with several assumptions to be met. The final fitted model of MLR produced eight significant variables: Population served, design capacity, water price, non-revenue water, non-domestic water, rainfall, relative humidity and temperature. However, modelling using a Multi-layer Perceptron (MLP) Neural Network and Radial Basis Function (RBF) Network outperformed the MLR model. The adoption of complex algorithms, specifically in the activation function, revealed that the model could predict better. The MLP model had the smallest prediction error (0.133) compared to the RBF Network (0.168). Similarly, the precision of the MLP model surpassed the RBF Network with the R-squared value of 0.974 and 0.940, respectively. The application of the MLP Neural Network in predicting domestic water demand seemed to fit the Malaysia dataset. This chosen technique was further confirmed by a previous study by Hassan (2013) that employed the MLP Neural Network model to predict domestic water demand in Malaysia using fewer variables. The model described the importance of the factors in relation to water demand. However, evidence based on the MLR model, water price is the most significant influential factor. The finding is supported by the study of Anang *et al*. (2019). As the population has increased over the years, the water treatment plant design capacity has also increased. Therefore, this variable was the second-highest contributing factor to the model. The research findings can assist decision-makers in strategic decision-making regarding water management.

Future researchers should use data monthly to ensure an adequate number of observations. When the number of observations is larger, it is easier for the error to be minimised due to possible less bias in the parameter estimation process. Furthermore, the researcher can include other advanced machine learning techniques, such as General Regression Neural Network (GRNN) and Support Vector Regression (SVR), for comparison. It is suggested that the water quality index (WQI) not be removed from the analysis as this variable might contribute more to the model. The best predictive model obtained in this study is recommended to be employed by responsible bodies in assisting the water management for the country as well as ensuring sustainability to meet Goal 6 of the United Nations Sustainable Development Goals (SDGs) 2030 target, which is to provide access to clean water and sanitation for all. Better management of this natural resource can prevent many disasters, including droughts and floods.

Acknowledgement

The authors would like to thank the Malaysian Water Association (MWA), Meteorological Department (MET), Department of Irrigation and Drainage (DID), Department of Statistics (DOSM) and Department of Environment (DOE) for providing data to be used in this study.

Conflict of Interest Statement

The authors declare that they have no conflict of interest.

References

Al-Zahrani, M. A., & Abo-Monasar, A. (2015). Urban residential water demand prediction based on artificial neural networks and time series models. *Water Resource Management, 29*(10), 3651-3662. https://doi.org/10.1007/s11269-015-1021-z

Anang, Z., Padli, J., Abdul Rashid, N. K., Alipiah, R. M., & Musa, H. (2019). Factors affecting water demand: Macro evidence in Malaysia. *Jurnal Ekonomi Malaysia*, *53*(1), 17-25. http://dx.doi.org/10.17576/JEM-2019-5301-2

Andey, S. P., & Kelkar, P. S. (2009). Influence of intermittent and continuous modes of water supply on domestic water consumption. *Water Resource Management*, *23*, 2555-2566. https://doi.org/10.1007/s11269-008-9396-8

Arouna,. A., & Dabbert, S. (2010). Determinants of domestic water use by rural households without access to private improved water sources in Benin: A seemingly unrelated Tobit approach. *Water Resource Management*, *24*, 1381-1398. https://doi.org/10.1007/s11269-009-9504-4

Chan, N. W. (2004). A critical review of Malaysia's accomplishment on water resources management under AGENDA 21. *Malaysian Journal of Environmental Management*, *5*, 55-78. https://core.ac.uk/download/pdf/11491284.pdf

Chandradevan, R. (2017). Radial basis functions neural networks - All we need to know, towards data science. Accessed on May 30, 2021 from https://towardsdatascience.com/radial-basis-functions-neural-networks-all-we-need-to-know-9a88cc053448.

Choudhary, A., & Mushtaq, A. (2023). From pollutant to valuable product: A novel reutilization strategy of wastewater. *International Journal of Chemical and Biochemical Sciences*, *23*(1), 31-37. https://doi.org/10.1016/j.jclepro.2022.134477

Corral - Verdugo, V., Frias - Armenta, M., Perez - Urias, F., Orduna - Cabrera, V., & Espinoza - Gallego, N. (2002). Residential water consumption, motivation for conserving water and the continuing tragedy of the commons. *Environmental Management*, *30*, 527-535. https://doi.org/10.1007/s00267-002-2599-5

Fan, L., Liu, G., Wang, F., Geissen, V., & Ritsema, C. J. (2013). Factors affecting domestic water consumption in rural households upon access to improved water supply: Insights from the Wei River Basin, China. *PLOS One*, *8*(8), 1-9. https://doi.org/10.1371/journal.pone.0071977.

