CHAPTER 5

RESULTS

In this chapter, the results of experiment in Chapter 4 were shown in 5.1-5.7, respectively.

5.1 Data descriptive statistics

According to raw data, 52-54 water quality parameters were collected for each record. Some example of missing value percentage of water quality parameters records of Chaophraya River between 2538 and 2556 was shown in Figure 5.1. According to the chart, parameter record missing ratio could roughly divided into three groups. First ,the group of parameters with less than 15% missing ratio, namely, DO, Temp, pH, turbidity, BOD, total coliform bacteria, fecal coliform bacteria, EC, salinity, SS, NO_3^-, TS, NO_2^-, NH_3, PO_4^{3-}, and TDS. Second, the group of parameters that more than 20% missing ratio and less than 70% missing ratio which were group of hardness and heavy metals. The last group were pesticides which had the missing value percentage more than 70%. Only the first group was used to the next step. Basic statistics of total 18 usable water quality parameters are shown in Table 5.1.

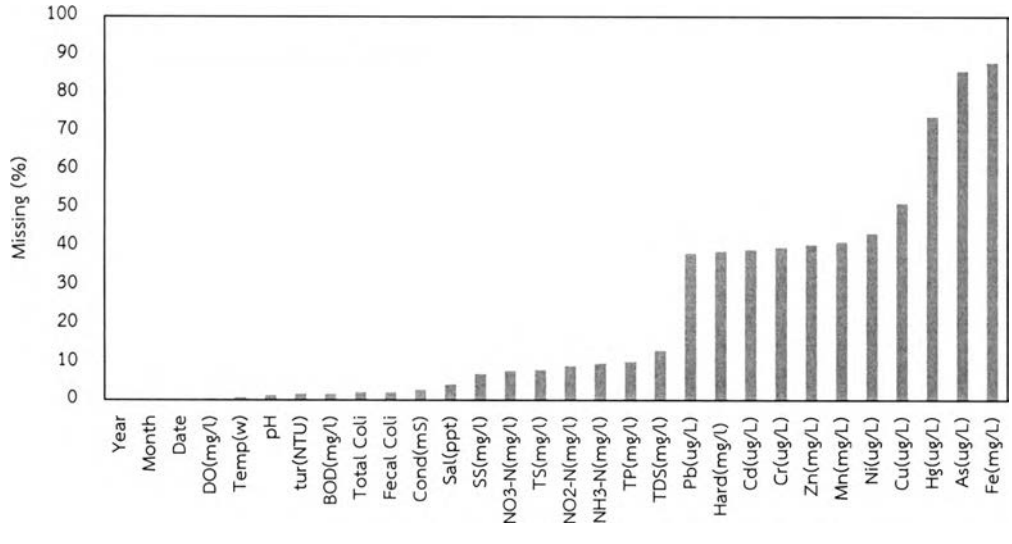

Figure 5.1 The example of missing value percentage of water quality parameters record of Chaophraya River during 2538-2556 B.E. (%)

Table 5.1 Basic statistics of the water quality parameters in Chaophraya River, Thailand during 2538-2556 B.E.

Parameter	Unit	Min	Max	Mean	SD
WT	°C	2.96	34.70	29.64	1.81
pH	-	5.70	9.00	7.22	0.55
Tur	NTU	0.20	5000.00	93.57	168.47
EC	µS/cm	0.00	38220.00	1341.11	4491.07
Sal	g/L	0.00	84.00	0.64	3.37
DO	mg/L	0.00	11.60	4.12	2.10
BOD	mg/L	0.10	12.90	2.09	1.78
TC	MPN/100 ml	2.00	24000000.00	156812.66	1400489.88
FC	MPN/100 ml	0.00	16000000.00	48801.20	566255.12
PO_4^{3-}	mg/L	0.00	3.80	0.13	0.22
NO_3^-	mg/L	0.00	12.60	0.86	1.55
NO_2^-	mg/L	0.00	3.00	0.12	0.32
NH_3	mg/L	0.00	5.00	0.39	0.54
SS	mg/L	1.00	464.60	54.83	49.50
TS	mg/L	4.98	42284.00	1215.75	3609.91
TDS	mg/L	0.00	42216.00	1243.09	3695.47
T	-	1.00	12.00	6.63	3.49
S	km	7.00	376.40	149.24	113.01

Next, Spearman correlation coefficient was used to show the pattern and relation of the parameters. Spearman correlation between water quality parameters and variables (monitoring year, monitoring month and monitoring station distance from

estuary selected parameters from) are shown in Table 5.2-5.5, respectively. Then, Spearman correlation between water temperature and other water quality parameters are shown in Table 5.5 to demonstrate the relation between parameters.

The monitoring year had significantly positive correlated with BOD and NO_3^-, which means there were the increasing trend of BOD and NO_3^- in Chaophraya River during 2533-2556 B.E. Moreover, the monitoring year had significantly negative correlated with pH, conductivity, total coliform, fecal coliform, PO_4^{3-} and SS, implied decreasing trends of those parameters in Chaophraya River over study period.

