CHAPTER 4

METHODOLOGY

In this chapter, the methodology is divided into three main sections: 1) data pre-processing, 2) modeling, and 3) the proposed prediction method. Data preprocessing is the first step that prepares data for modelling which can be divided into four parts: data collection and descriptive statistics, imputation, transformation, and normalization. Modelling is the core of water quality parameter prediction which consists of 2 sub-sections: parameter selection and model comparison. Finally, the new method which is used to select to the best model is proposed. The overview of the methodology is shown in Figure 4.1.

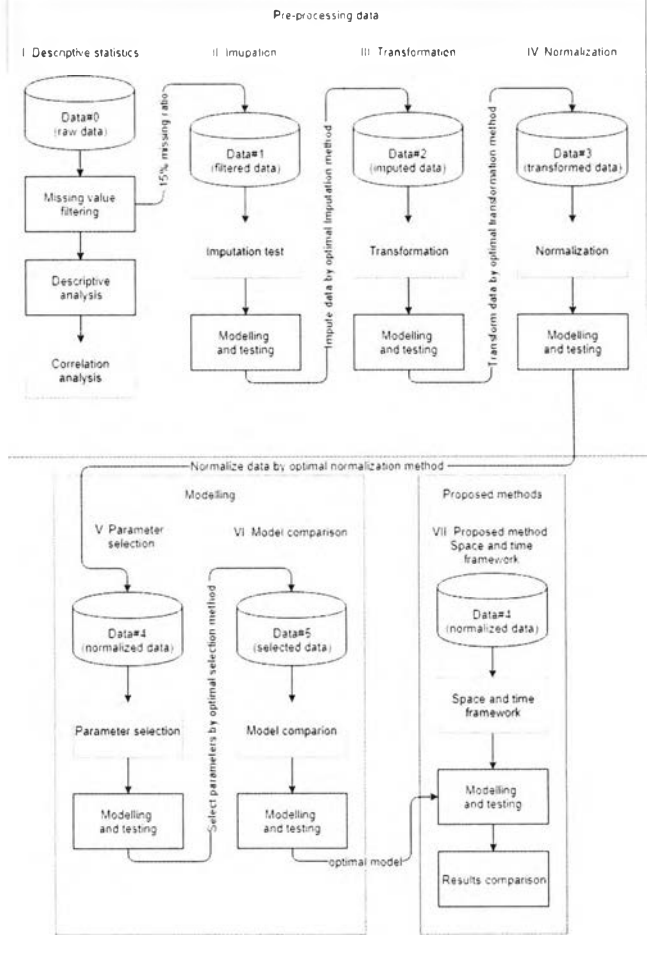

Figure 4.1 Overview and flowcharts of methods used in this dissertation

4.1 Data preprocessing

Data preprocessing section aims to prepare data for modelling which can be divided into four parts: data collection and descriptive statistics, imputation, transformation, and normalization. Each parts are explained in this Chapter.

4.1.1 Data collection and descriptive statistics

The water quality data of Chaophraya River collected by the Pollution Control Department, Ministry of Natural Resources and Environment are used to find missing value percentage and perform descriptive statistical analysis such as mean and standard deviation. After that Spearman correlation coefficient is used to determine linear relationship between water quality parameters, monitoring station location, monitoring month and monitoring year as shown in Figure 4.2.

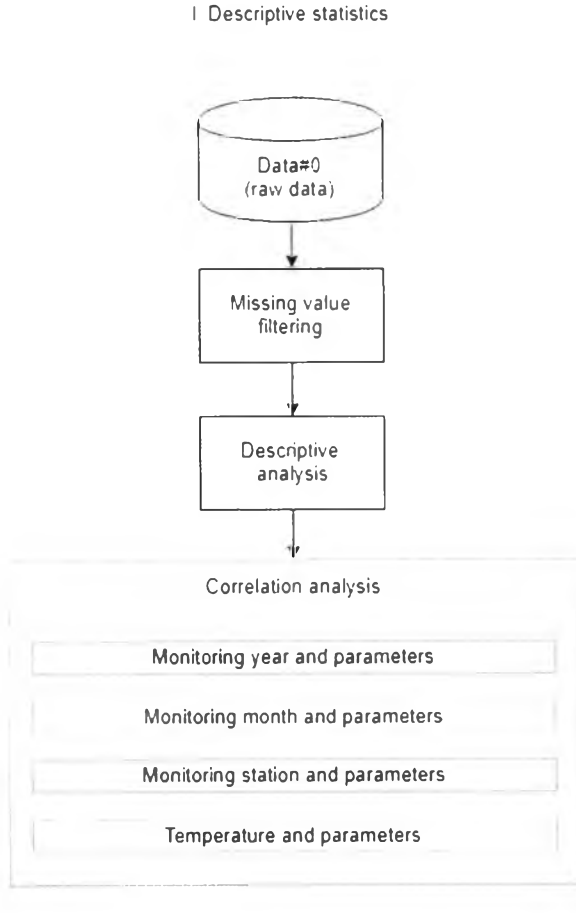

Figure 4.2 Data collection and descriptive analysis flowchart

Instead of using geographic coordinate latitude and longitude, the distance from estuary (Gulf of Thailand) is used as monitoring station location because latitude and longitude system cannot show relationship between stations. The distance from estuary shows the sequence and the distance between monitoring stations and distance from sea which could be useful for water quality parameter prediction. Location of 19 monitoring stations along Chaophraya River are shown in Figure 4.3 and Table 4.1.

