CHAPTER 3 METHODOLOGY

Gaps filling of monthly runoff data using singular spectrum analysis with a case study of the Bang Pakong river basin, which Bang Pakong river basin having 16 stations and Ban Kaeng Din So station was chosen for the case study because it has the complete data. We use idea applied missing data estimate in the basic steps of SSA:

- data of collection
- data preparing,
- check monthly runoff data and defragmenting,
- testing the efficiency of SSA method,
- filling gap of monthly runoff data for 10 cases in the Ban Kaeng Din So station.

3.1 Data of Collection

The monthly runoff data of Bang Pakong river basin was collected from the Royal Irrigation Department of Thailand (RID). There are 16 stations. After checking the availability of data only 16 stations were used in the study.

3.2 Data preparing

Preparing monthly runoff data of Ban Kaeng Din So station to test the efficiency of the gap filling missing data by choosing to study Ban Kaeng Din So station from the total of 16 stations of Bang Pakong river basin. The reason we chose Ban Kaeng Din So station because it is the only station that has completed data more than other stations of the Bang Pakong river basin. We used Ban Kaeng Din So station to prepare the data by choosing the data that are complete, which we choose the range during April 1968 to March 1993 for testing by random data and random cutting range data. The first test case is choosen cutting random data 10 case being 4%, 5%, 7%, 9%, 10%, 18%, 21%, 23%, 25% and 31. Then the second case is random cutting range of data 10 case being 4%, 5%, 7%, 9%, 10%, 18%, 21%, 23%, 25% and 31% of all data. Due to the QC data for each station, from total of 16 stations. It missing data is calculated as a percentage of the above. This study used Ban Kaeng Din So station to test the efficiency of missing data.

3.3 Check monthly Runoff data and Defragmenting

Monthly runoff data in the Bang Pakong river basin can be checked by statistical techniques: Checking missing data for each station from 16 stations of the Bang Pakong river by monthly runoff data of each station during 1941 to 2013 which are shown in Table 3.1

3.4 Testing efficiency of the SSA method

Testing efficiency of the SSA method is used to compute and consider the monthly runoff data of Ban Keang Din So station (KGT.15A) because Ban Kaeng Din So station has complete information, the 10 case reducing data and fill to gap by the SSA are tested. The main results of gap filling are compared with completed information data using Willmott's index of agreement.

Filling gaps of monthly runoff data using singular spectrum analysis: A case study of Ban Kaeng Din So station

The process and methods in this chapter are about the monthly runoff as follows:

1. Lead monthly runoff data of Ban Keang Din So station (300 months) of all the data and calculated data is anomaly

2. Defragmenting the monthly runoff data is in a row matrix, which data have 300. Assume a monthly runoff data of length is $N = 300$, $X_t = [x_1 \quad x_2 \quad x_3 \quad ... \quad x_N]_{1\times 300}$ be the original monthly runoff data and M is a window size $12 < M < 60$ to embed the initial time series.

3. Find the proper of window size by choosing $periods < M < \frac{N}{5}$, a window size of the case study data is 300 months. We used a window size $12 < M < 60$ and find proper a window size by testing Willmott's index of agreement.

4. Case study Ban Kaeng Din So station in this study design 2 cases are random data and random cutting range data from 300 months, which are shown in the Table 3.1

Table 3.1 Experiment 1 random data of Ban Kaeng Din So

Experiment 1 study	Missing data number	Missing (%)
Case1	12	4%
Case 2	15	5%
Case 3	21	7%
Case 4	27	9%
Case 5	30	10%
Case 6	54	18%
Case 7	63	21%
Case 8	69	23%
Case 9	75	25%
Case 10	93	31%

Table 3.2 Experiment 2 random cutting rage data of Ban Kaeng Din So

Experiment 2	Missing data number	Missing (%)
Case1	12	4%
Case 2	15	5%
Case 3	21	7%
Case 4	27	9%
Case 5	30	10%
Case 6	54	18%
Case 7	63	21%
Case 8	69	23%
Case 9	75	25%
Case 10	93	31%

3.5 Formatting of data

From Table 3.3 the monthly runoff data used in this study is from the Bang Pakong station from the hydrological division of the Royal Irrigation Department of Thailand, which Bang Pakong river basins has 16 stations and choosing Ban Kaeng Din So, Kabin Buri, Prachin Buri, (KGT.15A) station in a case study by consider data 300 months for each study 10 case.

The process of SSA method for estimation the monthly runoff data are as follows:

Step 1: Embedding

The study gaps filling of the monthly runoff data using this idea by the original data computing the unbiased value of the mean and set the missing data values to zero.

