Support Vector Regression with Missing Data Treatment Based Variables Selection for Water Level Prediction of Galas River in Kelantan Malaysia

Noraini Ibrahim and Antoni Wibowo
Department of Computer Science
Faculty of Computing
81310, UTM Johor Bahru, Johor, Malaysia
noraini87@live.utm.my and antoni@utm.my

Abstract: Rising in water level becomes an important issue in the state of Galas River in Kuala Krai-Kelantan Malaysia since it is one of important indicator toward to flooding when it achieves a certain level. The increasing of water level is influenced by some factors which called the predictor variables such as month, rainfall, temperature, relative humidity and surface wind. The data for this analysis including the predictors and water level as response were collected from Water Resources Management and Hydrology Division Department of Irrigation and Drainage Malaysia and Malaysian Meteorological Department. However, we noticed there are missing values in the collected data. The selection of suitable predictor variables useful for developing prediction model since the analysis data uses many variables. The suitable predictor variables are selected using Support Vector Regression (SVR) and Cross Validation to obtain an appropriate predictive water level of Galas River Kuala Krai. We take into account the K-fold cross-validation for determining of the dominant variables and best model. However, we need to perform pre-processing data of the datasets since the original data contain missing values. We perform two types of pre-processing data using mean (type I pre-processing data) and Ordinary Linear Regression (type II pre-processing data) to overcome the existing of the missing values. Our experimental `result` shows that the Gaussian kernel is the suitable kernel function with type I pre-processing data for the predicting water level in Galas River.

Key-Words: Galas River, Kelantan, support vector regression, missing value, water level, nonlinear regression, cross validation, variables selection.

1 INTRODUCTION

The rising of water in an existing waterway such as river stream or drainage ditch can caused flooding. The effect of the rising water level beyond the danger level can bring the long term floods which may last days or weeks which are common cited as being the most lethal of all natural disasters [7, 10, 18]. Floods in east coast of Peninsular Malaysia has worsened with more than 10,800 victims are evacuated to flood relief centres in December 2012. There are several districts in Kelantan that suffered from this flood in which Pasir Mas has the highest number of evacuees (2049), followed by Kuala Krai (355), Kota Bharu (264), Tanah Merah (217), Pasir Puteh (245), Machang (47) and Tumpat (41).

Kelantan is a state in the east coast of Peninsular Malaysia that suffers from flooding event which occurs between October and March every year during the northeast monsoon period. In order to facilitate the prediction of flooding in the river and the warning beforehand, this paper aims to build a model on the relation between the selected predictors and the water level of Galas River by adopting Support Vector Regression (SVR).

The Kelantan River is about 248 km long and it divides into the Galas and Lebir Rivers near Kuala Krai, about 100 km from the river mouth. It means that Kelantan River is the main river while Galas and Lebir Rivers are the tributary rivers. In this paper, we focused on one main tributary of Kelantan River which is Galas River in Kuala Krai, Kelantan.

Figure 1 Location of Kelantan, Malaysia

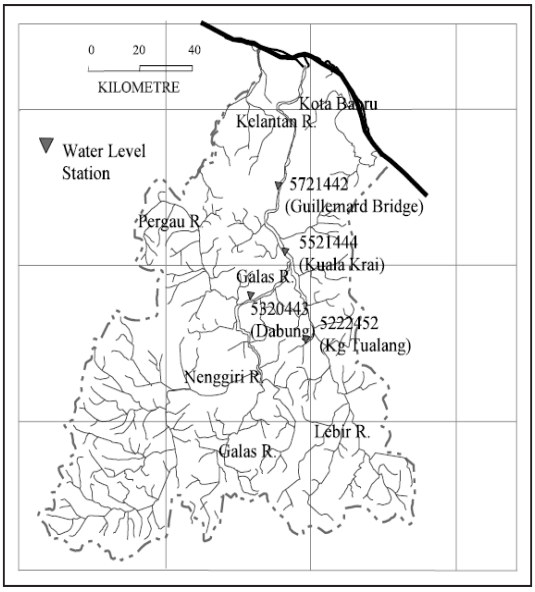

Figure 2 Location of Galas River

Figure 1 presents the location of Kelantan and Figure 2 shows the study area which is Galas River which is denoted as Galas R. in the map.

The datasets for this analysis are collected from Water Resources Management and Hydrology Division (WRMHD) and Malaysian Meteorological Department (MMD) in order to predict the water level of Galas River Kuala Krai. Department of Irrigation and Drainage (DID) Malaysia, is one department of WRMHD, has been measured some of flood characteristics which are water level, area inundation, peak discharge, volume of flow and duration. There are three categories of water level was introduced by DID including alert, warning and danger level. It is noticed that the original datasets contain missing data and we need to perform missing data treatment to overcome this problem. There are five factors that were identified and related to the rising level of the Galas River which can lead to the occurrence of flood phenomenon in Kuala Krai-Kelantan which are months from January until December for 11 years starting from 2001 until 2011, monthly mean of rainfall, monthly mean of temperature, monthly mean of relative humidity and monthly mean of surface wind.

