CHAPTER IV

WDC ALGORITHM

This chapter presents a pattern characterization approach for the imputation of missing samples of time-series data [23]. 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 subsequence similar to the missing sample subsequence and imputing the missing samples from this complete subsequence. 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, are used for evaluating our approach. The experimental results showed that the imputation accuracy of the proposed algorithm, namely varied window clustering (WDC) algorithm is comparable or better than the others traditional methods such as: the spline interpolation, the multiple imputation (MI), and the optimal completion strategy fuzzy c-means algorithm (OCSFCM) in case of the non-stationary time-series data.

4.1 Classifying the Characteristic of Incomplete Time-Series Data

The characteristic of incomplete time-series data can be described as a series of patterns. Supposing that a time-series is periodic. The periodical time-series data may be seasonal or cyclical. We denoted K to be the cycle of the time-series data. The data are

partitioned and arranged in a time-series form with a window size of K. Each segment of data is, then, stacked to form a matrix A as shown in equation (4.1).

$$A = \begin{bmatrix} x_1 & x_2 & \cdots & x_K \\ x_{K+1} & x_{K+2} & \cdots & x_{2K} \\ x_{2K+1} & x_{2K+2} & \cdots & x_{3K} \\ \cdots & \cdots & \cdots & \cdots \\ x_{N-K+1} & x_{N-K+2} & \cdots & x_N \end{bmatrix} \tag{4.1}$$

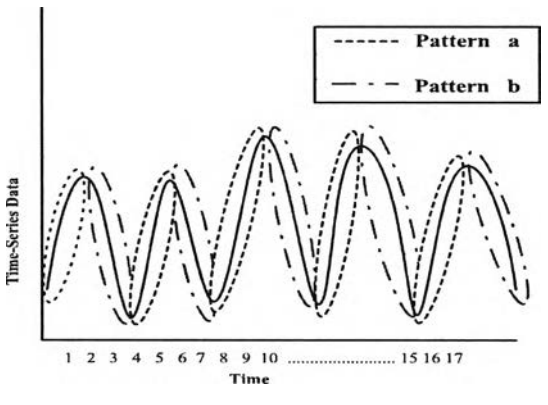


Figure 4.1: The figure shows the two groups of time-series data. The directions of subsequences of data in each circle of pattern a are similar. It is also true in each of pattern b.

The time-series data are plotted and illustrated in Figure 4.1. The data can be characterized as pattern a and pattern b. The gradient of pattern a is positive while the gradient of pattern b is negative. Notice that the directions of all the data points in each circle of pattern a are similar. It is also true in case of pattern b. All subsequences of length K are lined up in forms of a matrix as shown in equation 4.2. Each row of matrix A is of either pattern a or b.

$$A = \begin{array}{c} \begin{matrix} \textit{Data Format} & & & \end{matrix} \\ \begin{bmatrix} x_1 & x_2 & \cdots & x_K \\ x_{K+1} & x_{K+2} & \cdots & x_{2K} \\ & & \vdots & \\ x_{2K+1} & x_{2K+2} & \cdots & x_{3K} \\ x_{3K+1} & x_{3K+2} & \cdots & x_{4K} \\ & & \vdots & \\ x_{N-K+1} & x_{N-K+2} & \cdots & x_N \end{bmatrix} \end{array} \begin{array}{l} \textit{Pattern Group} \\ \leftarrow a \\ \leftarrow b \\ \vdots \\ \leftarrow a \\ \leftarrow b \\ \vdots \\ \leftarrow b \end{array} \tag{4.2}$$

It is obvious that if there are some missing x_m then x_m must be in either pattern a or pattern b. Suppose x_m belongs to a subsequence, $\mathbf{v}_\tau$, of pattern a. This subsequence $\mathbf{v}_\tau$ will be called *target subsequence*. The other subsequences besides the target subsequence will be called *reference subsequence*. The concept of imputing the value of x_m is to compare the target subsequence $\mathbf{v}_\tau$ with all reference subsequences of pattern a in the time-series data and to select the reference subsequence, $\mathbf{v}_q$, having the most similarity with the target subsequence $\mathbf{v}_\tau$. The value of x_m is, then, imputed from a subset of x_i in reference subsequence $\mathbf{v}_q$. For example, *Dat* is a periodical time-series data, whose segment size is four.

$$Dat = \{x_1, x_2, x_3, ..., x_{24}\} \tag{4.3}$$

$$A = \begin{array}{c} \begin{matrix} \textit{Data Format} & & & \end{matrix} \\ \begin{bmatrix} x_1 & x_2 & x_3 & ? \\ x_5 & x_6 & ? & x_8 \\ x_9 & x_{10} & x_{11} & ? \\ x_3 & x_{14} & x_{15} & x_{16} \\ x_{17} & x_{18} & x_{19} & x_{20} \\ x_{21} & x_{22} & x_{23} & x_{24} \end{bmatrix} \end{array} \begin{array}{l} \textit{Pattern Group} \\ \leftarrow a \\ \leftarrow b \\ \leftarrow a \\ \leftarrow b \\ \leftarrow a \\ \leftarrow b \end{array} \tag{4.4}$$

Suppose that x_4, x_7 and x_{12} are missing and are denoted by the symbol "?". The segmentation of *Dat* is formed by equation (4.2) and is represented as equation (4.4). Each row is considered as a subsequence. There are six subsequences. Assume that

subsequence 1 is of pattern a, subsequence 2 is of pattern b, subsequence 3 is of pattern a, subsequence 4 is of pattern b, subsequence 5 is of pattern a, and subsequence 6 is of pattern b. In equation 4.4, the missing x_4 and x_{12} appear in pattern a and missing x_7 also appears in pattern b. Assuming that each period has four data. Thus, the pattern data $\{x_1x_2x_3?\}$ in row 1 and the pattern data $\{x_9x_{10}x_{11}?\}$ in row 3 periodically appears again in row 5 as pattern data $\{x_{17}x_{18}x_{19}x_{20}\}$. Based on this observation, the missing x_4 in row 1 and x_{12} in row 3 can be estimated by using the data x_{20} in row 5 and $x7$ in row 2 can be estimated by using x_{15} in row 4 or x_{23} in row 6. In general, the real world time-series data are non-periodic and non-stationary. Fixing the value of length K to a constant may not be suitable in this situation.

4.1.1 Pattern Characterization with Varied Window Sizes

The important concept of our work is to find the proper length, K, of the subsequences that gives the maximum similarity between the target and reference subsequences. The hypothesis is that time-series data that are manifestations of natural phenomena often contain cycles within the series. Although all possible values of K must be in $[1, N]$, it can be quite difficult to determine an appropriate length K, especially when some of the values are missing. We measure the similarity correlation in terms of cosine, denoted by δ, between $\mathbf{v}_\tau$ and the other reference subsequences. The correlation value between the two vector representations of the two subsequences is in the range $[-1, 1]$. When $\delta < 0$, the two subsequences are partially in opposite phases. We would have to correct a reference subsequence before using it for estimating a missing value. Alternatively, we could discard these subsequences by setting their similarity values to zero. In our implementations, the sample values are normalized to the range $[0, 1]$, so that the correlation value is in the range $[0, 1]$. Accordingly, we define the distance between two

subsequences, β, in terms of δ as

$$\beta = \begin{cases} 1 - \delta & \text{if } \delta \geq 0 \\ 1 & \text{otherwise.} \end{cases} \tag{4.5}$$

We illustrate the effect of the choice of K in Figure 4.2 by a scatter diagram of the minimum similarity distance β as a function of the subsequence length. The β value increases and falls sharply and has higher variance at the longer subsequence length. Finding a good subsequence length by trying all possible lengths is computationally costly and time consuming. Some statistical analysis and experiments must be investigated to determine a feasible subsequence length.

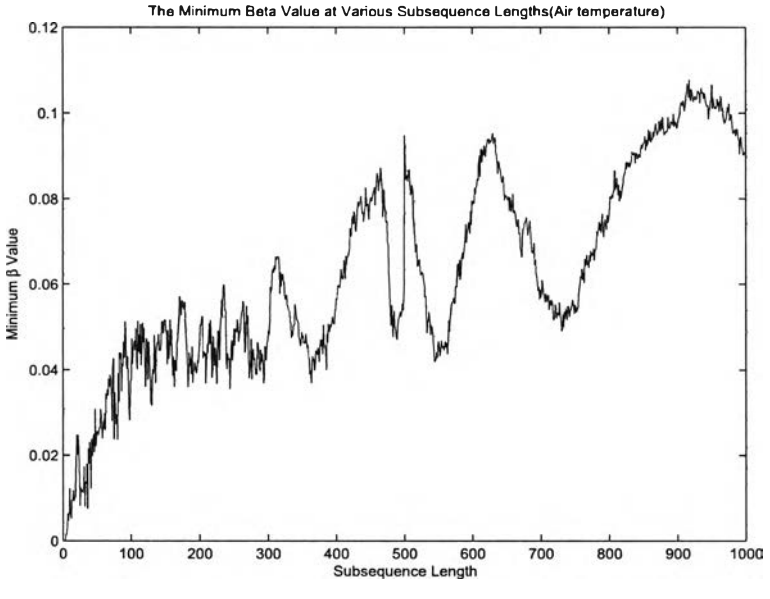


Figure 4.2: The scatter plot of the minimum β values at various fixed subsequence lengths of the *air temperature* data.

4.1.2 The Imputation of Missing Value

Let $x_m^{(\mathbf{v}_\tau)}$ be the considered missing value at time m of the target subsequence $\mathbf{v}_\tau$ and $x_m^{(\mathbf{v}_q)}$ the value at time m of the reference subsequence $\mathbf{v}_q$. The subsequences $\mathbf{v}_\tau$ and $\mathbf{v}_q$

are considered to have a very similar shape. Imputing the value of missing $x_m^{(\mathbf{v}_\tau)}$ with respect to $x_m^{(\mathbf{v}_q)}$ involves four neighboring values $x_{m-1}^{(\mathbf{v}_\tau)}$, $x_{m+1}^{(\mathbf{v}_\tau)}$, $x_{m-1}^{(\mathbf{v}_q)}$, and $x_{m+1}^{(\mathbf{v}_q)}$. Any two adjacent values of x_t and x_{t+1}, for any t, of any subsequence are assumed to be connected by a straight line. This assumption is valid by the following observation. If the limit of the difference between two adjacent times t and $t+1$ approaches zero and two adjacent values x_t and x_{t+1}, for any t, is connected by a straight line then all these piecewise linear lines form a continuous curve. Hence, the value of $x_m^{(\mathbf{v}_\tau)}$ can be computed from the value of either $x_{m-1}^{(\mathbf{v}_\tau)}$ or $x_{m+1}^{(\mathbf{v}_\tau)}$, and the difference between $\mathbf{v}_\tau$ and $\mathbf{v}_q$.

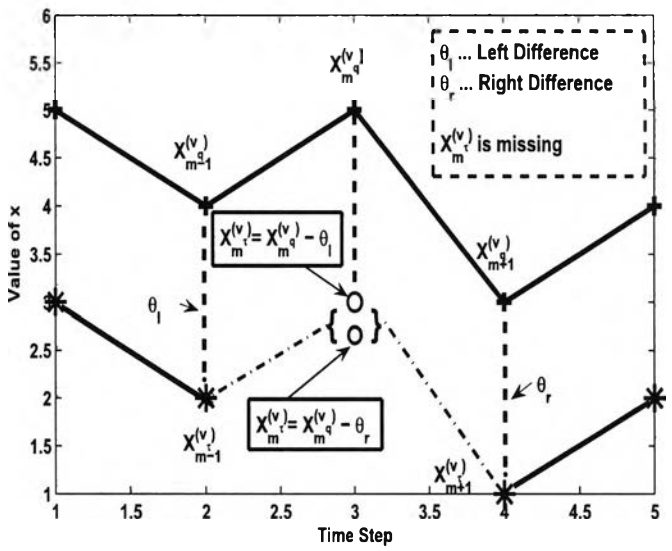


Figure 4.3: The subsequences $\mathbf{v}_\tau$ and $\mathbf{v}_q$ have a very similar shape. Since the value of $x_m^{(\mathbf{v}_\tau)}$ is missing, both left and right differences are used to impute the value of $x_m^{(\mathbf{v}_\tau)}$. The circle denotes the missing value.

From Figure 4.3, the difference between $x_{m-1}^{(\mathbf{v}_q)}$ and $x_{m-1}^{(\mathbf{v}_\tau)}$ is called *left difference*, denoted θ_l:

$$\theta_l = x_{m-1}^{(\mathbf{v}_q)} - x_{m-1}^{(\mathbf{v}_\tau)} \tag{4.6}$$

and the difference between $x_{m+1}^{(\mathbf{v}_q)}$ and $x_{m+1}^{(\mathbf{v}_\tau)}$ is called *right difference*, denoted θ_r:

$$\theta_r = x_{m+1}^{(\mathbf{v}_q)} - x_{m+1}^{(\mathbf{v}_\tau)}. \tag{4.7}$$

When the value of $x_m^{(\mathbf{v}_\tau)}$ is missing, both left and right differences are used to impute the value of $x_m^{(\mathbf{v}_\tau)}$. The estimated value of $x_m^{(\mathbf{v}_\tau)}$, denoted by $\hat{x}_m^{(\mathbf{v}_\tau)}$, can be computed from θ_l, θ_r, and $x_m^{(\mathbf{v}_q)}$ as a linear combination:

$$2\hat{x}_m^{(\mathbf{v}_\tau)} = 2x_m^{(\mathbf{v}_q)} - \theta_l - \theta_r. \tag{4.8}$$

From equations (4.6), (4.7), and (4.8),

$$\hat{x}_m^{(\mathbf{v}_\tau)} = \frac{1}{2}\left(2x_m^{(\mathbf{v}_q)} - \left(x_{m-1}^{(\mathbf{v}_q)} - x_{m-1}^{(\mathbf{v}_\tau)}\right) - \left(x_{m+1}^{(\mathbf{v}_q)} - x_{m+1}^{(\mathbf{v}_\tau)}\right)\right), \tag{4.9}$$

so that

$$\hat{x}_m^{(\mathbf{v}_\tau)} = x_m^{(\mathbf{v}_q)} - \frac{1}{2}\left(x_{m-1}^{(\mathbf{v}_q)} - x_{m-1}^{(\mathbf{v}_\tau)}\right) - \frac{1}{2}\left(x_{m+1}^{(\mathbf{v}_q)} - x_{m+1}^{(\mathbf{v}_\tau)}\right). \tag{4.10}$$

The resulting equation is very similar to the classical k-NN imputation. The difference is that the second and third terms of the right hand side of equation (4.10) are the correction terms based on the shape of subsequence. A comparison of the imputing values between the cubic spline interpolation and the linear interpolation from the similar subsequence $\mathbf{v}_q$ is shown in Figure 4.4. The result of the proposed linear interpolation from pattern characterization can be seen to be better than that of the cubic spline interpolation.

4.2 Proposed Algorithm Based on Similarity Measure

The concept of our proposed algorithm is as follows. First, all the given values x_i, for $1 \leq i \leq N$, are orderly partitioned into groups of equal size K. The value of K can be viewed as the width of the partitioning window. These groups form subsequences

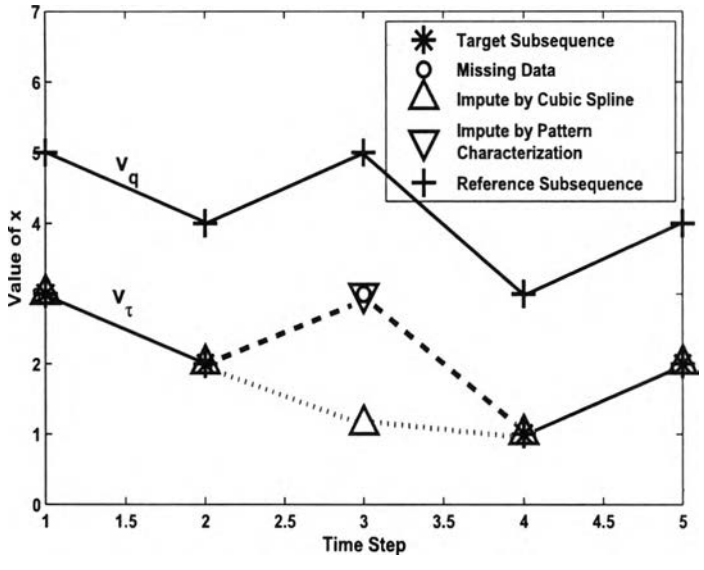
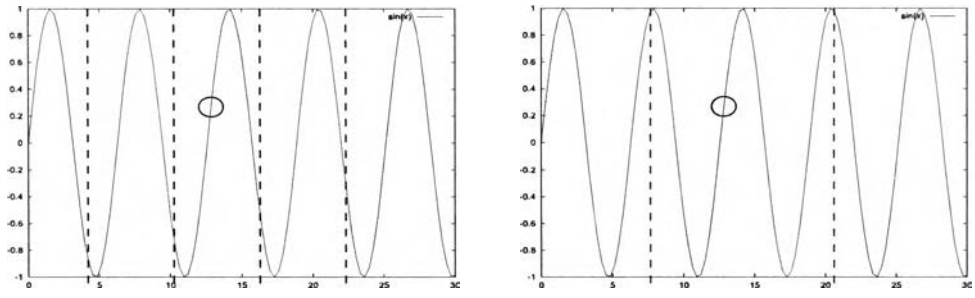


Figure 4.4: A comparison of imputing values between the cubic spline interpolation and the linear interpolation from the similar subsequence $\mathbf{v}_q$. The result of the proposed linear interpolation from characterization pattern is better than that of the cubic spline interpolation.

including the target subsequence, $\mathbf{v}_\tau$, and reference subsequences. Then, a reference subsequence $\mathbf{v}_q$ is selected with the maximum similarity to the target subsequence $\mathbf{v}_\tau$. Once a subsequence $\mathbf{v}_q$ based on a fixed value of K is selected, the missing value of x_m is estimated. For different values of K, the corresponding estimated values of x_m are computed. Finally, all the estimated values of x_m are averaged and set as the final estimated value of x_m. Figure 4.5 shows an example of the similarity measure with different window partitions. The missing value, x_m, is denoted by a circle. The width of the target subsequences in Figures 4.5 (a) and (b) are different. The dashed lines indicate the partitioning locations.