Table 5.2 Spearman correlation of monitoring year and water quality parameters

Significantly Correlated ($p < 0.05$)	Non-significantly Correlated
pH (-)	Temperature
Conductivity (-)	Turbidity
BOD (+)	Salinity
Total Coliform (-)	DO
Fecal Coliform (-)	NO_2^-
PO_4^{3-} (-)	NH_3
NO_3^- (+)	TS
SS (-)	TDS

The monitoring month was significantly positive correlated with suspended solid. This could be interpret that the suspended solid in the late of the year usually higher than the early of the year. The opposite pattern was found in many parameters which had significant negative correlation with monitoring month, namely, temperature, conductivity, salinity, DO, BOD, NO_2^-, NH_3, TS and TDS.

Table 5.3 Spearman correlation of monitoring month and water quality parameters

Significantly Correlated ($p < 0.05$)	Non-significantly Correlated
Temperature (-)	pH
Conductivity (-)	Turbidity

Significantly Correlated ($p < 0.05$)	Non-significantly Correlated
Salinity (-)	Total Coliform
DO (-)	Fecal Coliform
BOD (-)	PO_4^{3-}
NO_2^- (-)	NO_3^-
NH_3 (-)	
SS (+)	
TS (-)	
TDS (-)	

Some water quality parameters also related to the location where water sampling was collected. The upstream water normally has higher quality than the downstream water. This supported by the strong relationship between distance from monitoring station to estuary and TDS, TS, salinity and conductivity.

Table 5.4 Parameters and monitoring station relationship

Significantly Correlated ($p < 0.05$)	Non-significantly Correlated
pH (+)	Temperature
Conductivity (-)	**Turbidity**
Salinity (-)	Total Coliform
DO (+)	Fecal Coliform
BOD (-)	
PO_4^{3-} (-)	
NO_3^- (-)	
NO_2^- (-)	
NH_3 (-)	
SS (-)	
TS (-)	
TDS (-)	

Salinity was one of the parameter highly associated with the measuring location. The salinity parameter data measured by monitoring stations are shown in Figure 5.2. The monitoring stations were ordered by the location from the nearest to the farthest from estuary, hence the 1st station was the most downstream station and 32th station was the most upstream one. It is obvious that salinity value, regularly, is the highest at the downstream (close to the sea), and decreased gradually compared to upstream stations.

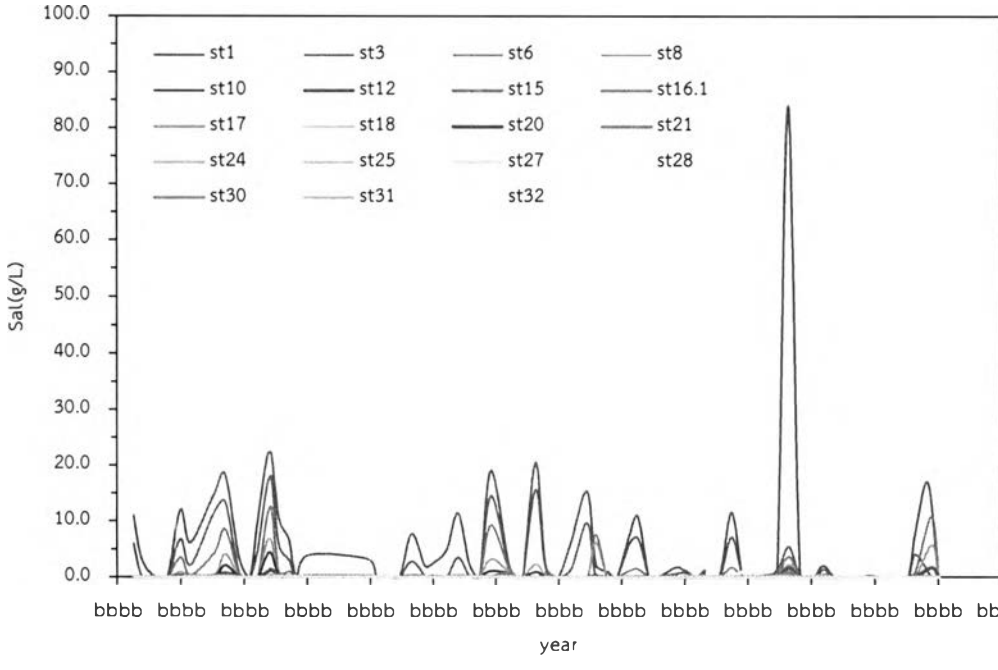

Figure 5.2 Historical data of salinity from monitoring stations along Chaophraya River during 2538-2556 B.E.

Some water quality parameters had no pattern; their value had low or no correlation with measuring time and location of the station. Turbidity value of Chaophraya River over study period was shown in Figure 5.4. The rest of historical data chart was shown in Appendix B.1.

