# CHAPTER V

## Conclusion

The dissertation showed the problem of incomplete time-series prediction by modelling the forecasting of several natural and social phenomena. The modeling consists of two main steps. The first step is to estimate the collected incomplete data. The second step is to predict the new data based on the nature of the data obtained from the first step. Our solution is to develop a new neural network model for forcasting incomplete time-series data and improve the accuracy of prediction. Two neural network models are proposed. There are FI-GEM network and RMD-FSE network. For FI-GEM network, various versions of EM-based algorithm and smoothing spline interpolation are used to preprocess the incomplete data sets. The individual networks are trained by supervised multilayer perceptron(MLP) with extended Kalman filtering [25]. The ensemble construction [24] is used for the combination of the individual networks. For RMD-FSE network, we present an approach that uses several EM-based algorithms as well as a smoothing spline interpolation and $k$-segment principal curves to fill in the missing data values. Each individual network uses a Finite Impulse Response model [26] to perform the prediction. The outputs of all individual neural networks are combined by the genetic algorithm-based selective neural network ensemble method (GASEN) [27]. We evaluated the accuracy of prediction with a performance index which measures the accuracy of prediction for the desired network with respect to the individual networks. We conducted our experiments using Mackey-Glass chaotic time-series, the annual sunspot

and the daily gauge height data collected at the Ban Luang gauging station, Mae Tun stream, Ping river, Thailand. Our results show that both of FI-GEM and RMD-FSE outperform each individual network, and both of them are proposed for highly accurate incomplete time-series prediction. Our cited reference [28] showed that our concept is pratical and be used as the real world applications.

There are some limitations of both of FI-GEM and RMD-FSE network. First, using an ensemble of networks increase the computational resources needed. Secondly, while the quality of the ensemble output is better than that of the individual networks, it can only be improved by having better individual networks. Thus, we have the further work for improving the accuracy of the imputation of incomplete data.

A new methodology (the WDC algorithm) for the pattern characterization, and the imputation of missing samples is proposed. The main idea is the time-series data are divided into separate subsequences of different sizes and, therefore, each subsequence can be viewed as a window. The imputation of missing samples is achieved by finding a complete sub- sequence similar to the missing sample subsequence and imputing the missing samples from this complete subsequence. This methodology has been applied to four cases studies such as: Mackey-Glass chaotic time-series, the sunspots data, the daily gauge height at Ban Luang gauging station, Mae Tun stream, Ping River, Thailand and the air temperature at Nakhon Ratchasima province, Thailand. We evaluated the accuracy of estimating missing values with an imputation performance index which measures the accuracy of estimating missing values for the WDC algorithm and the desired methods. Our experiments showed that the imputation accuracy of WDC algorithm can be comparable or better than the others traditional method such as: the spline interpolation, the multiple imputation (MI), and the OCSFCM algorithm. In case of the non-stationary time-series, especially the real-world problems, our results showed that WDC outperforms its competitors.

Two important issues concerning the proposed (WDC) algorithm is discussed. The first issue concerns the lengthy time for finding an appropriate partitioning window size. The second issue concerns the percentage of missing values and the appropriate size of partitioning window. The higher percentage of missing values implies the more robustness of the algorithm in terms of estimating precision is confirmed. Then, a solution for reducing time of the WDC algorithm is, an acceptable upper limit of segment length is observed from our experience.

Although those previously fill-in techniques are efficient, all of them must be based on several prior assumptions and predefined parameters such as the distribution function of data (in case of various version of EM technique), the number of centers (in case of fuzzy $C$-means), and the order of the interpolating polynomial (in case of fitting curve). A large number of trials must be conducted to obtain a set of acceptable assumptions and parameters. Since the actual nature of the data are not completely known in advance, making some assumptions on the given data is still a must. With this fact, the idea of our approach is based on only one assumption. The assumption is from the observation that nature phenomenon can repeat itself several times with similar characteristics. Hence, some missing data in a phenomenon can be imputed by searching and comparing with some other similar phenomena. This approach is appropriate for imputing the missing time-series data.

Our ongoing work includes the following. In Chapter 4, the way to reduce computational time of the proposed algorithm should be improved. There are two possible future improvements. First, the imputing missing values at various segment lengths should be done in parallel. Second is the mathematical simulation or the optimization technique should be used in finding an appropriate partitioning window size. Moreover, the on-line imputation of incomplete data for the real world application is really concerned especially the climate and hydrological applications.

Finally, the major advantage of this work is that we get a new technique of managing incomplete data and a new model for incomplete time series prediction with higher accuracy than the current prediction techniques. This work can be used in the real world applications.