The detail of the proposed algorithm (Appendix B) is given in the following steps. Let M be the set of all index values m of the missing values x_m.



(a) (b)

Figure 4.5: An example of the similarity measure and different window partitions. The circle denotes the missing value. (a) Partitioning with a width of K. (b) Partitioning with a width of $K + c$, for some constant c.

Proposed Algorithm

1. All the time-series values are normalized to values in the range [0,1].
   Initialize all missing values x_m, $m \in M$, to random values in the range [0,1].
2. Let c be a counter variable and initialize it to 0.
3. Let α be a positive threshold variable and set it close to 0.
4. **Repeat**
5. **For** all missing values x_m, $m \in M$.
6. Initialize the width of partitioning window K.
7. Let $\hat{x}_m$ be the estimated value of x_m. Initialize $\hat{x}_m$ to 0.
8. Let g be a counter variable and initialize it to 0.
9. Let *temp* be a variable and initialize it to 0.
10. Let *sum* be a variable that is initialized to 0.
11. **Repeat**
12. Create a target subsequence, $\mathbf{v}_\tau$, by including all values from $x_{m-\lfloor \frac{K}{2} \rfloor}$ to $x_{m+\lfloor \frac{K}{2} \rfloor}$
13. Partition from $x_{m-\lfloor \frac{K}{2} \rfloor+1}$ down to x_1 into groups of equal size of K.
14. Partition from $x_{m+\lfloor \frac{K}{2} \rfloor+1}$ up to x_N into groups of equal size of K.

15. Find a group, P, of reference subsequences, $\mathbf{v}_q$, with the least distance (β_j).
16. For each subsequence j in P, set a value of $\hat{x}_m^{(j)}$ to the imputed values of x_m computed from $\mathbf{v}_\tau$ and all subsequence in group P using equation(10).
17. Set $temp = \sum_j \left((1-\beta_j) \times \hat{x}_m^{(j)}\right) / \sum_j (1-\beta_j)$.
18. Compute $sum = sum + temp$.
19. Count the number of iterations by setting $g = g + 1$.
20. Increase the size of the partitioning window by setting $K = K + 1$.
21. **Until** the number of groups is equal to one.
22. Compute $\hat{x}_m = \frac{sum}{g}$.
23. Set $x_m = \hat{x}_m$
24. **EndFor**
25. Count the number of cycle by setting $c = c + 1$.
26. **Until** the maximum of all x_m, such that $\left|\hat{x}_m^{(c+1)} - \hat{x}_m^{(c)}\right| \leq \alpha$

4.3 Experimental Results and Discussion

The four time-series data sets (see Appendix 5.1), (a) the synthetic Mackey-Glass chaotic time-series data, (b) the monthly sunspots data from A.D. 1700 to A.D. 1994, (c) the daily gauge height at Ban Luang gauging station, Mae Tun stream, Ping river, Thailand, and (d) the daily air temperature at Nakhon Ratchasima province, Thailand, are used in our experiments. The original data are normalized in the range [0,1]. The missing at random (MAR) incomplete time series are created by randomly sampling the missing time steps from the four data sets. The levels of missing values considered in both of Mackey-Glass chaotic time-series and the air temperature data are set to 10%, 20%, 30%, 40%, 50%, 60%, and 70%. For monthly sunspots data and daily gauge height data,

the levels of missing are set to 10%, 20%, 30%, 40%, and 50%. At each level of missing, the experiments are repeated 10 runs by randomly selecting the missing locations run by run. The results of our experiment are explained in the following subsections. The starting missing values of those four case studies before WDC algorithm process are initialized with randomly values. K of those four case studies are initialized to 3. For Mackey-Glass data, α is set to 0.00006. α is set to 0.0008 for the sunspots, gauge height data and the air temperature data.

4.3.1 Mackey-Glass Chaotic Time-Series

This data set is selected because its behavior is almost periodic and almost stationary. Each data point can be mathematically generated with a constant variance. There are a total of 1,200 observations in our experiment. We can see from Table 4.1 that when the levels of missing are above 50%, the WDC algorithm yields the lowest MSE among all tested algorithms. The CORR value between missing value and the actual value is also shown in Table 4.1. We can see that the WDC algorithm yields the highest CORR at level of missing at 60% and 70%. The WDC algorithm also shows the highest CORR at the higher level of missing.

In Table 4.5, the simulation results show that, on the average, the minimum and maximum P_{Imp} are 1.55×10^{-5} and 3442.19×10^{-6}, respectively. All P_{Imp} values are also greater than zero. In case of Mackey-Glass chaotic time-series data, those results confirm that, on the average, WDC algorithm gives the better prediction performance than those from the spline, MI method and OCSFCM algorithm.

A scatter plot of the true values and the estimated values of missing at 60% missing are shown in Figure 4.8. The results of WDC algorithm are comparable to the spline interpolation. The MSE of the reconstruction by the four methods at different levels of

missing are shown in Figure 4.6(a). Because of the stationarity of the data, the OCSFCM method has the worst performance while the other three methods are comparable to each other at all levels of missing.

4.3.2 The monthly sunspots

This series represents the number of sunspots which has been recorded from the surface of the sun. The set of data is selected because it is a real world case study with a periodic behavior. Its variances are not stable in each period and the signal is more complex than the Mackey-Glass chaotic time series. There are a total of 1,070 observations used in our experiments.

In Table 4.2, we noticed that the WDC algorithm yields the lowest MSE and the highest CORR at every level of missing. This show WDC algorithm gives the best estimation of missing performance.

Further, in Table 4.5, the simulation results show that the minimum and maximum P_{Imp} are 2.09×10^{-3} and 8.07×10^{-3}, respectively. All P_{Imp} values are also greater than zero. Hence, WDC algorithm exhibit better performance than the spline, MI method and OCSFCM algorithm in case of sunspots data.

A scatter plot of the true value and the estimated value of missing at 30% missing are shown in Figure 4.9. The results of WDC algorithm show the best prediction of missing values. The MSE of the reconstruction by the four methods at different levels of missing are shown in Figure 4.6(b). The WDC algorithm has the best performance among all methods at all levels of missing.

4.3.3 The daily gauge height at Ban Luang gauging station, Mae Tun stream, Ping river, Thailand

The samples of the scatter plot of the true value and the estimated value of missing at 50% missing are shown in Figure 4.10. The results of WDC algorithm show the best prediction missing values.

This univariate time-series data is made available to us by the Royal Irrigation Department of Thailand. It is selected because it is less structured than either the Mackey-Glass chaotic time series or the monthly sunspots data. There are some high peaks at some time steps and there are some parts of data that decrease gradually. There are a total of 2,000 observations used in our experiments.

We can note from Table 4.3 that WDC algorithm yields the lowest MSE for every level of missing. This show WDC algorithm gives the better estimation of missing performance than the spline interpolation, the MI algorithm and OCSFCM algorithm. The WDC algorithm gives the highest CORR at every level of missing. Thus, the WDC algorithm gives the best prediction missing value.

In Table 4.5, the experimental results show that the minimum and maximum P_{Imp} are 8.99×10^{-4} and 45.48×10^{-4}, respectively. All P_{Imp} values are also greater than zero. Accordingly, for the gauge height data, the WDC algorithm gives the best imputation performance of missing.

A scatter plot of the true value and the estimated value of missing at 50% missing are shown in Figure 4.10. The results of WDC algorithm show the best prediction missing values. The MSE of the reconstruction by the four methods at different levels of missing are shown in Figure 4.7(a). The WDC algorithm has the best performance among all methods at all levels of missing.

4.3.4 The daily air temperature at Nakhon Ratchasima province, Thailand

This real-world data set is provided to us by the Meteorological Department of Thailand. It presents the most difficult problem in our case studies because of the sharp rises and falls in the series. We used a total of 2,000 observations in our experiments.

In Table 4.4, we see that the WDC algorithm yields the lowest MSE for every level of missing. This shows that WDC algorithm gives the better estimation of missing performance than the spline interpolation, the MI method and OCSFCM algorithm. The WDC algorithm gives the best prediction missing value, as indicated by the highest CORR of the WDC algorithm at every level of missing.

In Table 4.5, the experimental results show that the minimum and maximum P_{Imp} are 1.41×10^{-3} and 8.53×10^{-3}, respectively. All P_{Imp} values are also greater than zero. For the air temperature data, these results signify that WDC algorithm also gives the best prediction missing value performance.

A scatter plot of the true value and the estimated value of missing at 50% missing are shown in Figure 4.11. The WDC algorithm also gives the best prediction missing value. The MSE of the reconstruction by the four methods at different levels of missing are shown in Figure 4.7(b). The WDC algorithm has the best performance among all methods at all levels of missing.

4.4 Discussion

4.4.1 Appropriate Partitioning Window Size

We see from the previous subsections that the WDC algorithm outperforms other estimation methods in the four test cases. The limitation of the WDC algorithm is the

computation time required. The proposed algorithm may take a long time to impute all missing x_m when the amount of given data is large. This is because it must try all possible partitioning window sizes to find the reference subsequence having maximum similarity to the target subsequence.

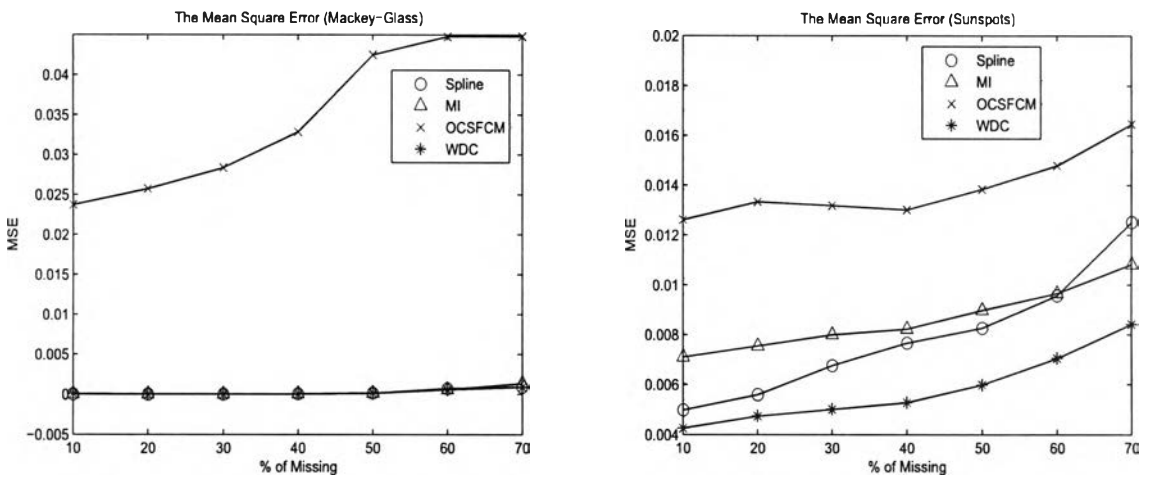
Then, a solution for reducing time of the WDC algorithm is, an acceptable upper limit of segment length is found from our experience. In our experiment, we tested with the partitioning window size from 3 through 1000 for the gauge height data and the air temperature, 3 through 600 for the mackey-glass data, and 3 through 535 for the sunspot data. In fact, the actual values of the missing values do not exist. The estimated missing values by spline interpolation are initially used as the virtual reference values for computing the reference MSE value. From our experiments, the reference MSE value at each window size of the four case studies: Mackey-Glass chaotic time-series, the monthly sunspots data, the gauge height data and the air temperature are shown in Figure 4.13 - Figure 4.20, Figure 4.21 - Figure 4.28, Figure 4.29 - Figure 4.36, and Figure 4.37 - Figure 4.44 respectively. The appropriate range of partitioning window size of the four case studies: Mackey-Glass chaotic time-series, the monthly sunspots data, the gauge height data and the air temperature are observed from those Figure 4.13 - Figure 4.44 and summarize in Table 4.6.

The Number of Most-Likely Subsequences

The number of most-likely subsequences of the four problem cases are also estimated by using the virtual referenced MSE value of spline interpolation. The appropriate numbers of most-likely subsequences of those case studies can be determined from Figures 4.12. From our experiment, the appropriate numbers of most-likely subsequences of four case studies: Mackey-Glass chaotic time-series, the monthly sunspots data, the gauge height data and the air temperature are 14, 3, 11 and 19, respectively.

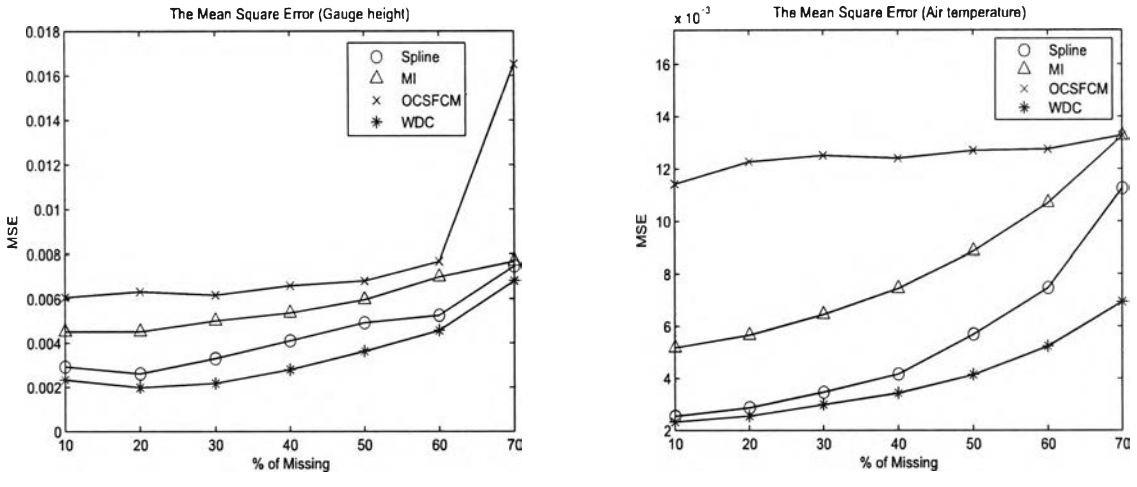
4.5 Summary

A new methodology (the WDC algorithm) for the pattern characterization, and the imputation of missing samples is presented. 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 signify that the imputation accuracy of the varied window clustering (WDC) algorithm can be comparable or better than the others traditional method such as: the spline interpolation, the multiple imputation (MI), and the optimal completion strategy fuzzy c-means (OCSFCM) algorithm. In case of the non-stationary time-series, especially the real-world problems, our results showed that WDC outperforms its competitors.



(a) (b)

Figure 4.6: The MSE of (a) Mackey-Glass data and (b) the sunspots data for spline interpolation, the MI method, the OCSFCM algorithm, and the WDC algorithm.



(a) (b)

Figure 4.7: The MSE of (a) gauge height data and (b) the air temperature for spline interpolation, MI method, the OCSFCM algorithm and the WDC algorithm.

Table 4.1: The mean square error (MSE) and the Pearson's correlation (CORR) of Mackey-Glass data for the spline interpolation method, the MI method, the OCSFCM algorithm, and the WDC algorithm.

