CHAPTER 3

THEORITICAL BACKGROUND

This chapter contains two parts of important background information in environmental science and computer science. Environmental science part explains the definition of water quality parameters and water quality standards of Chaophraya River according to Law and Regulations. Computer science part describes the fundamental concepts of techniques used in this dissertation.

3.1 Environmental science background

3.1.1 Water quality parameter description

The water quality can be determined by many parameters. Each of them has specific meaning. In this section, the definition of each parameter is described in detail.

1. Water temperature

The water temperature is a measure of the amount of solar energy that the water receives, including energy from soil and air in the surrounding area [120]. If water gets a lot of heat from the sun, its temperature will rise. In addition, the temperature of water from industrial plants can also be high. On the other hand, the evaporation of water on the surface can reduce the water temperature [121].

Temperature is one of the most important physical parameters of water because it has an impact on the aquatic ecosystems of living organisms in the water. It has a direct impact on aquatic plants, aquatic species and microorganisms growth rate [122]. Temperature also has negative effect on the amount of dissolved oxygen in the water [123]. Temperature also affects the rate of chemical reaction in the water and affects the odor and taste of the water. Naturally, the water temperature in the Chao Phraya River is about 29-30 degrees Celsius [124].

2. pH

Parameter pH also known as acidity and alkalinity. This value shows the potential activity of hydronium ion (H^+ or H_3O^+) in the water. The pH value is in the range of 0 to 14 in logarithmic scale. If the water sample has a pH lower than 7, it means the sample is in the acid condition. If the pH of water sample is greater than 7, it means the water is the basis condition. If the pH of water sample is equal to 7, it means the water is neutral.

The pH of water determines ability to dissolve and bioavailability of chemicals such as nutrients (phosphorus, nitrogen and carbon) [125] and heavy metals (lead, copper, cadmium, etc.) [126]. In the case of heavy metals, they tend to be more toxic at lower pH because they are more soluble [127]. Extremes in pH can make the river unhealthy. Low pH is harmful to fish and insects, while acidic water also accelerates the leaching of heavy metals that are harmful to organisms [128, 129].

1071371125

3. Turbidity

Water turbidity refers to the ability of water to absorb the amount of light that passes through it. The turbidity of water is caused by suspended solids that suspend in the water. The suspension can be either organic, inorganic or microorganisms [130].

Water turbidity may or may not affect the health or hygiene of the aquatic ecosystem [131]; however, it affects the water treatment system, such as the impact on the filter system, causing the filter to clog and lose speed. It affects the chlorine disinfection system because chlorine is encapsulated by the suspension [132]. As a consequence, chlorine cannot kill microbes. In addition, the turbidity in water also affects aquatic plant photosynthesis. Since turbidity blocks sunlight to go through the water, the amount of oxygen in the water will be decreased and the visibility of aquatic animals will also be affected [133].

4. Electrical conductivity

Electrical conductivity is a measure of the ability of water to pass through electricity. Electrical conductivity is normally caused by the dissolved inorganic compounds, such as anions of chloride, nitrate, sulfate and phosphate (anion is a negatively charged ion) or cations of sodium, magnesium, iron and aluminum (cation is a positive ion) [134]. On the other hand, it also indicates concentration of water soluble ions. The meaning is when the conductivity increases, it shows that the substance is more dissociable in water and vice versa [135].

Electrical conductivity also depends on dissolved organic compounds. Organic compounds, such as phenolic, alcohols and sugars are not very conductive by themselves and tend to decease conductivity of water when they dissolve in water [136]. Electrical conductivity also depends on temperature. If the temperature is higher, the conductivity tends to increase. For this reason, the conductivity is always reported at 25 °C.

10713711125

5. Salinity

Salinity is the concentration of water soluble salts, especially sodium chloride. Normally, salinity values are usually measured in grams per liter (g/L). The salinity of water is considered important especially on aquatic animals. The salinity of water affects the amount of water in the body as a result of the differences between osmotic pressure in the body of an aquatic animal and the outside [137]. In addition, the salinity of water also affects plants. It impacts on the transpiration system [138]. As same as in animal, the osmotic pressure is used to passively transport water into a tree. When salinity is too high, the water cannot osmosis to the root [139].

Salinity can be used to divide between fresh water, brackish water, and saline water. Salinity between 0 and 0.5 g/L is defined as freshwater, salinity between 0.5 and 30 g/L is defined as Brackish water and salinity of more than 30 g/L is saline (Sea water).

6. Dissolved oxygen

Dissolved oxygen (DO) is one of the most important parameters used to determine water quality in a surface water resource [123]. Good quality water standards generally has a DO value of 5-8 mg/L or an oxygen content of 5-8 mg/L, while wastewater has DO less than 3 mg/L. In general, most aquatic organisms can survive in a DO level of more than 3 mg/L. However, some reservoirs may have DO more than 10 mg/L at daytime, indicating that there may be abnormal growth of algae in the water. Too much dissolved oxygen may be harmful to aquatic animals by causing gas bubble disease, which will cause bubbles in the blood circulation [140, 141]. During nighttime, dissolved oxygen levels are low due to the respiration of the algae, resulting in a sudden lack of oxygen. This may cause the fish to die [141]. In addition, dissolved oxygen also depends on the temperature content. In a low temperature, oxygen dissolves better than in a high temperature [142].

