CHAPTER 2

LITERATURE REVIEW

In this chapter, several literatures related to modelling of water quality prediction are revised. The stage-of-the-art methods for developing the water quality prediction models are examined in order to point out the advantages and the disadvantages of each model.

Water quality prediction models are the effective tools to forecast and simulate the water quality parameters in the natural water resources. Prediction of water quality parameters from the collected data (such as metrological data, historical water quality data and waste water discharge data) usually uses the hydrologic models. Hydrologic models predict the water quality parameter value based on physical relationship between water and the pollutant diffusion. Its prediction yields high accuracy but the limitation is big data requirement.

Recently, another type of model called the empirical model, is the popular issue in many developing countries because of scanty water quality data that are not enough to use hydrologic model [4]. The empirical water quality is suitable for this problem. For more understanding of two water quality model types, it is necessary to provide a classification of model for water quality prediction and literature survey of both. The following types of models can be distinguished:

2.1 Hydrologic model

A hydrologic model is based on the behaviour of fluid by a physical law of fluid dynamic and interaction between components in a system. This model forecasts water quality parameter by calculating the physical distribution of pollutant in water or stream. Moreover, the hydrologic model can accept the remote sensing data to perform GIS-based model for large scale water basin management [10].

Hydrologic models have been developed for more than half a decade since Streeter and Phelps (1958) developed the first water quality model (S-P model) to control river pollution in Ohio state, USA [11]. Surface water quality models have made a big progress from single parameter of water quality to multi parameters of water quality, from a steady-state model to a dynamic model, from a point source model to the a coupling model of point and nonpoint sources, and from a zero-dimensional mode to one-dimensional, two-dimensional, and three-dimensional models [12]. The model continues to be developed continuously until now.

Recently, water quality modelling research mostly concentrate on three-dimensional hydrodynamic model and influences of sediments [13, 14]. The examples of the most advance hydrologic models are QUAL model [15-17], EFDC model [18], MIKE11 models [19-21] and WASP models [22-24]. These commercial models are often used for specific tasks that require special water quality monitoring, such as chemical spills and water supply systems. However, it may not be suitable for water sources that are generally monitored for regulation, because the collected data are not enough to develop a model.

2.2 Empirical model

An empirical model is a mathematical model based on the relationship between the existing parameter inputs without considering the knowledge of the hydrological system. This model is also called a data driven model. It involves mathematical equations derived from concurrent water quality parameter input and output. Thus, these models are valid only within the boundaries of observed input [5]. The framework of the empirical model are divided into six main steps which are descriptive statistics analysis, imputation, transformation, normalization, parameter selection and prediction [25]. The first four steps are pre-processing steps and the other two steps are the important parts which are focused in this review.

Descriptive statistical analysis is a basic step to be performed after data processing, which can indicate the overview of data with statistical criteria. It can roughly show the quality of data before starting to develop a model. The other five steps in the framework are reviewed in detail as follows:

2.2.1 Imputation

Full complete dataset always the first requirement of modelling; however, missing data are inevitable for a long period of water quality monitoring. Thus, missing value treatment is a must; this process is known as imputation. The issue of missing data can be solved in several ways. The most common solution is to simply remove any record from the dataset that contains a missing value for any of the parameters. This is referred to as a listwise deletion [26]. But the listwise deletion may ignore useful information and may change the structure of time series water quality data. To avoid data elimination, missing value replacement is more reasonable [27].

Most of water quality modelling research papers mentioned that mean replacement was used to solve this problem [26, 28, 29]. However, mean replacement can lead to a bias in modelling because the replacement value was calculated using all records. In fact, the characteristics of each parameter are only fluctuated for a while; thus, the replacement value should be calculated by a few previous records for more precision of imputation [30].

More recently, to avoid bias caused by the mean replacement, statistical based models were implemented for a missing value substitution. For example, autoregressive integrated moving average (ARIMA) time series model was used to impute the missing water flow in distribution network [31]. Linear interpolation method was reported as the well suit to filling missing parameter values [32]. The limitation of these two methods are that the monitoring interval must be equally fixed, which is not appropriate to Chaophraya River data.