In practice, water level is the variable that is considered when a flood warning is issued. Based on the previous researches, there are some approaches to forecast water level such as based on Muskingum routing model [13, 19] and using soft computing approaches including neural networks, a fuzzy logic model and genetic algorithms [12, 19]. Recently, the application of SVR has more attention in the field of hydrological engineering. We noticed that one-lead-day-rainfall forecasting and runoff forecasting has been performed using SVR in which the input data are pre-processed by Singular Spectrum Analysis [6, 19].

Besides, SVR is applied to forecast the flood stage in Dhaka, Bangladesh, and concluded that the accuracy of SVR exceeds that of Artificial Neural Network (ANN) in one-lead-day to seven-lead-day forecasting [19, 20]. Other application of SVR is in forecasting the real-time flood stage in Lan-Yang River, Taiwan and it shows that SVR can effectively predict the flood stage one-to-six-hours ahead [19]. SVR method has been widely used to forecast floods since last two decades. SVR also have been used in other application such as estimating sonic log distributios in Anadarko Masin, Oklahama [4], predicting the circulation rate in a vertical tube thermosiphon reboiler [21], analyzing urban atmosphere pollution [23] and predicting the colleges recruiting students [8].

The linear problems can be solved using linear SVR by finding the linear regression for the linear case but the linear function approximation is not

practical to be used in the real world problems. Hence, the nonlinear SVR is used to solve the nonlinear problems by mapping the input space into a higher dimensional feature space via a function. In this paper, SVR cannot be directly used in our case study since the original data consist of missing data. Therefore, we perform two types pre-processing data and data standardization in order to use SVR.

The following sections present an approach to the development of the water level models. Theories and methods are discussed in Section 2 while the evaluation of the prediction quality is described in Section 3. The datasets, experimental results and discussions are reported in Section 4. Finally, the conclusion is given in Section 5.

2 THEORIES AND METHODS

In recent years, there has been a lot of interest in studying support vector machines (SVMs) in the field of machine learning such as Optimization of Surface Roughness in End Milling using Potential Support Vector Machine [11]. SVMs are a class of supervised learning algorithms initially proposed by Vapnik [22]. In this paper, we conduct SVR (SVM for regression) for forecasting the water level of Galas River Kuala Krai - Kelantan.

2.1 Nonlinear Support Vector Regression

The nonlinear issues of forecasting can be handled by performing linear regression in the feature space. Figure 3 presents the concept of nonlinear SVR. Suppose $\mathbf{x}$ is mapped into a feature space by a nonlinear function $\phi(\mathbf{x})$, then the linear regression in feature space is given by

$$f(\mathbf{w},b) = \mathbf{w} \,.\, \phi(\mathbf{x}) + b \tag{1}$$

where $\mathbf{w}$ represents the parameter vector and b represents scalar of the function.

The smaller parameter vector $\mathbf{w}$ means a smoother and less complex approximating function. The tolerated errors within the extent of the ε-tube, as well as the penalized losses L_ε when data concern the outside of the tube, are defined by the so-called Vapnik's ε-intensive loss function as [19]

$$L_\varepsilon(\mathbf{y}_i) = \begin{cases} 0 & \text{for } |\mathbf{y}_i - (\mathbf{w} \,.\, \mathbf{x}_i + b)| \le \varepsilon \\ |\mathbf{y}_i - (\mathbf{w} \,.\, \mathbf{x}_i + b)| - \varepsilon & \text{for } |\mathbf{y}_i - (\mathbf{w} \,.\, \mathbf{x}_i + b)| > \varepsilon \end{cases} \tag{2}$$

Using the same manner of linear SVR, a nonlinear regression problem can be obtained by solving the following optimization model:

$$\begin{aligned} \min_{\mathbf{w},\mathbf{b},\xi,\xi^*} \quad & \frac{1}{2}\mathbf{w}^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ \text{subject to} \quad & \mathbf{y}_i - (\mathbf{w} \,.\, \phi(\mathbf{x}_i) + b) \le \varepsilon + \xi_i \\ & (\mathbf{w} \,.\, \phi(\mathbf{x}_i) + b) - \mathbf{y}_i \le \varepsilon + \xi_i \\ & \xi_i, \xi_i^* \ge 0, \; i = 1,2,\ldots,n \end{aligned} \tag{3}$$

where ξ_i and ξ_i^* represent the slack variables that specify the upper and the lower training errors subject to an error tolerance ε, and C represents the positive constant that determines the degree of penalized loss when a training occurs. The dual set of Lagrange multipliers, α_i and α_i^*, are introduced in order to enable the optimization problem to be solved more easily by adopting the standard quadratic programming algorithm. The dual form of the nonlinear SVR is given by

$$\begin{aligned} \min_{\alpha_i,\alpha_i^*} \quad & \frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\left\langle \phi(\mathbf{x}_i) \,.\, \phi(\mathbf{x}_j) \right\rangle \\ & + \varepsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) - \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) \\ \text{subject to} \quad & \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \\ & 0 \le \alpha_i \le C, \quad i = 1,2,\ldots,n \\ & 0 \le \alpha_i^* \le C, \quad i = 1,2,\ldots,n \end{aligned} \tag{4}$$

