CHAPTER 6

DISCUSSIONS

6.1 Data descriptive statistics

Missing data percentage that showed in 5.1 can occur for several reasons. The group which had the missing value percentage more than 90% was pesticide parameter group. According to Enhancement and Conservation of the National Environmental Quality Act 2535 B.E., pesticides were monitored every five years [174]. Thus, when recorded into the database, it appears to be missing data. Similarly, heavy metal parameter are scheduled to be monitored at least once a year. This causes both groups to have insufficient data for modeling to predict water quality.

The parameters with less than 15% missing ratio is a group of primary water quality parameters that are scheduled to be monitored four times a year. The disappearance of this data may be due to not collecting data at that time or occurring due to limitations of measurement methods or measurement errors. For these reasons, 16 water quality parameter data need to be imputed in the next step.

Basic statistics of the parameters in Table 5.1 illustrated the different units of measurement of parameters and the range of observed data between 2538-2556 B.E. This could be bias in the process of training model because some models focus on parameters with high value more than the parameters that are less. To prevent problems, the data must be transformed and normalized before the modeling process.

Spearman correlation coefficient between water quality parameter and monitoring year showed long term trend of each parameter overtime, as showed in Table 5.2. Long term trend roughly indicated slightly change of half of parameters over the year (8 of the 16 parameters). On the other side, 10 of the 16 parameters are significantly related to the month. This shows the relationship of seasons with most parameters. In this study, water quality model focus on prediction of parameters in less than one year in the future. Therefore, timestamp in purposed model was represented by months instead of years for more accurate in short term prediction.

The location of the monitoring stations, which were represented by the distance from the sea related to almost parameters. As shown in Table 5.4, TDS, TS, salinity and conductivity were significantly correlated parameters. It could be interpreted that the some parameters are related to the affected by sea water. The distance from the sea directly correlates with the salinity of river water. Salinity is the concentration of ions of dissolved salts in the water, which is proportional to the TDS and TS. Furthermore, high ionic salt in water also increase electrical conductivity as well.

Another group of parameters which also significant correlated with monitoring station location is DO, BOD, PO_4^{3-}, NO_3^-, NO_2^-, NH_3 and SS. Unlike the previous group. These parameters increases along the flow distance of water from upstream to downstream (except for DO, which decreases), due to the accumulation of organic matter from various sources, such as wastewater from industrial, household and agriculture. Degradation of organic matter could be observed by concentration of PO_4^{3-}, NO_3^-, NO_2^- and NH_3. It is directly related to the SS and BOD, which increases with the organic matter degradation activity in the water. When oxygen is used to degrade, it lowers the DO value. Thus, monitoring station was added to the model as one of the parameter which represented by using distance from estuary.

Spearman correlation coefficient between water temperature and other parameters was analysed to be an example of interaction among parameters. The result in Table 5.5, could indicate the limitation of Pearson correlation coefficient was that it only measures linear relationships between a parameter to another one. If the relationship was not linear then the result was inaccurate. This limitation was shown by non-significant correlated of temperature and DO which is non-linearity relationship [213, 214]. This was confirmation of complex and non-linear relationship between water quality parameters. Therefore, the models to be used for prediction must support this nonlinear relationship.

6.2 Imputation

Three imputation methods: mean replacement, K-nearest neighbor (K-nn) and artificial neural network (ANN) were implemented to fill the missing data. The mean replacement used only the value of a parameter to calculate the average without consider other parameters (known as univariate calculation). The missing parameter was calculated from average of that parameter measured in other stations and time. The results from several prediction model showed the highest error.

In contrast, artificial neural networks (ANN) calculated missing values from other parameters that were collected at the same time and same monitoring station. The results was better than mean replacement method. However, ANN was trained using an overview of the relationship of the parameters to be a single model that could be used for all missing value.

Unlike the previous methods, K-nearest neighbor (K-nn) clustered the similar events and used them to calculate missing value. Similarity of event was determined by Euclidean distance. The predicted results showed that the different k values had a significant effect on performance. The k values are two and seven gave the worst predictive performance in all trials. The value of k is equal to five got the highest predictive efficiency. Too small k value makes the calculated value depend on only the few closest events. On the other hand, with too large k values, the mean value becomes the mean of most events. At K equal to the number of record, this is not different from the mean replacement method. The imputed data by K-nearest neighbor with k=5 were used in the next step as the complete dataset.

6.3 Data transformation

The Osborne's transformation is an estimate of λ by measure the skewness as close to 0 as possible [189]. Table 5.7 showed skewness results of various λ transformation. Some parameters have been transform reasonably by logarithmic transformation, such as total coliform and fecal coliform, measured by the Most Probable Number (MPN) method. The MPN procedure is dilution of the water sample concentration by 10 times until the amount of bacteria cultured in the sample water is countable. Then, colony count was multiplied by the proportion of dilution. The MPN method is standard measurement for estimate the concentration of microorganisms. This is consistent with the exponential growth rate of microorganisms when there is no limiting factors. Therefore, these logarithmic transformation are reasonable.

