CHAPTER V

PROPOSED SOLUTION

5.1 Overview of Proposed Solution

Classification of rainfall types [10, 11] is important to the accuracy and reliability of data imputation process since the sparse data or zero data value relevant to this data imputation is applied only to the data collected during the convective rainfall duration, not the data collected during the stratiform rainfall duration. Determination of "missing" or "no rain" condition can be revealed according to the classification of rainfall type-convective or stratiform rainfall; the zero value obtained in the convective rainfall duration is considered "missing", while the zero value derived in the stratiform rainfall duration is called "no rain", because the amount of convective rainfall is almost twice as much as that of stratiform rainfall. An example of convective or stratiform rainfall type is shown in Figure 5.1. How do we know convective or stratiform rainfall type from rain gauge and radar data set? The rain rate of a cloud, particularly a convective cloud, is a function of the stage of its life cycle. Research [15, 18] into precipitation estimation techniques has revealed some basic properties of convective clouds. About convective clouds, there is a high correlation between a cloud area (as measured in radar reflectivity) and the total volume of rain per unit time (as measured in instantaneous of gauge rain intensity) falling from it.

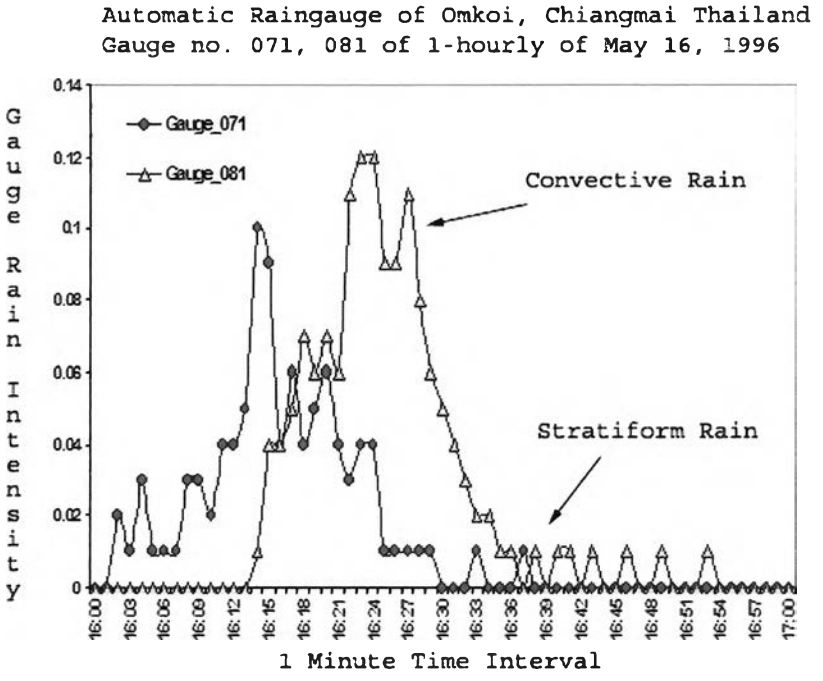

Figure 5.1: An example of a difference between convective and stratiform rain type.

Consider an example of convective rainfall type having high correlation is illustrated in Figure 5.2. This is an example of 2-hourly of radar-raingauge pairs of May 16, 1996. The parameters obtained for gauge no. 071 are a = 169, b = 1.62, correlation coefficient = 0.92 and for gauge no. 081 are a = 167, b = 1.61, correlation coefficient = 0.91.