7. Biochemical oxygen demand

The biochemical oxygen demand (BOD) is the amount of oxygen required by microorganism to degrade organic matter in water [143]. The BOD can indicate how much water is polluted by organic substances. In good quality water, BOD should not exceed 6 mg/l. Water resources with BOD more than 100 mg/l are classified as sewage or waste water. The industrial effluent standard defines that the waste water discharges into the natural water sources must have a BOD value not more than 20 mg/l [144].

High BOD indicates that microorganisms use a large amount of oxygen to decompose organic matter or sewage which cause the dissolved oxygen (DO) in water to be reduced and may become rotten [145]. Biochemical oxygen demand includes a carbonaceous oxygen demand and oxygen used in the inorganic oxidation, such as sulfide and ferrous ion [146]. This also includes the amount of oxygen used to oxidize nitrogen (nitrogenous demand) [147, 148].

8. Total coliform bacteria

Coliform bacteria is a group of bacteria mostly live in the intestines of mammals. But sometimes it can be found in other areas, such as plants, soil, seeds, and etc. [149]. The examination of this type of bacteria in a water resource represents the risk of contamination or spread of pathogens in the gastrointestinal tract in water such as cholera [143]. The total coliform count is measured in the Most Probable Number per 100 ml (MPN per 100 ml).

According to surface water quality standards, water sources are used in the production of tap water, as well as for swimming and water sports, the total amount of coliform bacteria should not exceed 5,000 MPN per 100 ml [150]. For water resources used in agricultural activities, the total coliform bacteria should not exceed 20,000 MPN per 100 ml.

9. Fecal coliform bacteria

Fecal coliform bacteria are found in human feces and warm-blooded animals [149]. Detection of this bacteria in water indicates that the water is likely to be contaminated or spread by pathogens that cause gastrointestinal disease [143]. Most fecal coliform bacteria are detected in household waste water that is crained directly into water resources [151]. The fecal coliform bacteria count has the same measurement as the total coliform count.

According to the surface water quality standards, water resources that are used in the production of tap water, as well as for swimming and water sports must have the total amount of fecal coliform bacteria no more than 1,000 MPN per 100 ml. For water resources used in agricultural activities, the fecal coliform bacteria should not exceeds 4,000 MPN per 100 ml.

10. Phosphate

Phosphorus is a nutrient usually found in all organisms and usually found in a form of phosphate (PO_4^{3-}). Phosphate compound is an essential nutrient for human, animals, and plants. In addition, phosphate is an essential ingredient in detergents, toothpaste, condensed milk, food, and beverages [152]. [152]. Phosphorus is mostly found in community waste water. This is due to the use of household detergents. Phosphate concentrations found in Thai community waste water are in the range of 2 -10 mg/l. Phosphorus is also found in both natural and waste water in a form of orthophosphate. Phosphate is often found as a solution in water and is called a soluble reactive phosphorus [153]. Organic phosphate in water may be in a form of a complex solution or in a suspended sediment.

10713711125

Phosphate in soluble water can be used by vegetable for growing and proliferation, especially the phytoplankton [154, 155]. It can grow very fast which will produce fertility to water resources. However, if there is too much phosphate, it will cause the deterioration of water in the river. Water-soluble phosphorus also come from fertilizer for agriculture [156].

11. Nitrate

Nitrate (NO_3^-) is an important nutrient for the growth of algae and aquatic plants [157]. Nitrate concentration in the water is depending on the amount of nitrate from the source to the water source. Nitrate concentration in water depends on the amount of nitrate from the source to water resource. Normally, nitrogen levels found in natural water resources are relatively low (less than 1 mg/L of nitrogen in the form of nitrates), resulting from the degradation process of animal waste and dead carcasses that the plants use them very quickly. In the water with high nitrogen levels, this may cause the eutrophication process [158]. Nitrogen levels may be higher due to natural or human activities. Nitrogen produced by human activities includes dumping waste into the river or washing chemical fertilizer into the river [159]. It may also contaminate groundwater [160] as well as the leachate from some animals and animal shelters.

Nitrate may be found in some areas of groundwater. When Nitrate is too high, it can cause disease called "Methemoglobinemia" [161, 162]. Waste water or water from any biological treatment system can be up to 30 mg of Nitrate. In some cases, it is used as growth limiting nutrient for plants due to its low level in natural water. On average, it is about 0.3 milligrams per liter, but not more than 10 milligrams per liter, and often less than 1 milligram per liter.

12. Nitrite

Nitrite (NO_2^-) found in natural water is mainly derived from the degradation of organic matter. Nitrite is the intermediate state in the nitrogen cycle [163] which is the process of oxidation of ammonia to nitrate and the reduction of nitrate to ammonia. Small amounts of nitrite are found in water, it is a result of biodegradation of proteins [164]. This is an indicator of the impurities by organic matter. At the surface water, Nitrite is usually found in very low concentration of 0.1 mg/l. It can be a corrosion inhibitor in the industrial process. In addition, nitrite is also the actual etiologic agent of the Methemoglobinemia disease [161, 162].

13. Ammonia

Ammonia is a nitrogen gas that is in the form of an ionized form (NH_4^+) or a un-ionized form (NH_3) [165]. Ammonia naturally found in surface water and groundwater. It is produced by the ammonia extraction process out of organic compounds call deamination. It can be produced by the hydrolysis of urea and it is also produced naturally by reducing nitrate under anaerobic conditions [166].