On the other hand, there were some transformed physical parameters that did not make sense when considering environmental aspect, such as logarithm of electrical conductivity or distance from the sea.

However, when these values are used to predicted the water quality parameters by several models. On average, the models generated from the transformed data are less effective than the unmodified data. Transformation to the data may change the pattern of the parameter relationship. The models could not be predicted effectively. In other words, the models used in this study can deal with non-normal distribution data. This is consistent with a number of studies that support this statement. It could be concluded that the data transformation is unnecessary for predicting water quality parameters in this study.

6.4 Normalization

According to the results showed in 5.4, Z-normalization gave the highest performance when compare with other methods, therefore it was used in the next step. However, after finished parameter selection step and model comparison step, genetic algorithm (GA) was proved to be an optimal parameter selection method and artificial neural network (ANN) was proved to be an optimal model (result of these two step would be discussed in detail on 6.5 and 6.6). The imputation, transformation and normalization were rechecked again to confirm validity of results. Normalization methods performance comparison of individual model were shown in Table 6.1. The result showed that there was no difference between normalization methods on GA-ANN model. This could be concluded that GA-ANN could adapt itself to different types of normalization equally. Instead of using Z normalization, range normalization was applied for the next step on 4.5, 4.6 and 4.7, because it could be easier to interpret the relationship of parameter without sign conversion.

Table 6.1 Normalization methods performance comparison show by individual model

Model code name	Normalization	Parameter Selection	model	RMSE	ρ
ZPcaAnn	Z	PCA	ANN	1.471	0.625
RPcaAnn	Range			1.472	0.667
PPcaAnn	Proportion			1.539	0.599
IPcaAnn	InterQuatile			1.615	0.446
ZGaAnn	Z	**Genetic Algorithm**		1.353	**0.673**
RGaAnn	**Range**			1.353	**0.673**
PGaAnn	Proportion			1.353	**0.673**
IGaAnn	InterQuatile			1.353	**0.673**
ZPcaSvm	Z	PCA	SVM	1.399	0.635
RPcaSvm	Range			1.464	0.584
PPcaSvm	Proportion			1.413	0.628
IPcaSvm	InterQuatile			1.762	0.292
ZGaSvm	Z	Genetic Algorithm		1.368	0.634
RGaSvm	Range			1.365	0.629
PGaSvm	Proportion			1.375	0.643
IGaSvm	InterQuatile			1.369	0.632

6.5 Parameter selection

Since principal component analysis (PCA) generates synthetic variables from the water quality parameter to use as inputs for predicting, instead of selecting the actual parameters, the results showed that principal component analysis (PCA) was the least predictive performance method, compared with forward selection (FS), backward elimination (BE) and genetic algorithm (GA). In mathematical aspect, PCA is another types of transformation, which affects to the predictive performance as discussed in 6.3. So this could be concluded that an advantage of PCA which is independent synthesized variables was not suitable for predicting water quality, because each parameter values are relevant and effect each other.

Selection of parameters using forward selection (FE) and backward elimination (BE) are classified as greedy algorithm, which input parameter are systematically added or removed to model, iteratively. Predictive performance was used as criteria to add/remove a parameter in each iteration without considering an overview of the prediction. So it is possible to miss some complex relationship of several parameters. For example, BOD is positively correlated with the concentration of nutrients in water (NO_3^- and PO_4^{3-}) as it demonstrates the biological degradation activity in water. However, when the temperature is not appropriate, Biodegradation activities were inhibited. These phenomena could not be detected by either FS or BE method, therefore both methods were not suitable for selection of parameters to predict the water quality.

Genetic algorithm (GA) is the global optimization method based on Charles Darwin's theory of natural evolution which was the dominant method for parameter selection, according to results in 5.5. The disadvantage of this method was longer calculation time compared with other methods. However, this study focuses on the effectiveness of prediction and when considering the actual application, supercomputer is often used for large environmental projects which could significantly shorten the calculation time, thus GA is used in the next step.

6.6 Prediction models comparison

The multiple linear regression (MLR) model was used as a benchmark model to compare with two machine learning models which were support vector regression (SVR) and artificial neural network (ANN). The results of the MLR predictive performance was significantly lower than SVR and ANN. This confirms the nonlinear relationship of water quality parameters. Therefore, it could be concluded that the MLR model is not suitable for predicting water quality, due to limitations in dealing with non-linear data.

Considering both machine learning models, it was found that water quality prediction performance was similar. Notice the standard deviation (SD) values of root mean square error (RMSE) and Spearman correlation with overlapping sections. The results table 5.11 shows the average of the predictive performance of all 16 water quality parameters. However, when considering each parameter, SVR provided some predicted parameters that close to the actual measured parameters more than ANN. In general, ANN has also been shown to be superior, so ANN was selected for using as the main structure for purposed method development.