Table 4.1 Detail of monitoring stations along Chaophraya River

Province	Detail of monitoring stations			
	Code	Address	Lat.	Long.
Samut Prakan	CH01	Phra Samut Chedi, Muang District	672495 N	1503718 E
	CH03	Phra Pradaeng District Office	666474 N	1510119 E
Bangkok	CH06	Bangkok Port (Fish Market), Yan Nawa District	669779 N	1515411 E
	CH08	Bangkok Bridge, Dao Khanong District	661571 N	1514712 E
	CH10	The Memorial Bridge, Samphanthawong District	662225N	1519063E
Nonthaburi	CH12	Rama VI Bridge, Bang Kruay District	664167 N	1527303 E
	CH15	Nonthaburi Bridge, Pakkred District	666174 N	1542211 E
Pathum Thani	CH16	Raw water pumping station for water supply, Muang District	667292 N	1551717 E
	CH17	Samkhok District	665076 N	1555811 E
Ayutthaya	CH18	Bangpa-in Paper Plant, Bang Pa-in District	668291 N	1569511 E
	CH20	Petch Fortress, Samphao Lom, Phra Nakhon Si Ayutthaya	670298 N	1586207 E
Ang Thong	CH21	Chaophraya River Bridge, Muang District	656788 N	1613207 E

Table 4.1 Detail of monitoring stations along Chaophraya River (cont'd)

Province	Detail of monitoring stations			
	Code	Address	Lat.	Long.
Sing Buri	CH24	Chaophraya River Bridge, Muang District	651186 N	1647307 E
	CH25	Central market, Indra Buri District	643068 N	1660098 E
Chai Nat	CH27	Chao Phraya Dam, Muang District	627201N	1675915E
	CH28	City Hall, Muang District	620791N	1678621E
Nakhon Sawan	CH30	Somdet Phra Wan Rat Bridge, Tha Nam Oi, Phayuha Khiri	622189 N	1705401 E
	CH31	Wat Maneewong, Yan Matsi, Phayuha Khiri	619792 N	1719201 E
	CH32	Dechatiwong Bridge, Muang District	620681 N	1734498 E

Variables of water quality parameters are defined in Table 4.2. There are 16 water quality parameters and other two variables which are monitoring time and monitoring station location are collected as one record. To avoid confusion, all 18 variables are called "parameter" in this dissertation.

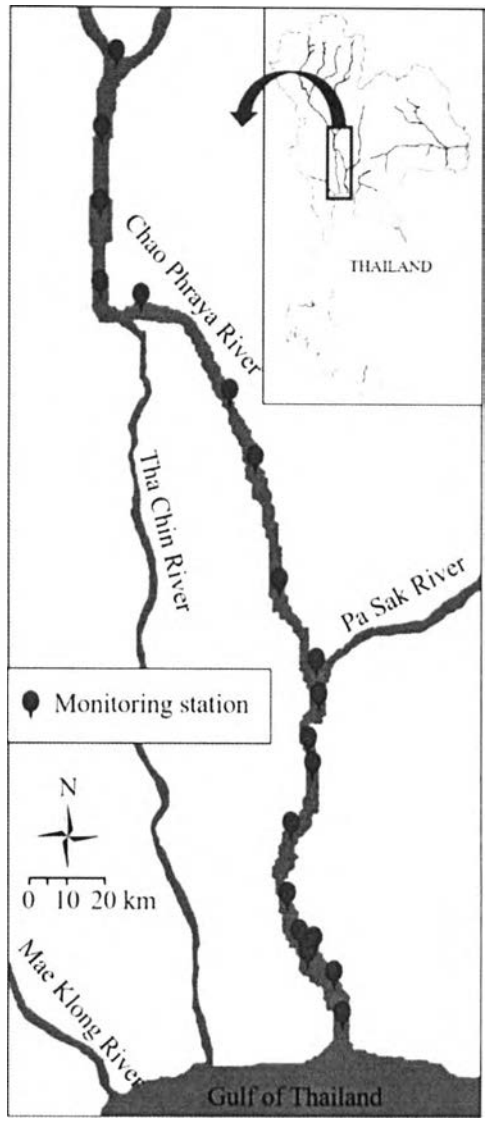

Figure 4.3 Chaophraya River and location of 19 monitoring stations along the river

Table 4.2 List of water quality parameters in Chaophraya River, Thailand

Variable name	Parameter code	Parameter name	Unit
WQ_1	WT	water temperature	°C
WQ_2	pH	acidity or basicity of water	-
WQ_3	Tur	turbidity	NTU
WQ_4	EC	electrical conductivity	µS/cm
WQ_5	Sal	salinity	g/L
WQ_6	DO	dissolved oxygen	mg/L
WQ_7	BOD	biochemical oxygen demand	mg/L
WQ_8	TC	total coliform bacteria	MPN/100 ml
WQ_9	FC	fecal coliform bacteria	MPN/100 ml
WQ_{10}	PO_4^{3-}	phosphate concentration	mg/L
WQ_{11}	NO_3^-	nitrate concentration	mg/L
WQ_{12}	NO_2^-	nitrite concentration	mg/L
WQ_{13}	NH_3	ammonia concentration	mg/L
WQ_{14}	SS	suspended solid	mg/L
WQ_{15}	TS	total solid	mg/L
WQ_{16}	TDS	total dissolved solid	mg/L
WQ_{17}	T	monitoring month	-
WQ_{18}	S	distance from estuary	km

The missing value percentage shows the possibility to impute data under acceptable error. The threshold value is defined at 15% which means that any parameters that have more than 15% missing ratio are discarded. Only parameters with less than 15% missing ratio are used in the next step.