where $\left\langle \mathbf{x}_i \,.\, \mathbf{x}_j \right\rangle$ is the inner product of $\mathbf{x}_i$ and $\mathbf{x}_j$, α_i and α_i^* are determined after the Lagrange multipliers while C is a positive constant that determines the degree of penalized loss when training error occurs. The remaining nonzero coefficients $-(\alpha_i + \alpha_i^*)$ which are outside the ε-intensive tube are involved in the final decision function and the data that have nonzero Lagrange multipliers are called the support vectors.

The objective function of (4) can be replaced by a kernel function $K(\mathbf{x}_i, \mathbf{x})$ according to Mercer's condition [19, 22]. The decision function of nonlinear SVR can be expressed as follows.

$$f(\mathbf{x}) = \sum_{i=1}^{n} (-\alpha_i + \alpha_i^*) K(\mathbf{x}_i, \mathbf{x}) + b \quad (5)$$

The computing of *b* can be done by exploiting the so called Karush–Kuhn–Tucker (KKT) conditions which states that the point of the product solution between dual variables and constrains has to vanish [2].

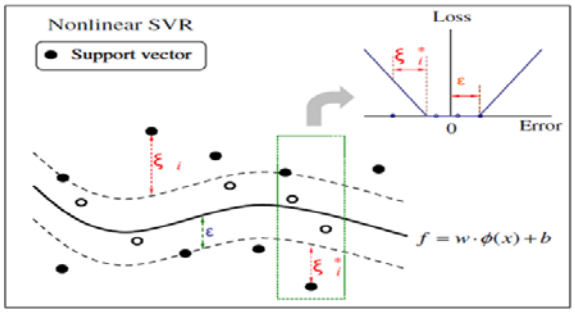

Figure 3 Nonlinear SVR with Vapnik's ε - intensive loss function [19].

In this paper, we used four types of kernel function as follows.

(a) Linear kernel:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j \quad (6)$$

(b) Polynomial kernel:

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j + t)^d \quad (7)$$

where *t* is the intercept and *d* is the degree of the polynomial.

(c) Gaussian(RBF) kernel:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2} \right) \quad (8)$$

with σ is the parameter of Gaussian kernel.

(d) Sigmoid kernel:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh\left[-c + \frac{\mathbf{x}_i \mathbf{x}_j}{\sigma^2} \right] \quad (9)$$

with $c \geq 0$ and σ^{-1} is the scaling vector.

3 EVALUATING THE QUALITY OF THE PREDICTION

Cross-validation (CV) is a statistical method that can be used to evaluate the performance of machine learning based prediction algorithms [17]. In this paper, CV is used in selection of predictors and model selection for predicting water level of Galas River. The CV is a statistical method to evaluate the algorithms by dividing the data into two segments which are for training and validation and the basic form of CV is *K-fold* CV. Machine-learning-based prediction algorithms are trained by the training subsets and tested by the validation subset.

Stratified 10-fold CV was recommended as the best model selection method since it tends to provide less biased estimation of the accuracy compared to regular cross-validation, leave-one-out CV and bootstrap methods [15, 16]. The selection the types of CV can be based on the size of the datasets. The *K-fold* CV is used for this research since it can reduce the computation time and maintain the accuracy of the estimation. For this analysis, we used *10-fold* CV because it can give accurate performance estimation and suitable for small samples of performance estimation. The performance of the prediction algorithms can be estimated by the mean squared error of cross-validation (MSECV). We used CV to choose an appropriate model by comparing the value of *mean squared error of cross-validation* (MSECV) using four types of kernel function and two types of pre-processing data.

The data are divided into *K* segments of roughly equal size and the inner sum of MSECV is taken over the observations in the *kth* segment [1, 3, 15]. For each of *K* experiments, the *K-fold* CV uses *K*-1 folds for training and the remaining is used for testing. There is an advantage of using *K-fold* CV which is all the examples in the dataset are eventually used for both training and testing. We will select a better model according to lowest value of MSECV and it is a measure of how well the model fits the data. Figure 3 presents the

example of the procedure for three-fold cross-validation. We apply this same procedure for 10-fold cross-validation.

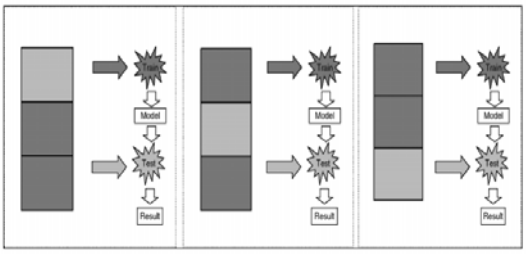

Figure 4: The procedure for three-fold cross-validation [16].