6.7 Space and time neural network

Predicted EC value from three models (space and time neural network, time delay neural network, and distance neural network) were analyzed to clarify the predictive efficiency. Relationship between observed EC and predicted EC from three models is show in Figure 6.1. Ideally, the perfect prediction could be showed diagonal line between predicted data and real data and every point are exactly on the line. The data point of space and time neural network (STNN) is closest to the fitted line compared with other two models. A closely followed pattern of variation by real data and STNN mocel computed EC in the river water were shown in Figure 6.2. While the time delay model tents to overestimate and distance model was under estimator, this could be confirmed by Figure 6.3 which showed the histogram of prediction error.

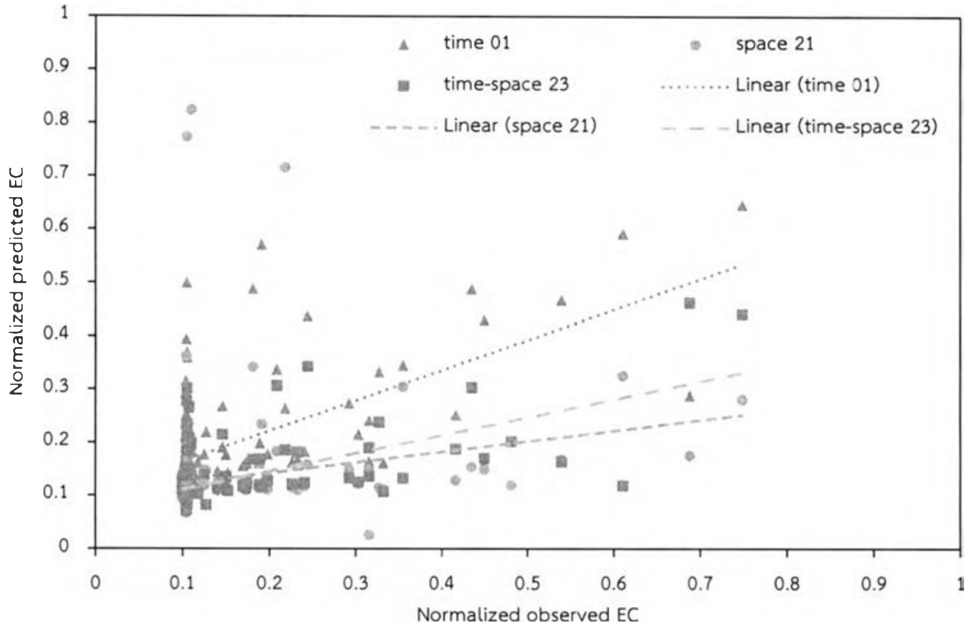

Figure 6.1 Relationship between normalized observed EC and normalized predicted EC from three models

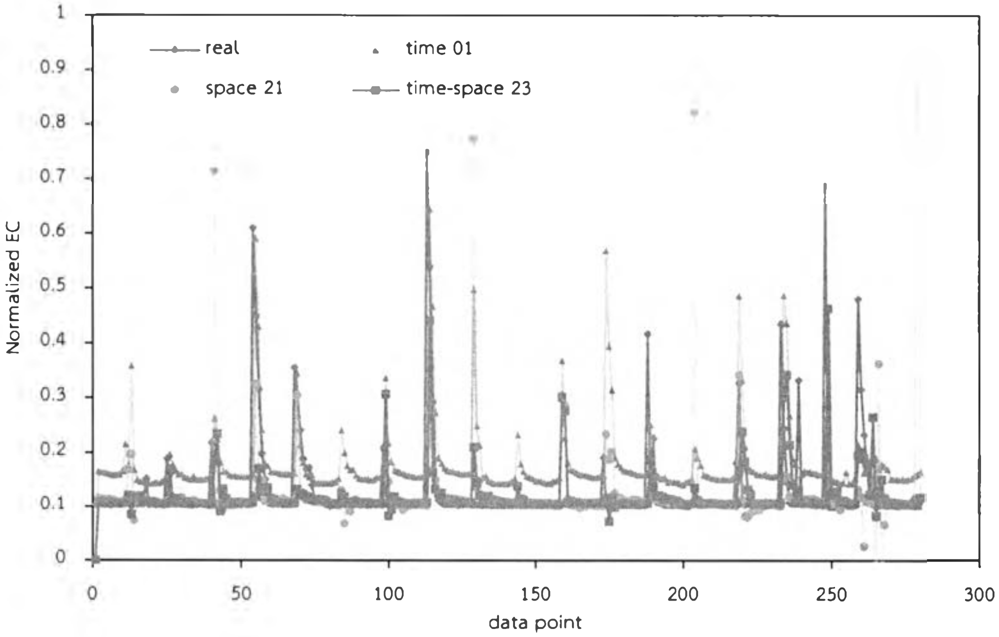

Figure 6.2 Comparisons of the models computed and measured EC in Chaophraya River during 2552 to 2556 B.E.