Recently, machine learning technique was applied for imputation as well [33]. Many novel methods were developed, such K-nearest neighbor based imputation [34-38], support vector machines based imputation [39-41] and neural network based imputation [42-46] but none of them was used with water quality data before. These new models also require testing for using with water quality data. Some methods will be tested to verify the appropriate model for the Chaophraya River in this research.

2.2.2 Data transformation

In most traditional statistical models, the data are normally distributed before the model coefficients can be estimated efficiently [25]. If this is not the case, suitable data transformations to normality must be performed. It has been suggested in the many researchers that some machine learning techniques can overcome this problem (i.e. artificial neural network and support vector regression), as the probability distribution of the input data does not have to be known [47].

On the other hand, it has been pointed out that the data need to be normally distributed in order to obtain optimal prediction results even when the machine learning techniques were used because the training process can be biased by non-normal distributed data [48]. In some water quality modelling researches, data were transformed based on expertise's knowledge without any verification. For example, total coliform bacteria were transformed by the logarithmic function before feeding to the model [49, 50].

In summary, the necessity of transformation has not been confirmed by empirical trials when the model fits were the same regardless of whether raw or transformed data were used [51]. Clearly, this issue requires further investigation. Thus, in this dissertation, the transformation was tested with the Chaophraya River data.

2.2.3 Normalization

In order to ensure that all parameters have equal weight during the modelling process, they should be normalized. In water quality modelling, range normalization was widely used for data preparation [52]. The data were scaled in the range of specific value. For example, range between 0 and 1 was recommended for water resource modelling [53]. In addition, some researcher applied different ranges to the model and reported better results (such as a range from -1 to 1 [54] or a range from 0.1 to 0.9 [55, 56]).

However, the disadvantage of range normalization is that it is outlier (extreme value) sensitive. Alternative normalization approaches were proposed to solve this problem. For example, an interquartile normalization [57-59], a Z normalization and proportion normalization [60, 61]. Difference method might effect to modelling performance depended on data characteristic. These became the optional testing in this dissertation.

2.2.4 Parameter selection

Selecting an appropriate subset of input parameters is an important step in the empirical model development process. Water quality parameters in natural aquatic is complex and have effects on each other. If an important parameter is excluded from the model, this could make it impossible to predict accurately. On the other hand, if the input was too large, this would increase model training time and lead to overfitting problems [8]. A number of published literatures selected parameter input by only expertise knowledge [25]. This can result in either too few or too many input, which is undesirable.

Nowadays, many parameter selection techniques are available for finding potential relationship between input parameters and output parameters. These methods can be classified into three types of algorithms which are 1) filter algorithm, 2) wrapper algorithm, and 3) embedded algorithm. Wrapper algorithm relies on the development of a number of prediction models with different input parameters to determine which candidates should be the optimal. The significant disadvantage of this algorithm is time complexity which depends on the number of prediction models that have to be developed. However, this algorithm can guarantee the predictive performance because the candidate parameter subsets are tested with the real model.

Many methods were used to select the combination of input parameters that optimizes the predictive performance. For examples, stepwise methods whose input are systematically added (known as forward selection, FS) [62-64] or removed (known as backward elimination, BE) [62-64] from a model and global optimization methods, such as genetic algorithm (GA) [65-67] which adapted the concepts from natural and evolutionary principles to select the appropriate input for the prediction.

In contrast to the wrapper algorithm, the filter algorithm does not rely on the predictive performance of the model, but develops a model with a set of input that are selected by some statistical analysis [25]. A statistical measure is generally used to determine the strength of the relationship between input parameters and output parameters. The most commonly used measures of statistical dependence for input selection are correlation [68-72] and sensitivity analysis [55, 73-75]. However, correlation and sensitivity analysis only measure a pair of parameter relationship, not all parameter interaction. Principle component analysis (PCA) can handle this problem by generating new variables from all parameters and used them as inputs [62, 76-79]. The new variables are independent; thus they can overcome the problem of input redundancy.