1. Consider the following time series of each case in Ban Kaeng Din So station by raw data anomaly stored in the vector X_t :

$$X_t = [x_1 \quad x_2 \quad x_3 \quad \ldots \quad x_{300}]_{1\times 300}$$

And transpose matrix X_t

$$X_t^T = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{300} \end{bmatrix}_{300\times 1}$$

2. Create the matrix D , which is written as follows:

$$D = \begin{bmatrix} x_1 & x_2 & \cdots & x_{38} \\ x_2 & x_3 & \cdots & x_{39} \\ \vdots & \vdots & \ddots & \vdots \\ x_{262} & x_{263} & \cdots & x_{299} \\ x_{263} & x_{264} & \cdots & x_{300} \end{bmatrix}_{263\times 38}$$

And transpose matrix D as follows:

$$D^T = \begin{bmatrix} x_1 & x_2 & \cdots & x_{262} & x_{263} \\ x_2 & x_3 & \cdots & x_{263} & x_{264} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{38} & x_{39} & \cdots & x_{299} & x_{300} \end{bmatrix}_{38\times 263}$$

Step 2: Calculating the covariance matrix C_X

$$C_X = \frac{1}{263} D^T D$$

$$C_X = \frac{1}{263}\begin{bmatrix} x_1 & x_2 & \cdots & x_{262} & x_{263} \\ x_2 & x_3 & \cdots & x_{263} & x_{264} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{38} & x_{39} & \cdots & x_{299} & x_{300} \end{bmatrix}_{38\times 263} \begin{bmatrix} x_1 & x_2 & \cdots & x_{38} \\ x_2 & x_3 & \cdots & x_{39} \\ \vdots & \vdots & \ddots & \vdots \\ x_{262} & x_{263} & \cdots & x_{299} \\ x_{263} & x_{264} & \cdots & x_{300} \end{bmatrix}_{263\times 38}$$

$$C_X = \frac{1}{263}\begin{bmatrix} c_1 & c_2 & c_3 & \cdots & c_{38} \\ c_2 & c_3 & c_4 & \cdots & c_{39} \\ c_3 & c_4 & c_5 & \cdots & c_{40} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{38} & c_{39} & c_{40} & \cdots & c_{75} \end{bmatrix}_{38\times 38}$$

Find the eigenvalue λ and eigenvector E of $M \times M$ covariance matrix C_X as follow:

$$C_X E = \lambda E$$
$$(C_X E) - (E\lambda) = \bar{0}$$
$$(C_X - \lambda I)E = \bar{0}$$
$$(C_X - L)E = \bar{0}$$

Where E is eigenvector,
λ is eigenvalue,
$\bar{0}$ is zero vector,
I is identity matrix,
L is diagonal matrix.

Find the eigenvalues and eigenvector by

$$C_X - L = 0$$

$$\frac{1}{263}\begin{bmatrix} c_1 & c_2 & c_3 & \cdots & c_{38} \\ c_2 & c_3 & c_4 & \cdots & c_{39} \\ c_3 & c_4 & c_5 & \cdots & c_{40} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{38} & c_{39} & c_{40} & \cdots & c_{75} \end{bmatrix}_{38\times 38} - \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ 0 & 0 & \lambda_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_{38} \end{bmatrix}_{38\times 38} = \bar{0}$$

$$\frac{1}{263}\begin{bmatrix} c_1 - \lambda_1 & c_2 & c_3 & \cdots & c_{38} \\ c_2 & c_3 - \lambda_2 & c_4 & \cdots & c_{39} \\ c_3 & c_4 & c_5 - \lambda_3 & \cdots & c_{40} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{38} & c_{39} & c_{40} & \cdots & c_{75} - \lambda_{38} \end{bmatrix}_{38\times 38} = \bar{0}$$

$M \times M$ covariance matrix is the system of homogenous equations, which has nontrivial obtains $\lambda_1, \lambda_2, \lambda_3, ..., \lambda_{38}$ where the determinant of this covariance matrix is zero

$$\left|C_X - L\right| = 0$$

$$\left|C_X - L\right| = \begin{vmatrix} c_1 - \lambda_1 & c_2 & c_3 & \cdots & c_{38} \\ c_2 & c_3 - \lambda_2 & c_4 & \cdots & c_{39} \\ c_3 & c_4 & c_5 - \lambda_3 & \cdots & c_{40} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{38} & c_{39} & c_{40} & \cdots & c_{75} - \lambda_{38} \end{vmatrix}_{38\times 38} = 0$$

where the eigenvalues $\lambda_1, \lambda_2, \lambda_3, ..., \lambda_{38}$ correspond to eigenvector $E_1, E_2, E_3, ..., E_{38}$

$$\begin{array}{cccc} \lambda_1 & \lambda_2 & \ldots & \lambda_{38} \\ \Downarrow & \Downarrow & \Downarrow & \Downarrow \\ E_1 & E_2 & \ldots & E_{38} \end{array}$$