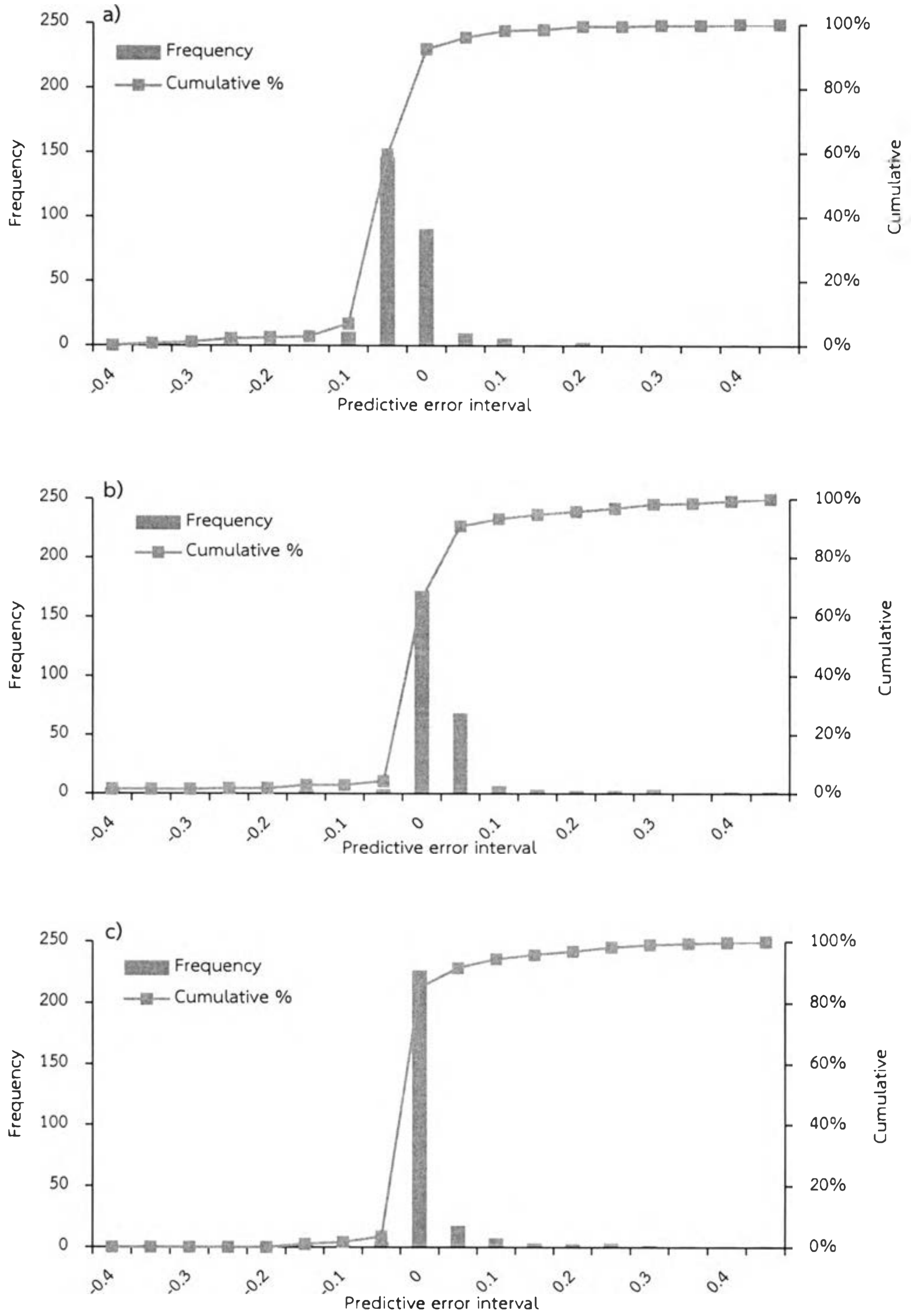

Figure 6.3 Histogram of difference in EC predicted (observed data – predicted data) from a) time delay model, b) distance model and c) space and time model.

Predicted TDS value from three models (space and time neural network, time delay neural network, and distance neural network) were analyzed to clarify the predictive efficiency. Relationship between observed TDS and predicted TDS from three models is show in Figure 6.4. The data point of space and time neural network (STNN) is closest to the fitted line compared with other two models. A closely followed pattern of variation by real data and STNN model computed TDS in the river water were shown in Figure 6.5. While time delay model and distance model tent to underestimate, this could be confirmed by Figure 6.6 which showed the histogram of prediction error.