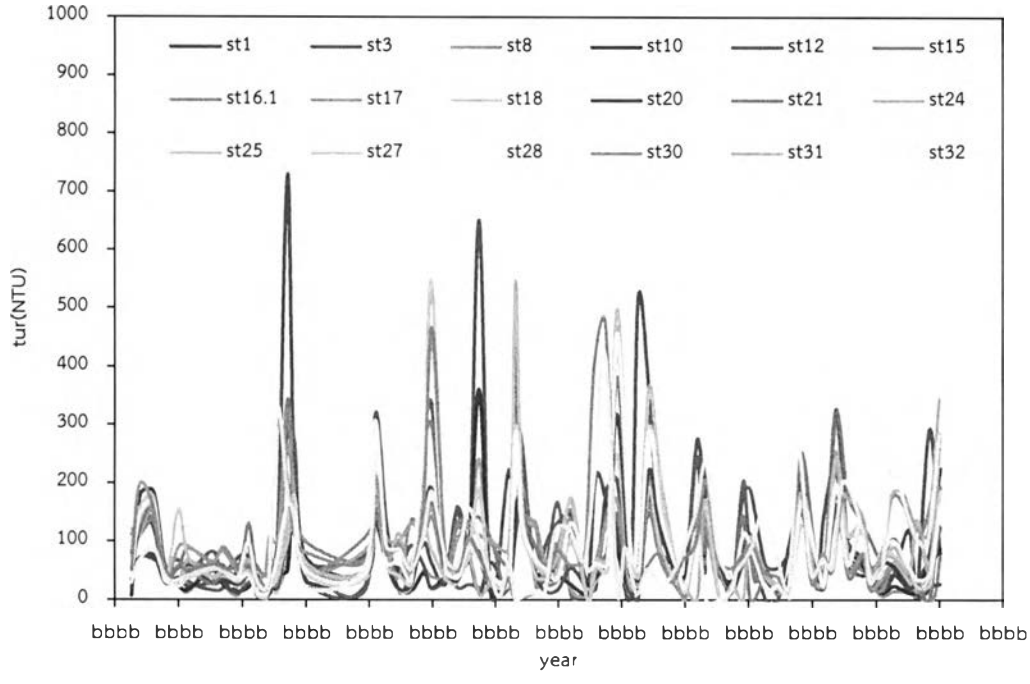

Figure 5.3 Historical data of turbidity from monitoring stations along Chaophraya River during 2538-2556 B.E.

Although some water quality parameter did not have any pattern, they may relate to other water quality parameters. In this part, water temperature was select to be an example as it is fundamental effect which relate to many parameters. Water temperature was significantly positive correlated with BOD, NO_3^- , NO_2^- , TS and SS. However, relation between temperature and many parameter could not be found by Spearman correlation, such as DO, Conductivity and Total Coliform.

Table 5.5 Spearman correlation of water temperature and other water quality parameters

Significantly Correlated (p > 0.05)	Non-significantly Correlated
BOD (+)	pH
NO_3^- (+)	Turbidity
NO_2^- (+)	Conductivity
SS (+)	Salinity
TS (+)	DO
	Total Coliform
	Fecal Coliform
	PO_4^{3-}
	NH_3
	TDS

5.2 Imputation results

In this part, selected parameters from 5.1 with missing value was imputed by three methods: mean replacement, K-nearest neighbor (K-nn) and artificial neural network (ANN). The imputed data were used to predict water quality parameters and the performances were shown in Table 5.6 (complementary results was shown in Table B.1). After comparison, K-nearest neighbor with k=5 gave the highest performance considered to lowest RMSE and high Spearman correlation. The imputed data by K-nearest neighbor with k=5 were used in the next step.

Table 5.6 Three imputation methods performance evaluation

Imputation method	Argument	RMSE	Spearman correlation
mean replacement	-	1.418±0.001	0.642±0.029
ANN	-	1.389±0.002	0.660±0.021
K-nn	k=2	1.451±0.078	0.667±0.041
	k=3	1.398±0.007	0.666±0.030
	k=4	1.412±0.028	0.666±0.030
	k=5	**1.362±0.045**	**0.668±0.035**
	k=6	1.478±0.124	0.669±0.029
	k=7	1.668±0.021	0.612±0.021

5.3 Data transformation results

The imputed data from 5.2 were transform by Equation 3.5 with λ = -1, -0.5, 0, 0.5 and 1. Then, the transform data were measured skewness as shown in Table 5.7. The skewness showed symmetry of distribution. Ideally, normal distribution will provide a near zero skewness. Therefore, the transformation function which provided closest to zero skewness is selected for each parameters. Noted that according to Equation 3.5, λ =1 mean transformed data is original data.