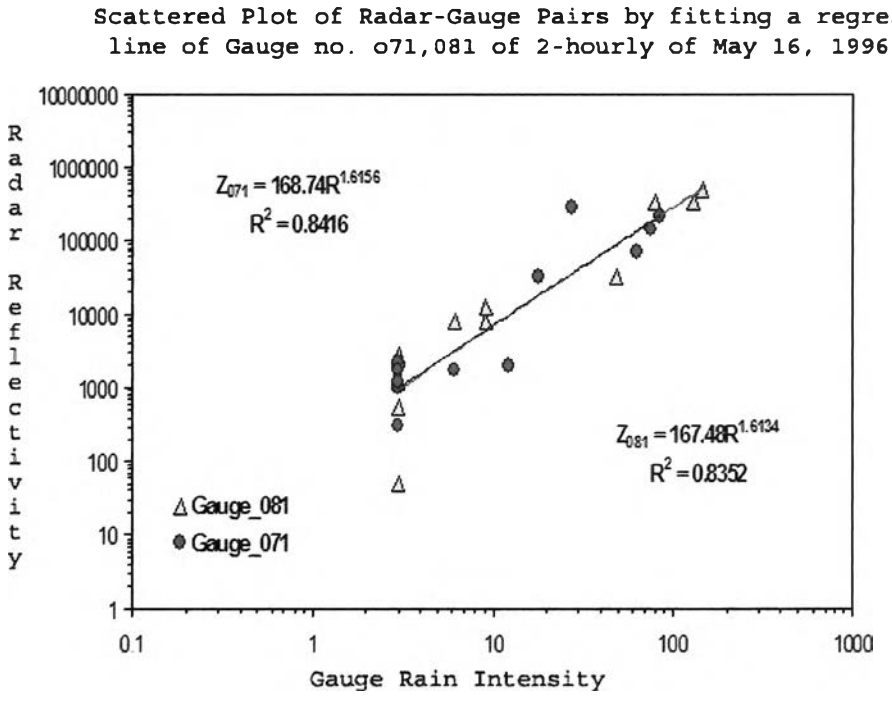

Figure 5.2: An example of convective rainfall type having high correlation.

The missing data in our study occur in both input patterns and targets, and, also, form a long missing data string scattering along the collected data sequence. This creates a complex situation and a mixture of several solutions to fill-in these data must be introduced. A new solution called similarity manifold matching is proposed and used in the first step. This step is to cut a piece of time data with some missing values and find the similarity of the manifold of these data with the manifold of the data in some other time sequences. If the exact match cannot be found then the techniques of expectation maximization (EM) and generalized reduced gradient (GRG) algorithms are applied to fill-in the missing data. The overview of our solution is the following.

Let $X_i = [x_{i,1}, x_{i,2}, ..., x_{i,n}]^T$ be an input pattern composing of features $x_{i,j}$ and d_i the corresponding target of x_i at time sequence i . There are three possible cases of missing

data. The first case is when the value of d_i is missing. The second case is when the values of some $x_{i,j}$ are missing. The third case is the combination of the first and second cases. Suppose X_i is the input pattern with some missing values. The definition of similarity measure will be given in the next subsection.

Algorithm for Estimating Missing Data

1. Let $\mathbf{X}_{i,p} = \left[X_{i-p},\ldots,X_{i+p}, d_i\right]^T$ be the feature manifold with the window size of $2p$, where p is a positive constant.
2. Slide $\mathbf{X}_{i,p}$ through the time data sequence and find $\mathbf{X}_{g,p}$ having maximum similarity, $S_{i,g}$.
3. Apply GRG algorithm to estimate the elements of $\mathbf{X}_{i,p}$ and $\mathbf{X}_{g,p}$, whose values are equal to zero, with the objective of maximizing $S_{i,g}$ by using the element from $\mathbf{X}_{i,p}$ and $\mathbf{X}_{g,p}$.
4. Apply EM algorithm once more to $\mathbf{X}_{i,p}$ and $\mathbf{X}_{g,p}$ to estimate those elements with zero values.
5. **Repeat** step 2 to 4 **until** all missing $x_{i,j}$ and d_i are estimated.

5.2 Similarity Measure

Let $\mathbf{X}_{i,p}$ and $\mathbf{X}_{k,p}$, $i \neq k$, be two feature manifolds. The similarity, $S_{i,k}$, of these two manifolds is defined as follows.

$$S_{i,k} = \frac{\sum_{l=-p}^{p}\left(\sum_{j=1}^{n} x_{i+l,j}\; x_{k+l,j} + d_{i+l}\; d_{k+l}\right)}{\sqrt{\sum_{l=-p}^{p}\left(\sum_{j=1}^{n} x_{i+l,j}^2 + d_{i+1}^2\right)}} \times \frac{1}{\sqrt{\sum_{l=-p}^{p}\left(x_{k+l}^2 + d_{k+l}^2\right)}}$$

This measure is better than Euclidean distance measure because it preserves the significance of local gradients of both manifolds when they are similar. In the case of Euclidean measure, two vectors with opposite gradients or the same gradients give the same Euclidean distance. The value of $S_{i,k}$ is in between 0 and1 .