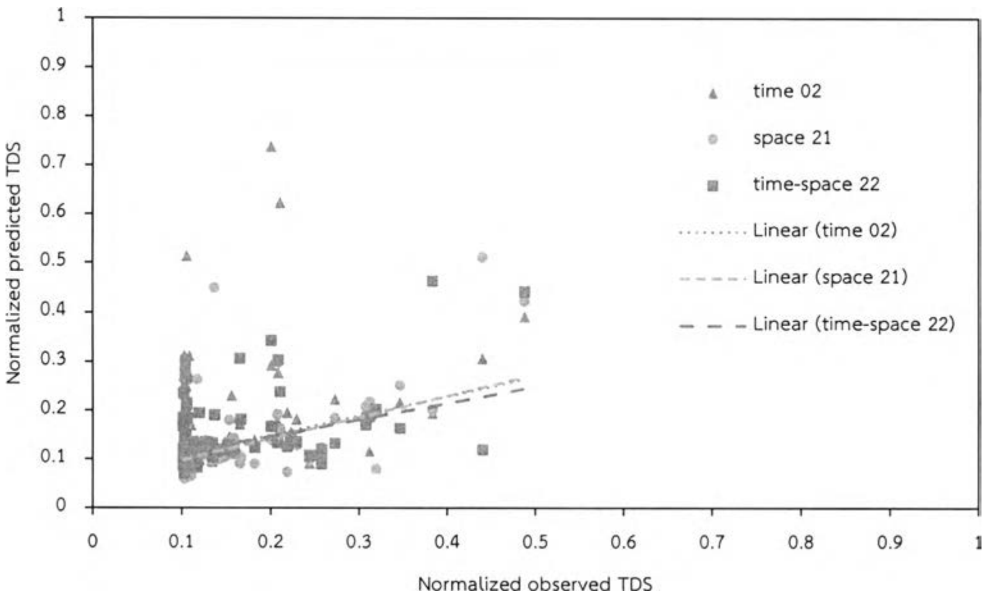

Figure 6.4 Relationship between observed TDS and predicted TDS from three models

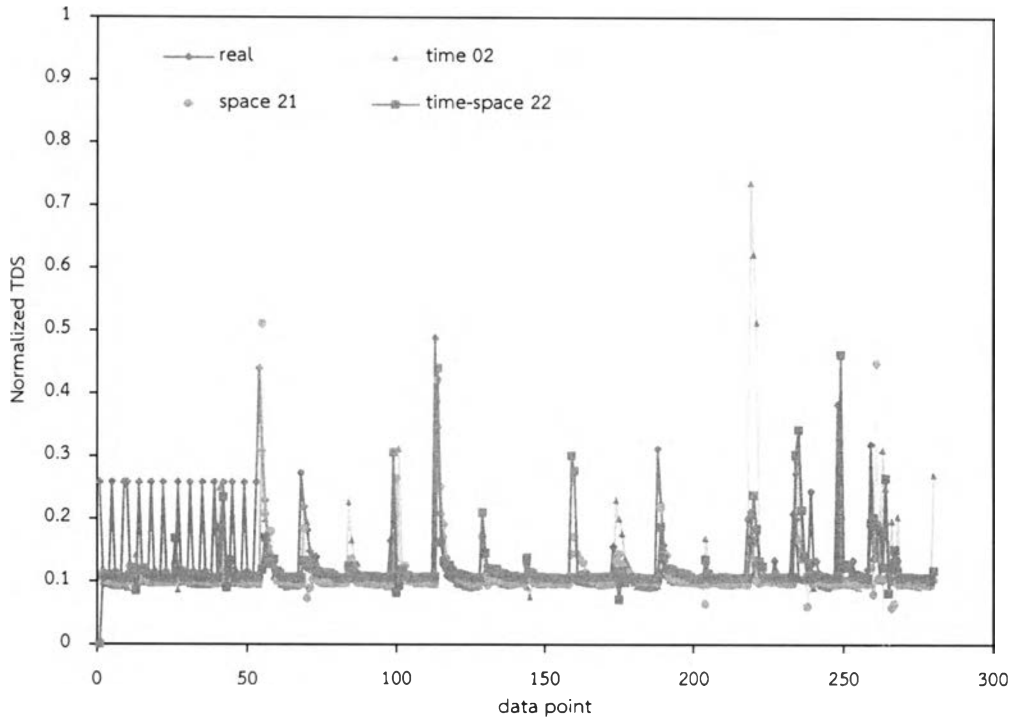

Figure 6.5 Comparisons of the models computed and measured TDS in Chaophraya River during 2552 to 2556 B.E.