Nitrogen content in a form of ammonia is important to identify the impurities of water resources generated by wastes or wastes containing nitrogen, such as inorganic proteins, organophosphorus compounds, fertilizers, manure, etc., especially from farm effluents [167]. If the water source is found high ammonia-nitrogen content, it indicates that the water source is contaminated by high pollution and may be toxic to aquatic organisms. According to the surface water quality standards, ammonia-nitrogen should not exceed 0.5 mg/L.

The measured ammonia called the total ammonia consists of two types of ammonia: ionized form ammonia (NH_4^+) and un-ionized form (NH_3). NH_4^+ is non-toxic to aquatic animals but NH_3 is toxic to aquatic animals. The proportion of both types of ammonia depends on pH of water and its temperature. High pH tends to increase NH_3 form which will make the ammonia to be more toxic [168].

14. Suspended solids

Suspended solids (SS) are insoluble particles that suspend in water which can be clearly separated. SS causes color and turbidity. Some of these substances are commonly found in waste water from various resources such as industrial factory and community waste water [169].

The amount of suspended solids is measured in mg/l. Suspended solids in the range of 25-80 mg/L can be used as a good source of fishery but if they are in the range of 80-400 mg/L, the yield is reduced [170].

1071371125

15. Total Dissolved Solids

Total Dissolved Solids (TDS) refers to the amount of solids dissolved in water, including metal ions, salts and dissolved metals [171, 172]. Normally, TDS in water expressed in milligram per unit volume of water (mg/l). TDS is directly related to the purity of water and the quality of a water purification system. It can be used to estimate the amount of sediment that will be discharged by the settling tank and indicate the settling tank efficiency. It also indicates the amount of salt in waste water, such as chlorine.

16. Total Solids

Total Solids refer to all solids in a water sample. All solids are suspended solids and dissolved solids [173]. The amount of solids or all the substances in water can be measured by the remaining of the evaporated water sample at 103-105 °C. It is very useful in determining the suitability of water to be consumed.

3.1.2 Surface water quality standards in Thailand

Section 32 of the Enhancement and Conservation of the National Environmental Quality Act 2535 B.E. enforced the national environment board to set the standards for environmental quality as the goal of maintaining environmental quality to the appropriate criteria. The environmental quality standards require that the basic scientific principles must take into account the possible social economic and technology. Water quality standards are the kind of environmental quality standards that aim to control and maintain the water quality that is suitable for using and securing the health of the public and conserves natural environment [174].

The Water Quality Management Bureau of the Pollution Control Department proposed two water quality standards (surface water quality standards and coastal water quality standards) to National Environment Board that the Prime Minister as Chairman of the board signed on January 20, 2537 B.E.

Principles for determining water quality standards, including the standard configuration to maintain water quality suitable for use by organized manner, make use of water resources, and establish rules and procedures for monitoring water quality. Criteria to determine the water quality standards are as follows:

1. Appropriate to apply the benefit of the activities or the types of events that water has been exploited. In the case that water has be used for many aspects (multiple purposes), focus only on the main benefit is significant. The water quality standards are not in conflict with the use of multiple purposes.

2. Water quality, water resource situation in the country, and trends of water quality may be changed due to various developments in the future.

3. Consider the health and safety of humans and aquatic life.

4. Satisfaction in the quality of water in various areas of habitats in local basin and the public.

However, improving the standards in the future must consider the appropriateness of the level of investment and economic conditions in the basin. The plan development is in progress as well as the possibility of applying new technology in the treatment of waste and toxins from the origin of the waste, including activities arising from the economic and social development planning.

The purposes of determining the quality of water resources are to guide the treatment of water quality in water bodies to remain well-suited to the use of various benefits and restoration of degraded water quality in water resources to be better.

There are three goals of setting surface water quality standards. The first goal is to provide a water resource with a reasonable and consistent standard assortment with the use of water resources. Second, in order to have water quality standards and audit procedures that are primarily for placing projects that take into account the water resource is important. The final goal is to maintain the quality of natural upstream water resources without any contamination from any activities [174].

Surface water resources in Thailand are classified into five classes by the objectives and usage. Class 1 is the highest quality and class 5 is the lowest quality, as shown in Table 3.1. The Chaophraya River is divided into three classes as shown in Table 3.2.

1071371125

Table 3.1 Classification and objectives of classification of water quality standard

Classification	Objectives/Condition and Beneficial Usage
Class 1	Extra clean fresh surface water resources are used for: (1) conservation not necessary pass through water treatment process require only ordinary process for pathogenic destruction (2) ecosystem conservation where basic organisms can breed naturally
Class 2	Very clean fresh surface water resources are used for : (1) consumption which requires ordinary water treatment process before use (2) aquatic organism of conservation (3) fisheries (4) recreation
Class 3	Medium clean fresh surface water resources are used for : (1) consumption, but passing through an ordinary treatment process before using (2) agriculture
Class 4	Fairly clean fresh surface water resources are used for : (1) consumption, but requires special water treatment process before using (2) industry
Class 5	The sources which are not classification in class 1-4 and used for navigation.