Hassan, F. A. (2013). *Analysis of domestic water consumption in Malaysia* [Master of Engineering (Civil - Hydraulics & Hydrology), Universiti Teknologi Malaysia]. http://eprints.utm.my/id/eprint/33076/5/FilzahAliHassanMFKA2013.pdf

Jorgensen, B., Graymore, M., & O'Toole, K. (2009). Household water use behavior: An integrated model. *Journal of Environmental Management*, *91*, 227-236. https://doi.org/10.1016/j.jenvman.2009.08.009

Kim, H. Y., & Kang, D. W. (2020). South Kore's experience with smart infrastructure services: Smart water management. https://publications.iadb.org/publications/english/

document/South-Koreas-Experience-with-Smart-Infrastructure-Services-Smart-Water-Management.pdf.

Kutner, M. H., Natchtsheim, C. J., Neter, J., & Li, W. (2004). *Applied linear statistical models* (5th ed.). New York: McGraw Hill.

Lee, D., & Derrible, S. (2019). Predicting residential water demand with machine-based statistical learning. *Journal of Water Resource and Planning Management, 146*(1), 1-14. https://doi.org/10.1061/(asce)wr.1943-5452.0001119

Li, H., Zhang, Z., & Liu, Z. (2017). Application of artificial neural networks for catalysis: A review. *Catalysts*, *7*(10), 1-19. https://doi.org/10.3390/catal7100306

Liu, J. (2013). *Radial Basis Function (RBF) Neural network control for mechanical systems*. Springer. https://link.springer.com/content/pdf/10.1007/978-3-642-34816-7.pdf

Machingambi, M., & Manzungu, E. (2003). An evaluation of rural communities' water use patterns and preparedness to manage domestic water sources in Zimbabwe. *Physics Chemistry and Earth, Parts A/B/C*, *28*(20-27), 1039-1046. https://doi.org/10.1016/j.pce.2003.08.045

Mahmud, Z., Shaadan, N., & Aimran, A. N. (2023). A handbook on research methodology: A simplified version, UiTM Press, UiTM.

Markopoulos, A. P., Georgiopoulos, S., & Manolakos, D. E. (2016). On the use of back propagation and radial basis function neural networks in surface roughness prediction, *Journal of Industrial Engineering International*, *12*(3), 389-400. https://doi.org/10.1007/s40092-016-0146-x.

Mikaiil, A. T., Rouki, M., Azarbakhsh, B., & Yousef, G. (2023). Urban development and its relationship with the water crisis in the next twenty years (2042) in the city of Zanjan, Iran. *Asian Journal of Science, Technology and Society*, *2*(1), 1-16.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis*. Hoboken, New Jersey: A John Wiley & Sons, Inc.

Muhammad, A. U., Li, X., & Feng, J. (2019). Artificial intelligence approaches for urban water demand forecasting: A review. In Zhai, X., Chen, B., Zhu, K. (Eds.), *Machine learning and intelligent communications*. MLICOM 2019. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering (Vol. 294). Cham: Springer. https://doi.org/10.1007/978-3-030-32388-2_51

Nur Syuhada, C. I., Mahirah, K., & Roseliza, M. A. (2020). Dealing with attributes in a discrete choice experiment on valuation of water services in East Peninsular Malaysia. *Utilities Policy*, *64*, 101037. https://doi.org/10.1016/j.jup.2020.101037

Payus, C. (2020). Impact of extreme drought climate on water security in North Borneo: Case study of Sabah'. *Water* (Switzerland), *12*(4), 1-19. https://www.mdpi.com/2073-4441/12/4/1135

Qi, Y., Chan, F. K. S., Thorne, C., O'Donnel, E., Quagliolo, C., Comino, E., Pezzoli, A., Li, L., Griffiths, J., Sang, Y., & Feng, M. (2020). Addressing challenges of urban water management in Chinese sponge cities via natured- based. *Water*, *12*(10), 2788. https://doi.org/10.3390/w12102788

Renwick, M. E., & Green, R. D. (2000). Do residential water demand side management policies measure up? An analysis of eight California water agencies. *Journal of Environment Economic and Management*, *40*, 37-55.

Shuang, Q., & Zhao, R. T. (2021). Water demand prediction using machine learning methods: A case study of the Beijing–Tianjin–Hebei Region in China. *Water*, *13* (310), 1-16. https://doi.org/10.3390/w13030310

Syme, G. J., & Shao, Q., & Po, M., & Campbell, E. (2004), Predicting and understanding home garden water use. *Landscape Urban Planning*, *68*, 121-128 .

Yaacob, M. R., Radam, A., & Samdin, Z. (2011). Willingness to pay for domestic water service improvements in Selangor, Malaysia: A choice modeling approach. *International Business Management, 2,* 30-39. http://dx.doi.org/10.17576/JEM-2018-5203-4