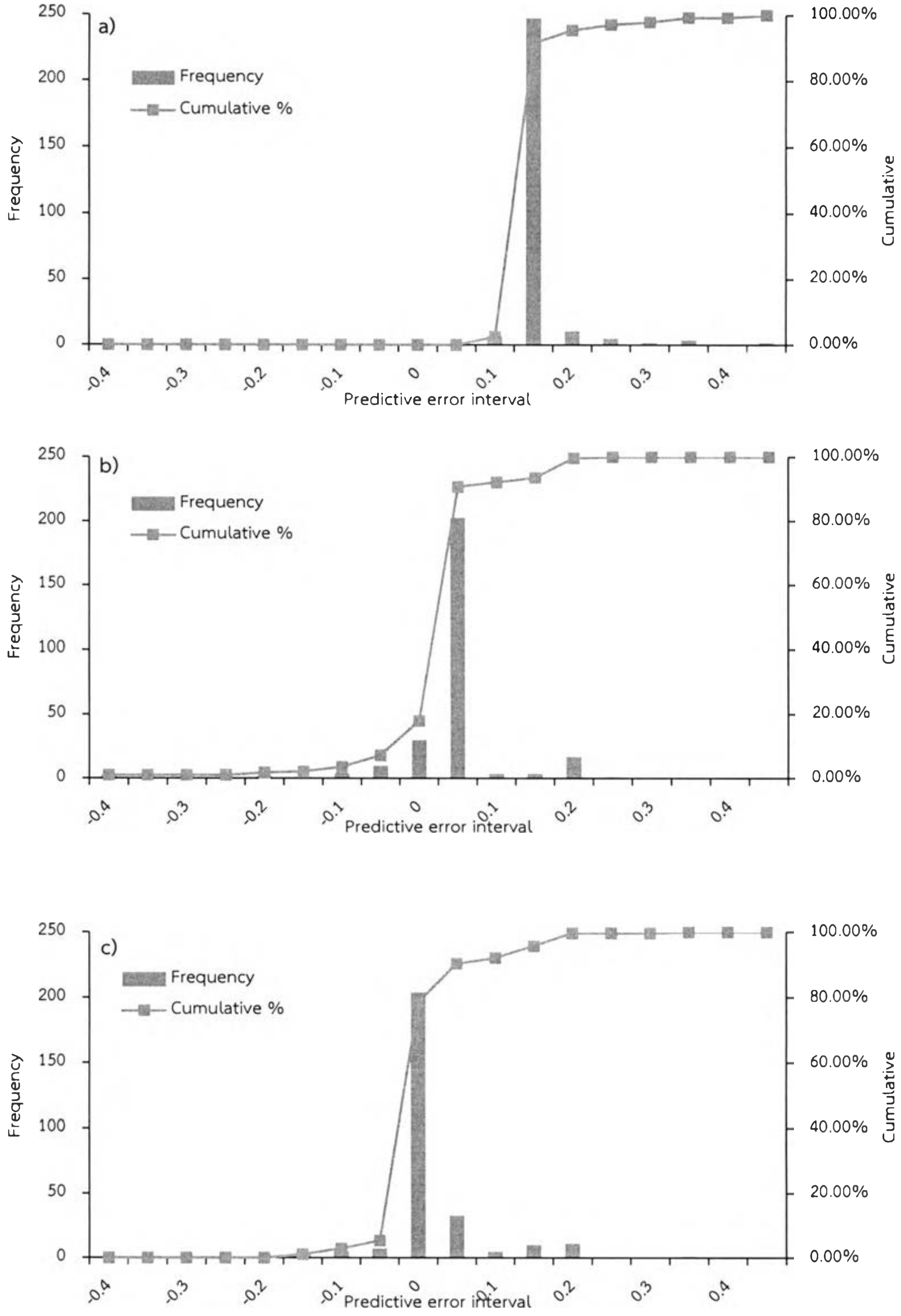

Figure 6.6 Histogram of difference in TDS predicted (observed data – predicted data) from a) time delay model, b) distance model and c) space and time model.

Predicted phosphate concentration from three models (space and time neural network, time delay neural network, and distance neural network) were analyzed to clarify the predictive efficiency. Relationship between observed phosphate and predicted phosphate from three models is show in Figure 6.7. The data point of three models were mostly cluster closed to the minimum value. A followed pattern of variation by real data and three models computed phosphate in the river water were shown in Figure 6.8. In this case, all models tent to underestimate, this could be confirmed by Figure 6.9 which showed the similar pattern of histogram of prediction error.