In our experiments, we used MSECV for parameter selection as follows:

$$\text{MSECV} = \frac{1}{K}\frac{1}{T}\sum_{k=1}^{K}\sum_{i=1}^{T}(\mathbf{y}_{ki} - \hat{\mathbf{y}}_{ki})^2 \quad (10)$$

and used mean square error (MSE) for validation of the best model as follows:

$$MSE_k = \frac{1}{T}\sum_{i=1}^{T}(\mathbf{y}_{ki} - \hat{\mathbf{y}}_{ki})^2, \quad (11)$$

where $k = 1, 2, ..., 10$, $T = 12$ and $K = 10$. $\mathbf{y}_{ki}$ is the actual value and $\hat{\mathbf{y}}_{ki}$ is the predicted value using SVR.

3.1 Datasets

As predictors in predicting water level of Galas River, months (x_1), average monthly rainfall (x_2), temperature (x_3), relative humidity (x_4) and surface wind (x_5) were identified and related to the occurrence of flood phenomenon in Kuala Krai Kelantan. Observed predictors and response for the period 2001-2011 were extracted from WRMHD Kuala Lumpur and DID Selangor. It is noted that the data consist of missing values for rainfall and water level and we performed cleaning data to replace these missing values. The data are separated into two sub datasets which are 120 data for developing models and variables selection using *10-fold* CV and 12 data for validating the models. The data that were used in this analysis are shown in Table 1.

3.1.1 Original data

The data set is cover from January until December for 11 years and yet it has shown a total of 132 data. Table 1 describes the summary of the collected data from WRMHD and MMD, while Table 2 is a snapshot of original data. From Table 2, the 47th row represents the item data November 2004. We can see that there is NA value in this row which means that there is a missing data of rainfall in November 2004.

3.1.2 Pre-processing data

Data pre-processing is the process that was performed to the original data in order to prepare it for next processing procedure. Thus, it will transform the data into the format that more effective according to our purpose of analysis. Data pre-processing is important since the real world data normally are noisy which are containing errors and outliers. There are five types of performing data preprocessing which are data cleaning, data integration, data transformation, data reduction and data discretization. For this analysis, we performed two types of data cleaning which are using mean and OLR to replace the missing values of rainfall and water level.

Table 1: Details of the data

Station	Period	Data
Kuala Krai	2001-2011	Monthly 24 h MeanTemperature Monthly 24 h Mean relative humidity Monthly mean surface wind
Dabong	2001-2011	Monthly mean rainfall and water level

Table 2: The snapshot of original data of Galas River

Month	Rainfall	Temperature	Relative humidity	Surface wind	Water level
37	9.840	25.7	86.5	0.5	27.58
38	0.810	26.4	81.5	0.7	27.00
39	6.510	27.4	82.8	0.7	26.84
40	1.930	27.8	82.1	0.5	26.29
41	4.120	28.0	83.4	0.5	26.48
42	12.450	27.6	83.1	0.6	26.17
43	1.875	26.7	83.5	0.7	26.08
44	7.390	27.2	83.6	0.7	26.12
45	13.180	26.5	85.6	0.6	28.26
46	9.600	25.9	88.5	0.3	28.49
47	NA	26.0	88.9	0.2	27.76
48	NA	25.0	89.3	0.2	29.12

Table 3: The snapshot of pre-processing data of galas river using type I pre-processing data

Month	Rainfall	Temperature	Relative humidity	Surface wind	Water level
37	9.840	25.7	86.5	0.5	27.58
38	0.810	26.4	81.5	0.7	27.00
39	6.510	27.4	82.8	0.7	26.84
40	1.930	27.8	82.1	0.5	26.29
41	4.120	28.0	83.4	0.5	26.48
42	12.450	27.6	83.1	0.6	26.17
43	1.875	26.7	83.5	0.7	26.08
44	7.390	27.2	83.6	0.7	26.12
45	13.180	26.5	85.6	0.6	28.26
46	9.600	25.9	88.5	0.3	28.49
47	12.750	26.0	88.9	0.2	27.76
48	163.280	25.0	89.3	0.2	29.12

Table 4: The snapshot of pre-processing data of Galas River using type II pre-processing data

Month	Rainfall	Temperature	Relative humidity	Surface wind	Water level
37	9.840	25.7	86.5	0.5	27.58
38	0.810	26.4	81.5	0.7	27.00
39	6.510	27.4	82.8	0.7	26.84
40	1.930	27.8	82.1	0.5	26.29
41	4.120	28.0	83.4	0.5	26.48
42	12.450	27.6	83.1	0.6	26.17
43	1.875	26.7	83.5	0.7	26.08
44	7.390	27.2	83.6	0.7	26.12
45	13.180	26.5	85.6	0.6	28.26
46	9.600	25.9	88.5	0.3	28.49
47	6.570	26.0	88.9	0.2	27.76
48	6.570	25.0	89.3	0.2	29.12

Table 5: The snapshot of standardized data of Galas River using type I pre-processing data