Methodology		**Spline**	**MI**	**OCSFCM**	**WDC**
Level of Missing					
10 %	MSE $\times 10^{-6}$	8.77	**0.8**	23740	92.56
	CORR $\times 10^{-1}$	9.99905	**9.99992**	9.32662	9.8969
20 %	MSE $\times 10^{-6}$	12.71	**4.69**	25722	14.18
	CORR $\times 10^{-1}$	9.99859	**9.99947**	9.25783	9.9847
30 %	MSE $\times 10^{-6}$	**12.13**	13.83	28351	27.95
	CORR $\times 10^{-1}$	**9.99863**	9.99842	9.17196	9.99690
40 %	MSE $\times 10^{-6}$	77.61	**50.91**	32858	58.71
	CORR $\times 10^{-1}$	9.99134	**9.99420**	8.91467	9.99355
50 %	MSE $\times 10^{-6}$	153.12	**140.59**	42507	170.57
	CORR $\times 10^{-1}$	9.98289	**9.98412**	4.70478	9.98115
60 %	MSE $\times 10^{-6}$	730.91	589.01	44784	**555.65**
	CORR $\times 10^{-1}$	9.91888	9.93367	4.52367	**9.93788**
70 %	MSE $\times 10^{-6}$	936.58	1356.20	44813	**903.80**
	CORR $\times 10^{-1}$	9.89363	9.84686	4.51896	**9.89907**

Table 4.2: The mean square error (MSE) and the Pearson's correlation (CORR) of the sunspots data for the spline interpolation method, the MI method, the OCSFCM algorithm, and the WDC algorithm.

Methodology		**Spline**	**MI**	**FCM**	**WDC**
Level of Missing					
10 %	MSE ($\times 10^{-3}$)	4.98	7.11	12.62	**4.25**
	CORR ($\times 10^{-1}$)	9.45	9.15	8.45	**9.48**
20 %	MSE ($\times 10^{-3}$)	5.59	7.55	13.34	**4.73**
	CORR ($\times 10^{-1}$)	9.32	9.09	8.30	**9.42**
30 %	MSE ($\times 10^{-3}$)	6.76	8.00	13.19	**5.01**
	CORR ($\times 10^{-1}$)	9.19	9.04	8.34	**9.38**
40 %	MSE ($\times 10^{-3}$)	7.67	8.23	13.02	**5.28**
	CORR ($\times 10^{-1}$)	9.11	9.03	8.43	**9.36**
50 %	MSE ($\times 10^{-3}$)	8.27	8.98	13.84	**5.99**
	CORR ($\times 10^{-1}$)	9.04	8.93	8.33	**9.28**
60 %	MSE ($\times 10^{-3}$)	9.58	9.67	14.79	**7.05**
	CORR ($\times 10^{-1}$)	8.90	8.85	8.27	**9.15**
70 %	MSE ($\times 10^{-3}$)	12.53	10.83	16.46	**8.42**
	CORR ($\times 10^{-1}$)	8.63	8.73	8.16	**9.00**

Table 4.3: The mean square error (MSE) and the Pearson's correlation (CORR) of the gauge height data for the spline interpolation method, the MI method, the OCSFCM algorithm, and the WDC algorithm.

Methodology		**Spline**	**MI**	**OCSFCM**	**WDC**
Level of Missing					
10 %	MSE ($\times 10^{-3}$)	2.94	4.51	6.07	**2.34**
	CORR ($\times 10^{-1}$)	9.02	8.50	7.85	**9.20**
20 %	MSE ($\times 10^{-3}$)	2.62	4.52	6.33	**2.00**
	CORR ($\times 10^{-1}$)	9.15	8.53	7.92	**9.33**
30 %	MSE ($\times 10^{-3}$)	3.31	5.01	6.17	**2.18**
	CORR ($\times 10^{-1}$)	8.94	8.37	7.99	**9.26**
40 %	MSE ($\times 10^{-3}$)	4.11	5.37	6.59	**2.80**
	CORR ($\times 10^{-1}$)	8.71	8.28	7.90	**9.06**
50 %	MSE ($\times 10^{-3}$)	4.92	5.96	6.81	**3.64**
	CORR ($\times 10^{-1}$)	8.46	8.06	7.92	**8.79**
60 %	MSE ($\times 10^{-3}$)	5.25	6.99	7.68	**4.56**
	CORR ($\times 10^{-1}$)	8.55	7.81	7.92	**8.56**
70 %	MSE ($\times 10^{-3}$)	7.47	7.68	16.5	**6.81**
	CORR ($\times 10^{-1}$)	7.65	7.48	3.02	**7.89**

Table 4.4: The mean square error (MSE) and the Pearson's correlation (CORR) of the air temperature data for the spline interpolation method, the MI method, the OCSFCM algorithm, and the WDC algorithm.

Methodology		**Spline**	**MI**	**OCSFCM**	**WDC**
Level of Missing					
10 %	MSE ($\times 10^{-3}$)	2.55	5.18	11.43	**2.33**
	CORR ($\times 10^{-1}$)	9.47	8.93	7.30	**9.50**
20 %	MSE ($\times 10^{-3}$)	2.88	5.66	12.27	**2.55**
	CORR ($\times 10^{-1}$)	9.43	8.88	7.24	**9.48**
30 %	MSE ($\times 10^{-3}$)	3.47	6.45	12.51	**2.99**
	CORR ($\times 10^{-1}$)	9.30	8.71	7.14	**9.38**
40 %	MSE ($\times 10^{-3}$)	4.17	7.44	12.41	**3.44**
	CORR ($\times 10^{-1}$)	9.17	8.51	7.15	**9.29**
50 %	MSE ($\times 10^{-3}$)	5.70	8.86	12.70	**4.15**
	CORR ($\times 10^{-1}$)	8.88	8.22	7.10	**9.14**
60 %	MSE ($\times 10^{-3}$)	7.47	10.70	12.76	**5.24**
	CORR ($\times 10^{-1}$)	8.56	7.86	7.08	**8.89**
70 %	MSE ($\times 10^{-3}$)	11.27	13.26	13.28	**6.94**
	CORR ($\times 10^{-1}$)	7.95	7.34	6.99	**8.52**

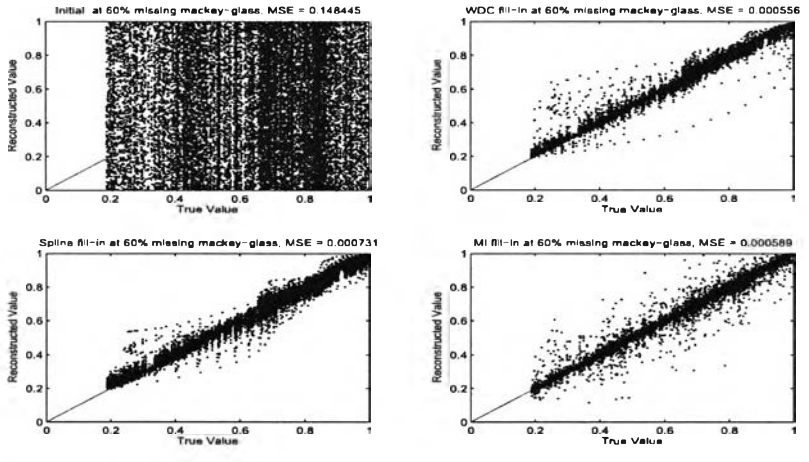


Figure 4.8: Scatter plots of reconstructed versus true values of the missing data for the 60% missing case using the Mackey-Glass data set. Missing values were initially set to random values (*top left*) and filled-in by the WDC algorithm (*top right*). The WDC output can be compared to the reconstructions by spline interpolation (*lower left*) and by the MI method (*lower right*).

Table 4.5: Average imputation performance index P_{Imp} for Mackey-Glass data, sunspots data, gauge height data and and air temperature data. We found that all values of P_{Imp} are greater than zero.

Fill-In Methodology	**Mackey-Glass** $\times 10^{-5}$	**Sunspots** $\times 10^{-3}$	**Gauge Height** $\times 10^{-4}$	**Air Temperature** $\times 10^{-3}$
Spline	1.55	2.09	8.99	1.41
MI	4.75	2.80	22.45	4.27
OCSFCM	3442.19	8.07	45.48	8.53

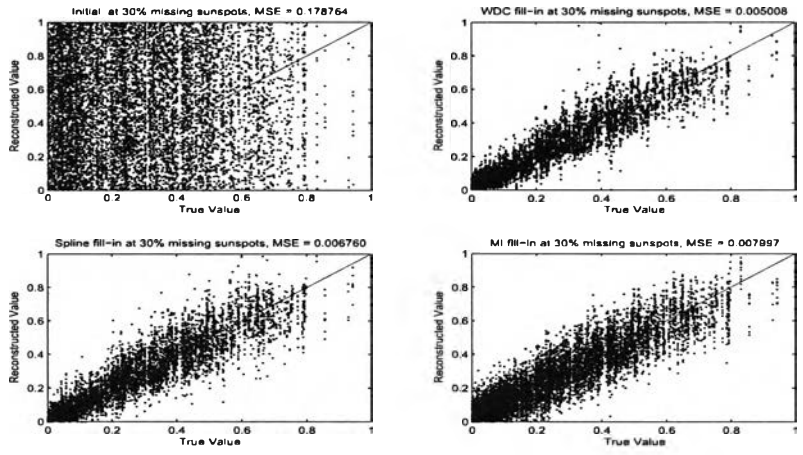


Figure 4.9: Scatter plots of reconstructed versus true values of the missing data for the 30% missing case using the sunspot data set. Missing values were initially set to random values (*top left*) and filled-in by the WDC algorithm (*top right*). The WDC output can be compared to the reconstructions by spline interpolation (*lower left*) and by the MI method (*lower right*).

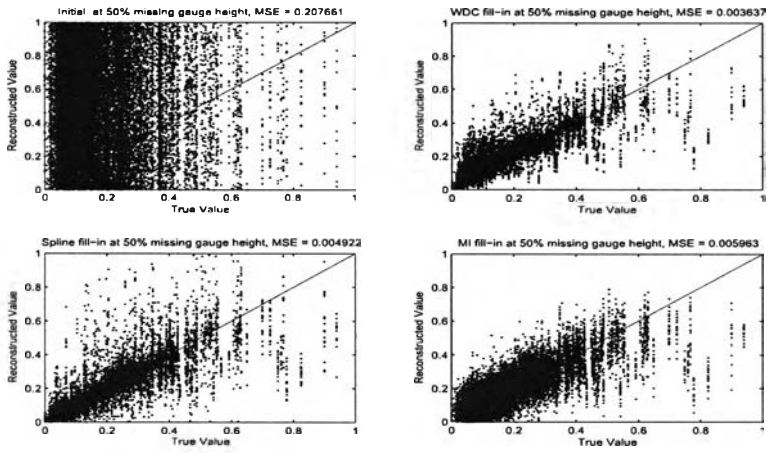


Figure 4.10: Scatter plots of reconstructed versus true values of the missing data for the 50% missing case using the gauge height data set. Missing values were initially set to random values (*top left*) and filled-in by the WDC algorithm (*top right*). The WDC output can be compared to the reconstructions by spline interpolation (*lower left*) and by the MI method (*lower right*).

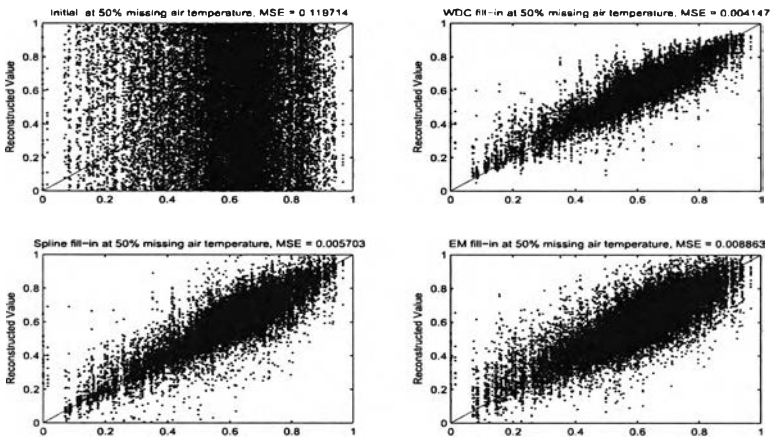


Figure 4.11: Scatter plots of reconstructed versus true values of the missing data for the 50% missing case using the air temperature data set. Missing values were initially set to random values (*top left*) and filled-in by the WDC algorithm (*top right*). The WDC output can be compared to the reconstructions by spline interpolation (*lower left*) and by the MI method (*lower right*).

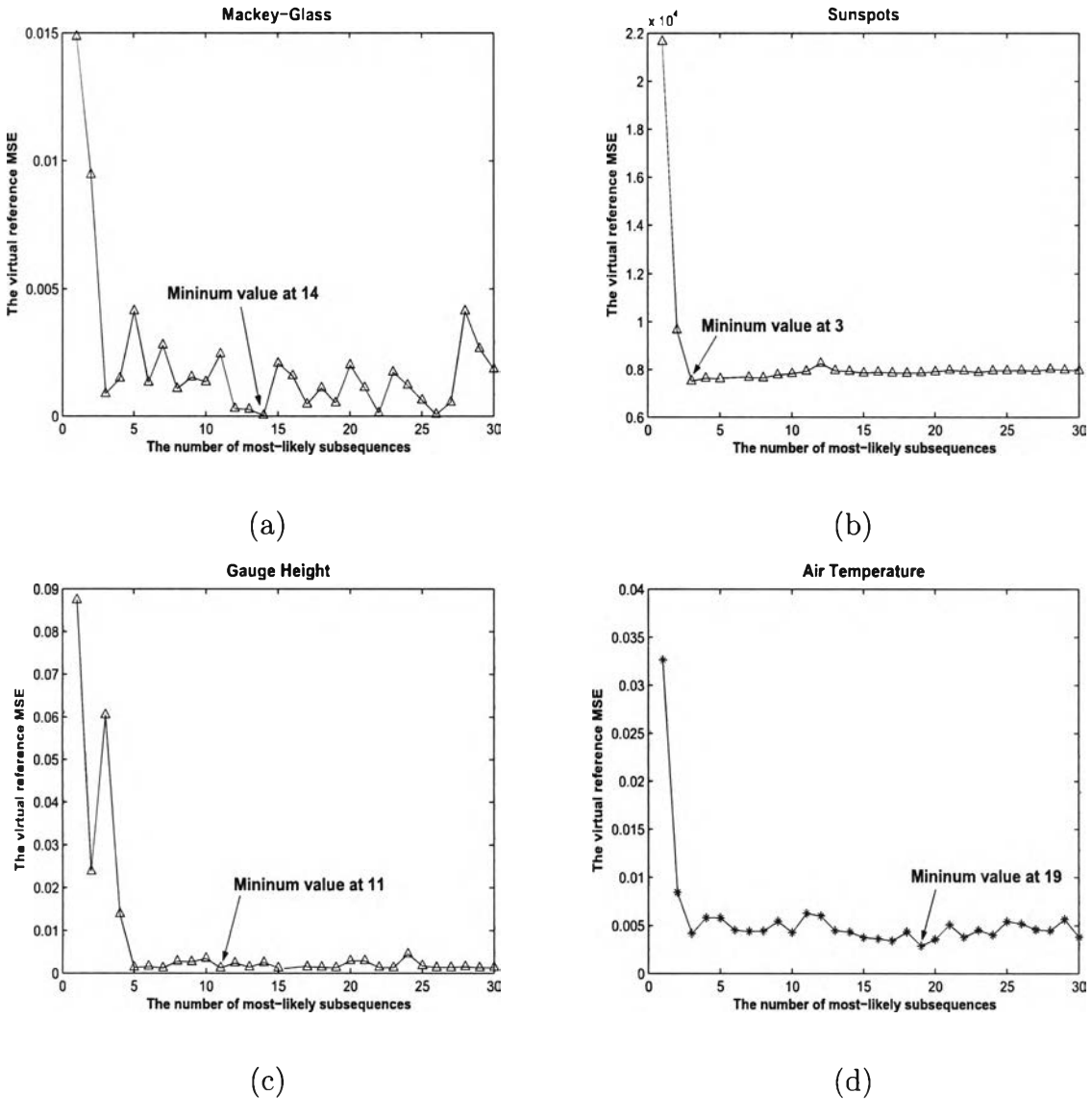


Figure 4.12: For each scatter plot, the number of most-likely subsequences of (a) Mackey-Glass (b) sunspots (c) gauge height data and (d) the air temperature are plotted.