1071371125

Table 3.2 Classification of Chaophraya River

Part	Control Areas (km. from River Mouth)	Standards of Water Classification
1	From Pra Samutchedi Samutprakarn Province to the Old Nontaburi City Hall (Km. 7 to 62)	4
2	From the Old Nontaburi City Hall to Pompetch in Ayutthaya (Km. 62 to 142)	3
3	From Pompetch in Ayutthaya to the begin of Chaophraya River in Nakhornsawan Province (Km.142 to 379)	2

Surface water quality standards in Thailand define 28 parameters to be monitored and controlled [174], as shown in Table 3.3. Each parameter has specific measuring method. The defined standard methods for examination and statistic of each parameter are shown in Table 3.4.

1071371125

Table 3.3 Surface water quality standards

Parameter[1]	Units	Standard Value for Class[2]				
		1	2	3	4	5
1. Colour, Odour and Taste	-	n	n'	n'	n'	-
2. Temperature	°C	n	n'	n'	n'	-
3. pH	-	n	5-9	5-9	5-9	-
4. Dissolved Oxygen (DO)[2]	mg/L	n	6	4	2	-
5. BOD (5 days, 20°C)	mg/L	n	1.5	2	4	-
6. Total Coliform Bacteria	MPN/100 ml	n	5,000	20,000	-	-
7. Fecal Coliform Bacteria	MPN/100 ml	n	1,000	4,000	-	-
8. NO_3^-	mg/L	n	5			-
9. NH_3	mg/L	n	0.5			-
10. Phenols	mg/L	n	0.005			-
11. Copper (Cu)	mg/L	n	0.1			-
12. Nickle (Ni)	mg/L	n	0.1			-

Table 3.3 Surface water quality standards (cont'd)

Parameter[1]	Units	Standard Value for Class[2]				
		1	2	3	4	5
13. Manganese (Mn)	mg/L	n	1			-
14. Zinc (Zn)	mg/L	n	1			-
15. Cadmium (Cd)	mg/L	n	0.005* 0.05**			-
16. Chromium Hexavalent	mg/L	n	0.05			-
17. Lead (Pb)	mg/L	n	0.05			-
18. Total Mercury (Total Hg)	mg/L	n	0.002			-
19. Arsenic (As)	mg/L	n	0.01			-
20. Cyanide (Cyanide)	mg/L	n	0.005			-
21. Radioactivity						
- Alpha	Becquerel/l	n	0.1			-
- Beta			1			
22. Total Organochlorine Pesticides	mg/L	n	0.05			-
23. DDT	µg/L	n	1			-
24. Alpha-BHC	µg/L	n	0.02			-
25. Dieldrin	µg/L	n	0.1			-
26. Aldrin	µg/L	n	0.1			-
27. Heptachlor & Heptachlorepoxide	µg/L	n	0.2			-
28. Endrin	µg/L	n	None			-

Note:

[1] Determined only in the standard class 2-4. The first class parameters configure to be natural, and class 5 are not configured.

[2] DO is the minimum standard.

n naturally

1071371125

n' naturally but changing not more than 3°C

* when water hardness not more than 100 mg/L as $CaCO_3$

** when water hardness more than 100 mg/L as $CaCO_3$

Table 3.4 Methods for Examination and statistic of each parameter

Parameter	Statistics	Methods for Examination
1. Colour, Odour and Taste	-	-
2. Temperature	-	Thermometer
3. pH	-	Electrometric pH Meter
4. Dissolved Oxygen (DO)[2]	P20	Azide Modification
5. BOD (5 days, 20°C)	P80	Azide Modification at 20°C , 5 days
6. Total Coliform Bacteria	P80	Multiple Tube Fermentation Technique
7. Fecal Coliform Bacteria	P80	Multiple Tube Fermentation Technique
8. NO_3^-	-	Cadmium Reduction
9. NH_3	-	Distillation Nesslerization
10.Phenols	-	Distillation,4-Amino antipyrene
11.Copper (Cu)	-	Atomic Absorption-Direct Aspiration
12.Nickle (Ni)	-	Atomic Absorption-Direct Aspiration
13.Manganese (Mn)	-	Atomic Absorption-Direct Aspiration
14.Zinc (Zn)	-	Atomic Absorption-Direct Aspiration
15.Cadmium (Cd)	-	Atomic Absorption-Direct Aspiration
16.Chromium Hexavalent	-	Atomic Absorption-Direct Aspiration
17.Lead (Pb)	-	Atomic Absorption-Direct Aspiration
18.Total Mercury (Total Hg)	-	Atomic Absorption-Cold Vapour Technique
19.Arsenic (As)	-	Atomic Absorption-Direct Aspiration
20.Cyanide (Cyanide)	-	Pyridine-Barbituric Acid
21.Radioactivity		
- Alpha	-	Gas-Chromatography
- Beta		
22.Total Organochlorine Pesticides	-	Gas-Chromatography

1071371125

Table 3.4 Methods for Examination and statistic of each parameter (cont'd)

Parameter	Statistics	Methods for Examination
23.DDT	-	Gas-Chromatography
24.Alpha-BHC	-	Gas-Chromatography
25.Dieldrin	-	Gas-Chromatography
26.Aldrin	-	Gas-Chromatography
27.Heptachlor & Heptachlorepoxide	-	Gas-Chromatography
28.Endrin	-	Gas-Chromatography

10713711125

3.2 Computer science background

As mentioned in Chapter 2, the water quality modeling framework normally consists of six main steps which are 1) descriptive statistical analysis, 2) imputation, 3) transformation, 4) normalization, 5) parameter selection, and 6) prediction. There are many different methods that can be used in each step. Some of the existing methods that are used in the proposed framework are fully explained here.