Spearman correlation coefficient is the simple method that can be used to show relation between two parameters (see 3.2.1 for detail of Spearman correlation analysis). In this part, four relationships are analysed roughly in order to determine which parameter should be used as input to predict a parameter.

Spearman correlation coefficient between water quality parameter and monitoring year are determined to show long term trend of each parameter over time. For example, if positive correlation exists, the parameter has upward direction over the long-term. On the other hand, negative correlation means downward trend is shown on that parameter over the long-term.

Spearman correlation coefficient between water quality parameters and monitoring months indicate seasonal effect on each parameter. For example, some of them may be affected by rainfall in rainy season, some parameters may be affected by human activities in each season such as agricultural season.

Spearman correlation coefficient between water quality parameters and distances from monitoring stations and estuary can show upstream and downstream effects along the river. Low value of the distance means the station is a downstream station, positive correlation indicates that a increment of parameter depends on location and it tends to be higher on downstream. Another possible factor is the sea water effect, some parameters such as salinity is directly affected by the distance from the sea.

Spearman correlation coefficient between each water quality parameter can show interaction among parameters. In this part, correlation between water temperature and other parameters are shown as examples of interactions. Temperature is fundamental effect which relates to many parameters.

4.1.2 Imputation

The selected parameters from 4.1.1, which are less than 15% missing value percentage are imputed in this part. Three imputation methods are used to calculate imputed data which are mean replacement, K-nearest neighbor (K-nn) and artificial neural network (ANN) as showed in Figure 4.4. Detail of each methods are mentioned in 3.2.2.

There are some argument in the imputation method need to initialize. Initialized argument of each method are shown in Table 4.3 [207]. Then, the imputed data are used to predict water quality parameters and evaluated performance. The performance evaluation tested to find the suitable imputation method by various model. The model setting is mentioned in details on 4.2. Performance of the models generated by three imputed data is evaluated by RMSE and Spearman correlation. The optimal imputed data will be used in the next step.

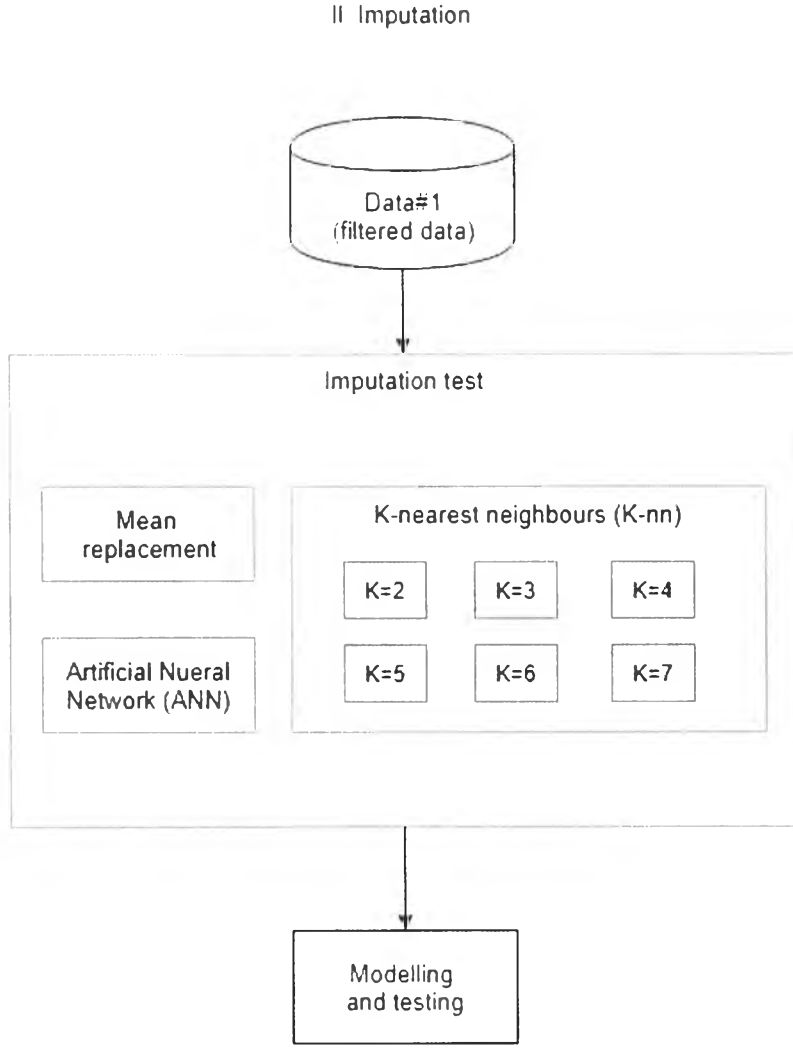

Figure 4.4 Imputation experiment flowchart

Table 4.3 Three imputation methods argument setting

Imputation method	Argument setting
Mean replacement	no argument
K-nn	k = 2,3,...,7
	(totally 6 experiments)
ANN	training cycle (epoch) = 500
	learning rate = 0.3
	momentum = 0.2