Table 5.7 Skewness of parameters at different λ value according to Osborne's transformation

Model input	λ				
	-1	-0.5	0	0.5	1
Distance	1.03	0.6	**-0.18**	0.2	0.53
Month	1.73	1.11	-0.61	-0.2	**0.15**
WT	30.99	23.8	-13.87	-6.57	**-3.09**
pH	2.71	1.6	-0.84	-0.28	**0.17**
Tur	-1.8	-2.6	**4.79**	10.72	21.09
EC	-3.02	-3.29	**3.67**	4.21	4.92
Sal	-4.32	-4.82	**5.9**	8.68	14.89
DO	1.73	1.34	-0.94	-0.52	**-0.09**
BOD	-0.2	-0.63	**1.11**	1.69	2.38
TC	-7.67	-9.12	**11.05**	13.1	14.83
FC	11.88	-14.08	**16.74**	19.84	23.19
PO_4^{3-}	-1.44	-1.94	**2.82**	4.6	8.16
NO_3^-	-1.28	-1.75	**2.37**	3.19	4.18
NO_2^-	-1.99	-2.42	**3.06**	4.03	5.35
NH_3	-0.64	-0.98	**1.44**	2.1	3.13
SS	-0.13	-0.5	**0.97**	1.61	2.5
TS	-2.96	-3.22	**3.6**	4.17	5.11
TDS	-2.75	-2.99	**3.35**	3.92	4.89

The skewness pattern in Table 5.7 were plotted to represent the pattern of change as shown in Figure 5.4. Figure 5.4 a) showed a group of parameter of S shape pattern. This normally happen when the original data is skewed right or positive skewness. To convert to be normal distribution, these parameter should be transform by λ =0, which is logarithmic transformation. Another group showed in Figure 5.4 b) showed U shape pattern which mean these parameter is already in normal distribution, no need to transform. Therefore, 14 parameters were transform by logarithmic

function, namely, turbidity, salinity, total coliform, fecal coliform, PO_4^{3-}, NO_2^-, NO_3^-, NH_3, SS, TS, TDS, EC, BOD and distance from sea.

After transformation, the transformed data and non-transformed data form 5.2 were used to train several model for water quality prediction. The predictive performance were shown in Table 5.8 (complementary results was shown in Table B.2). After comparison, model which were trained by non-transformed data gave the highest performance considered to high Spearman correlation. Noted that in this experiment, RMSE is bias because of the scale of transformed data were shrunk, Spearman correlation was used as only one criteria in this study. Therefore, non-transformed data were used in the next step.