3.2.1 Descriptive statistics

Descriptive statistics are the basic information analysis that describes general characteristics of water quality parameters. The main statistics that are useful in this dissertation are mean, standard deviation, missing value percentage, and correlation of parameters.

The missing values are commonly found in water quality monitoring that are caused by the environmental phenomena, human errors, or monitoring technical errors. A missing value is a common problem in environmental modelling system, thus, it needs to be determined before starting the modelling process.

Pearson correlation coefficient (R) is a parametric measure of the linear correlation between two variables [175]. It has a value between +1 and −1, where 1

means total positive linear correlation, 0 means no linear correlation, and −1 means total negative linear correlation.

Suppose there are two sets of parameter $\{x_1, x_2 \ldots, x_n\}$ and $\{y_1, y_2 \ldots, y_n\}$. Each collection contains n values; then, the correlation is calculated by Equation (3.1).

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.1}$$

where n is number of sample, x_i and y_i are two parameters measure in the same time. $\bar{x}$ is average value of parameter x; and analogously for $\bar{y}$.

10713711125

Unlike Pearson correlation coefficient, Spearman's rank correlation coefficient or called as Spearman' rho (ρ) is a non-parametric measure of rank correlation between two variables. The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not). Spearman correlation has a value between +1 and -1, where 1 is total positive correlation, 0 is no correlation, and −1 is total negative correlation.

The Spearman correlation coefficient is defined as the Pearson correlation coefficient between the ranked variables. For a sample of size n, the n raw data x_i and y_i are converted to ranks rg_{x_i} and rg_{y_i}. The Spearman correlation coefficient is computed from:

$$\rho_{rg_x, rg_y} = \frac{\operatorname{cov}(rg_x, rg_y)}{\sigma_{rg_x}\sigma_{rg_y}} \tag{3.2}$$

where ρ_{rg_x, rg_y} denotes the usual Pearson correlation coefficient, but applied to the rank variables, $\operatorname{cov}(rg_x, rg_y)$ is the covariance of the rank variables and σ_{rg_x} and σ_{rg_y} are the standard deviations of the rank variables.

3.2.2 Imputation techniques

Water quality modelling always requires a complete dataset which is quite impossible in the real situation. Most of the collected datasets contain missing values; however, these missing values can be replaced by synthesized the values to reconstruct a complete dataset, this process is called imputation. The replaced values can be computed in many ways which were developed for specific types of data. Water quality parameter imputation mainly aims to conserve the characteristics of data with minimum bias. Three imputation methods used in this dissertation are mean replacement, K-nearest neighbor (K-nn), and artificial neural network (ANN).

The mean replacement is simple and straightforward method that replaces any missing value with the mean value of that parameter [176, 177]. This method has the benefit of not changing the original mean of that parameter. However, mean imputation attenuates any correlations involving the parameters that are imputed. This is because, this method guarantees that there will be no relationship between the imputed parameters and any other measured parameters. Thus, mean replacement has some attractive properties for the univariate analysis but becomes problematic for the multivariate analysis.

1071371125

K-nearest neighbor (K-nn) is a clustering algorithm that is useful for matching a record with its closest k neighbors in a multi-dimensional space [178]. In K-nn perspective, the water quality data can act like a set of points in multiple dimensions. Each point represents each value. The idea of K-nn imputation is to group the k-closest points and define them as sub-groups and the mean value of each sub-group is computed. Instead of replacing a missing value by the mean value of the overall data, K-nn replaces the missing value by the mean value of that sub- group. The algorithm of K-nn imputation is shown in Figure 3.1 [179, 180]. However, the argument k needs to be initialized before the algorithm starts. The optimal k depends on each dataset.

Procedure: Complete-Case kNN Imputation

input: Dataset $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{n}$, where $\mathbf{x}_i = (x_{im_1}, \ldots, x_{im_{\alpha(i)}}, x_{io_1}, \ldots, x_{io_{\beta(i)}})$ has $\alpha(i)$ missing attribute values $x_{im_1}, \ldots, x_{im_{\alpha(i)}}$ and $\beta(i)$ observed values $x_{io_1}, \ldots, x_{io_{\beta(i)}}$. Further, $\mathcal{D} = \mathcal{C} \cup \mathcal{M}$, where $\mathcal{C}$ is the set of complete examples ($\mathcal{C} = \{\mathbf{x}_j \in \mathcal{D} \mid \alpha(j) = 0\}$) and $\mathcal{M} = \mathcal{D} \setminus \mathcal{C}$; number of nearest neighbors k. The w attributes are denoted $x_1, \ldots, x_w$, and we assume $w = \alpha(i) + \beta(i)\ \forall\ \mathbf{x}_i \in \mathcal{D}$.

output: Dataset $\widehat{\mathcal{D}} = \mathcal{C} \cup \widehat{\mathcal{M}}$ where $\widehat{\mathcal{M}}$ is the set of imputed examples.