According to Table 2.1, a number of published literatures were reviewed. The proportion of research using wrapper algorithm is 40%, while another 60% used filter algorithm. Considering individual methods, the expert selection method is the most popular one as it appeared in 25 papers, followed by the principal component analysis and correlation, which were used in four papers each. The wrapper algorithms such as genetic algorithm, forward selection, and backward elimination were used by four, three, and three literatures, respectively.

Table 2.1 showed the research related to the prediction of water quality parameters. Input and output parameters in each article were listed. Table alsc outlines how to select the parameters and predictive methods that are used in the research. All research articles are sorted by year of publication.

Table 2.1 Detail of reviewed water quality modelling research

Year	Input	Output	Parameter selection method	Prediction model	Ref.
1998	1. BOD 2. treatment cost 3. BOD assimilate	BOD allowable loading	Expert selection	ANN	[83]
2001	1. TOC 2. surfactants 3. pH 4. ammonia 5. salinity 6. total phenolic 7. organic acids 8. DO1 9. DO2 10. DO3 11. air flow 12. bioreactor level 13. flow rate	1. TOC 2. surfactants 3. pH 4. ammonia 5. salinity 6. total phenolic	Expert selection	ANN	[84]
2001	1. Flow rate 2. DO 3. pH 4. SV 5. SS 6. NH_4^+	TKN	PCA	ANN	[77]

Table 2.1 Detail of reviewed water quality modelling research (cont'd)

Year	Input	Output	Parameter selection method	Prediction model	Ref.
	7. NO_3^{--} 8. NO_3^- 9. temperature 10. TSS 11. COD				
2002	rainfall	Wastewater inflow rate	-	ANN	[102]
2002	Hydrological parameters	1. DO 2. BOD 3. NO_3^-	Expert selection	1. QUAL2E 2. QUAL2K	[103]
2002	1. DO 2. BOD 3. COD 4. NH_3 5. SS	Water quality index	Expert selection	1. ANN 2. MNN 3. RBF	[85]
2003	1. BOD 2. SOD	DO	Genetic algorithm	QUAL2E	[65]
2003	Hydrological parameters	DO	Expert selection	1. SIMCAT 2. TOMCAT 3. QUAL2E 4. QUASAR 5. MIKE-11 6. ISIS	[104]
2005	1. NO_3^- 2. EC	1. NO_3^- 2. EC	Expert selection	ANN	[105]

Table 2.1 Detail of reviewed water quality modelling research (cont'd)

Year	Input	Output	Parameter selection method	Prediction model	Ref.
	3. DO 4. HCO_3^- 5. SO_4^{2-} 6. Cl^- 7. Na^+ 8. Mg^{2+} 9. Ca^{2+} 10. NH_4^+ 11. PO_4^{3-} 12. Q	3. DO 4. Na^+ 5. Mg^{2+} 6. Ca^{2+}			
2008	Time lagged Cl^-	Cl^-	Partial mutual information	ANN	[106]
2009	1. temperature 2. pH 3. DO 4. SS 5. TKN 6. NH_3-N 7. NO_3-N 8. PO_4^{3-} 9. total coliform	BOD	Expert selection	1. ANN 2. MLR	[107]
2009	1. TDS 2. EC 3. turbidity	1. TDS 2. EC 3. turbidity	Expert selection	ANN	[108]

Table 2.1 Detail of reviewed water quality modelling research (cont'd)

Year	Input	Output	Parameter selection method	Prediction model	Ref.
2009	1. pH 2. DO 3. BOD	Classify Water Quality	Expert selection	ANN	[109]
2009	1. COD 2. NH_3 3. Chlorophyll a 4. NO_2^- 5. NO_3^- 6. DO 7. temperature	BOD	Sensitivity analysis	ANN	[55]
2009	1. pH 2. TS 3. Total-Alk 4. Total-Hard 5. chloride 6. PO_4^{3-} 7. K^+ 8. Na^+ 9. NH_4^+ 10. NO_3^- 11. COD	1. DO 2. BOD	Expert selection	ANN	[54]
2009	water quality$^{(t)}$	Water Quality$^{(t+1)}$	Expert selection	1. SVR 2. ANN 3. ARIMA 4. GML	[89]
2010	1. solid waste$^{(t)}$	solid waste$^{(t+1)}$	1. PCA	ANN	[78]