4.1.3 Data transformation

The imputed data from 4.1.2 are checked whether which kind of transformation is suitable for each parameters for used in water quality prediction model. First, all parameters are transformed with all possible transformation equations which are mentioned on 3.2.3. Then, skewness of transformed data are calculated. Finally, the transformation which give skewness closest to zero is selected for each parameter and set them as transformed data.

Transformed data and non-transform data (imputed data from 4.1.2) are used to predict water quality parameters and evaluated performance. The performance evaluation tested to find the suitable imputation method by various models. The model setting is mentioned in detail on 4.2. Performance of models are evaluated by Spearman correlation. The optimal one will be used in the next step. Overview of transformation experiment is shown in Figure 4.5.

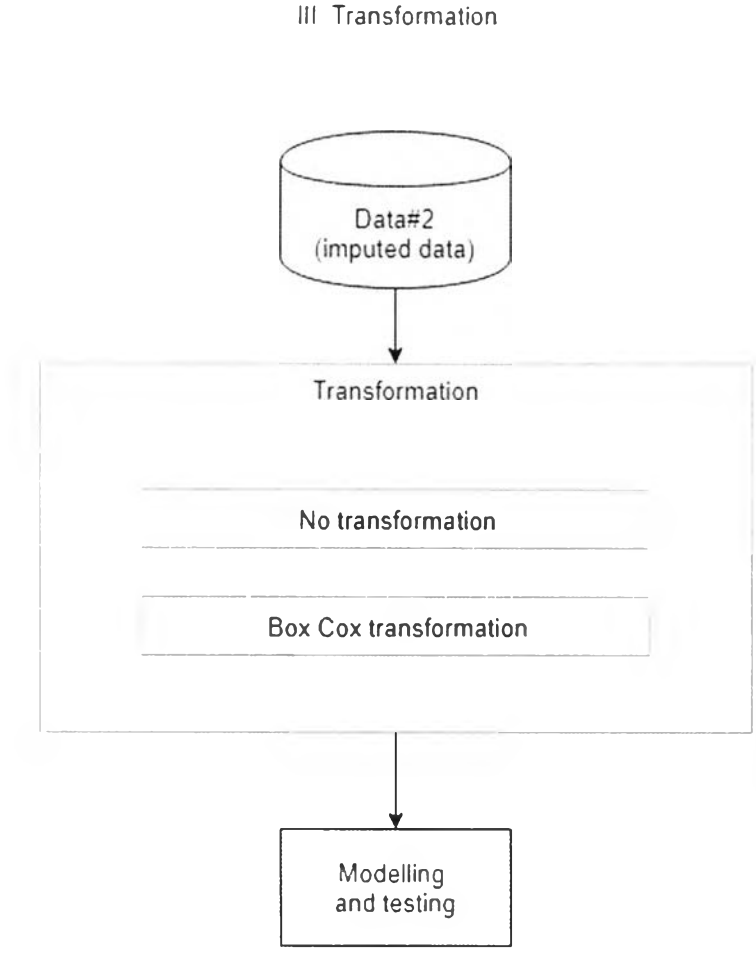

Figure 4.5 Transformation experiment flowchart

4.1.4 Normalization

Transformed data from 4.1.3 are normalized with four methods. As showed in Figure 4.6, Z normalization, range normalization, proportion normalization and interquartile normalization are implement to normalize data (see 3.2.4 for detail of each method). There are two arguments in range normalization which are x_{min} and x_{max}. To avoid the saturation effect [55, 56, 208, 209], they are set as 0.1 and 0.9, respectively. Another three imputation need no initialized argument. Then, normalized data are used to predict water quality parameters and evaluated performance. The performance evaluation tested to find the suitable normalization method by various model. The model setting is mentioned in detail on 4.2. Performance of models that

are generated by four normalized data are evaluated by RMSE and Spearman correlation. The optimal normalized data will be used in the next step.