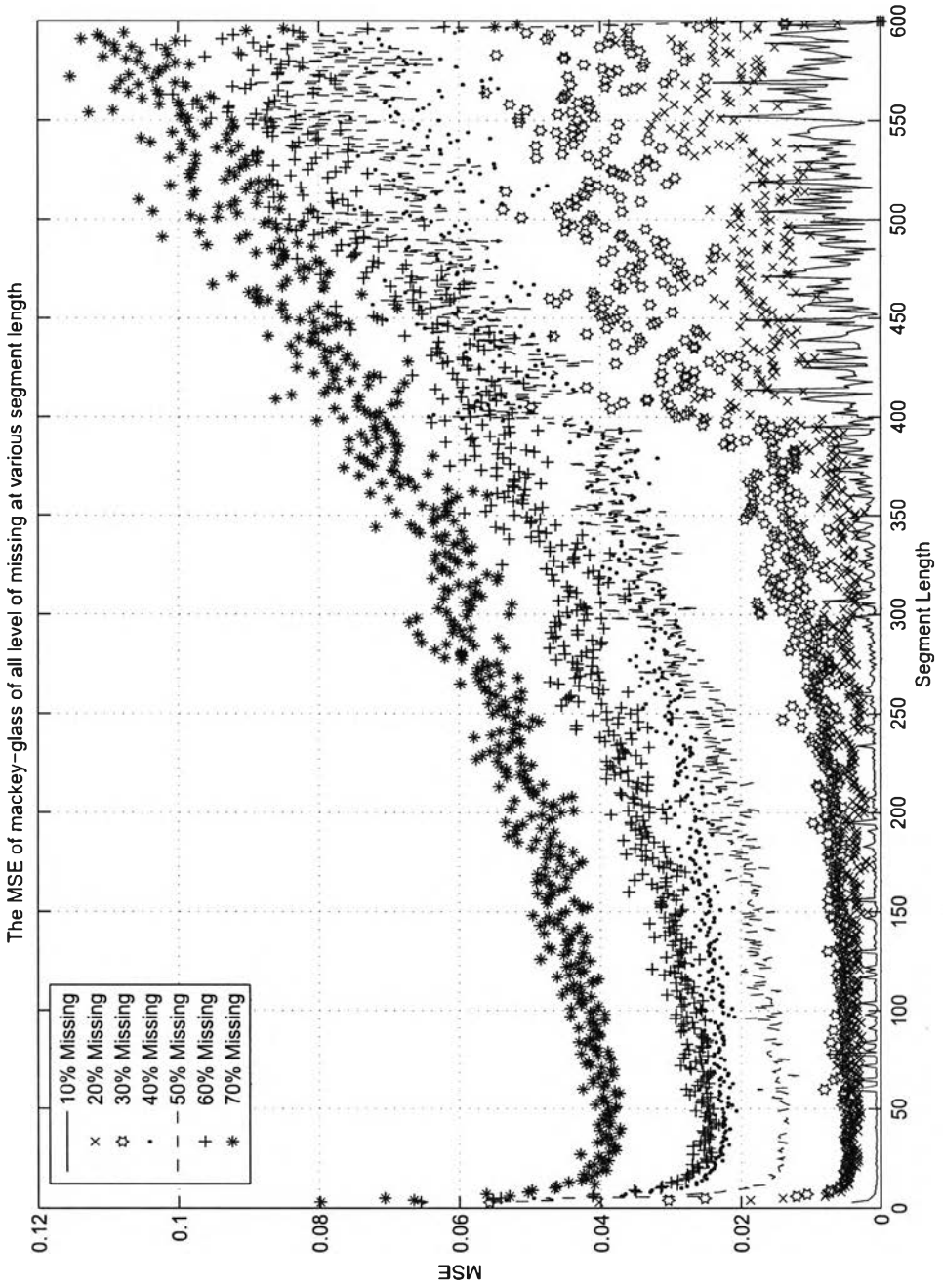


Figure 4.13: Scatter plots of the reference MSE values versus segment length for all levels of missing of the Mackey-glass data set.

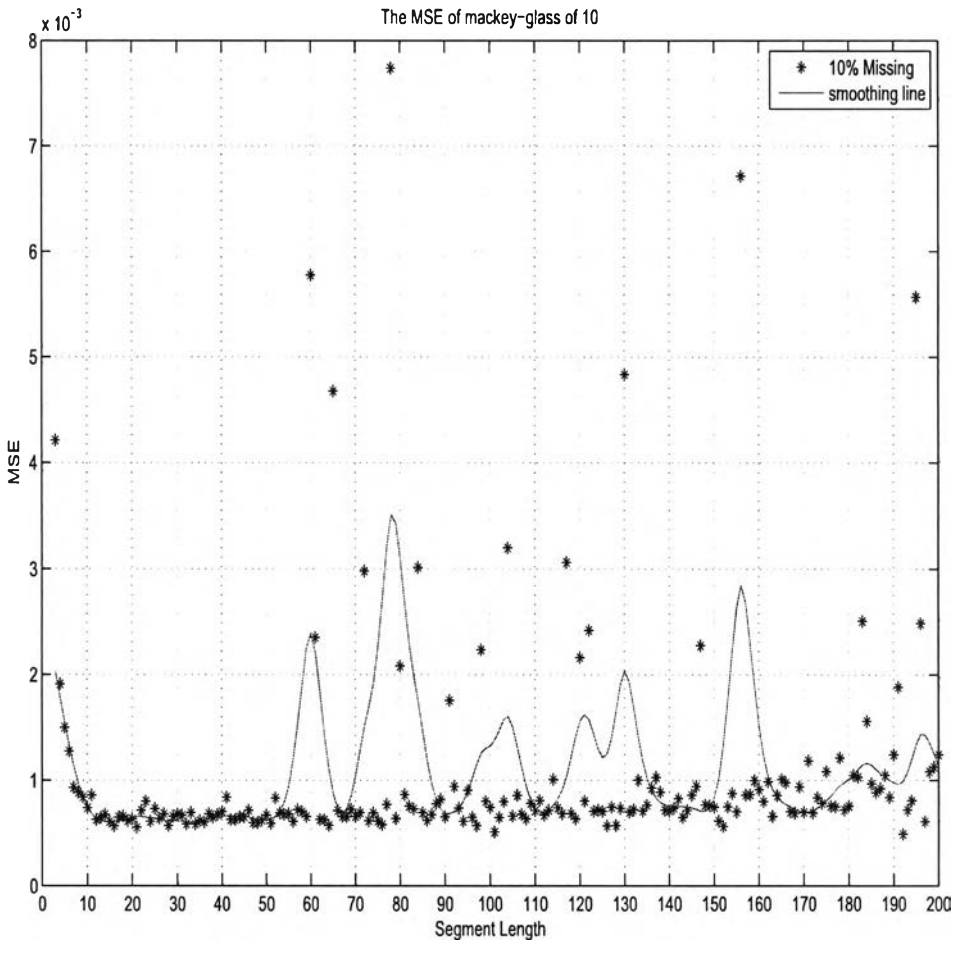


Figure 4.14: Scatter plots of the reference MSE values versus segment length at 10% missing of the Mackey-glass data set.

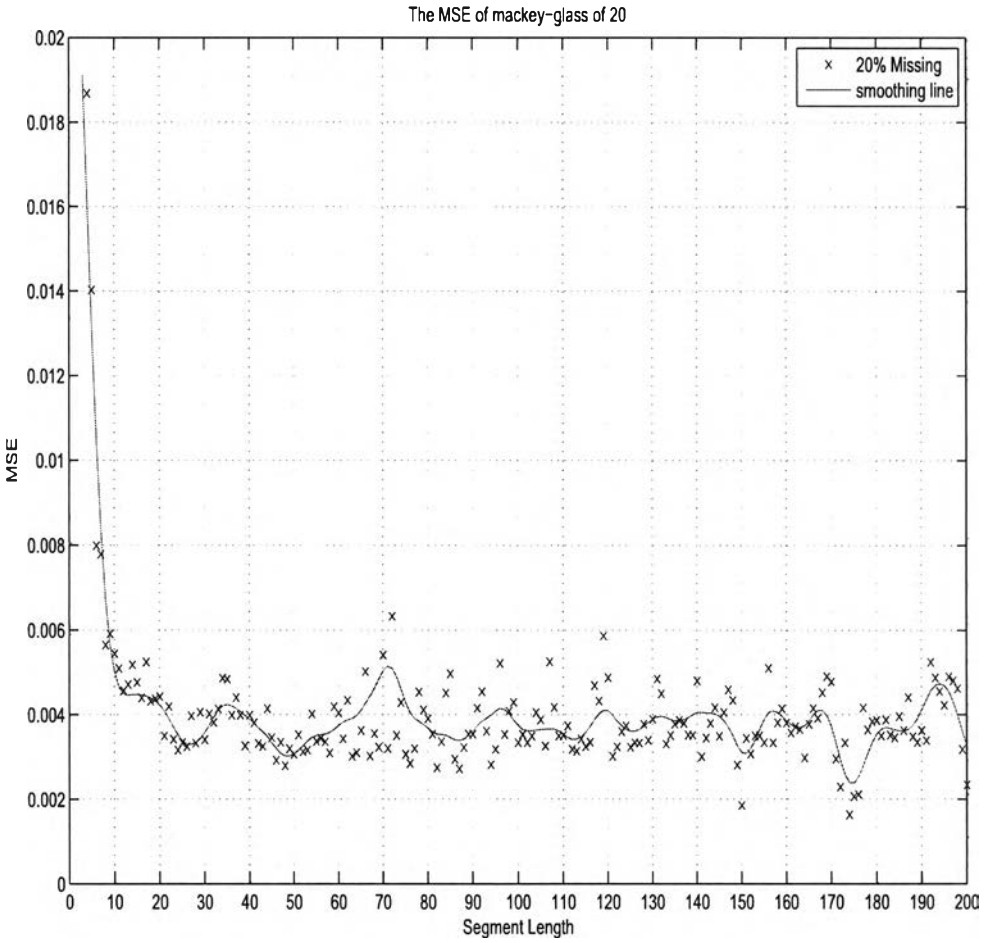


Figure 4.15: Scatter plots of the reference MSE values versus segment length at 20% missing of the Mackey-glass data set.

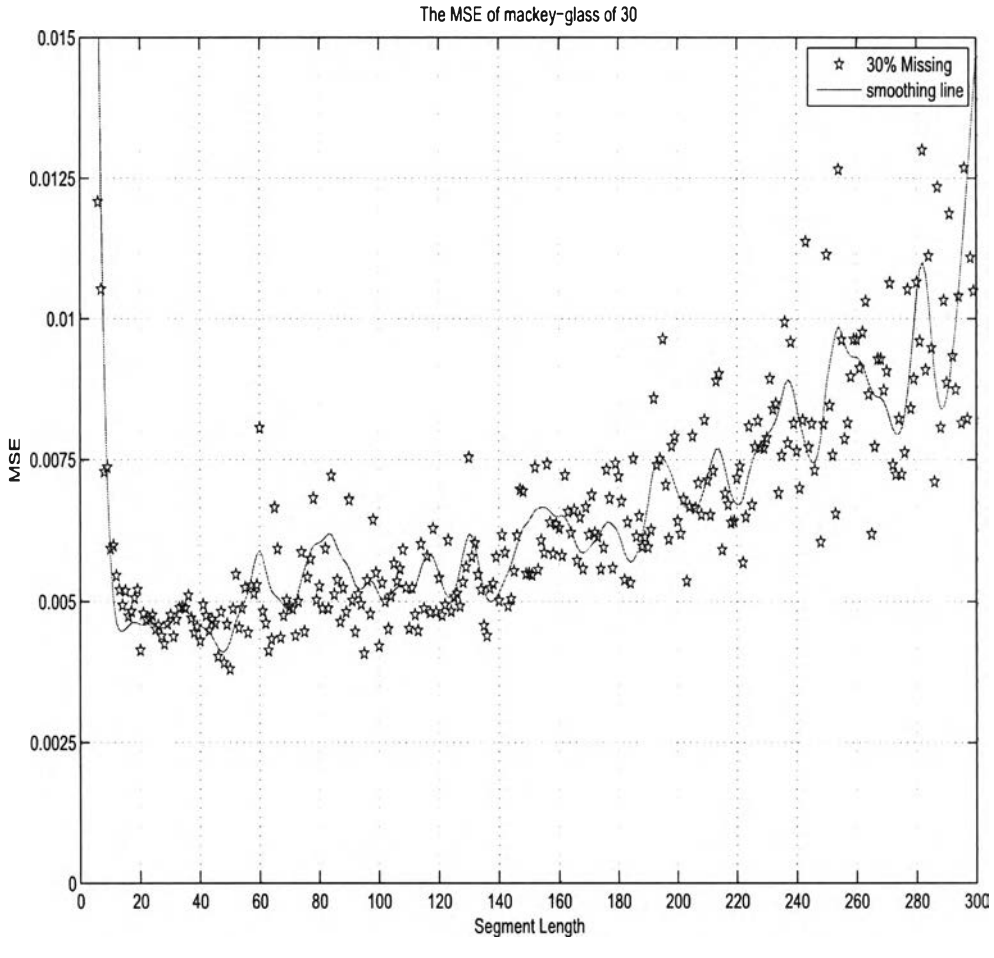


Figure 4.16: Scatter plots of the reference MSE values versus segment length at 30% missing of the Mackey-glass data set.

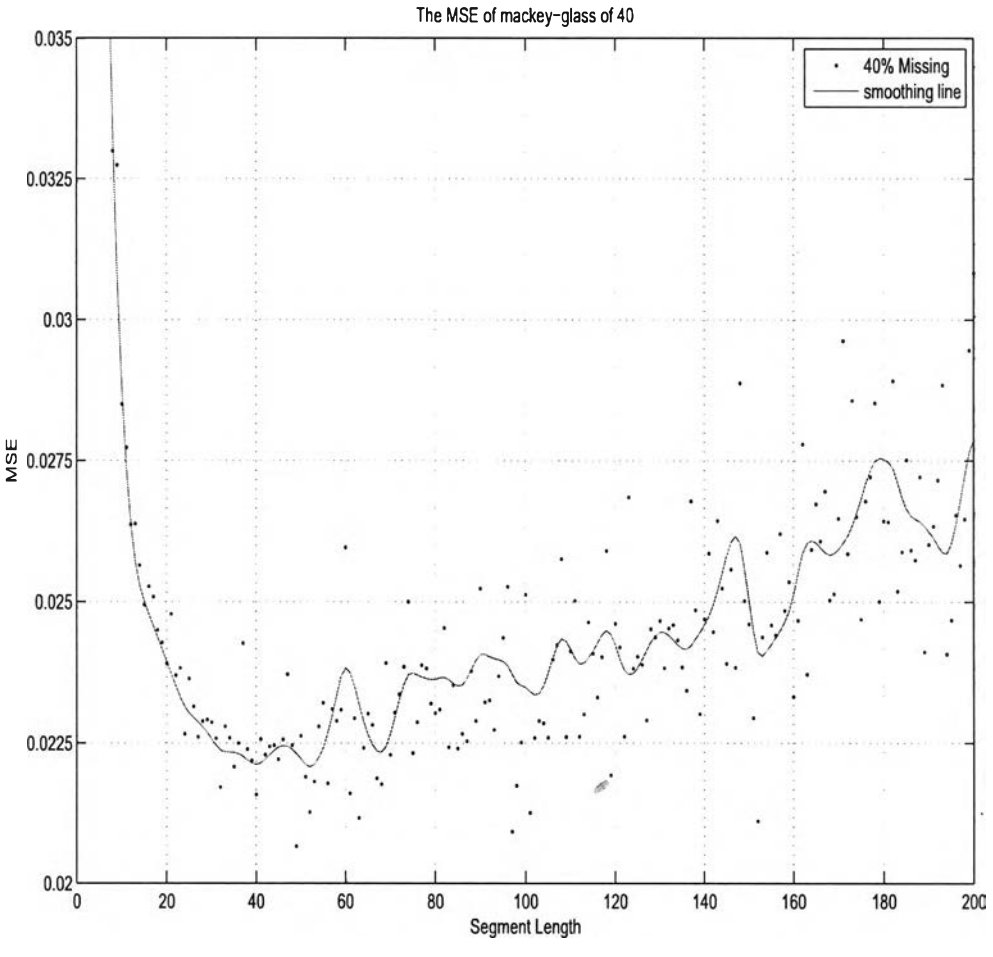


Figure 4.17: Scatter plots of the reference MSE values versus segment length at 40% missing of the Mackey-glass data set.