Table 2.1 Detail of reviewed water quality modelling research (cont'd)

Year	Input	Output	Parameter selection method	Prediction model	Ref.
	2. solid waste$^{(t-1)}$ 3. ... 13. solid waste$^{(t-12)}$		2. Gamma test		
2010	1. pH 2. DO 3. BOD 4. NO_3^- 5. NH_3 6. total coliform	Classify Water Quality	Expert selection	1. ANN 2. CART	[110]
2010	1. temperature 2. DO 3. Boron	1. temperature 2. DO 3. Boron	Expert selection	1. ANN 2. ARIMA 3. Hybrid	[111]
2010	$DO^{(t)}$	$DO^{(t+1)}$	-	1. ANN 2. SVR	[112]
2011	1. pH 2. BOD 3. COD 4. SS 5. TKN 6. NH_3 7. NO_2^- 8. NO_3^- 9. PO_4^{3-} 10. total coliform	DO	Expert selection	ANN	[56]
2011	1. Rainfall 2. discharge	Stream flow	1. PCA	1. ANN 2. SVR	[62]

Table 2.1 Detail of reviewed water quality modelling research (cont'd)

Year	Input	Output	Parameter selection method	Prediction model	Ref.
	3. sun radiation 4. temperature		2. Forward selection 3. Gamma test		
2011	1. NH_3 2. NO_3^- 3. detrital nitrogen 4. dissolved organic nitrogen 5. PO_4^{3-} 6. detrital phosphorus 7. dissolved organic phosphorus 8. phytoplankton concentration 9. DO 10. BOD 11. detrital carbon 12. salinity	1. NH_3 2. NO_3^- 3. DO 4. total chlorophyll-a	Expert selection	WASP model	[113]
2011	1. DO 2. COD 3. BOD 4. SS 5. NH_3 6. pH	Water Quality Index	PCA	MLR	[79]

Table 2.1 Detail of reviewed water quality modelling research (cont'd)

Year	Input	Output	Parameter selection method	Prediction model	Ref.
2012	1. DO 2. COD 3. NH_3 4. NO_3^- 5. total coliform	BOD	Expert selection	1. ANN 2. ANFIS	[88]
2012	1. temperature$^{(t-1)}$ 2. pH$^{(t-1)}$ 3. EC$^{(t-1)}$ 4. salinity$^{(t-1)}$	1. temperature$^{(t)}$ 2. pH$^{(t)}$ 3. EC$^{(t)}$ 4. salinity$^{(t)}$	1. Forward selection 2. Backward elimination	1. ANN 2. SVR	[64]
2012	1. pH 2. DO 3. BOD 4. NH_3 5. NO_3^- 6. total coliform	Classify Water Quality	Expert selection	ANN	[114]
2013	1. EC 2. TDS 3. turbidity 4. TS	1. Conductivity 2. TDS 3. Turbidity	Expert selection	1. ANN 2. RBF	[86]
2013	1. temperature 2. pH 3. DO 4. H_2S 5. BOD 6. SS 7. TKN	COD	Expert selection	ANN	[115]

Table 2.1 Detail of reviewed water quality modelling research (cont'd)

Year	Input	Output	Parameter selection method	Prediction model	Ref.
	8. NH_3 9. NO_2^- 10. NO_3^- 11. PO_4^{3-} 12. total coliform				
2013	1. pH 2. DO 3. EC 4. water temperature 5. SR 6. air temperature 7. WS	1. $DO^{(t+1)}$ 2. $WT^{(t+1)}$	Genetic algorithm	1. RGA–SVR 2. SVR 3. ANN	[66]
2013	1. temperature 2. pH 3. H_2S 4. DO 5. BOD 6. COD 7. SS 8. TKN 9. NH_3 10. NO_2^- 11. NO_3^- 12. PO_4^{3-} 13. total coliform	$DO^{(t+1)}$	Expert selection	1. Hybrid (K-mean &ANN) 2. Hybrid (Fuzzy c-means &ANN)	[116]