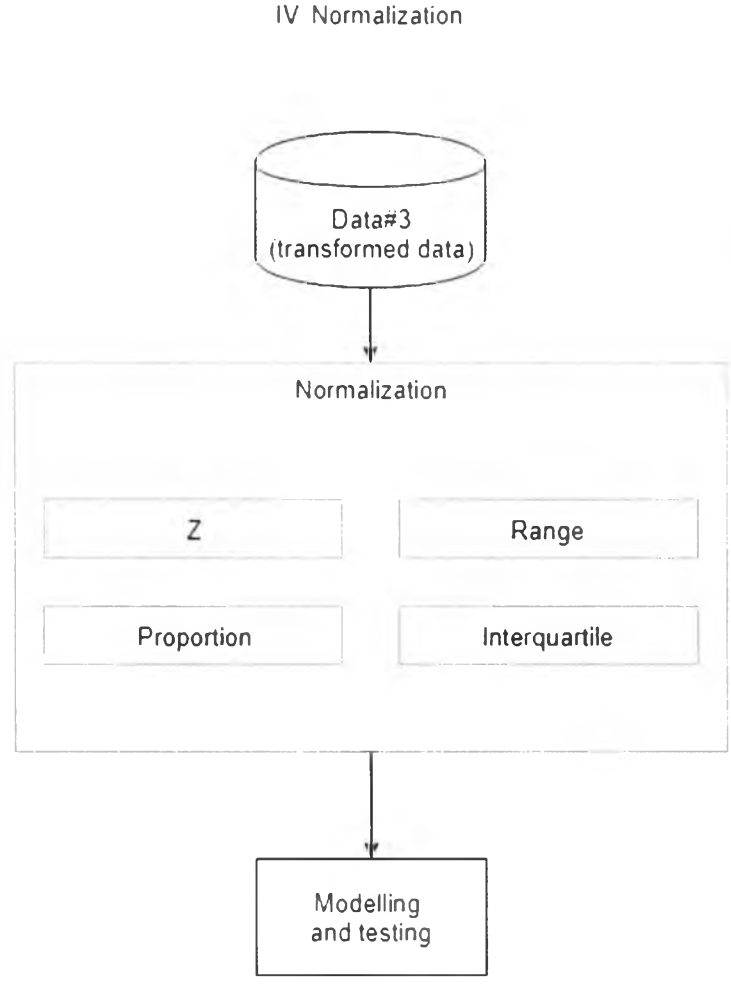

Figure 4.6 Normalization experiment flowchart

4.2 Modelling

Modelling section aims to predict water quality parameter which consists of two parts: parameter selection and model comparison. Each parts are explained in this Chapter.

4.2.1 Parameter selection

Four parameter selection methods which consist of forward selection (FS), backward elimination (BE), principal component analysis (PCA) and genetic algorithm

(GA) are implement with various model for water quality parameters prediction. The normalized parameters from 4.1.4 are used as input of parameter selection methods. After selection, the selected parameters are fed into models. Models are trained and evaluated performance. The performance of each methods are compared to find the suitable parameter selection method. The optimal methods will be used in the next step. Overview of parameter selection experiment is shown in Figure 4.7.

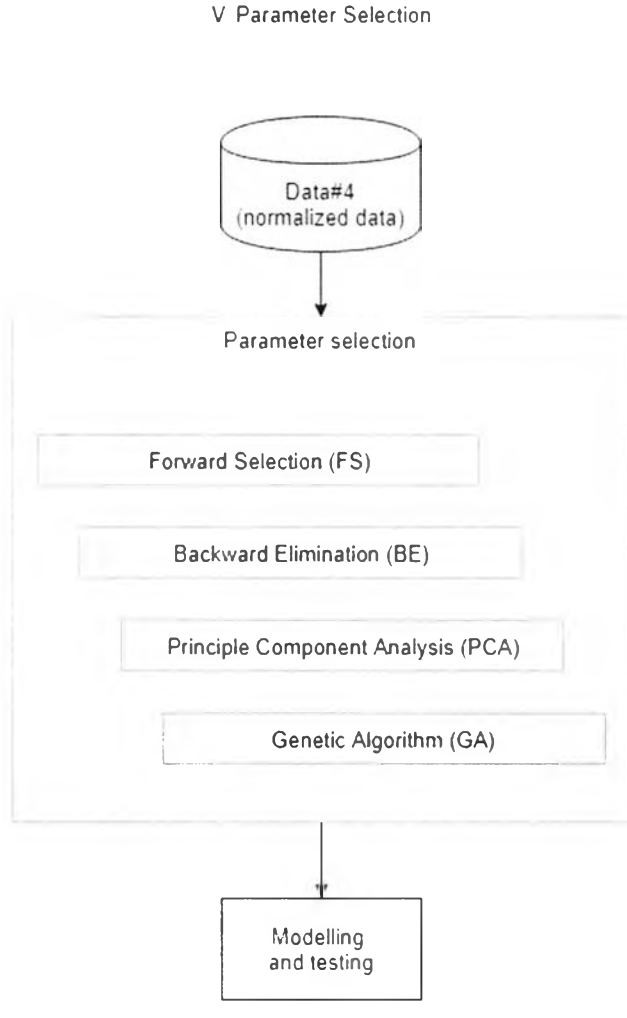

Figure 4.7 Parameter selection experiment flowchart

According to 3.2.5, some methods need pre-defined argument. In this study, the argument setting is shown in Table 4.4.

Table 4.4 Argument setting for parameter selection methods

Parameter selection	Argument setting
Forward selection	-
Backward elimination	-
PCA	Variance threshold = 95%
Genetic Algorithm	Number of children (C) = 20 Number of generation (g) = 200 Mutation probability = $\frac{1}{n}$ where n is the total number of parameters. Crossover probability = 0.5

4.2.2 Model comparison

In this part, Support vector regression (SVR), artificial neural network (ANN) and multiple linear regression (MLR) are used to predict water quality parameter. For modelling and testing, normalized Chaophraya River quality parameters of 17 years (1250 records x 18 parameters) are divided (by monitoring time) into two subsets which are training set and testing set. Data which collected in 2539 – 2551 B.E. are set as training set and the rest for testing set (2552 – 2556 B.E.). The training set and testing set comprised of 945 (70%) and 375 (30%) records, respectively. The prediction output is the water quality parameter which corresponding to the input parameters belong to the same water sample, thus collected in same time and station. The model can call as "Nowcasting model". For example, BOD normally takes 5 days to determine is predicted by other 17 parameters which collected in the same time and station.