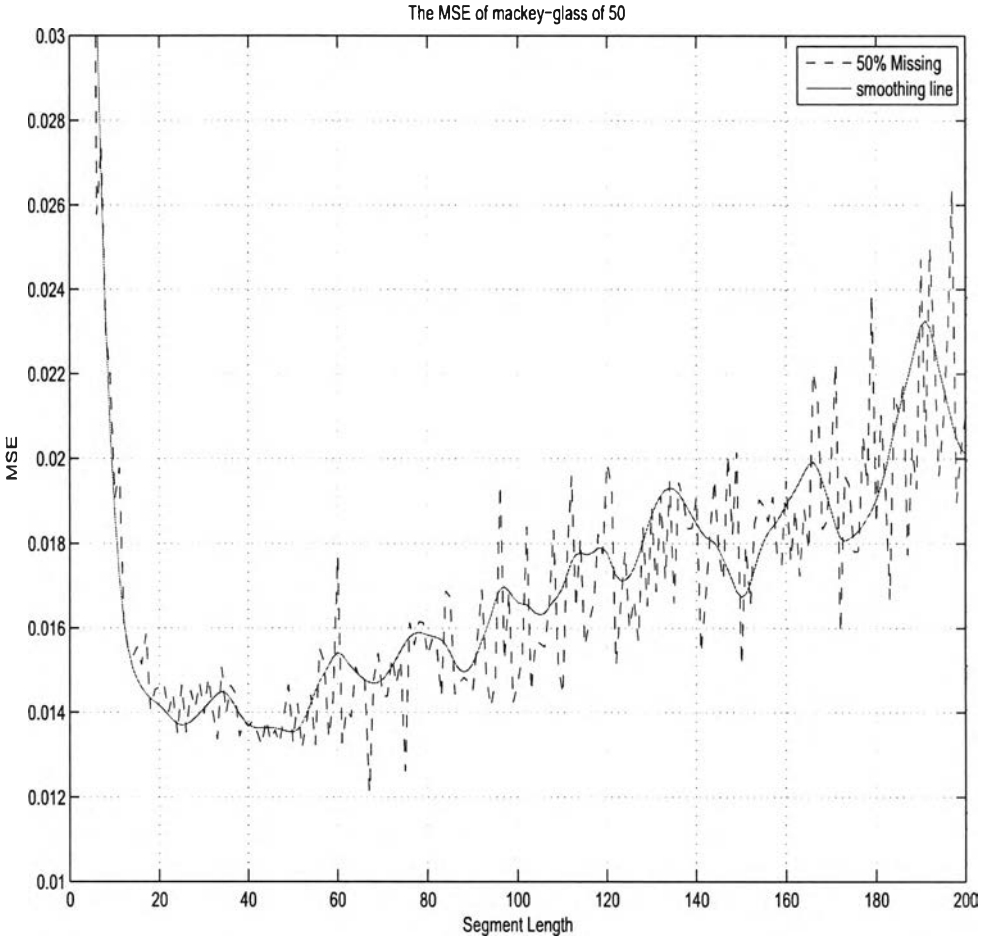


Figure 4.18: Scatter plots of the reference MSE values versus segment length at 50% missing of the Mackey-glass data set.

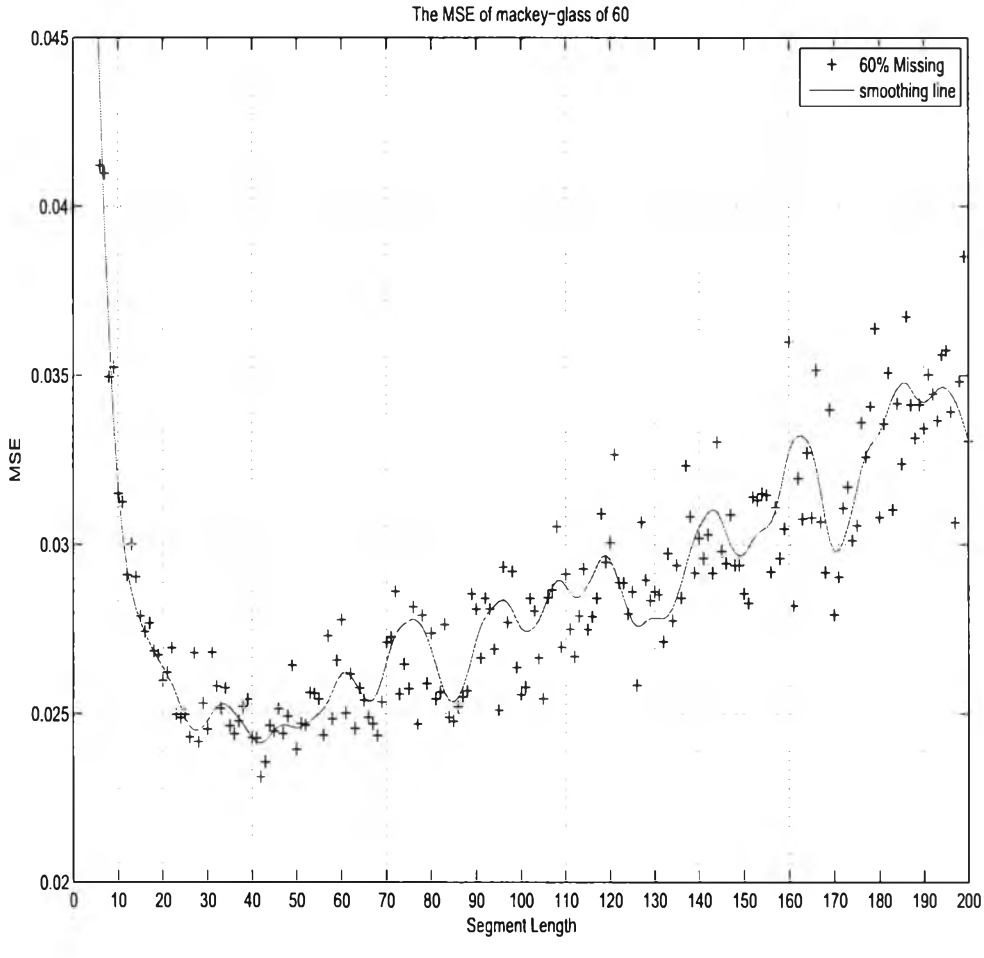


Figure 4.19: Scatter plots of the reference MSE values versus segment length at 60% missing of the Mackey-glass data set.

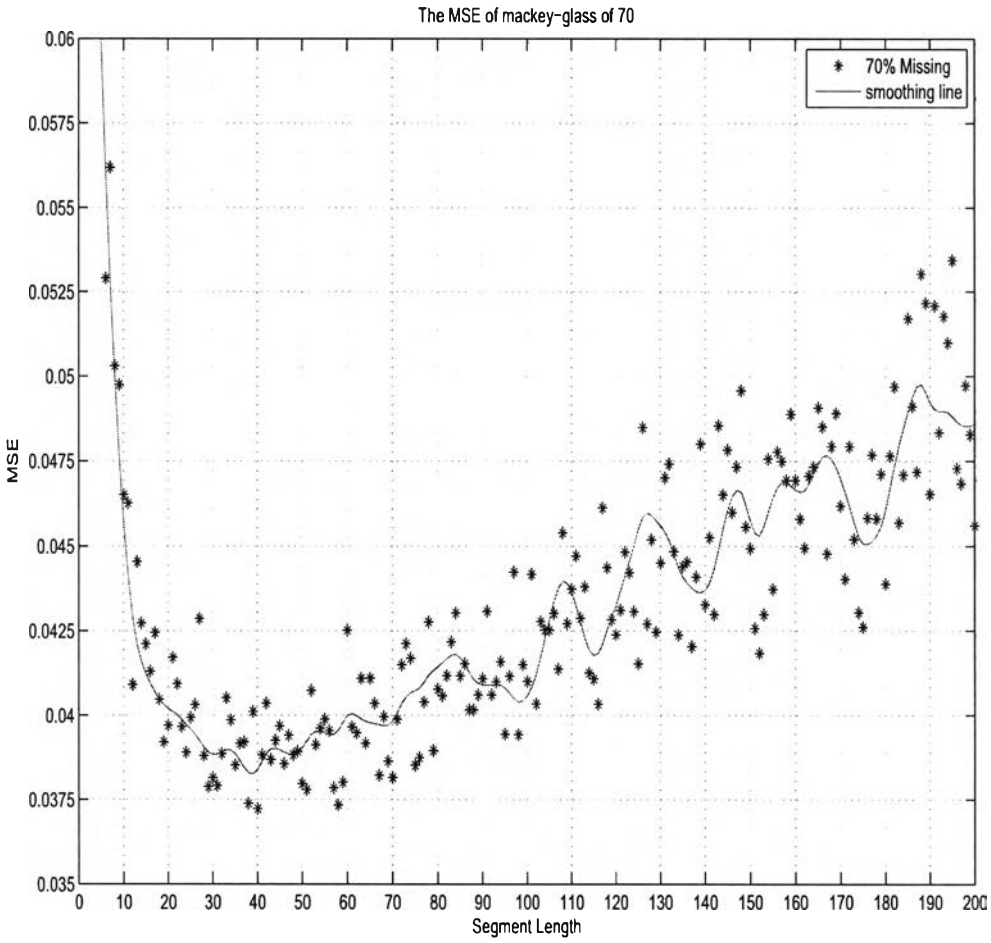


Figure 4.20: Scatter plots of the reference MSE values versus segment length at 70% missing of the Mackey-glass data set.

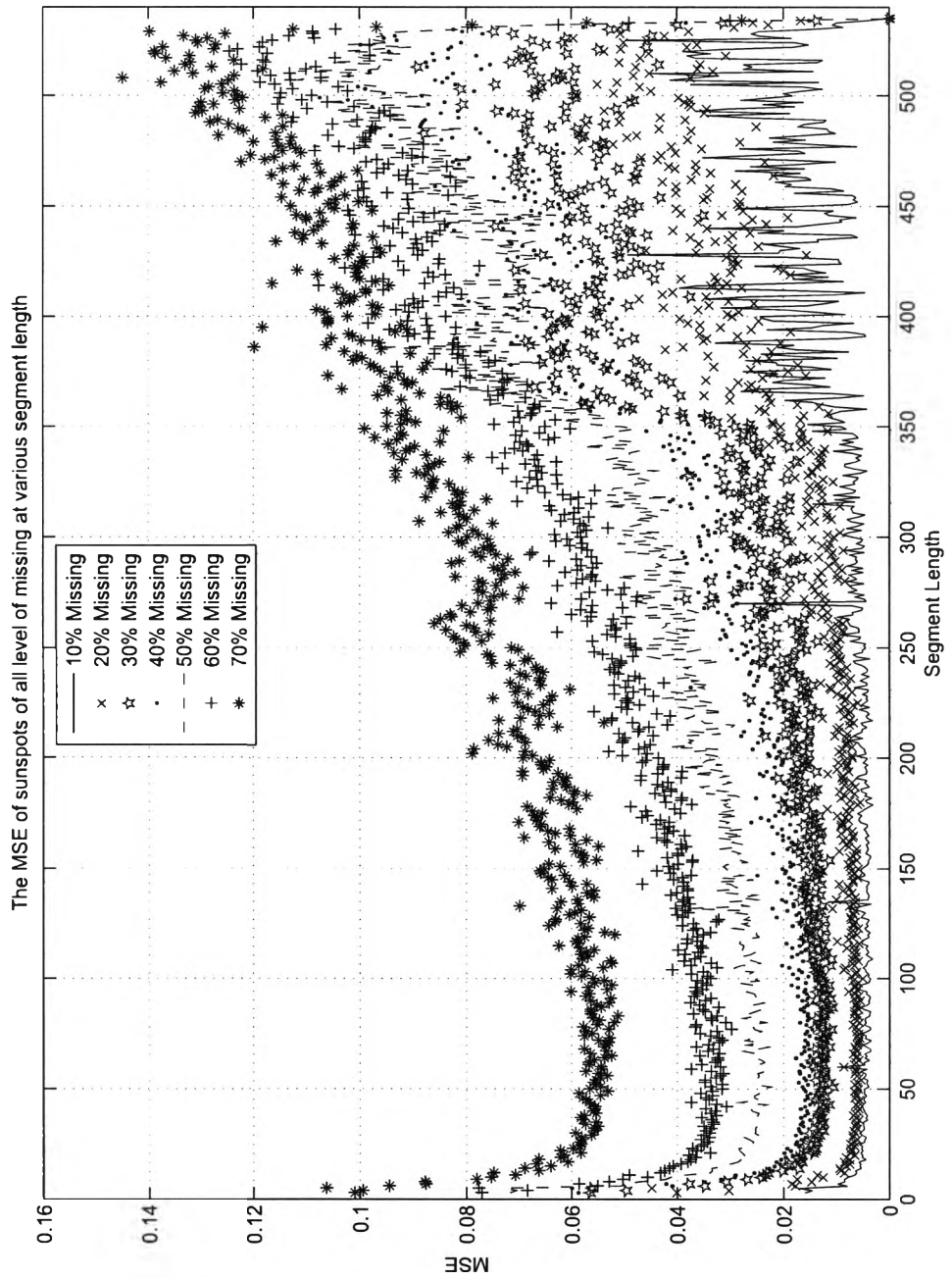


Figure 4.21: Scatter plots of the reference MSE values versus segment length for all levels of missing of the sunspots data set.

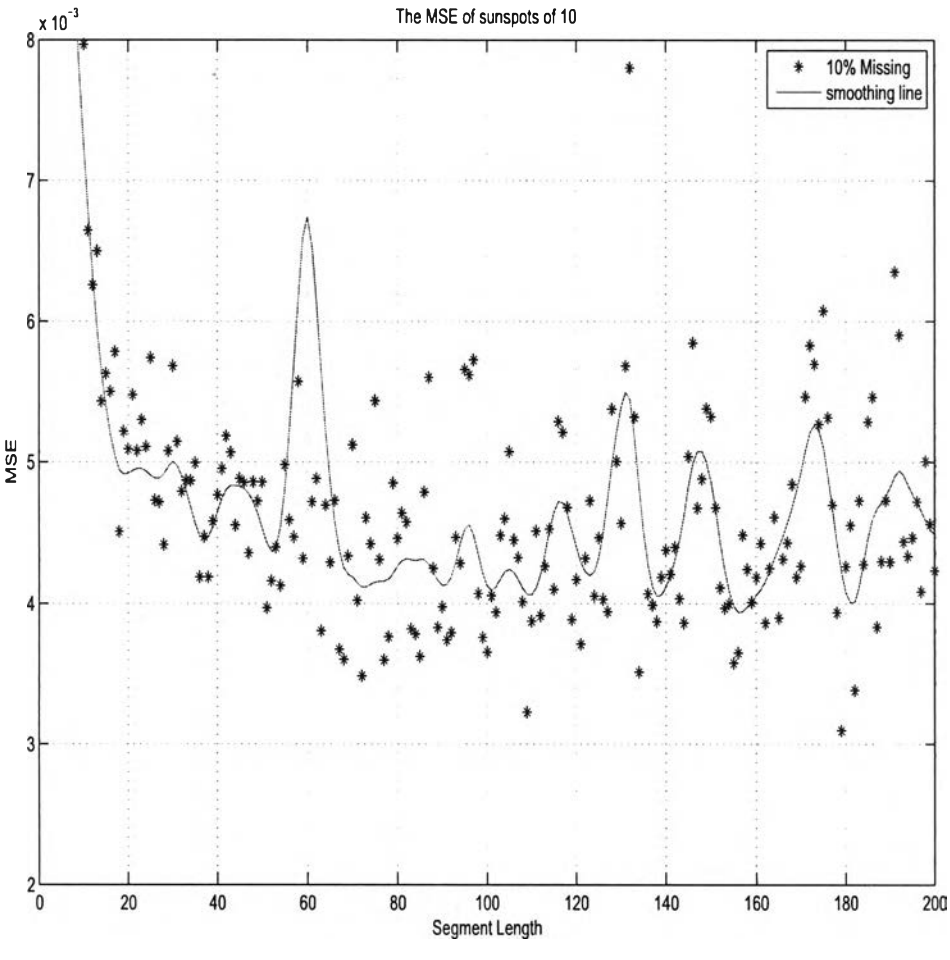


Figure 4.22: Scatter plots of the reference MSE values versus segment length at 10% missing of the sunspots data set.