1. do $\forall\ \mathbf{x}_i \in \mathcal{M}$:
2. $\quad \forall\ \mathbf{x}_j \in \mathcal{C}$, find $d_j = d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{l=1}^{\beta(i)} (x_{io_l} - x_{jo_l})^2}$ where $o_l = o_l$.
3. $\quad$ Let $\mathcal{K}_i = \{\mathbf{x}_j \in \mathcal{C} \mid d_j \leq d_q\ \forall\ \mathbf{x}_q \notin \mathcal{K}_i\}$; $k = |\mathcal{K}_i|$.
4. $\quad$ For $l = 1, \ldots, \alpha(i)$, $\hat{x}_{im_l} = k^{-1} \sum_{\mathbf{x}_p \in \mathcal{K}_i} x_{po_l}$; $o_l = m_l$.
5. $\quad \hat{\mathbf{x}}_i = (\hat{x}_{im_1}, \ldots, \hat{x}_{im_{\alpha(i)}}, x_{io_1}, \ldots, x_{io_{\beta(i)}})$.
6. end do
7. $\widehat{\mathcal{M}} = \{\hat{\mathbf{x}}_i \mid \mathbf{x}_i \in \mathcal{M}\}$, $\widehat{\mathcal{D}} = \mathcal{C} \cup \widehat{\mathcal{M}}$.
8. Return $\widehat{\mathcal{D}}$.

Figure 3.1 K-nn imputation algorithm pseudocode

1071371125

Artificial neural network (ANN) is a data prediction technique that can be applied to imputation task [43, 181, 182]. In ANN assumption, missing data are related to other parameters that are collected in the same time period. The ANN structure basically consists of three layers, which are input layer, hidden layer, and output layer [183, 184]. In each layer, there are several nodes which are fully connected to nodes in the adjacent layer as shown in Figure 3.2 [185]. The input layer consists of other parameters that are collected in the same time period and the output layer consists of the missing value.

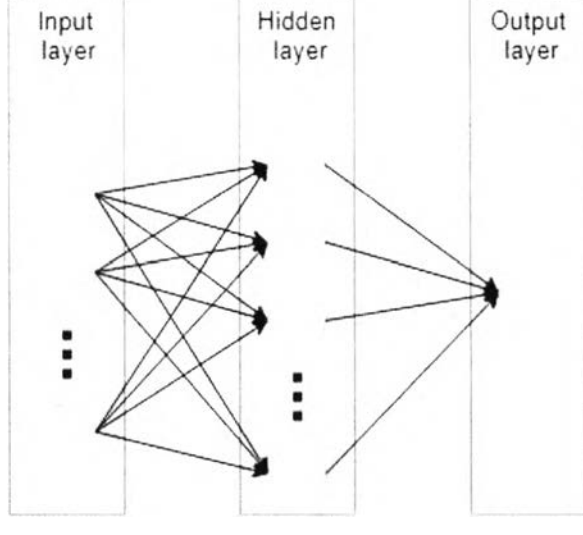

Figure 3.2 Artificial neural network structure

Hidden nodes in the hidden layer are calculated by the summation of the product of weight and input values as shown in Equation (3.3).

$$h_j = \sum_{i=1}^{n} \beta_{i,j} x_i \tag{3.3}$$

where h_j is jth hidden nodes, x_i is the ith input node which is the parameter value that is related to the missing value, $\beta_{i,j}$ is the connection weight from x_i to h_j and n is number of input nodes.

The output node is calculated by the summation of the product of the weight and the hidden nodes as shown in Equation (3.4).

$$y = \sum_{j=1}^{m} \alpha_j h_j \tag{3.4}$$

1071371125

where h_j is jth hidden nodes, y is the output node which is the missing value, α_j is the connection weight from h_j to y and m is number of hidden nodes.

The ANN imputation model starts from randomly initializing all connection weights in the model then training the network by Back Propagation algorithm (BP) [186, 187]. The Back Propagation algorithm works from the output backward; i.e., the output is first calculated by Equation (3.3) and (3.4), then the output are used to compare to the observed data and find the errors. Finally, each connection weight is subtracted by the percentage of the gradient which is calculated from the prediction error and the weight. The process is repeated until some conditions, which are normally defined by the maximum iteration and the acceptable prediction error, are met [187].

3.2.3 Data transformation method

Data transformation is an optional step for making the data more like normal distribution, which may be useful for prediction models. The simplest transformation is the power transformation which was proposed by Box and Cox in 1964 [188]. The Equation of Box-Cox transformation is shown in Equation (3.5).

$$x' = \begin{cases} x^{\lambda} & ;\lambda \neq 0 \\ \log x & ;\lambda = 0 \end{cases} \tag{3.5}$$

where x' is a transformed data, x is the original data and λ is the power argument. The Equation is simple but the way to estimate λ is complex. However, Box-Cox transformation was improved by Osborne (2010) to make it easier to estimate λ argument [189]. Osborne proposed the new way to apply Box-Cox transformation by the following steps. First, the data are transformed using various λ and then the transformed data are calculated and the skewness is compared. The skewness of transformed data that is the closest to zero yields the optimal transformation.

3.2.4 Normalization methods

Normalization is basically aimed at adjusting the difference of the measurement unit to the same scale or dimensionless [190]. There are various types of normalization depending on each specific purpose. Some popular normalization techniques are described in this section.