Month	Rainfall	Temperature	Relative humidity	Surface wind	Water level
0.37	0.0984	0.257	0.865	0.5	0.2758
0.38	0.0081	0.264	0.815	0.7	0.2700
0.39	0.0651	0.274	0.828	0.7	0.2684
0.40	0.0193	0.278	0.821	0.5	0.2629
0.41	0.0412	0.280	0.834	0.5	0.2648
0.42	0.1245	0.276	0.831	0.6	0.2617
0.43	0.0188	0.267	0.835	0.7	0.2608
0.44	0.0739	0.272	0.836	0.7	0.2612
0.45	0.1318	0.265	0.856	0.6	0.2826
0.46	0.0960	0.259	0.885	0.3	0.2849
0.47	0.1275	0.260	0.889	0.2	0.2776
0.48	1.6328	0.250	0.893	0.2	0.2912

Table 6: The snapshot of standardized data of Galas River using type II pre-processing data

Month	Rainfall	Temperature	Relative humidity	Surface wind	Water level
0.37	0.0984	0.257	0.865	0.5	0.2758
0.38	0.0081	0.264	0.815	0.7	0.2700
0.39	0.0651	0.274	0.828	0.7	0.2684
0.40	0.0193	0.278	0.821	0.5	0.2629
0.41	0.0412	0.280	0.834	0.5	0.2648
0.42	0.1245	0.276	0.831	0.6	0.2617
0.43	0.0188	0.267	0.835	0.7	0.2608
0.44	0.0739	0.272	0.836	0.7	0.2612
0.45	0.1318	0.265	0.856	0.6	0.2826
0.46	0.0960	0.259	0.885	0.3	0.2849
0.47	0.0657	0.260	0.889	0.2	0.2776
0.48	0.0657	0.250	0.893	0.2	0.2912

3.1.2.1 Pre-processing data using mean of the corresponding months

For this subsection, we used mean which is represented by type I pre-processing data to replace these missing values. For example, NA value of rainfall in November 2004 is replaced by the mean of the non missing rainfall data in November from 2001 until 2011. Table 3 presents the snapshot of the pre-processing data using mean of the corresponding months for Galas River.

3.1.2.2 Pre-processing data using ordinary linear regression

The second type of cleaning data that we used is OLR and we represent it as type II pre-processing data. We performed OLR to replace the missing values of the dataset in Galas River. Table 4 shows the snapshot of the pre-processing data using OLR for Galas River.

The model to replace the missing value of the water level for Galas River is given as [14]:

$$f_{OLR_{WL1}}(x_1) = 26.5767 + 0.0135x_1 \quad (12)$$

The model to replace the missing values of the rainfall for Galas River is represented by [14]:

$$f_{OLR_{RF1}}(x_1) = 6.4736 + 0.0021x_1 \quad (13)$$

3.1.3 Data Standardization

Data standardization is the process to make the datasets internally consistent in regression when there is a big difference in the values of the datasets. We standardized the datasets by multiply the column of Month, Rainfall, Temperature, Relative Humidity and Water Level by 0.01 in order to make the value of datasets is in the same range. The snapshot of standardized datasets is shown in Table 5 and Table 6 using two types of pre-processing data.

4 EXPERIMENTAL RESULTS AND DISCUSIONS

In this section, experiments of two real datasets which are type I and type II pre-processing data have been carried out using SVR to obtain lowest MSECV. As we mentioned before, we adopted four types of kernel which are linear, Gaussian, polynomial and sigmoid kernel. All the

calculations are carried out with programs developed in MATLAB 7.14 [9]. In this paper, we used SVR and the input variables are month, rainfall, temperature, relative humidity and surface wind while the response is water level. For this analysis, we use *10-fold* CV in SVR and some parameters have to be determined before running SVR. Table 7 gives an overview of parameters setting for SVR.

4.1 Selection of predictors

The selection of appropriate predictors is one of the most important steps in predicting the water level of Galas River. The predictors are chosen based on the smallest value of MSECV and the result is compared between two types of pre-processing data which are type I pre-processing data and type II pre-processing data.

4.2 Tuning parameters with SVR

The parameters of SVR are determined according to previous research on parameter optimization in SVR [5]. According to Table 7, there are four parameters in all, i.e. C, ε, σ, and d to be tuned. Parameter C denotes the penalty (cost) parameter of the training error in the RBF kernel function, parameter d represents the degree of polynomial kernel function and ε denotes the epsilon-insentive value in epsilon-SVR [5]. The interesting point in this analysis is the fact that the local optimal values can be found in only a few iterations (in this case 30 iterations).

We tried to increase the number from 30 iterations to 100 iterations, but the forecasting error did not increase significantly. The results are given in Table 8,Table 9 and Table 10.Table 8 shows the result of MSECV using Gaussian RBF kernel for type I and type II pre-processing data while Table 9 and Table 10 present the results of the four types of kernel function to select the dominant predictors. From these results, we can see that the value of MSECV using RBF kernel of type II pre-processing data with five predictors which are x_1, x_2, x_3, x_4,x_5 is smaller than that of type I pre-processing data with five predictors which are x_1, x_2, x_3, x_4, x_5.The comparison results for different types of kernel function are shown in Table 9 and Table 10.