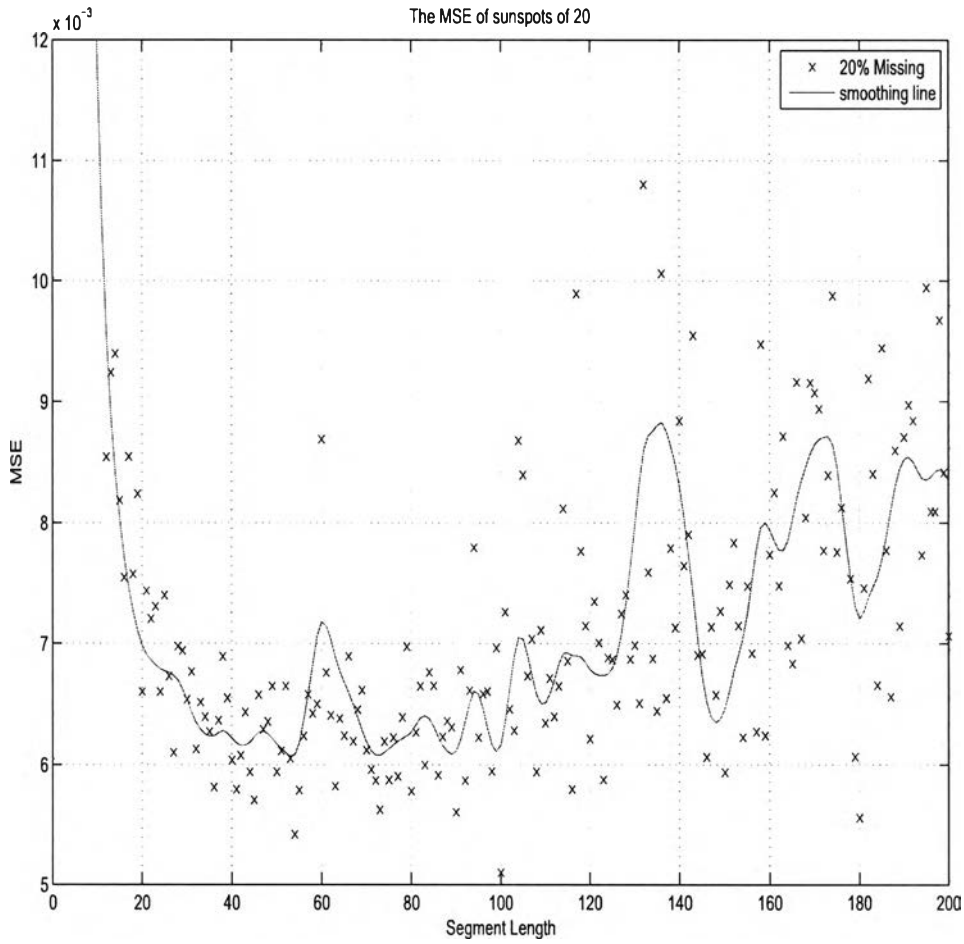


Figure 4.23: Scatter plots of the reference MSE values versus segment length at 20% missing of the Sunspots data set.

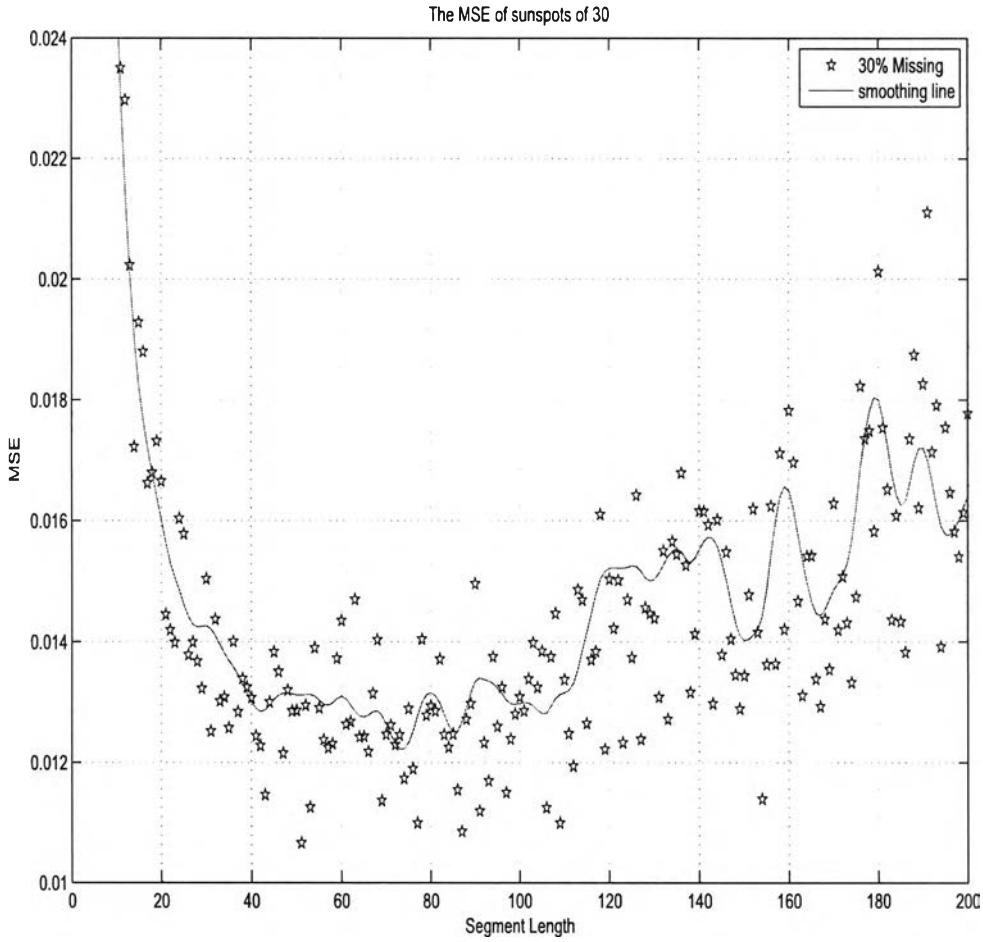


Figure 4.24: Scatter plots of the reference MSE values versus segment length at 30% missing of the sunspots data set.

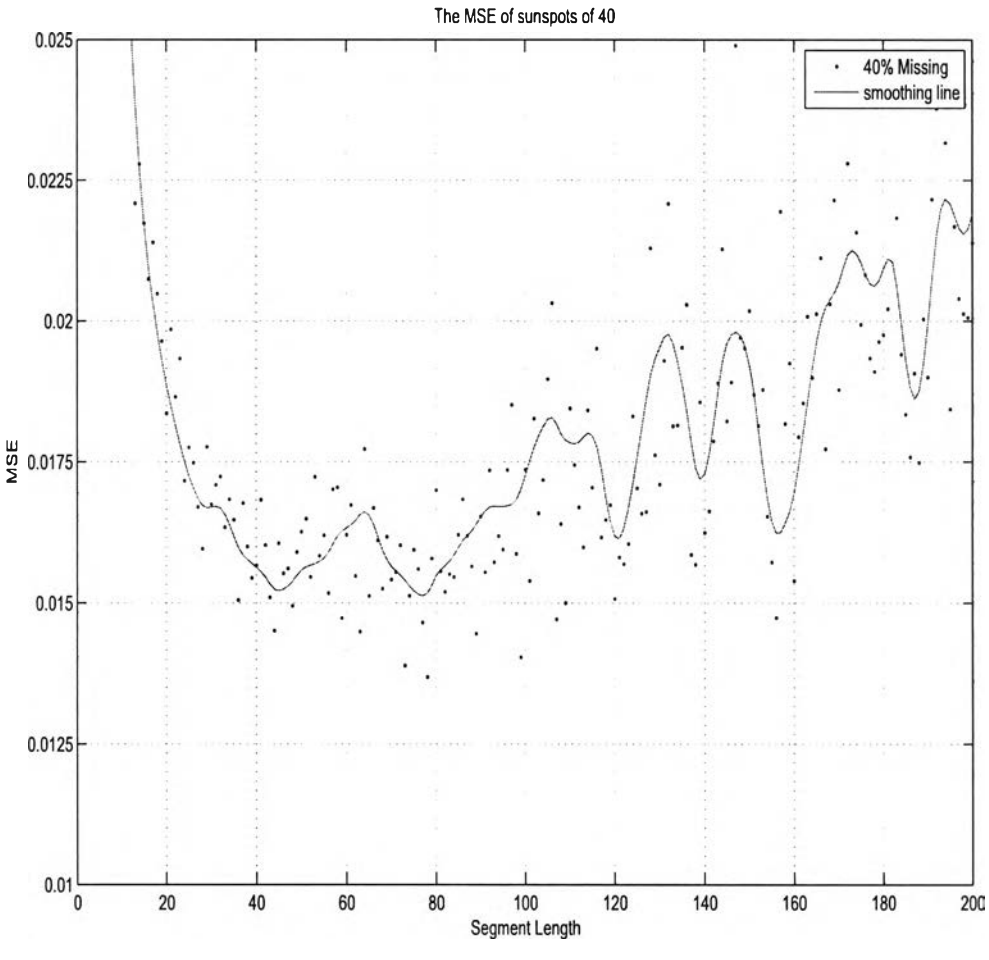


Figure 4.25: Scatter plots of the reference MSE values versus segment length at 40% missing of the Sunspots data set.

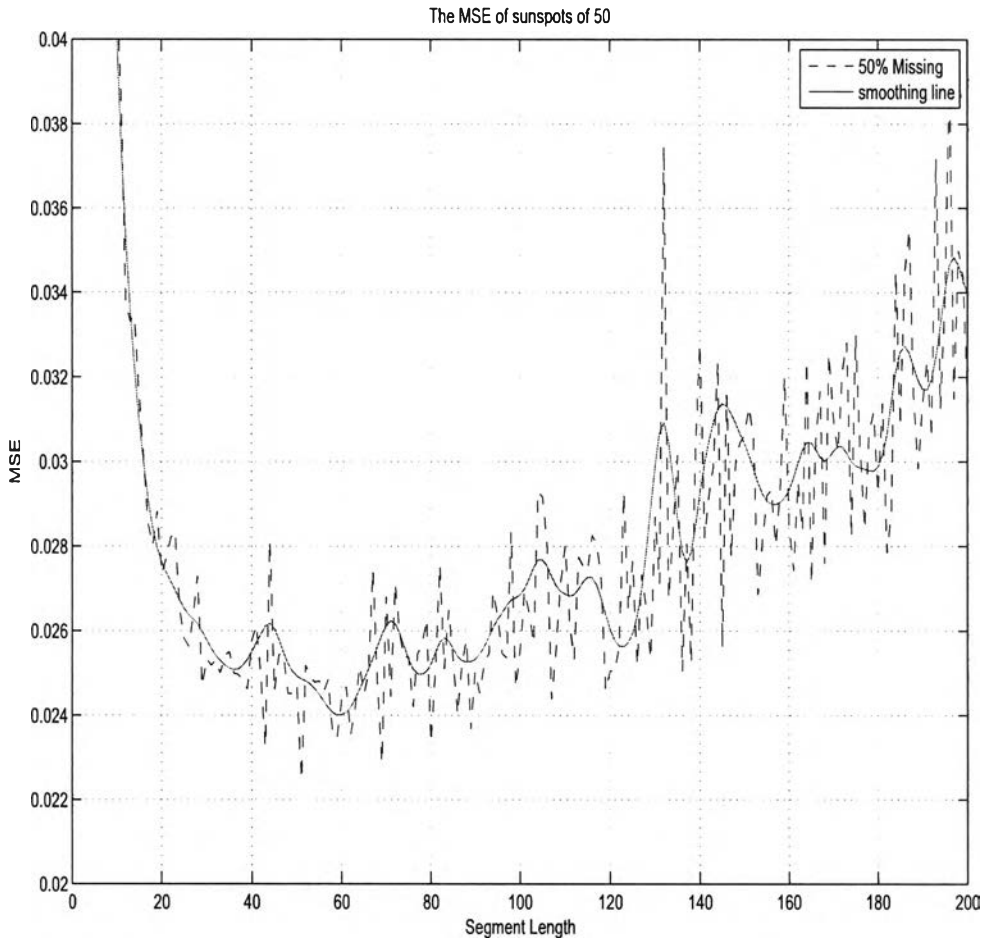


Figure 4.26: Scatter plots of the reference MSE values versus segment length at 50% missing of the sunspots data set.

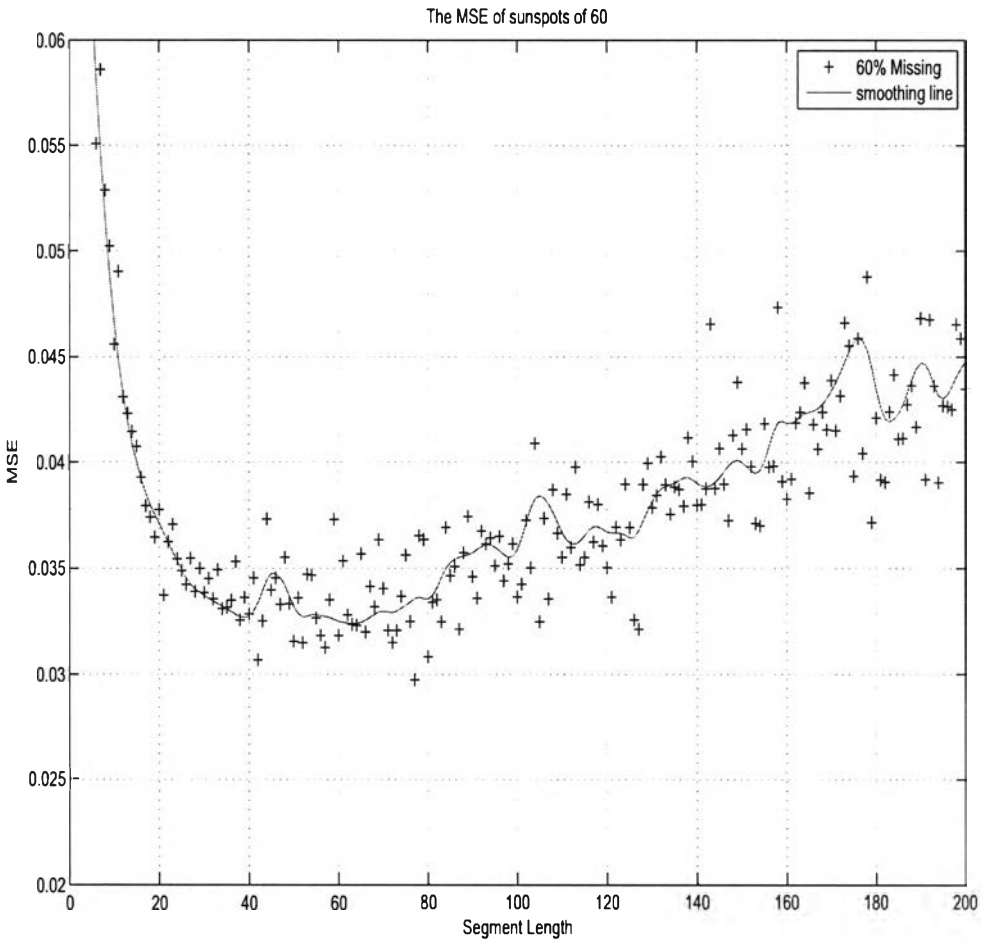


Figure 4.27: Scatter plots of the reference MSE values versus segment length at 60% missing of the sunspots data set.

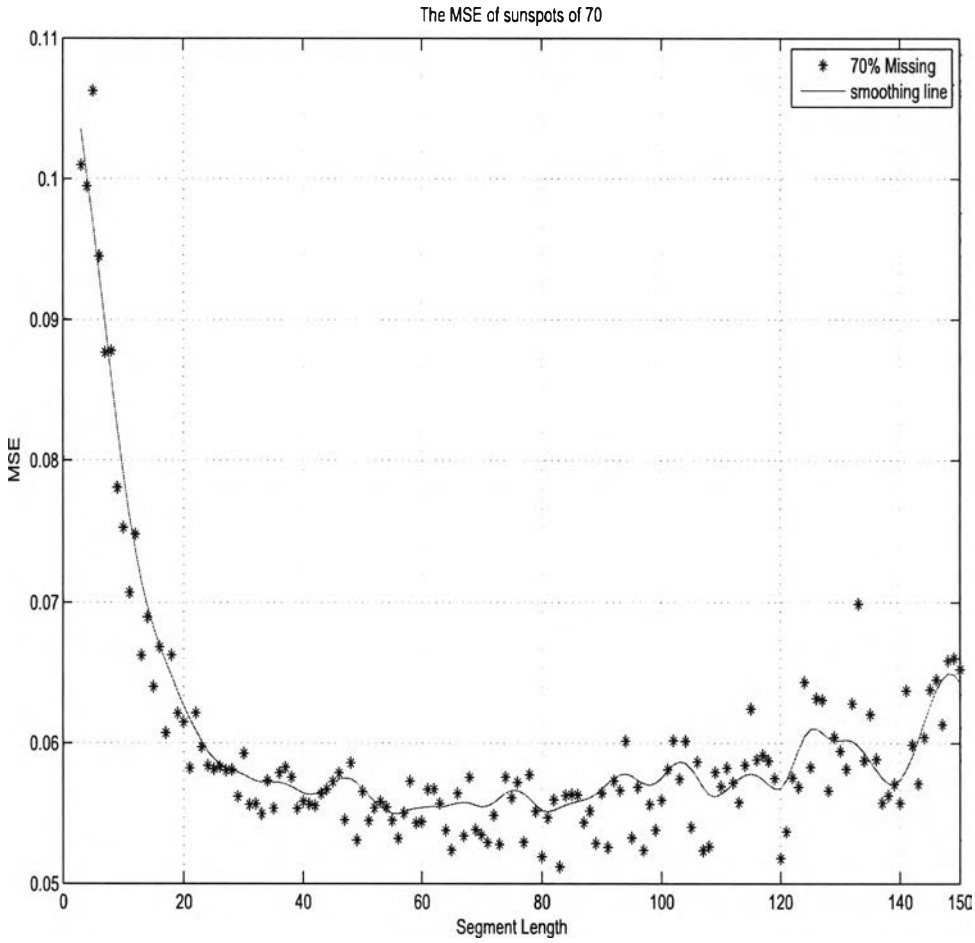


Figure 4.28: Scatter plots of the reference MSE values versus segment length at 70% missing of the Sunspots data set.