Standard normalization or Z normalization ensures that all data are transformed into the new values whose mean is approximately zero and standard deviation is close to one. Observed data above the mean become positive normalized data, while values below the mean become negative normalized data. Z normalization is calculated by Equation (3.6).

$$x' = \frac{x - \bar{x}}{s} \tag{3.6}$$

where x' is the normalized data, x is the original data, $\bar{x}$ is mean of all original data and s is the standard deviation of x.

Range normalization or feature scaling bring all data into the specific range between a anc b. The general formula is given as:

$$x' = a + \frac{(x - x_{min})(b - a)}{x_{max} - x_{min}} \tag{3.7}$$

10713711125

where x' denotes the normalized data, x denotes the original data, a is minimum of normalized data, b is maximum of normalized data and x_{max} are the maximum and x_{min} minimum values in the data set, respectively.

Proportion normalization is based on the proportion of parameter values. This means that each value is divided by the total sum of those parameter values. As its name, proportion normalization can keep the proportion of parameter values to be the same as the original data. The calculation is shown in Equation (3.8)

$$x_i' = \frac{x_i}{\sum_{x=1}^{n} x_i} \tag{3.8}$$

where x' denotes the normalized data, x denotes the original data.

Interquartile normalization is performed using the interquartile range. The interquartile range is the difference between the 25th and 75th percentiles, which are also called lower and upper quartiles, or Q_1 and Q_3. They can be calculated by first sorting the data and then the median is the 50th percentile or Q_2, so it is the value that separates the sorted values into half. The formula for the interquartile normalization is:

$$x' = \frac{x - Q_2}{|Q_3 - Q_1|} \tag{3.9}$$

where x' is the normalized data, x is the original data, Q_1 , Q_2 and Q_3 is the 25th 50th and 75th percentile of x, respectively. This normalization method is less influenced by the extreme value.

3.2.5 Parameter selection methods

Parameter selection or feature selection is a process of selecting a subset of relevant parameter values for using in the model as input. This is a useful technique for filling unnecessary or noisy parameter values, simplified the model to make them easier to interpret and shorten training time of the model [191]. The main objective of parameter selection is to minimize the prediction error. The parameter selection

method can be classified into three types based on the combination of selection algorithms and model structures which are filter algorithm, wrapper algorithm, and embedded algorithm [192, 193].

A filter algorithm selects parameters based on general relationships between input parameters and output parameters. Some relationship measurements are used to score the usefulness of input parameters, such as Pearson correlation coefficient, mutual information, and principle component analysis (PCA). After scoring, the useful parameters are selected according to user's criteria. Then, the selected parameter subset is used to generate the model. This type of parameter selection is effective in term of computation time; however, a filter algorithm does not consider interaction between input parameters which can possibly lead to missing some useful parameters or selecting redundant parameters.

A wrapper algorithm selects parameters based on performance of prediction. This algorithm generates a number of possible parameter subsets. Then, all subsets are used to generate a model and evaluated model performance [194]. The optimal modelling performance also indicates the suitable parameter subset. This type of parameter selection is effective in term of prediction accuracy and possibility of detecting the interaction between input parameters [195]. However, the wrapper algorithm takes the significant computation time compared to the filter algorithm. The examples of wrapper algorithms are forward selection (FS), backward elimination (BE), and genetic algorithm (GA).

An embedded algorithm tries to keep the advantage of filter and wrapper algorithms by selecting parameters during the training process of the model. The embedded model is the modification of the prediction model which adds an option to select or eliminate some parameters [196]. For example, when the weight of a parameter is less than the threshold value, that parameter is eliminated.

Several methods are tested and predictive performances are compared. They are explained in detail as follows:

Forward selection (FS) process starts with no parameter in a model. In the first iteration, each parameter is added as a single input of a model. Then, model performance is tested and the most performance improvement parameter is kept. After that, each of the rest of parameters is added to a model and a model is tested, iteratively. This process only stops when adding parameters can no longer improve performance [197, 198].

On the other hand, backward elimination (BE) process starts with all parameters in a model. In the first iteration, each parameter is eliminated from a model one at a time, then model performance is tested and the most performance improvement parameter subset is kept. After that, each of parameters is eliminated from a model and tested, iteratively. This process only stops when eliminating parameters can no longer improve performance [197, 198].

1071371125

Principal component analysis (PCA) is a transformation process which is popularly used as dimensional reduction technique. The idea of PCA is to obtain some usefulness of information from the variance of data. Mathematically, the original parameter data which are set as a vector of parameter are transformed into a new vector called eigenvector [199]. The number of components in eigenvector is equal to the number of the original parameters. If we consider a parameter matrix $X_{n\times p}$ where n is the number of data records and p is the number of parameters, then the transformation weight defined as $\vec{w}_{(k)} = (w_1,\ldots,w_p)_{(k)}$ can map each row of vector $\vec{x}_{(i)}$ of X to a new vector of the principal component scores $\vec{t}_{(i)} = (t_1,\ldots,t_l)_{(i)}$ by

$$\vec{t}_{k(i)} = \vec{x}_{(i)} \bullet \vec{w}_{(k)} \qquad \text{for} \quad i = 1,\ldots,n \text{ and } k = 1,\ldots,l \tag{3.10}$$

The first component is the maximum variance of the data set, the weight of the first component $\vec{w}_{(1)}$ has to satisfy

$$\vec{w}_{(1)} = \arg\max_{\|w\|=1}\left\{\sum_i (\vec{t}_1)^2_{(i)}\right\} = \arg\max_{\|w\|=1}\left\{\sum_i (\vec{x}_{(i)} \bullet w)^2\right\} \tag{3.11}$$

Another component which is perpendicular to the existing component can be found by subtracting the first $k-1$ principal components from X :

$$\hat{X}_k = X - \sum_{s=1}^{k-1} X\vec{w}_{(s)}\vec{w}_{(s)}^T \tag{3.12}$$

and then a new vector is calculate by extracting the maximum variance from this new matrix again.