From these results, we can see that the value of MSECV using RBF kernel of type II pre-processing data with five predictors which are x_1, x_2, x_3, x_4, x_5 is smaller than that of type I pre-processing data with five predictors which are x_1, x_2, x_3, x_4, x_5 . The best model according to four types of kernel function is marked in bold. In all models, the best model is the RBF kernel function for type I pre-processing data with five predictors are chosen and 6.7073E-06 MSECV. Based in the previous researches, the RBF seemed to be the best choice for the type of SVR kernel for non-linear forecasting [5]. According to this claim, it shows that our analysis also achieved RBF for the best choice of kernel function in prediction.

Based on the results obtained by SVR in Table 9, we found that the suitable kernel function type of SVR is RBF and the optimal parameters are C=1, $\varepsilon = 0.001$ and $\sigma = 4$. After the *10-fold* CV, five predictors that have the lowest MSECV which are x_1, x_2, x_3, x_4, x_5 were selected to predict the water level of Galas River, Kuala Krai-Kelantan. In order to test the reasonabilty of SVR model, we apply the real data of water level measured at Galas River in 2011. The performance of the selected final model is evaluated using MSE by taking the data of real water level at Galas River from January until December 2011. The result of MSE between the predicted data and the real data is 5.8565E-05 by using the optimal parameters of SVR.