Table 2.1 Detail of reviewed water quality modelling research (cont'd)

Year	Input	Output	Parameter selection method	Prediction model	Ref.
2014	1. temperature 2. DO 3. NH_3 4. PO_4^{3-} 5. phytoplankton concentration 6. zooplankton concentration	1. Temperature^{t+1} 2. DO^{t+1} 3. NH_3^{t+1} 4. $PO_4^{3-\ t+1}$ 5. Phytoplankton concentration^{t+1} 6. Zooplankton concentration^{t+1}	Expert selection	ISSADM	[117]
2014	1. temperature 2. pH 3. NO_3^- 4. NH_3	DO	Expert selection	1. ANFIS 2. ANN	[87]
2014	1. DO 2. volatile phenol 3. COD 4. NH_3	Water Quality Index	Expert selection	1. PSO-ANN 2. ANN	[73]
2014	1. pH 2. COD 3. BOD 4. TSS 5. TDS 6. EC 7. Cl^- 8. HCO_3^- 9. SO_4^{2-}	Alkalinity	Correlation	Linear Regression	[68]

Table 2.1 Detail of reviewed water quality modelling research (cont'd)

Year	Input	Output	Parameter selection method	Preciction model	Ref.
	10. Ca^{2+} 11. Mg^{2+} 12. Na^{+} 13. K^{+}				
2014	1. BOD 2. COD 3. TSS 4. NH_3 5. SS 6. total coliform 7. pH 8. DO 9. EC 10. temperature	1. BOD 2. COD 3. TSS 4. NH_3 5. SS 6. Total coliform	Expert selection	QUAL2E model	[118]
2014	53 variables included meteorological data and water parameters	1. DO 2. Temperature 3. PO_4^{3-} 4. Chlorophyll a	Genetic algorithm	1. AHGA 2. NSHGA	[67]
2014	1. bioaugmentation 2. treatment 3. reactor 4. day 5. TSS 6. R 7. Dy 8. Fo	1. COD 2. NH_3 3. NO_3^- 4. 3-CA	Sensitivity analysis	SVR	[73]

Table 2.1 Detail of reviewed water quality modelling research (cont'd)

Year	Input	Output	Parameter selection method	Prediction model	Ref.
2014	variables included physical/chemical data of water and water parameters	1. DO 2. BOD 3. NO_3^- 4. PO_4^{3-}	Expert selection	OpenMI	[119]
2014	1. total nitrogen 2. PO_4^{3-} 3. DO 4. rainfall 5. turbidity 6. stream flow 7. temperature 8. flow travel time	1. total nitrogen 2. PO_4^{3-}	Sensitivity analysis	1. SVR 2. ANN	[74]
2014	23 water quality parameters	Water quality index	PCA	ANN	[76]
2014	1. TSS 2. TS 3. pH 4. temperature	1. BOD 2. COD	Correlation	1. MLR 2. ANN	[69]
2015	1. pH 2. PO_4^{3-} 3. DO 4. COD 5. SS 6. stream flow 7. temperature	Water quality index	Sensitivity analysis	1. SVR 2. ANN 3. RBF	[75]

Table 2.1 Detail of reviewed water quality modelling research (cont'd)

Year	Input	Output	Parameter selection method	Prediction model	Ref.
	8. EC 9. NH_3 10. NO_2^- 11. NO_3^-				
2016	1. NH_3 2. TKN 3. temperature 4. Total Coliform 5. Fecal Coliform 6. pH	COD	Correlation	1. SVR 2. MARS 3. M5Tree	[70]
2016	1. Ca^{2+} 2. Mg^{2+} 3. Na^+ 4. SO_4^{2-} 5. Cl^- 6. EC	salinity	Correlation	1. ANFIS 2. ANN 3. Hybrid	[71]
2017	DO, SS, pH, NH_3, Temp, EC, Tur, TDS, TS, NO_3^-, Cl^-, PO_4^{3-}, As, Zn, Ca^{2+}, Fe, K^+, Mg^{2+}, Na^+, OG, E-Coli, Coliform, Cd, Cr, Pb	Water quality index	1. Forward selection 2. Backward elimination	1. ANN 2. Multiple ANN	[63]
2017	1. pH 2. Alkalinity 3. DO 4. NO_3^-	BOD	Correlation	ANFIS	[72]