PCA can be used for parameter selection by setting variance threshold arguments. According to Equation (3.10) - (3.12), the principal component is ranked by the variances. The first n components are selected when summation of variances reaches the threshold. This method can reduce the number of input variables.

Genetic algorithm (GA) is a stochastic search technique that can be used as parameter selection based on the concept of natural genetics and the evolutionary principle [200]. The concept of this method, which was inspired by the theory of natural evolution, was first proposed by Holland (1975) [201]. The genetic algorithm works with a population of individual strings (chromosomes), each position of a chromosome represents a possible parameter selection. Each chromosome represents a selected parameter subset whose goodness is evaluated by performance of prediction. High performance chromosomes are given more opportunities to reproduce and the offsprings share characteristics taken from their parents. The genetic algorithm is a powerful tool for finding the optimal parameter subset for a prediction model.

1071371125

The procedure of the genetic algorithm for input selection can be summarized in the following steps. First, a parameter subset is randomly generated by binary strings (chromosomes) to represent the selected and unselected input as 1 and 0, respectively. Next, the models are trained and the prediction error for each chromosome in the population is calculated. The high performance (low prediction error) chromosomes are kept as the new parents and used to generate the offspring (children) by genetic operators: crossover and mutation. The process is repeated until the stopping conditions are met. The children from the final iteration with the optimal performance are included in the optimal parameter subset.

There are four arguments of the genetic algorithm; i.e., the number of children for each generation, the maximum number of generation (stopping condition), the

mutation probability, and the crossover probability. Crossover and mutation are the two main operators in genetic algorithm. Crossover is the process of producing a new child from a pair of parents by randomly selecting each position in a chromosome from both parents. Mutation is the process of randomly changing some positions in a chromosome.

3.2.6 Prediction models

Prediction is the core step in water quality prediction framework. The water quality parameter prediction is classified as a regression problem which is the estimation of parameter values from relationships among them. Many techniques can solve this problem in many different ways. The simplest one is the linear regression [202]. The linear regression aim to represent relationship between two parameters x (input) and y (output) by a line which shows in Equation (3.13).

$$y = \beta x + m \tag{3.13}$$

where m is the intersection and β is the slope of the line. Equation (3.14) can be generalized to a multiple input parameter system of Equations as follows:

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + m \tag{3.14}$$

where x_n is n th input parameter which is called a Multiple linear regression (MLR). The argument β can be optimized by least square method [203]. However, this simple technique tends to be bias by the available parameters. It is suitable for a large number of observed data. It will provide poor results when the number of observed data is small.

Alternatively, there is another approach that can solve regression problem in different way which is called machine learning. Machine learning tries to teach and learn the pattern or characteristics of data without directly training it [204]. The concept of machine learning is by trial-and-error approach which is the same as human learning approach. Firstly, a computer sees data and guesses the output without any clue. Then, it starts to learn the pattern from the error of the first guess and make a

10713711 25

new guess which is getting closer to the correct answer. The process is repeated until the acceptable output is predicted [205]. According to this concept, many techniques were developed and applied to many fields.

According to literature review in 2.2.5, there are two techniques which are widely used in water quality modelling. Artificial neural network (ANN) and support vector regression (SVR) are powerful learning models for water quality parameter prediction. Artificial neural network was already explained in 3.2.2. Instead of setting the missing value as output, the target parameter is set for prediction.

Support vector regression (SVR) is another popular model in water quality prediction field. Support vector regression is the extension of support vector machine (SVM) which aims to solve the classification problem [206]. The idea of support vector regression is to calculate the output parameter y by an error of prediction no greater than ξ for each input parameter x. The equation is the same as Equation (3.14), but the way of optimizing β is different; i.e., the prediction error is minimized by the objective function:

$$f(\beta)=\frac{1}{2}\|\beta\|^2+c\sum_{i=1}^{n}(\xi_i+\xi_i^*) \tag{3.15}$$

where β is the weight vector of input parameter, ξ_i and ξ_i^* are the ith slack arguments which are the flexible terms of fitting and c is the constant of slack arguments. The equation is based on three conditions which are:

$$\forall n: y-(\beta x_i+m)\le\varepsilon+\xi_i \tag{3.16}$$

$$\forall n: (\beta x_i+m)-y\le\varepsilon+\xi_i^* \tag{3.17}$$

$$\forall n: \xi_i,\xi_i^*\ge 0 \tag{3.18}$$

In the next chapter, dissertation methodologies are explained which focus on the experiment to find the optimal method in each step of the framework and the proposed method.