Table 7 : SVR parameter settings

Parameter	Value
C	1, 5, 10, 20, 30, infinite
ε	0.001, 0.005
σ	0.01-15
d	1-10

MSECV													
		ε =0.001						ε =0.005					
Pre-processing Data	Predictors	C=1	C=5	C=10	C=20	C=30	C=Infinite	C=1	C=5	C=10	C=20	C=30	C=Infinite
Type I	x_1, x_2, x_3, x_4, x_5	8.7754E-06	1.2628E-05	1.2628E-05	1.2628E-05	1.2628E-05	1.2628E-05	1.1220E-05	1.0047E-05	1.2133E-05	1.2628E-05	1.2628E-05	1.2628E-05
	Parameter	(4.5)	(1)	(1)	(1)	(1)	(1)	(11.5)	(13)	(13.5)	(1)	(1)	(1)
	x_1, x_2, x_3, x_4	1.0099E-05	1.2188E-05	1.2188E-05	1.2188E-05	1.2188E-05	1.2188E-05	1.0399E-05	2.0047E-05	2.0047E-05	2.0047E-05	2.0047E-05	2.0047E-05
	Parameter	(8.5)	(1)	(1)	(1)	(1)	(1)	(3)	(1)	(1)	(1)	(1)	(1)
	x_1, x_2, x_4, x_5	1.0743E-05	1.2628E-05	1.2628E-05	1.2628E-05	1.2628E-05	1.2628E-05	1.7336E-05	1.7582E-05	1.7492E-05	1.7713E-05	1.7713E-05	1.7713E-05
	Parameter	(15.5)	(1)	(1)	(1)	(1)	(1)	(4)	(13)	(15.5)	(1)	(1)	(1)
	x_1, x_2, x_3, x_5	1.0946E-05	1.2628E-05	1.2628E-05	1.2628E-05	1.2628E-05	1.2628E-05	1.0415E-05	1.1024E-05	1.1024E-05	1.1024E-05	1.1024E-05	1.1024E-05
	Parameter	(15.5)	(1)	(1)	(1)	(1)	(1)	(15.5)	(1)	(1)	(1)	(1)	(1)
	x_1, x_2, x_3	4.2098E-05	4.2223E-05	4.9894E-05	6.5618E-05	8.1989E-05	1.4053E-03	9.9560E-06	1.2085E-05	2.3121E-05	8.6327E-05	1.7675E-04	8.1764E-03
	Parameter	(13)	(13)	(14)	(15)	(14.5)	(15.5)	(11.5)	(15.5)	(15.5)	(15.5)	(10)	(15.5)
	x_1, x_2, x_4	2.4689E-05	7.1403E-05	1.6152E-04	3.5278E-04	6.4134E-04	2.5798E-03	2.0844E-05	4.5900E-05	1.0619E-04	3.7651E-04	8.0716E-04	5.9850E-03
	Parameter	(15.5)	(15.5)	(15.5)	(15.5)	(15.5)	(15.5)	(15.5)	(15.5)	(15.5)	(15.5)	(15.5)	(15.5)
	x_1, x_2, x_5	1.1219E-05	6.0341E-05	2.4271E-04	5.9873E-04	0.0017	4.5363E-03	1.0947E-05	1.1024E-05	1.1024E-05	1.1024E-05	1.1024E-05	1.1024E-05
	Parameter	(15.5)	(15.5)	(15.5)	(6)	(11.5)	(15.5)	(15.5)	(1)	(1)	(1)	(1)	(1)
	x_1, x_2	1.1812E-05	9.8075E-06	4.1550E-05	1.0511E-05	3.9483E-05	4.5406E-03	1.1246E-05	1.2089E-05	2.5467E-05	7.5259E-05	1.4558E-04	2.8234E-03
	Parameter	(11.5)	(6.5)	(8)	(12)	(11.5)	(15.5)	(8.5)	(15.5)	(15.5)	(15.5)	(15.5)	(15.5)
Type II	x_1, x_2, x_3, x_4, x_5	**6.7073E-06**	1.0964E-05	1.0964E-05	1.0964E-05	1.0964E-05	1.0964E-05	7.8688E-06	1.0964E-05	1.0964E-05	1.0964E-05	1.0964E-05	1.0964E-05
	Parameter	(4)	(1)	(1)	(1)	(1)	(1)	(6.5)	(1)	(1)	(1)	(1)	(1)
	x_1, x_2, x_3, x_4	7.6923E-06	9.0606E-06	9.0606E-06	9.0606E-06	9.0606E-06	9.0606E-06	1.0617E-05	1.1038E-05	1.1038E-05	1.1038E-05	1.1038E-05	1.1038E-05
	Parameter	(11.5)	(1)	(1)	(1)	(1)	(1)	(14.5)	(1)	(1)	(1)	(1)	(1)
	x_1, x_2, x_4, x_5	7.5024E-06	1.0964E-05	1.0964E-05	1.0964E-05	1.0964E-05	1.0964E-05	2.9817E-05	3.0484E-05	3.7931E-05	4.7600E-05	4.7600E-05	4.7600E-05
	Parameter	(14)	(1)	(1)	(1)	(1)	(1)	(6.5)	(14)	(15.5)	(1)	(1)	(1)
	x_1, x_2, x_3, x_5	2.7589E-05	2.9794E-05	3.1558E-05	3.1151E-05	4.1014E-05	4.7600E-05	1.0964E-05	1.0964E-05	1.0964E-05	1.0964E-05	1.0964E-05	1.0964E-05
	Parameter	(4)	(7.5)	(10)	(14)	(15)	(1)	(1)	(1)	(1)	(1)	(1)	(1)
	x_1, x_2, x_3	8.0130E-06	3.9046E-05	1.0772E-04	4.9479E-04	3.7454E-04	5.0625E-03	1.0958E-05	2.7579E-05	3.3473E-05	1.4151E-04	1.3385E-04	3.6923E-03
	Parameter	(12.5)	(15.5)	(15.5)	(15.5)	(13.5)	(15.5)	(15.5)	(15.5)	(5.5)	(10.5)	(10.5)	(15.5)
	x_1, x_2, x_4	7.7914E-06	1.6094E-05	4.5878E-05	2.0028E-04	3.4629E-04	5.6544E-05	1.0824E-05	2.7082E-05	5.7199E-05	3.4992E-04	9.0617E-05	3.9022E-03
	Parameter	(12.5)	(15.5)	(15.5)	(15.5)	(15.5)	(15.5)	(6.5)	(13.5)	(15.5)	(1)	(1)	(1)
	x_1, x_2, x_5	7.2149E-05	1.3492E-05	2.9715E-05	1.6053E-04	3.4001E-04	4.1703E-03	3.1288E-05	3.1592E-05	3.6082E-05	4.7600E-05	4.7600E-05	4.7600E-05
	Parameter	(11.5)	(15.5)	(15.5)	(15.5)	(10)	(15.5)	(6.5)	(13.5)	(15.5)	(1)	(1)	(1)
	x_1, x_2	2.4449E-05	8.3850E-05	1.0283E-04	2.9346E-04	2.6392E-04	1.4122E-03	2.3304E-05	7.4962E-05	2.2349E-04	7.4946E-04	0.0014	3.1767E-03
	Parameter	(15.5)	(15.5)	(8)	(11.5)	(15)	(15.5)	(15.5)	(15.5)	(15.5)	(15.5)	(15)	(15)

Table 8: MSECV for variables selection of Galas River using Gaussian RBF kernel

Table 9: MSECV for variable selection of galas river using type I pre-processing data

		Kernel Function			
Optimal values	Optimal predictors	RBF	Poly	Sigmoid	Linear
Optimal CVMSE	x_1, x_2, x_3, x_4, x_5	8.7754E-06	9.2112E-06	8.8217E-06	0.0432
Optimal C		1	1	30	1
Optimal ε		0.001	0.001	0.001	0.001
Optimal σ		4.5	-	2	-
Optimal d		-	4	-	-

Table 10: MSEV for variable selection of Galas River using type II pre-processing data

		Kernel Function			
Optimal values	Optimal predictors	RBF	Poly	Sigmoid	Linear
Optimal CVMSE	x_1, x_2, x_3, x_4, x_5	**6.7073E-06**	9.0606E-06	9.0884E-06	0.0051
Optimal C		1	1	30	1
Optimal ε		0.001	0.005	0.005	0.001
Optimal σ		4	-	1	-
Optimal d		-	4	-	-

5. CONCLUSIONS

The conclusion of this paper is summarized as follows:

i. We presented the use of SVR to obtain the dominant parameters in selecting the best predictors for water level forecasting according to lowest MSECV using *10-fold* CV.

ii. Missing data treatment have been done to replace the missing value of the data sets by adopting two types of pre-processing data which are type I and type II pre-processing data.

iii. Four types of kernel functions such as Gaussian kernel, polynomial kernel, linear kernel and sigmoid kernel are used in SVR in order to obtain and compares the MSECV.

iv. The performance of this method has been evaluated by carrying out the experiments of two types pre-processing data which are type I and type II pre-processing data.

v. Based on this analysis, the results show that RBF kernel is the suitable kernel with lowest MSECV and five predictors are selected as the predictors to predict the water level of Galas River in Kuala Krai.