The predicted parameter are compared with observed parameter to determine the performance of models. Two different criteria are used which are root mean square error (RMSE) and Spearman correlation coefficient. RMSE represents the error associated with the model and can be computed as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \quad (4.1)$$

where $\hat{y}_i$ and y_i are predicted and observed value of parameter, and n is the number of records. RMSE can measure the goodness of fit and describes an average measure of the error in predicting the parameter.

Table 4.5 Argument setting for models

Model	Argument setting
Support vector regression (SVR)	$c = 0$ $\varepsilon = 0.00001$ Kernel type = dot product (linear) Maximum epoch = 100000
Artificial neural network (ANN)	$\varepsilon = 0.00001$ Learning rate = 0.3 Momentum = 0.2 Maximum epoch = 100-1000 Number of hidden layer = 1 Number of hidden neuron = $\lfloor (n+1)/2 \rfloor + 1$ where n is number of input node
Multiple linear regression (MLR)	-

Argument of each model are reviewed and tested by trial and error process, until the optimal the optimal condition is found and show in Table 4.5 [207, 210-212]. Overview of this part is shown in Figure 4.8. All model are implemented by Rapid Miner software.

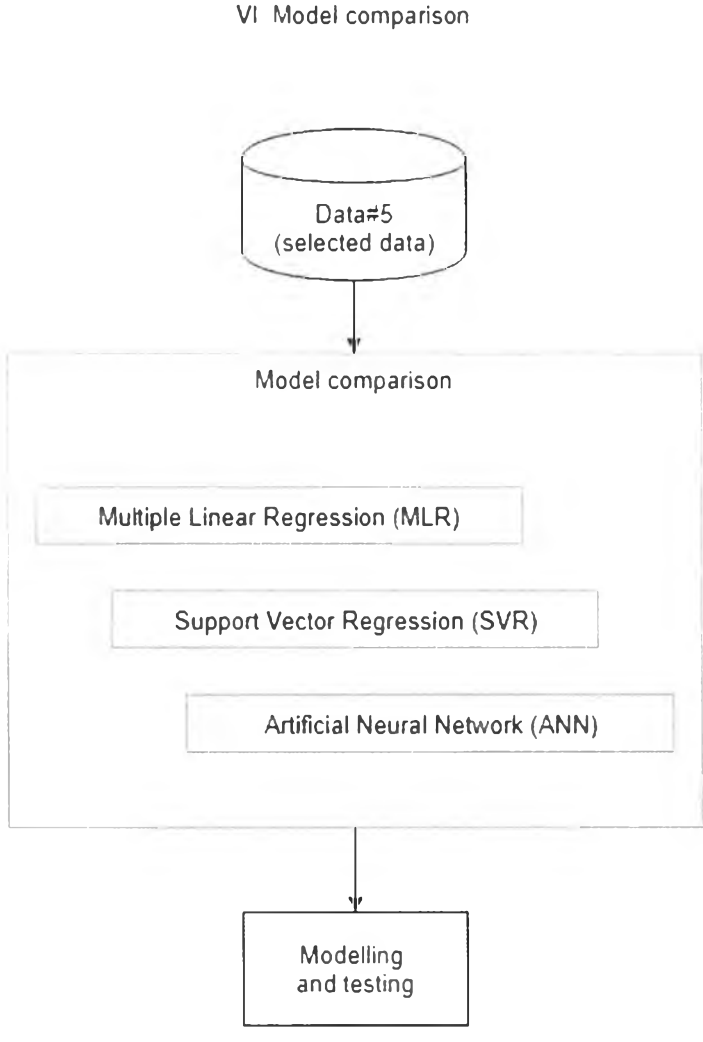

Figure 4.8 Model comparison experiment flowchart

4.3 Purposed method: Space and time neural network

Unlike 4.2 which aim to predict water quality parameter in the same time and same monitoring station, this part aim to develop the new model to forecast the water quality parameter in the future. Traditional artificial neural network accept only one dimensional data, which is not suitable for water quality parameter forecasting. Thus Space and Time Neural Network (STNN) is proposed for water quality parameter forecasting in river. The water quality parameter in multi-dimensional form, which

include parameter in different space and time is acceptable in the proposed model. The model can gain more information from upstream data and historical data simultaneously.

For example, suppose that the DO at 10th station (Phra Phuttha Yodfa Bridge station, Bangkok) in February 2562 B.E. is the goal of the forecasting. The relevant data for this prediction will consist of three major dimensions:

1) DO value at 12th station (Rama VI bridge station, Bangkok) and 15th station (Nuanchawee station, Nonthaburi), which is upstream station of target value. The number of upstream station that used to forecast a single parameter is defined as "space lag".

2) DO value at 10th station which measured in November and August, 2561 B.E. The number of timestamp delay that used to forecast a single parameter is defined as "time lag".