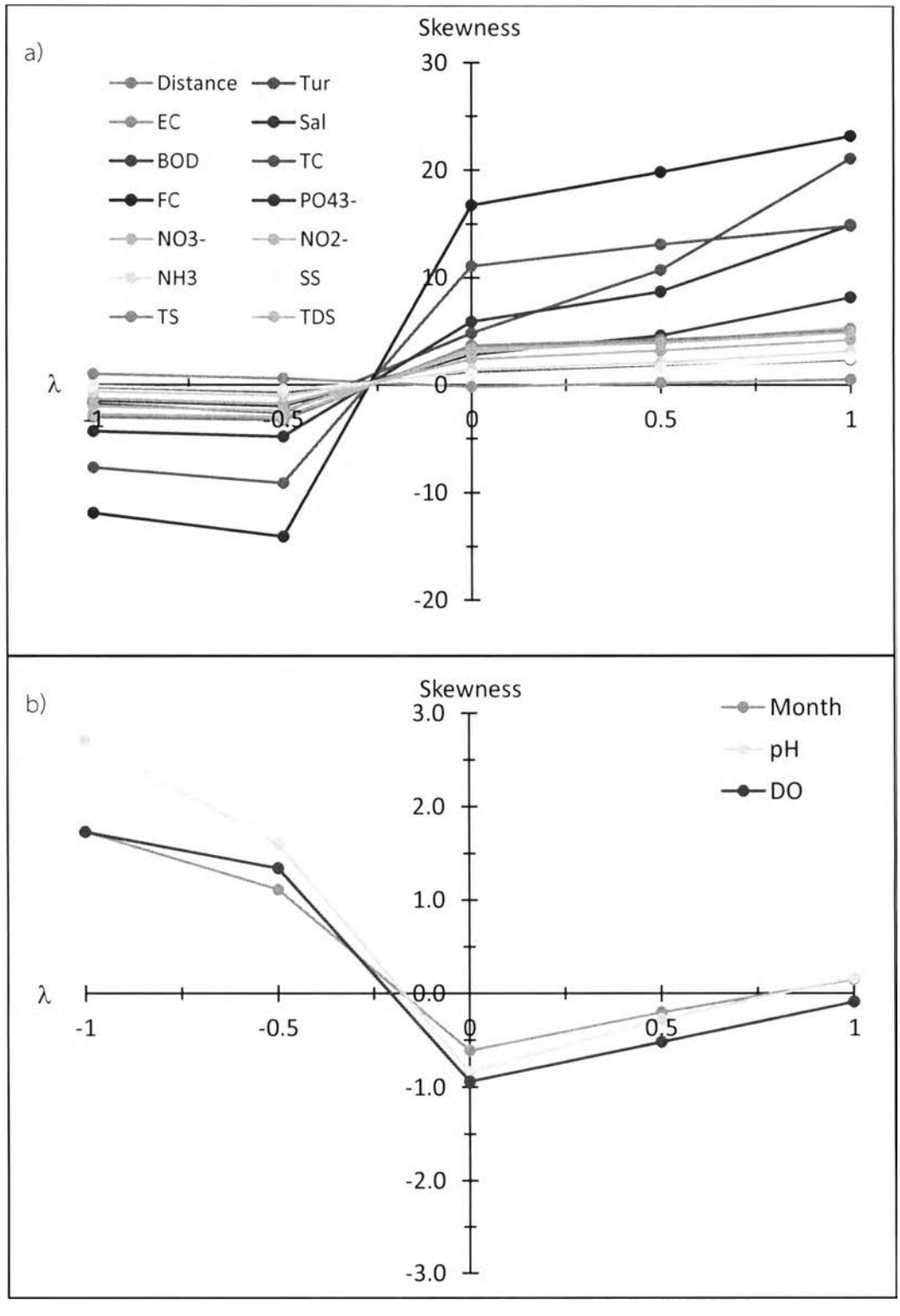

Figure 5.4 Relationship between λ and skewness of each parameter, a) and b) showed different pattern of relationship.

Table 5.8 Performance comparison of transformed data and non-transformed data

Data	RMSE	Spearman correlation
Transformed	0.087±0.006	0.641±0.043
Non-transformed	**0.129±0.007**	**0.653±0.015**

5.4 Normalization results

In this part, transformed data were normalized by four methods: Z normalization, range normalization, proportion normalization and interquartile normalization. Then, normalized data were used to predict water quality parameters to evaluated suitable normalized method. To predict water quality, two parameter selection algorithm and two models were used with four normalized data.

The average of model performance by each normalized method are shown in Table 5.9. On average, Z-normalization gave the highest performance considered to lowest RMSE and highest R. Therefore, Z-normalization was used in the next step.

Table 5.9 Four normalization methods performance comparison

Normalization	RMSE	Spearman correlation
Z	**1.398±0.052**	**0.642±0.021**
Range	1.414±0.063	0.638±0.041
Proportion	1.420±0.083	0.636±0.031
InterQuatile	1.525±0.198	0.511±0.176

5.5 Parameter selection results

The normalized data are used to check whether parameter selection algorithm is suitable for water quality prediction. As mentioned in 4.2.1, 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 to evaluate the predictive performance which are shown in Table 5.11 (complementary results was shown in Table B.3 - B.5). According to RMSE and R value, genetic algorithm method gave the highest performance.

Table 5.10 Four parameter selection methods performance comparison

Parameter selection method	RMSE	Spearman correlation
Forward selection	1.440±0.049	0.680±0.019
Backward elimination	1.393±0.182	0.723±0.011
Principal component analysis	1.635±0.048	0.417±0.055
Genetic Algorithm	**1.355±0.149**	**0.735±0.006**

5.6 Prediction models comparisons

As mentioned in 4.6, three model which were support vector regression (SVR), artificial neural network (ANN) and multiple linear regression (MLR) were used to predict water quality parameter. Those models were implemented and tested to evaluate the predictive performance which are shown in Table 5.11. According to RMSE and Spearman correlation value, artificial neural network gave the highest performance. Thus, ANN were chosen to be a core of proposed model in the next part.

Table 5.11 SVM, ANN, MLR performance comparison

Model	RMSE	Spearman correlation
SVM	1.539±0.129	0.574±0.039
ANN	**1.642±0.096**	**0.638±0.024**
MLR	1.501±0.218	0.566±0.076

5.7 Space and time neural network results

The space and time neural network (STNN) were developed handle the multi-dimensional water quality data. The experiment was set to determine the predictive performance of space and time neural network by compare with time delay neural network (TDNN) which was set argument s_{max} = 0 and distance neural network (DNN) which was set argument t_{max} = 1. Three model types were used to predict all water quality parameter. The result showed some example of parameter, namely, electrical

conductivity (EC), total dissolved solids (TDS) and phosphate concentration (PO_4^{3-}) in Table 5.12-5.14, respectively.