6. ACKNOWLEDGEMENT

This project is funded by the Fundamental Research Grant Scheme (FRGS) (vot number: 4F084). The authors would like to thank the Research Management Centre for supporting this research and Drainage and Irrigation Department of Malaysia for general assistant.

References:

[1] A. C. Davison, and D. V. Hinkley, *Bootstrap methods and their application, Cambridge Series in Statistical and Probabilistic Mathematics*, Cambridge University Press, Cambridge, UK, 1997.

[2] A. J. Smola and B. Scholkopf, *A tutorial on support vector regression,* NeuroCOLT2 Technical Report Series, 2003.

[3] B. H. Mevik, and H. R. Cederkvist, Mean squared error of prediction (MSEP) estimates for principle component regression (PCR) and partial least squares regression (PLSR), *Journal of Chemometrics*, 2004, 18(9): 422-429.

[4] C. Cranganu and M. Breaban, Using support vector regression to estimate sonic log distributions: A case study from the Anadarko basin. *Oklahoma, Journal* of Petroleum science and Engineering, 2013, 103: 1-13.

[5] C. H. Wu, G. H. Tzeng and R. H. Lin, A novel hybrid genetic algorithm for kernel function and parameter optimization in support vector regression, *Expert Systems with Applications*, 2009, 36:4725-4735.

[6] C. Sivapragasam, S. Y. Liong and M. F. K. Pasha, Rainfall and runoff forecasting with SSA–SVM approach, *Journal of Hydroinformatics* 3 (3), 2001, 141–152.

[7] D. Alexander, *Natural disasters*, UCL Press, London, 1993.

[8] E. Ying, Application of support vector regression algorithm in colleges recruiting students prediction, *International Conference on Computer Science and Electronics Engineering, 2012.*

[9] http://www.codeforge.com/read/7180/ svmSim.m_html.

[10] J. G. French and K. W. Holt, Floods, In M.B. Gregg (ed.) T*he public health consequences of disasters, US Department of Health and Human Services, Public Health Service*, CDC, Atlanta, GA, pp. 69–78, 1989.

[11] K. Kadirgama, M. M. Noor and M. M. Rahman, Optimization of surface roughness in end miling using potential support vector machine, *Arab J Sci Eng*, 2012, 37:2269-2275.

[12] L. See and S. Openshaw, Applying soft computing approaches to river level forecasting, *Hydrological Sciences Journal* 44 (5), 1999, 763–778.

[13] M. Franchini and P. Lamberti, A flood routing Muskingum type simulation and forecasting model based on level data alone. *Water Resources Research*, 30 (7), 1994, 2183–2196.

[14] N. Ibrahim and A. Wibowo, Partial least squares regression based variables selection for water level predictions, *American Journal of Applied Science,* 2013, 10 (4): 322-330.

[15] N. Ibrahim, and A. Wibowo, Predictions of water level in Dungun River Terengganu using partial least squares regression, *International Journal of Basic and Applied Sciences* IJBAS/IJENS, Vol: 12 No:02, 2012.

[16] P. Refaeilzadeh, L. Tang and H. Liu, *Cross-validation*, Arizona State University, 2008.

[17] S. Cheng and M. Pecht, Using cross-validation for model parameter selection of sequencial probability ratio test, *Expert Systems with Applications* 39, 2012, 8467-8473.

[18] S. N. Jonkman, and I. Kelman, An analysis of the causes and circumstances of flood disaster deaths, *Disasters*, 2005, 29(1): 75−97.

[19] S. P. Yu, S. T. Chen and I. F. Chang, Support vector regression for real-time flood stage foreasting*, Journal of Hydrology*, 328, 2006, 704-716.

[20] S. Y. Liong, and C. Sivapragasam, Flood stage forecasting with support vector machines, *Journal of the American Water Resources Association* 38 (1), 2002, 173–196.

[21] S. Zaidi, Development of support vector regression (SVR)-based model for prediction of circulation rate in a vertical tube thermosiphon reboiler, *Chemical engineering Science*, 2012, 69: 514-521.

[22] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 1995.

[23] Y. Xue, L. Yu, K .Cao and P. Xu, Using support vector regression to analyze urban atmosphere pollution with optical remote sensing data, *Journal of Earth Science and Engineering*, 2013, 3:180-189.