3) Other parameters such as BOD, pH etc. measured in dimensions 1 and 2.

STNN can calculate any parameters at any station and at any time by applies the same concept.

4.3.1 Model construction

Firstly, Water quality parameter are set in three dimension which are distance dimension, time dimension and parameter dimension to construct three dimension input layer as showed in Figure 4.9. Input layer can show in matrix form as Equation (4.2).

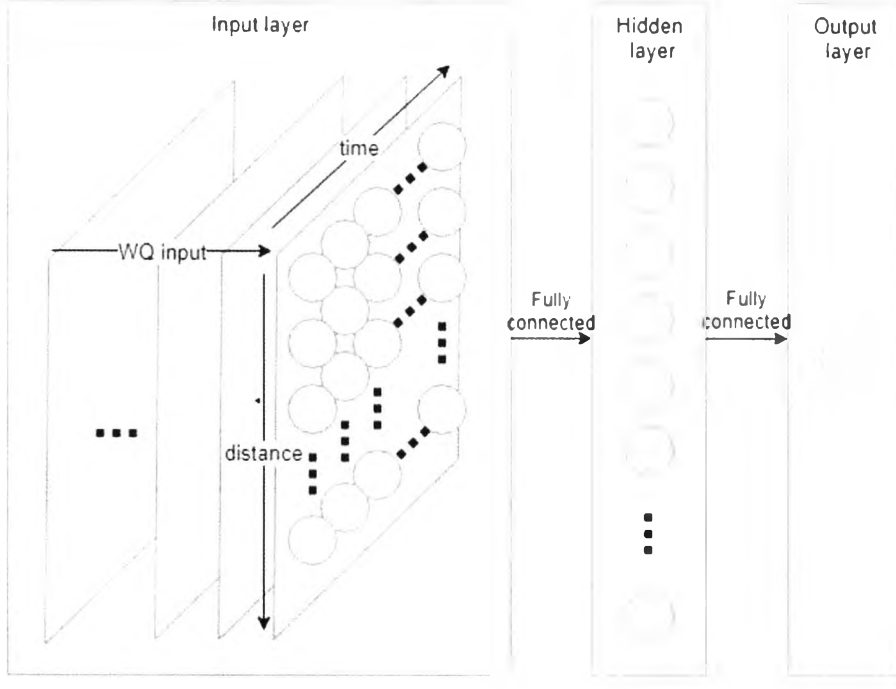

Figure 4.9 Multi-dimensional artificial neural network structure

$$WQ_{i,j,k} = f\left[\begin{bmatrix} WQ_{1,j,k-1} & WQ_{1,j,k-2} & \cdots & WQ_{1,j,k-t} \\ WQ_{1,j-1,k-1} & WQ_{1,j-1,k-2} & \cdots & WQ_{1,j-1,k-t} \\ \vdots & \vdots & \ddots & \vdots \\ WQ_{1,j-s,k-1} & WQ_{1,j-s,k-2} & \cdots & WQ_{1,j-s,k-t} \end{bmatrix} \begin{bmatrix} WQ_{2,j-s,k-1} & WQ_{2,j-s,k-2} & \cdots & WQ_{2,j-s,k-t} \end{bmatrix} \ddots \begin{bmatrix} WQ_{18,j-s,k-1} & WQ_{18,j-s,k-2} & \cdots & WQ_{18,j-s,k-t} \end{bmatrix}\right] \quad (4.2)$$

where $WQ_{i,j,k}$ is ith water quality parameter at jth monitoring station and kth timestamp as target.

Secondly, all node in input layer is fully connected to hidden layer. Each connected path is weight between a particular input node to a particular hidden node. Hidden node is calculated by summation of production of weight and input node in three dimension as show in Equation (4.3).

$$N_l = \sum_{i=1}^{18}\sum_{t=1}^{t_{max}}\sum_{s=0}^{s_{max}} \beta_{i,j-s,k-t,l} WQ_{i,j-s,k-t} \qquad ;l = 1,2,3,...,\lfloor (n+1)/2 \rfloor + 1 \tag{4.3}$$

where N_l is hidden nodes, $WQ_{i,j-s,k-t}$ is the i th water quality parameter at $j-s$ th monitoring station and $k-t$ th timestamp as input node, $\beta_{i,j-s,k-t,l}$ is the connection weight from the i th water quality parameter at $j-s$ th monitoring station and $k-t$ th timestamp to l th hidden node, t is number of timestamp delay, t_{max} is number of maximum of timestamp delay, s is number of upstream monitoring station, s_{max} is number of maximum of upstream monitoring station and n is number of input node in all dimension.

Finally, all hidden node is fully connected to output node, which is predicted parameter. Each connected path is weight between a particular hidden node to the output node. Output node which is the forecasted parameter is calculated by summation of production of weight and hidden node in as show in Equation (4.4).

$$WQ_{i,j,k} = \sum_{l=1}^{\lfloor (n+1)/2 \rfloor + 1} \alpha_l N_l \tag{4.4}$$

where n is the number of input nodes, N_l is hidden nodes, $WQ_{i,j,k}$ is i th water quality parameter at $j-s$ th monitoring station and $k-t$ th timestamp as target output, α_l is the connection weight from l th hidden node to the i th water quality parameter at j th timestamp and k th monitoring station and n is number of input node in all dimension.