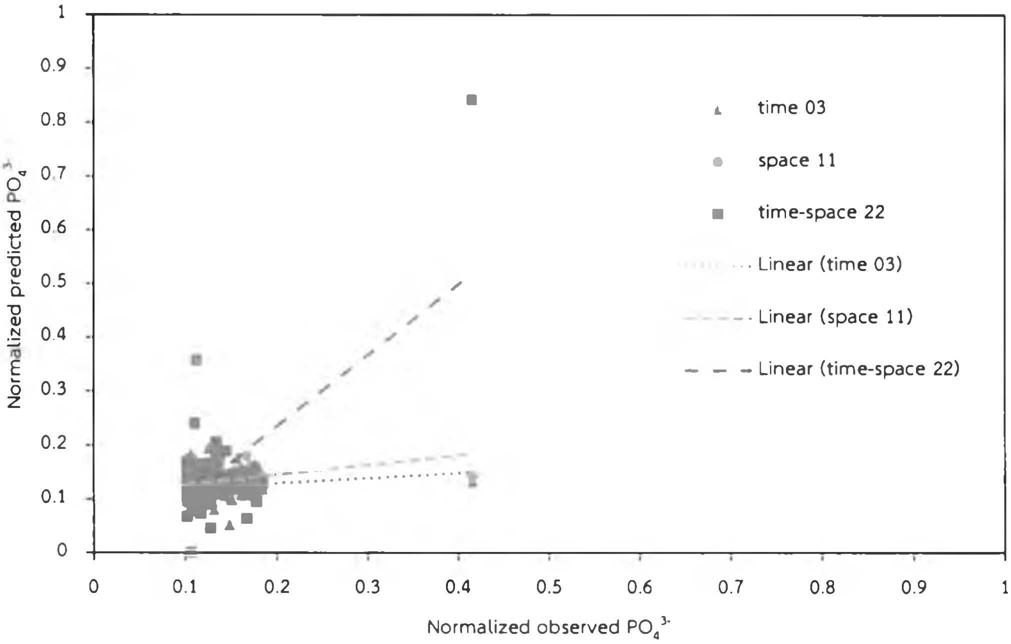

Figure 6.7 Relationship between observed PO_4^{3-} and predicted PO_4^{3-} from three models

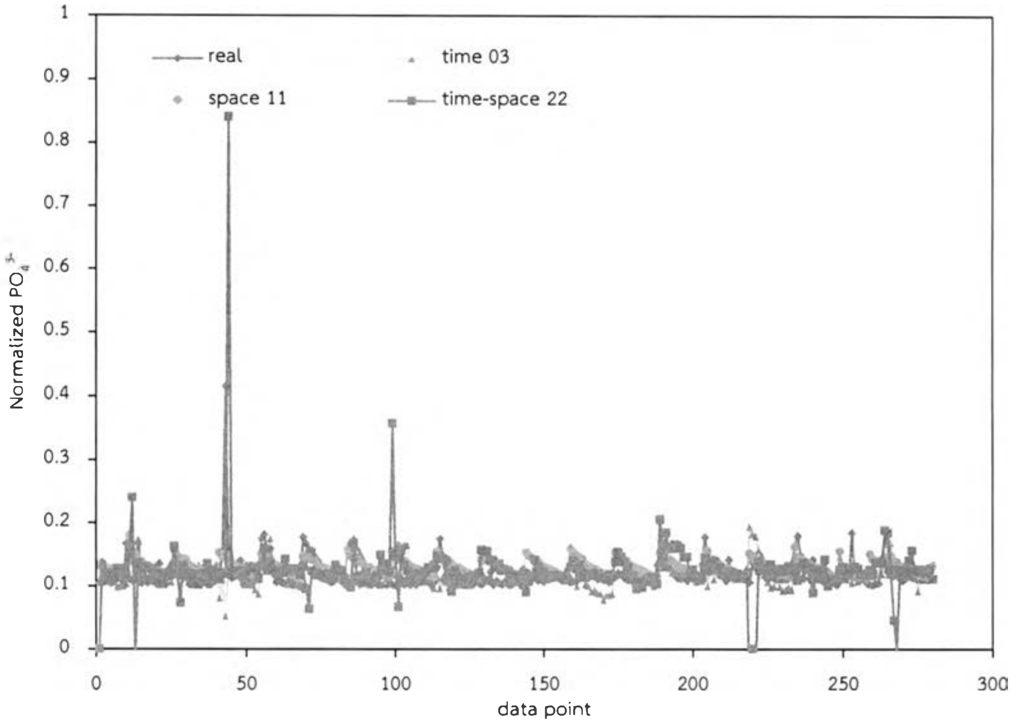

Figure 6.8 Comparisons of the models computed and measured TDS in Chaophraya River during 2552 to 2556 B.E.