Table 2.1 Detail of reviewed water quality modelling research (cont'd)

Year	Input	Output	Parameter selection method	Prediction model	Ref.
	5. TDS				
	6. TS				
	7. Hardness				
	8. K^+				
	9. turbidity				

As mentioned before, the parameter selection methods have their own advantages and disadvantages. The chosen method depends on the purpose of research and characteristics of data and a model. Thus, several methods which are both filter and wrapper algorithms will be tested to find the most suitable method for Chaophraya River quality modelling.

2.2.5 Prediction model

A prediction model is the core of framework that uses regression and correlation models to find the functional relationship between input and output. Linear regression and autoregressive integrated moving average (ARIMA) model are the basic statistical models for capturing the time series data characteristics. However, in case of water quality prediction, those models may be not enough [80, 81]. The important reason is that linear regression and ARIMA model cannot handle nonlinear relationships which are commonly found in water quality parameters [82].

Recently, machine learning techniques have been developed to capture nonlinear pattern in this area. Artificial neural network (ANN) [56, 83-85], radial basis functions (RBF) [85, 86], adaptive neuro-fuzzy inference systems (ANFIS) [87, 88] and support vector regressions (SVR) [64, 70, 73, 74] are examples of the machine learning techniques used in hydroinformatics.

The reviewed literatures (shown in Table 2.1) indicate that various water quality models have been developed and evaluated over the period 1998-2017. Multiple linear regression (MLR) and ARIMA models are not very popular, due to the limitations mentioned above, whereas artificial neural network (ANN), which has been used traditionally in applications related to hydrology and water resources [25], is the most popular model. Other alternative machine learning techniques, such as SVM, ANFIS, and RBF were applied in only a few literatures between 2 and 10 papers, compared with 34 papers where ANN was used.

Comparing the predictive performance of each model from the reviewed literatures, it cannot be definitely conclude which model is the most suitable model. For example, Khuan *et al.* (2002) reported that the model developed by ANN was the most suitable model for use to determine the water quality index, in terms of accuracy and fast learning time, after compare with the RBF model [85]. Unlike the results from Xiang and Jiang (2009), it was found that the SVM model outperformed the ANN and ARIMA models in terms of forecasting accuracy of water quality [89]. On the other hand, Najah *et al.* (2014) showed that the ANFIS model was capable of providing greater accuracy compared with ANN, particularly in the case of extreme events [87]. These results indicate that each individual water resource needs a specific prediction model development. Therefore, several techniques could be tested and optimized to determine an appropriate model for Chaophraya River quality modelling.

2.2.6 Space and time prediction model

The techniques available for water quality forecasting depend on the scale of interest. Both space and time scales must be considered and depended on the particular application [90]. In the case of water quality in the river, it is clear that the quality of the upstream will affect the downstream in the same river and the water quality which was monitoring at a particular time is also affected by the water in the earlier time as well [91].

Space and time prediction models were developed and successfully applied to various research areas. For examples, rainfall forecasting [90, 92], air quality forecasting [93-97], meteorology [98-100] and oceanology [101]. However, space and time model has never been developed for water quality forecasting application before.

From all of literature reviewed, experiments are set to find the suitable method for water quality prediction in each step which consist of imputation, transformation, normalization, parameter selection and modelling algorithm. After the state-of-the-art methods are tested and found the suitable framework for Chaophraya River, the framework are extended and improved the performance by developed to the new model which can handle space and time input parameters. Since this space and time model was show the effectively success in many research fields and it performed better than a single dimensional model (either space or time). The proposed method potentially perform better than the traditional one. Moreover, the most probable subset of input parameters for predicting water quality are determined by the parameter selection method. This relation between input parameters and output parameters is usefully for water quality management.

In the next chapter, the theoretical background of water quality parameter and the method in each step of framework are fully explained.