A matrix E has eigenvector, which can be as follows:

$$E = \begin{bmatrix} E_1 & E_2 & E_3 & \ldots & E_{38} \end{bmatrix}$$

$$E = \begin{bmatrix} e_1 & e_2 & e_3 & \cdots & e_{38} \\ e_2 & e_3 & e_4 & \cdots & e_{39} \\ e_3 & e_4 & e_5 & \cdots & e_{40} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ e_{38} & e_{39} & e_{40} & \cdots & e_{75} \end{bmatrix}_{38\times 38}$$

where E is the matrix of eigenvectors in which each column matrix is the eigenvector of $M \times M$ covariance matrix C_X . The eigenvector of each column is extracted from the raw data anomaly.

To compute the principal components time series of each EOF by defined matrix A, matrix A is the principal components time series as follows:

$$A = D_{263\times 38} E_{38\times 263} .$$

Can be written in matrix form as:

$$A = \begin{bmatrix} x_1 & x_2 & \cdots & x_{38} \\ x_2 & x_3 & \cdots & x_{39} \\ \vdots & \vdots & \ddots & \vdots \\ x_{262} & x_{263} & \cdots & x_{299} \\ x_{263} & x_{264} & \cdots & x_{300} \end{bmatrix}_{263\times 38} \begin{bmatrix} e_1 & e_2 & e_3 & \cdots & e_{38} \\ e_2 & e_3 & e_4 & \cdots & e_{39} \\ e_3 & e_4 & e_5 & \cdots & e_{40} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ e_{38} & e_{39} & e_{40} & \cdots & e_{75} \end{bmatrix}_{38\times 38}$$

$$A=\begin{bmatrix} a_1 & a_2 & \cdots & a_{38} \\ a_2 & a_3 & \cdots & a_{39} \\ \vdots & \vdots & \ddots & \vdots \\ a_{262} & a_{263} & \cdots & a_{299} \\ a_{263} & a_{264} & \cdots & a_{300} \end{bmatrix}_{263\times 38}$$

where A is the principal components time series (PC) for each column and principal components time series of each EOF is $PC\,(PC_1, PC_2, PC_3, ..., PC_M)$.

Step 3: Calculating the reconstruction

To compute the reconstruction of missing data of the raw data anomaly in each case of Ban Kaeng Din So station, defined matrix R is the reconstruction as follows:

$$R = A_{263\times 38} E_{38\times 38}.$$

can be written in matrix form as

$$R=\begin{bmatrix} a_1 & a_2 & \cdots & a_{38} \\ a_2 & a_3 & \cdots & a_{39} \\ \vdots & \vdots & \ddots & \vdots \\ a_{262} & a_{263} & \cdots & a_{299} \\ a_{263} & a_{264} & \cdots & a_{300} \end{bmatrix}_{263\times 38} \begin{bmatrix} e_1 & e_2 & e_3 & \cdots & e_{38} \\ e_2 & e_3 & e_4 & \cdots & e_{39} \\ e_3 & e_4 & e_5 & \cdots & e_{40} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ e_{38} & e_{39} & e_{40} & \cdots & e_{75} \end{bmatrix}_{38\times 38}$$

Such as there reconstructed by using linear combinations of these principal components and EOFs, which provide the reconstructed components (RCs) the principal component time series.

Filling gap in missing data of each case study in Ban Kaeng Din So station by reconstructed matrix R and written in matrix from as

$$RC=\begin{bmatrix} r_1 & r_2 & \cdots & r_{38} \\ r_2 & r_3 & \cdots & r_{39} \\ \vdots & \vdots & \ddots & \vdots \\ r_{262} & r_{263} & \cdots & r_{299} \\ r_{263} & r_{264} & \cdots & r_{300} \end{bmatrix}_{263\times 38}$$

Find the average of each diagonal Matrix R and construct matrix R as following:

$$RC=\begin{bmatrix} r_1 & r_2 & \cdots & r_{38} \\ r_2 & r_3 & \cdots & r_{39} \\ \vdots & \vdots & \ddots & \vdots \\ r_{262} & r_{263} & \cdots & r_{299} \\ r_{263} & r_{264} & \cdots & r_{300} \end{bmatrix}_{263\times 38}$$

3.6 Efficiency to test the results by Willmott's index of agreement

For cross validation SSA method in the case study Ban Kaeng Din So station we use cross validation produced by Willmott's index of agreement for finding the optimal of window size and SSA component.