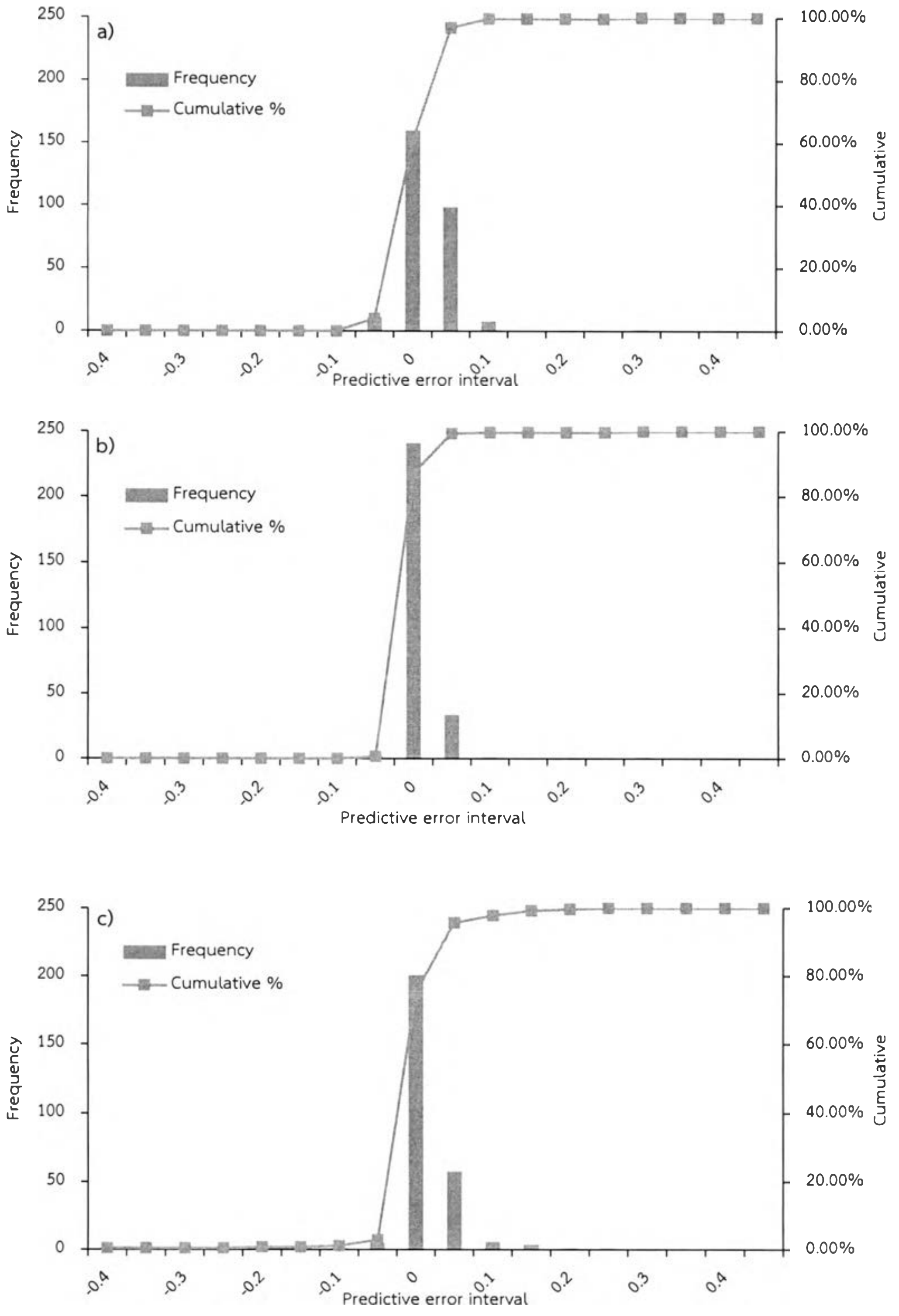

Figure 6.9 Histogram of difference in PO_4^{3-} predicted (observed data – predicted data) from a) time delay model, b) distance model and c) space and time model.

According to the important rank of the model which showed in Table 5.15 - 5.17, the same parameter could be summed up to indicated the relationship of parameter with space and time regardless. The parameter importance on EC and TDS model were shown in Table 6.2 and Table 6.3, respectively. These two parameters theoretically related to each other; this is consistent with relative importance that calculated from the purposed model. Furthermore, EC and TDS shared three common parameters in top five importance, namely, distance from sea, fecal coliform and turbidity. Distance from sea was the most importance parameter for predicting both EC and TDS, this could be interpreted that EC and TDS is strongly related to location of Chaophraya River. High negative decomposed weight of distance means EC and TDS values are increased by the flow of the river. It had the lowest value at the upstream and the highest at the downstream which might be affect from accumulative of pollution along the river or an effect from the sea water. However, this strong effect could not be clearly indicated the reason behind which need more study to determine this relationship in detail.

Table 6.2 Top five important parameters on EC model.

Importance rank	Parameter	Decomposed weight	Relative importance (%)
1	Distance	-41.40	16.69%
2	Fecal Coliform	39.13	15.78%
3	Turbidity	25.20	10.16%
4	Total Coliform	20.56	8.29%
5	TDS	18.63	7.51%

Table 6.3 Top five important parameters on TDS model.

Importance rank	Parameter	Decomposed weight	Relative importance (%)
1	Distance	-27.63	20.55%
2	NH_3	18.56	13.81%
3	EC	13.55	10.08%
4	Fecal Coliform	10.30	7.67%
5	Turbidity	9.75	7.25%

Top five important parameter on PO_4^{3-} model was shown in Tāble 6.4. Unlike EC and TDS, Fecal coliform, BOD, salinity and PO_4^{3-} (in the past or upstream) have positive effect to predicted PO_4^{3-}. Fecal coliform and BOD can indicate organic pollution in the water, which proportional related to phosphate concentration in water.

Table 6.4 Top five important parameters on PO_4^{3-} model.

Importance rank	Parameter	Decomposed weight	Relative importance (%)
1	Fecal Coliform	10.45	13.21%
2	Temperature	-8.18	10.34%
3	BOD	8.08	10.22%
4	Salinity	7.75	9.80%
5	PO_4^{3-}	7.61	9.62%

This relative importance of each parameter modelling are very useful for used as the recommendation of water management regulations. Hopefully, this dissertation is useful for the Chaophraya River quality management in future.