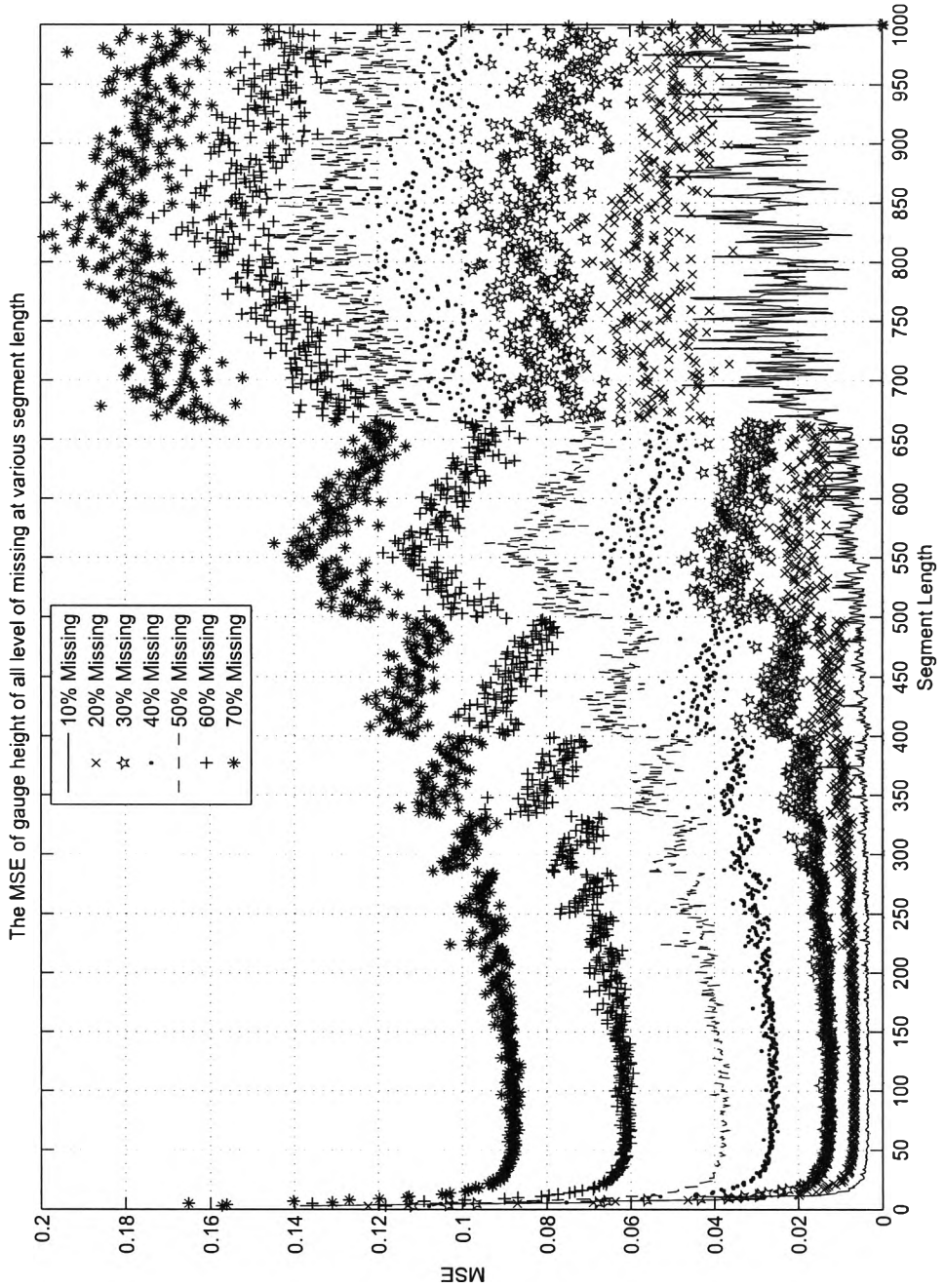


Figure 4.29: Scatter plots of the reference MSE values versus segment length for all levels of missing of the gauge height data set.

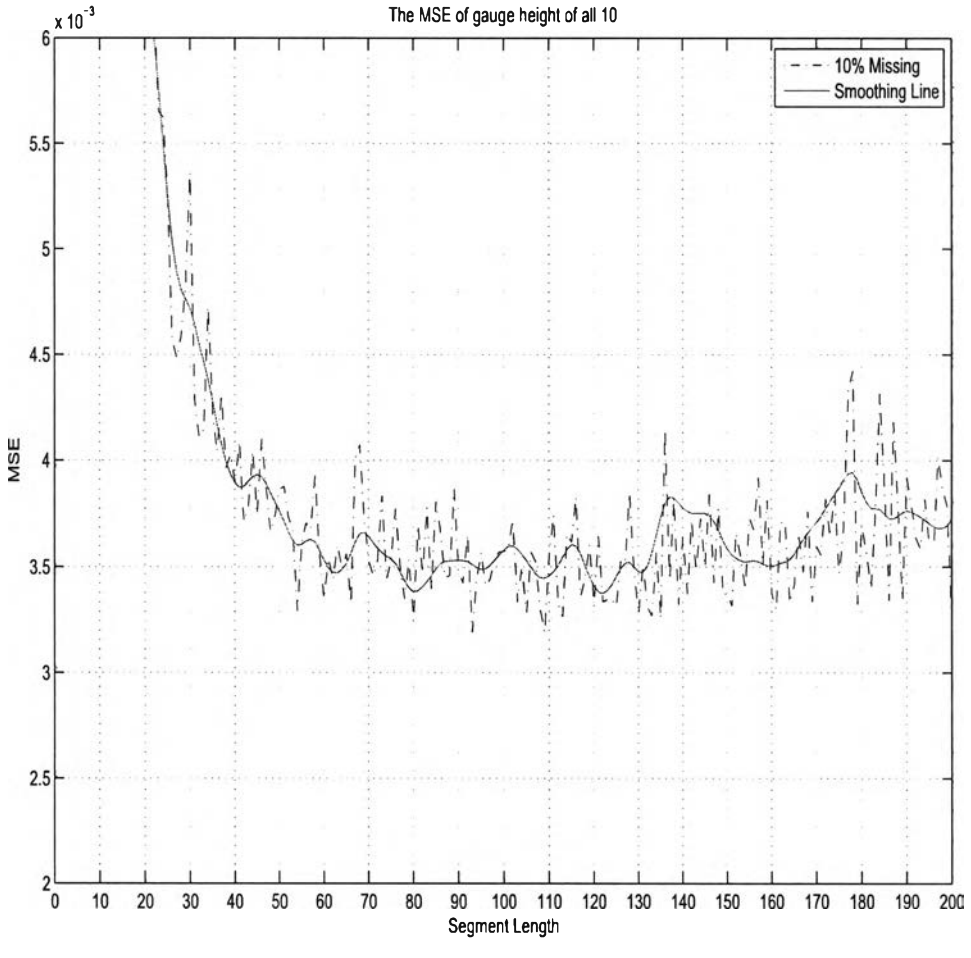


Figure 4.30: Scatter plots of the reference MSE values versus segment length at 10% missing of the gauge height data set.

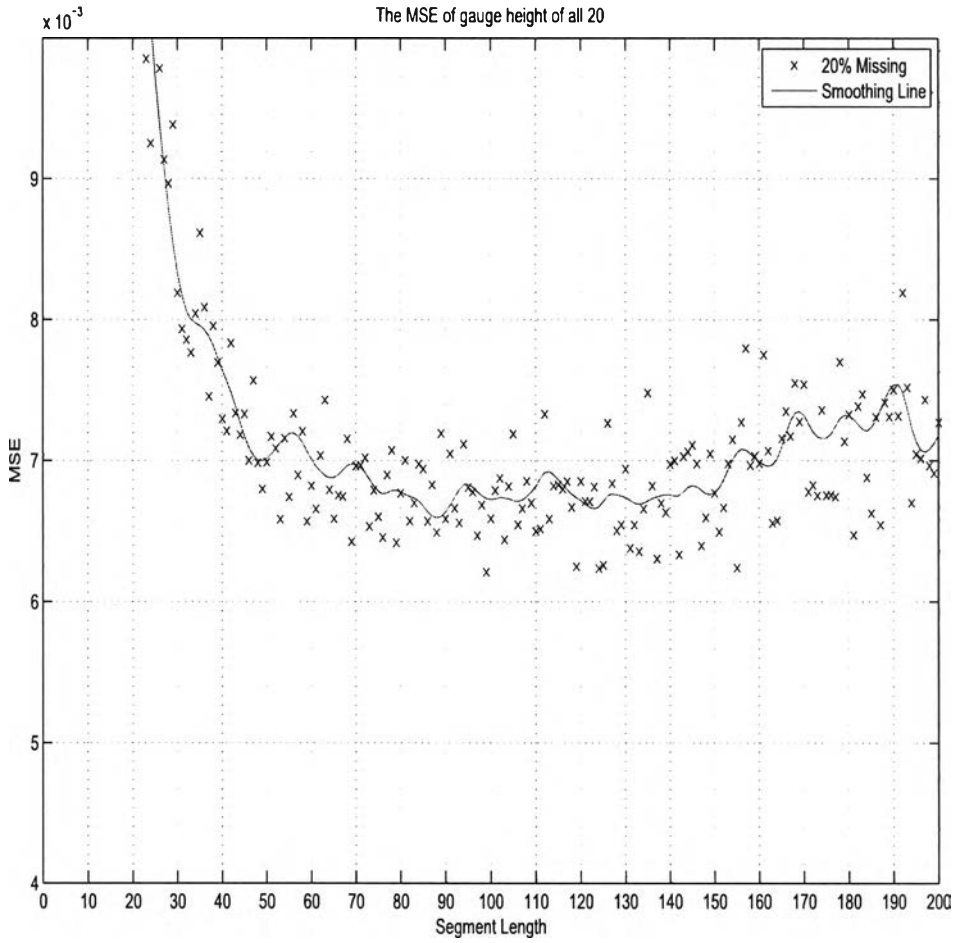


Figure 4.31: Scatter plots of the reference MSE values versus segment length at 20% missing of the gauge height data set.

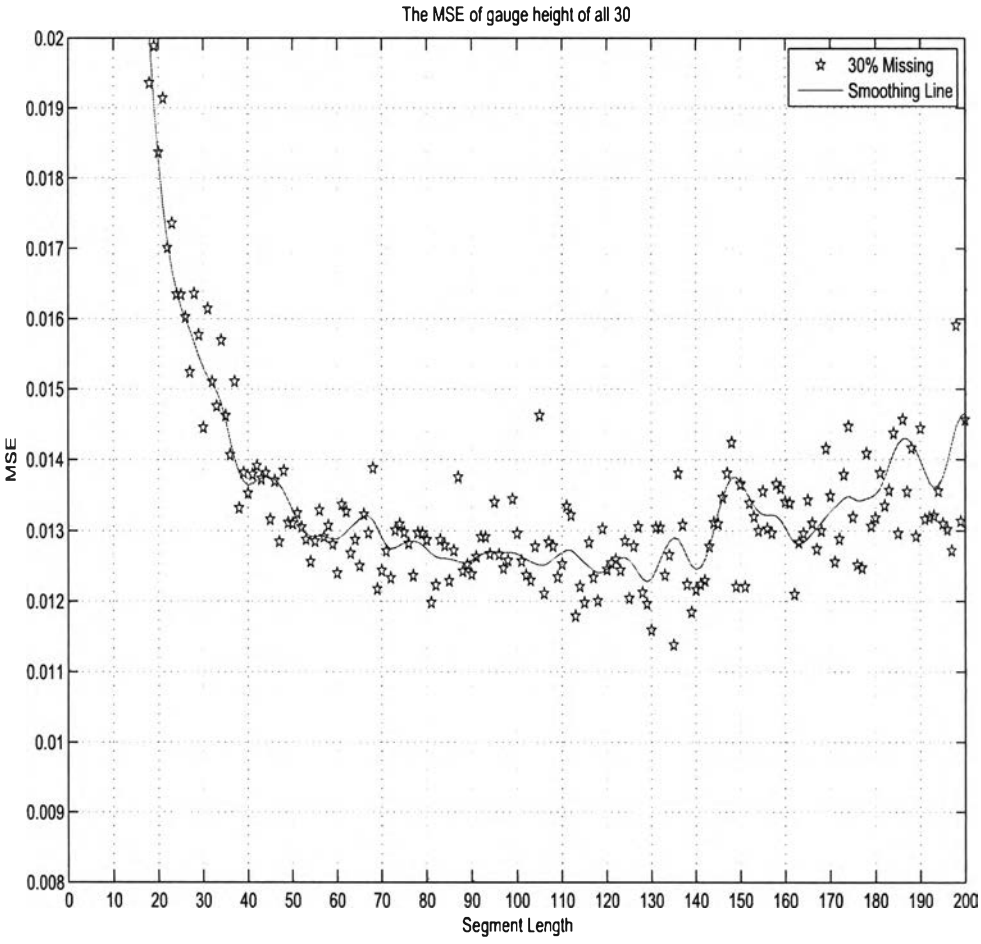


Figure 4.32: Scatter plots of the reference MSE values versus segment length at 30% missing of the gauge height data set.

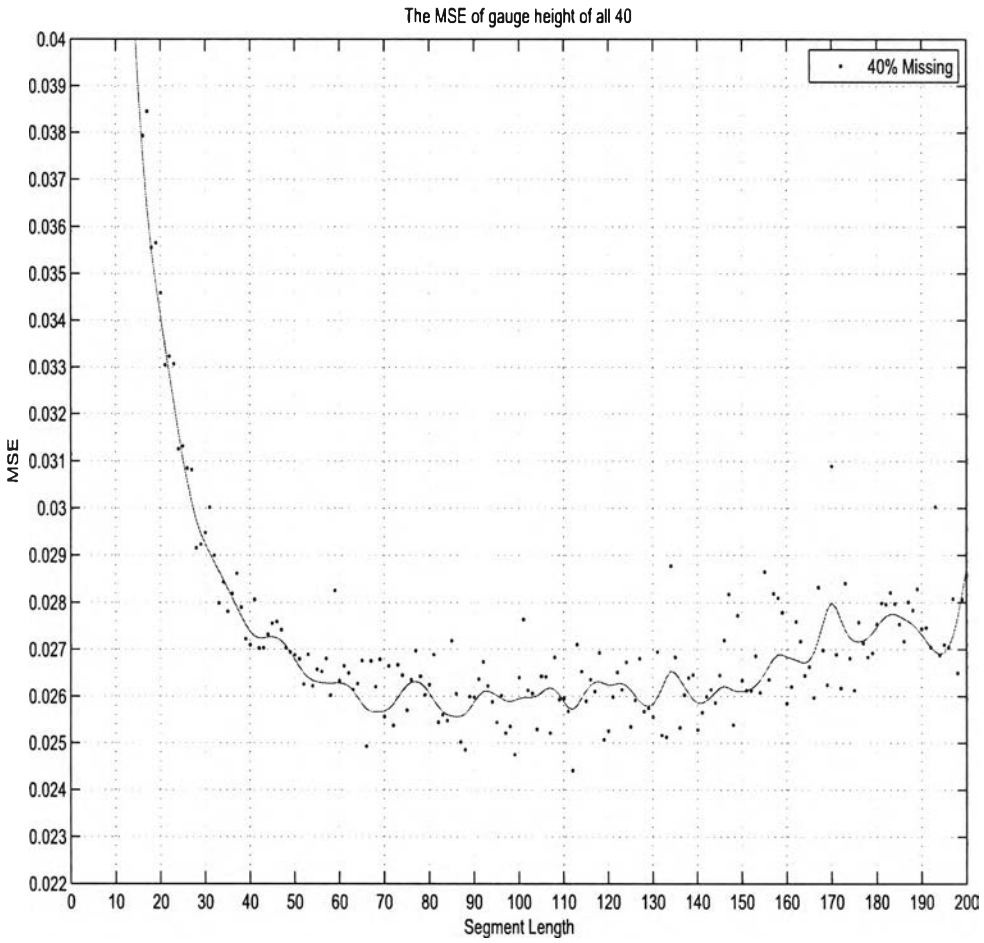


Figure 4.33: Scatter plots of the reference MSE values versus segment length at 40% missing of the gauge height data set.