Table 5.12 Evaluation of model fits to EC observations

EC models	Upstream station (s_{max})	Time lag (t_{max})	Initial input	Neural network structure	RMSE	Spearman correlation
Time delay NN	0	**1**	**20**	**12-8-1**	**0.053**	**0.776**
		2	38	19-11-1	0.059	0.726
		3	56	19-11-1	0.057	0.751
Distance NN	1	**1**	38	17-10-1	0.058	0.732
	2		**56**	**25-14-1**	**0.057**	**0.744**
	3		74	35-19-1	0.067	0.628
Space and Time NN	1	2	74	41-22-1	0.060	0.728
		3	110	46-25-1	0.054	0.774
	2	2	110	60-32-1	0.055	0.782
		3	**164**	**77-40-1**	**0.049**	**0.816**
	3	2	146	86-45-1	0.065	0.658
		3	218	127-65-1	0.068	0.591

According to Table 5.12, the results of electric conductivity (EC) modelling showed that the STNN model was fittest predictor (RMSE = 0.049 and Spearman correlation = 0.816) compared with TDNN model and DNN model (RMSE = 0.053, Spearman correlation = 0.776 and RMSE = 0.057, Spearman correlation = 0.774, respectively). The optimal arguments setting of STNN were s_{max} = 2 and t_{max} = 3 which means that the prediction is based on the historical data three timestamps recorded in the past (roughly 9 months) and two upstream stations. Input data has a total of 164 parameters. After parameter selection step, the optimal 77 selected parameters were used to train models (shown in Appendix B.4).

The EC model simulator snapshot from Rapid miner studio software was shown in Figure 5.5. The left hand side is normalized input parameters value setup part, which including 77 parameters. The prediction results are shown on the right hand side.

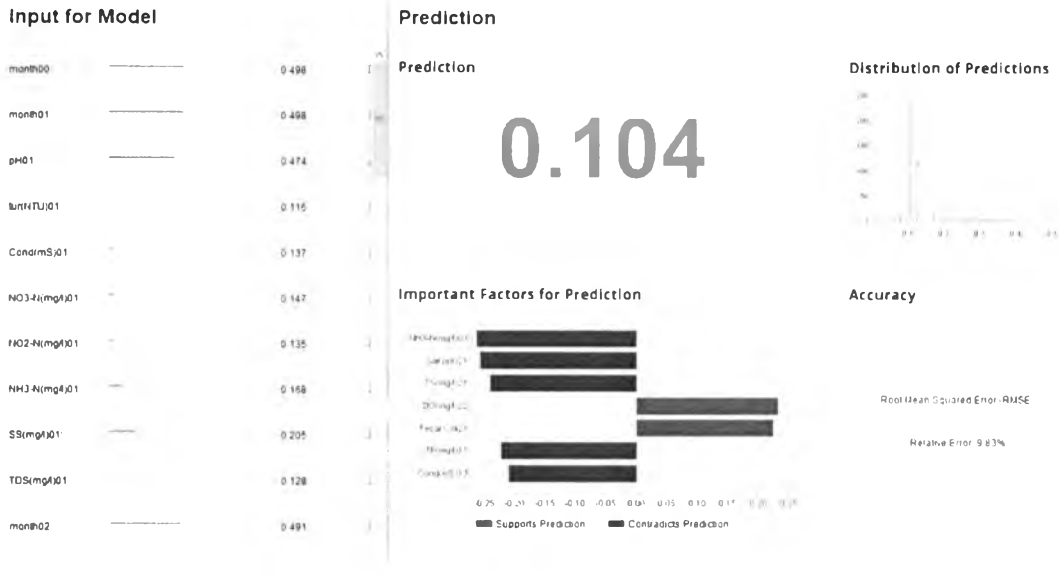

Figure 5.5 EC model simulator snapshot from Rapid miner studio

According to Table 5.13, the results of total dissolved solids (TDS) modelling showed that the STNN model was fittest predictor (RMSE = 0.044 and Spearman correlation = 0.659) compared with TDNN model and DNN model (RMSE = 0.046, Spearman correlation = 0.599 and RMSE = 0.047, Spearman correlation = 0.594, respectively). The optimal arguments setting of STNN were s_{max} = 2 and t_{max} = 2 which means that the prediction is based on the historical data two timestamps recorded in the past (roughly 6 months) and two upstream stations. Input data has a total of 110 parameters. After parameter selection step, the optimal 52 selected parameters were used to train models (shown in Appendix B.4).

Table 5.13 Evaluation of model fits to TDS observations

TDS models	Upstream station (S_{max})	Time lag (t_{max})	Initial input	Neural network structure	RMSE	Spearman correlation
Time delay NN	0	1	20	8-6-1	0.046	0.596
		2	**38**	**18-11-1**	**0.046**	**0.599**
		3	56	27-15-1	0.050	0.578
Distance NN	**1**	**1**	**38**	**15-9-1**	**0.047**	**0.594**
	2		56	25-14-1	0.050	0.527
	3		74	32-18-1	0.050	0.534
Space and Time NN	1	2	74	34-19-1	0.053	0.503
		3	110	64-34-1	0.050	0.529
	2	**2**	**110**	**52-28-1**	**0.044**	**0.659**
		3	164	86-45-1	0.052	0.485
	3	2	146	77-40-1	0.049	0.553
		3	218	87-45-1	0.052	0.515