5.3 The Expectation Maximization (EM) Algorithm

Referring to "The EM Algorithm and Extensions" textbook [2], let Y be the random vector corresponding to the observed data y , having probability density function (p.d.f.) postulated as $g(y;\Psi)$. $\Psi = (\Psi_1,...,\Psi_d)^T$ is a vector of unknown parameters with parameter space Ω . Let x denote the vector containing complete data and z denote the vector containing missing data. Let $g_c(x;\Psi)$ denote the p.d.f. of the random vector X corresponding to the complete-data vector x . Then, the complete data log likelihood function that could be formed for Ψ if x was fully observable is given by

$$\log L_c(\Psi) = \log g_c(x;\Psi).$$

Formally, we have two samples spaces χ and γ and a many-to-one mapping from χ to γ . Instead of observing the complete-data vector x in χ , we observe the incomplete-data vector $y = y(x)$ in γ . It follows that

$$g(y;\Psi) = \int_{x(y)} g_c(x;\Psi)\, dx ,$$

where $\chi(y)$ is the subset of χ determined by the equation $y = y(x)$.

The EM algorithm approaches the problem of solving the incomplete-data likelihood equation indirectly by proceeding iteratively in term of the complete-data log likelihood function, $\log L_c(\Psi)$. As it is unobservable, it is replaced by its conditional expectation given y, using the current fit for Ψ.

The EM algorithm

1. Let $\Psi^{(0)}$ be some initial value for Ψ.
2. On the first iteration, the **E-step** requires the calculation of

$$Q(\Psi; \Psi^{(0)}) = E_{\Psi^{(0)}} \{\log L_c(\Psi) | y\}.$$

3. The **M-step** requires the maximization of $Q(\Psi, \Psi^{(0)})$ with respect to Ψ over the parameter space Ω. That is, we choose $\Psi^{(1)}$ such that

$$Q(\Psi^{(1)}; \Psi^{(0)}) \geq Q(\Psi; \Psi^{(0)}) \quad \text{for all } \Psi \in \Omega.$$

4. This time with $\Psi^{(0)}$ replaced by the current fit $\Psi^{(1)}$.
5. The E-step and M-step are repeated, until the difference of $L(\Psi^{(1)}) - L(\Psi^{(0)})$ convergence.

The Generalized Reduced Gradient (GRG) Algorithm

The GRG (Generalized Reduced Gradient) method, used in the standard Excel Solver since 1990, assumes that the objective function and constraint are smooth nonlinear functions of the variables. A smooth nonlinear function has a smooth graph with no sharp. The GRG method is quite accurate and quite fast than genetic or evolutionary algorithm and yields a locally optimal solution.

We briefly summarized overall procedures of our research as follows:

1. Euclidean distance with topography is considered for selecting neighboring gauge.
2. Determination of "missing" or "no rain" condition can be modeled by classification concept and the selection of targets designed by user which "no rain" is +1 whereas "missing" is 0. Then, the accuracy was tested by Backpropagation neural networks with a data set of radar-raingauge pairs consist of 60% of training set and 40% of testing set.
3. For our proposed method, Similarity Manifold Matching (SMM), this step cut a piece of time data with some missing value and finds the similarity of the manifold of these data with the manifold of the data in some other time sequence with objective of maximizing similarity measure (SM).
4. Imputing missing data by comparison the proposed missing data estimation (SM) with the other techniques; Expectation Maximization (EM) and Neural Network (NN), two aspects; robustness and estimation technique, are considered by gradually increasing the percentage of missing data, and measuring correlation coefficient between the complete data manifold and incomplete data manifold.
5. Designing threshold value by filtering method for filling in missing of gauge data.
6. Filling-in missing data if exact match of some missing data (zero value) of two manifolds can be found – fill-in the missing data by the technique of EM, if not, fill-in the missing data by SM with generalized reduced gradient (GRG) method.
7. Perform similarity alignment, only 2-dimensional space, to solve collocation and synchronization error.
8. Achieve estimation of $Z_e - R$ relation of synchronous Z_e and gauge rain intensity (G) datasets which illustrated in both scattered plot of radar-gauge pairs by fitting a regression line and comparison of accumulated radar-gauge rainfall.