4.3.2 Modelling

As same as 4.6, Input data are separated into two parts call training set and testing set. Training set which is data collected in 2539-2551 B.E. (approximately 70% of all data) is use to calibrate model. Data collected in 2552-2556 B.E. are used as testing set.

To start training process, the connection weight $\beta_{i,j-s,k-t,l}$ and α_l are randomly initialized. Then, the first record is fed to space and time neural network (STNN), calculated the forecasting value and compared to the observed parameter.

Comparison show the error of prediction, it will be used to update the connection weight $\beta_{i,j-s,k-t,l}$ and α_l. They are updated by Back propagation learning algorithm which mention in detail in 3.2.6. After that, the second record are fed as same as the first one. The process is continue until the last record. When the whole training set are fed, call one "epoch". Learning process are start again until reach 500th epoch. The training set of each epoch is the same, but the order of feeding record is shuffled. In theoretical aspect, the error is decrease until the optimal condition are met.

The trained model are used to calculate parameter on testing set and compared to the observed parameter. Two different criteria are used which are root mean square error (RMSE) and Spearman correlation coefficient.

4.3.3 Performance comparisons

The most important arguments of STNN are s_{max} and t_{max} which indicate space lag and time lag of model, respectively. Suppose $s_{max} = 0$ means that the model does not consider the dimension of the station. The model is time dependent model or time delay neural network (TDNN). On the other hand, $t_{max} = 1$ means that the model does not consider the dimension of the time. The model is space dependent model or distance neural network (DNN). These two arguments need to be carefully adjusted. If there is too high value, the input data is increasing a lot. The learning process takes a long time. But if they are too small, input data may not be sufficient to predict the parameters.

The arguments s_{max} and t_{max} are tested by different values and then evaluate performance. The argument s_{max} is set in range of 0 to 3 and t_{max} is set in range of 1 to 3. Other arguments are set as follows: $\varepsilon = 0.00001$, Learning rate = 0.3, Momentum = 0.2, Maximum epoch = 500, Number of hidden layer = 1, Number of hidden neuron = $\lfloor (n+1)/2 \rfloor + 1$, where n is number of input node.

All three type of models which are time delay neural network (TDNN), distance neural network (DNN) and space and time neural network (STNN) will be evaluated to compare the performance. Overview of this part is shown in Figure 4.10.

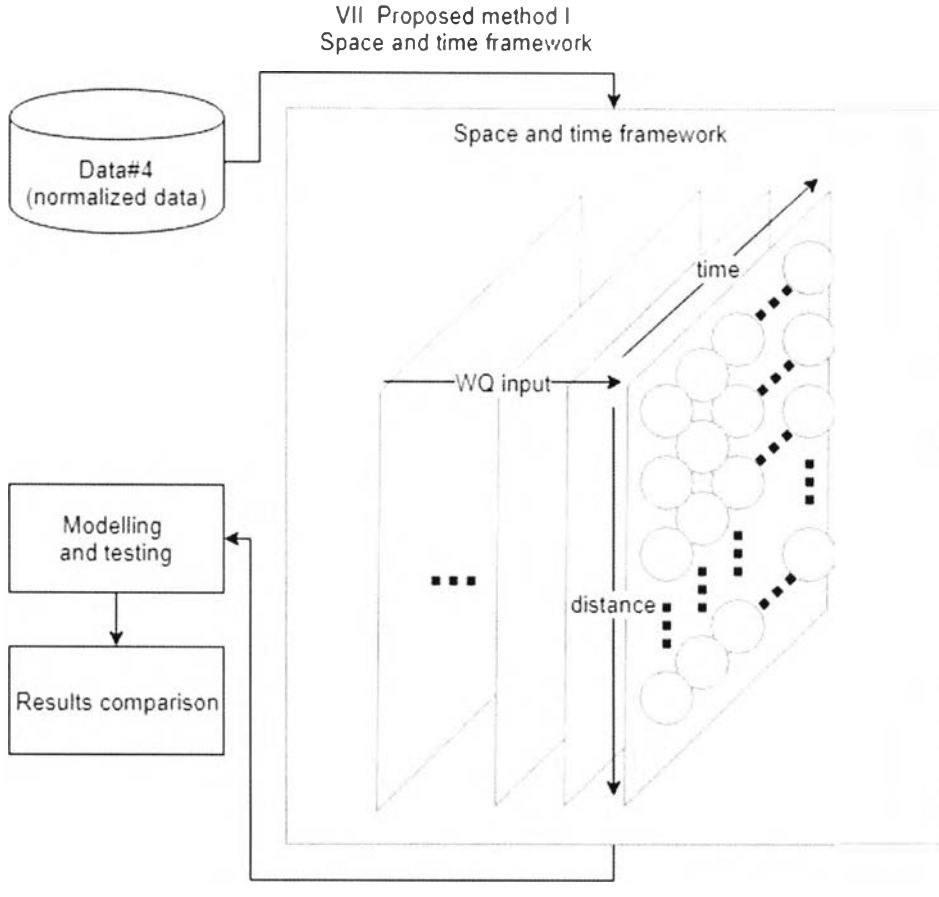

Figure 4.10 Space and time neural network experiment flowchart