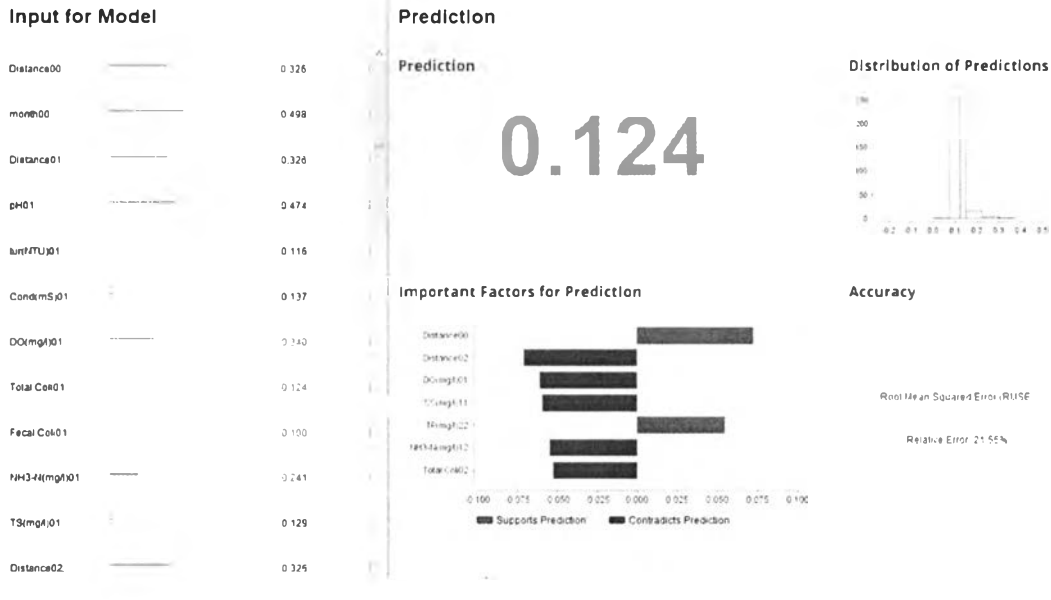

Figure 5.6 TDS model simulator snapshot from Rapid miner studio

The TDS model simulator snapshot from Rapid miner studio software was shown in Figure 5.6. The left hand side is normalized input parameters value setup part, which including 52 parameters. The prediction results are shown on the right hand side.

Table 5.14 Evaluation of model fits to PO_4^{3-} observations

PO_4^{3-} models	Upstream station (s_{max})	Time lag (t_{max})	Initial input	Neural network structure	RMSE	Spearman correlation
Time delay NN	0	1	20	4-4-1	0.023	0.365
		2	38	13-8-1	0.023	0.407
		3	**56**	**21-12-1**	**0.020**	**0.621**
Distance NN	**1**	**1**	**38**	**13-8-1**	**0.023**	**0.371**
	2		56	23-13-1	0.024	0.272
	3		74	29-16-1	0.030	0.213
Space and Time NN	1	2	74	35-19-1	0.024	0.351
		3	110	48-26-1	0.026	0.306
	2	**2**	**110**	**38-21-1**	**0.025**	**0.710**
		3	164	80-42-1	0.022	0.625
	3	2	146	79-41-1	0.025	0.445
		3	218	108-56-1	0.025	0.186

According to Table 5.14, the results of phosphate (PO_4^{3-}) modelling showed that the STNN model was fittest predictor (RMSE = 0.025 and Spearman correlation = 0.710) compared with TDNN model and DNN model (RMSE = 0.020, Spearman correlation = 0.621 and RMSE = 0.023, Spearman correlation = 0.371, respectively). The optimal arguments setting of STNN were s_{max} = 2 and t_{max} = 2 which means that the prediction is based on the historical data two timestamps recorded in the past (roughly 6 months) and two upstream stations. Input data has a total of 110 parameters. After parameter selection step, the optimal 38 selected parameters were used to train models (shown in Appendix B.4).

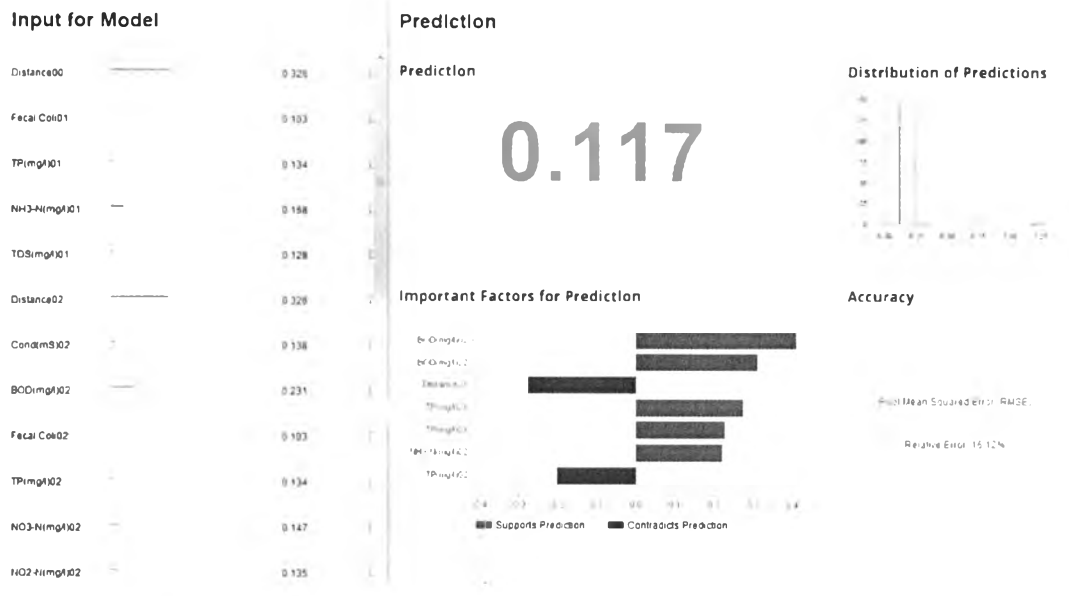

Figure 5.7 PO_4^{3-} model simulator snapshot from Rapid miner studio

The PO_4^{3-} model simulator snapshot from Rapid miner studio software was shown in Figure 5.7. The left hand side is normalized input parameters value setup part, which including 38 parameters. The prediction results are shown on the right hand side.

After predictive performance evaluation, the selected parameter of each model was analyzed to estimate the importance of input parameter to predicted parameter. Decomposed weight was calculated from summation of connected weights inside the trained neural network to show the direction and magnitude of relationship between input and output parameter. Relative importance was calculated from proportion of input magnitude to total magnitude (summation of all magnitude), which could indicated the impact of input parameter on output parameter.

Noted that parameter name was followed by two number, the first is number of upstream monitoring station and the second is the time delay. For example, EC01 mean EC parameter which was measured at the same monitoring station (S = 0) on the previous monitoring period (t = 1).

According to the optimal model which showed in Table 5.12, 77 selected parameters were analyzed and ranked the importance on EC model. As shown in Table 5.15, the most important parameters for calculating EC were turbidity21, fecal

coliform21, distance11, distance13, EC01, $NH_3$01, distance03, fecal coliform22, BOD21 and total coliform11, respectively.

Table 5.15 Top 10 dimensional parameter importance on EC model.

Importance rank	Parameter	Decomposed weight	Relative importance (%)
1	Turbidity21	10.55	3.27%
2	Fecal Coliform21	10.32	3.20%
3	Distance11	-9.77	3.03%
4	Distance13	-9.70	3.01%
5	EC01	9.51	2.95%
6	$NH_3$01	9.29	2.88%
7	Distance03	-8.94	2.77%
8	Fecal Coliform22	8.54	2.65%
9	BOD21	-8.50	2.64%
10	Total Coliform11	8.46	2.62%

According to the optimal model which showed in Table 5.13, 52 selected parameters were analyzed and ranked the importance on TDS model. As shown in Table 5.16, the most important parameters for calculating TDS were $NH_3$02, TDS02, turbidity11, $NH_3$01, distance12, pH21, distance11, distance01, distance02 and distance00, respectively.

Table 5.16 Top 10 dimensional parameter importance on TDS model.

Importance rank	Parameter	Decomposed weight	Relative importance (%)
1	$NH_3$02	8.35	5.54%
2	TDS02	-7.32	4.86%
3	Turbidity11	6.97	4.62%
4	$NH_3$01	6.14	4.07%
5	Distance12	-5.91	3.92%
6	pH21	5.87	3.89%
7	Distance11	-5.87	3.89%
8	Distance01	-5.36	3.55%

Importance rank	Parameter	Decomposed weight	Relative importance (%)
9	Distance02	-5.27	3.50%
10	Distance00	-5.23	3.47%

According to the optimal model which showed in Table 5.14, 38 selected parameters were analyzed and ranked the importance on PO_4^{3-} model. As shown in Table 5.17, the most important parameters for calculating PO_4^{3-} were temperature22, salinity22, distance21, PO_4^{3-}01, BOD02, fecal coliform02, month11, EC22, total coliform11 and TDS01, respectively.

Table 5.17 Top 10 dimensional parameter importance on PO_4^{3-} model.

Importance rank	Parameter	Decomposed weight	Relative importance (%)
1	Temperature22	-8.18	6.29%
2	Salinity22	7.75	5.96%
3	Distance21	-6.48	4.98%
4	PO_4^{3-}01	6.06	4.66%
5	BOD02	6.02	4.63%
6	Fecal Coliform02	5.94	4.57%
7	month11	5.93	4.56%
8	EC22	5.68	4.37%
9	Total Coliform11	5.37	4.13%
10	TDS01	-4.74	3.65%