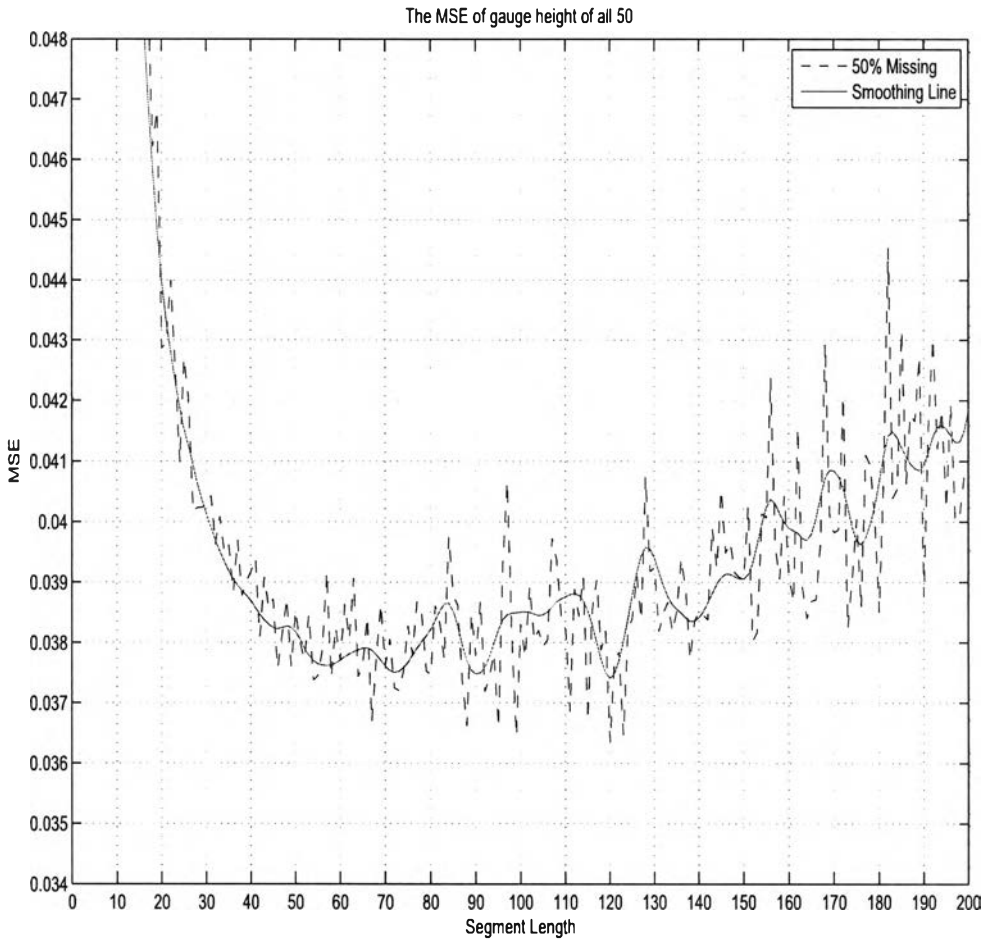


Figure 4.34: Scatter plots of the reference MSE values versus segment length at 50% missing of the gauge height data set.

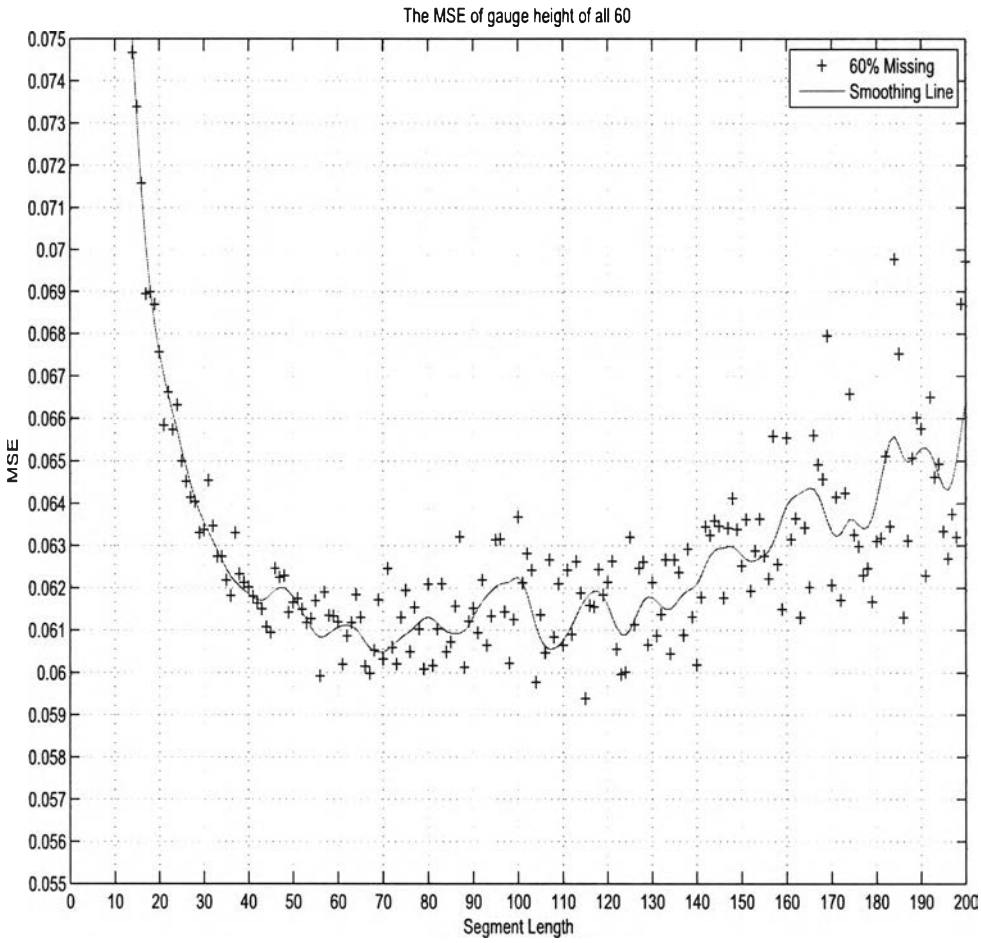


Figure 4.35: Scatter plots of the reference MSE values versus segment length at 60% missing of the gauge height data set.

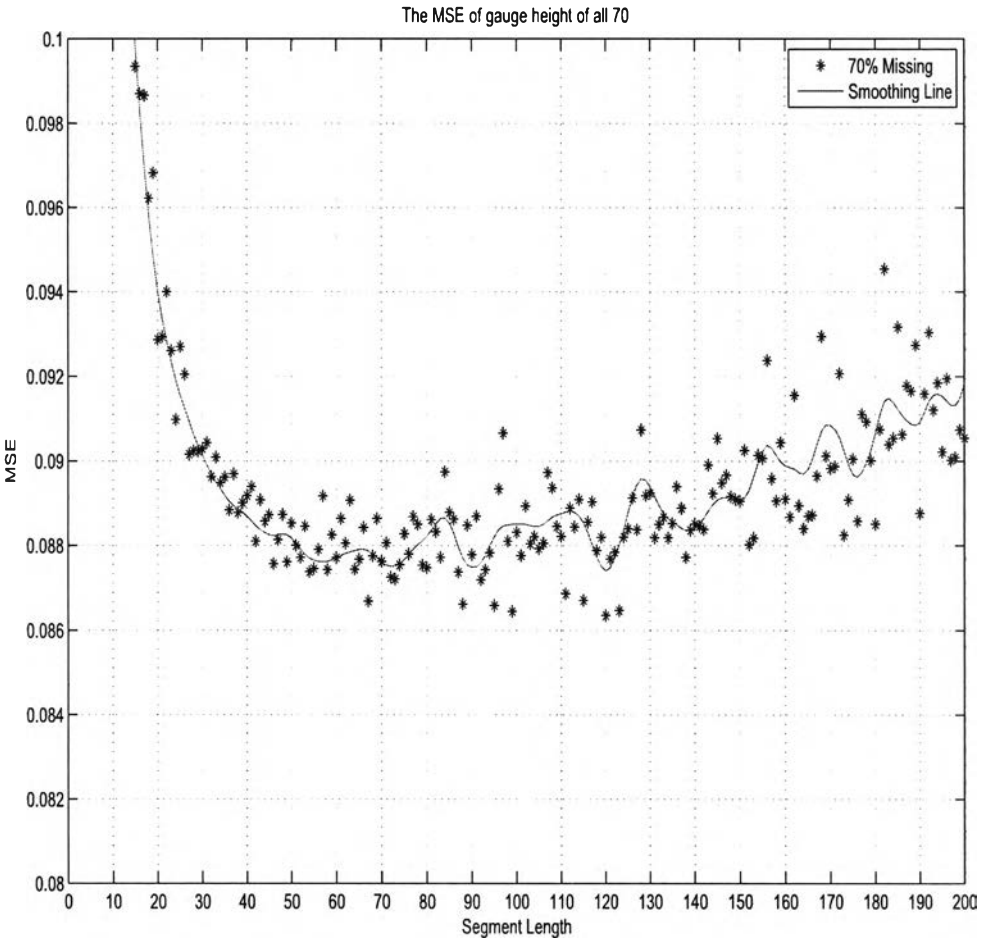


Figure 4.36: Scatter plots of the reference MSE values versus segment length at 70% missing of the gauge height data set.

Table 4.6: The appropriate range of partitioning window size of the four case studies: Mackey-Glass chaotic time-series, the monthly sunspots data, the gauge height data and the air temperature.

Study Case **Level of Missing**	**Mackey-glass**	**Sunspots**	**Gauge Height**	**Air Temp.**
10 %	10-50	20-50, 70-110	60-120	60-120
20 %	23-50	30-52, 70-100	60-140	60-130
30 %	24-50	40-100	53-120	65-120
40 %	30-50	42-100	67-125	70-120
50 %	35-54	39-90	58-120	60-120
60 %	35-56	39-85	55-123	60-150
70 %	30-60	40-120	55-120	70-110

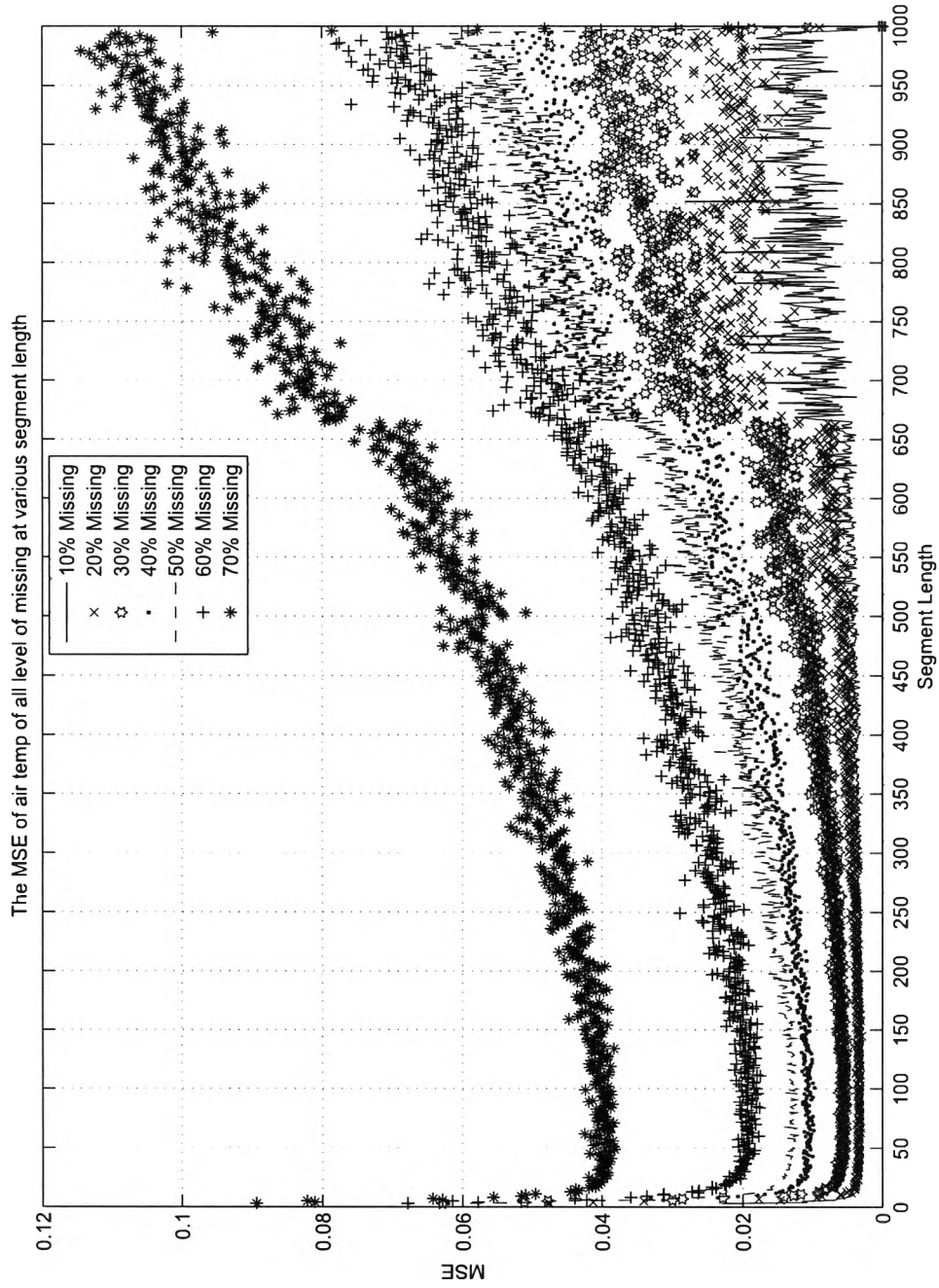


Figure 4.37: Scatter plots of the reference MSE values versus segment length for all levels of missing of the air temperature data set.

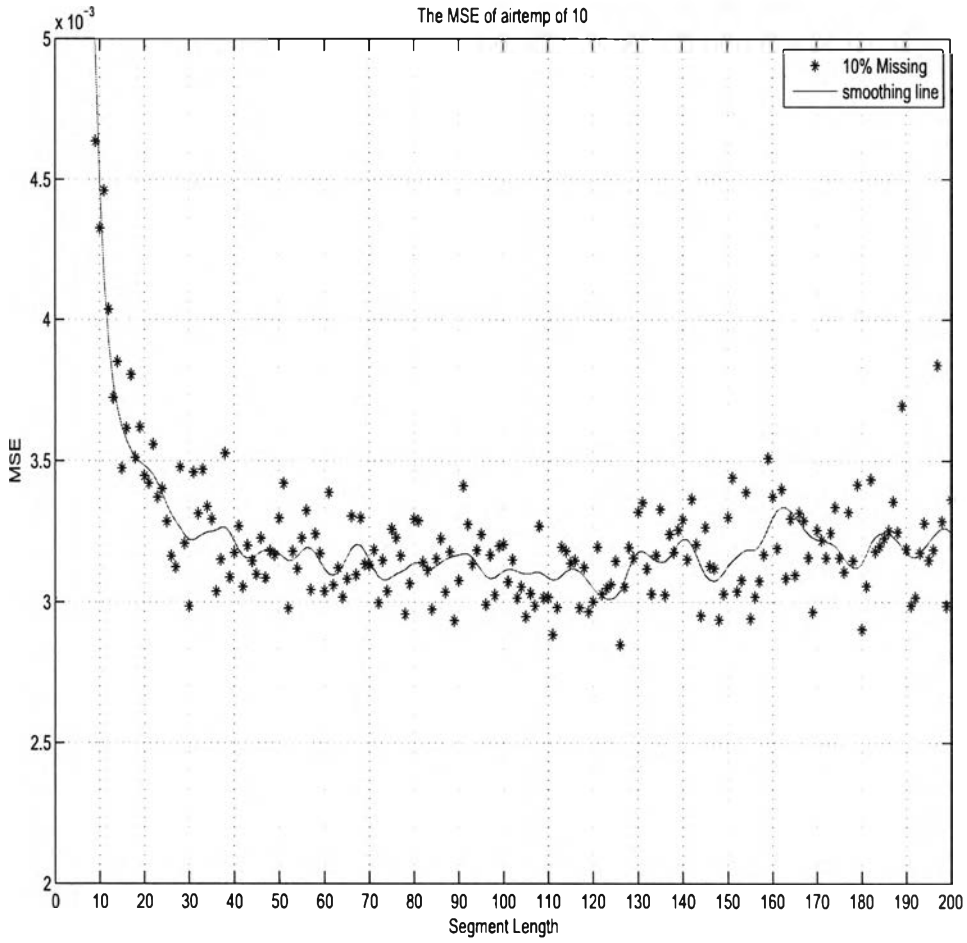


Figure 4.38: Scatter plots of the reference MSE values versus segment length at 10% missing of the air temperature data set.

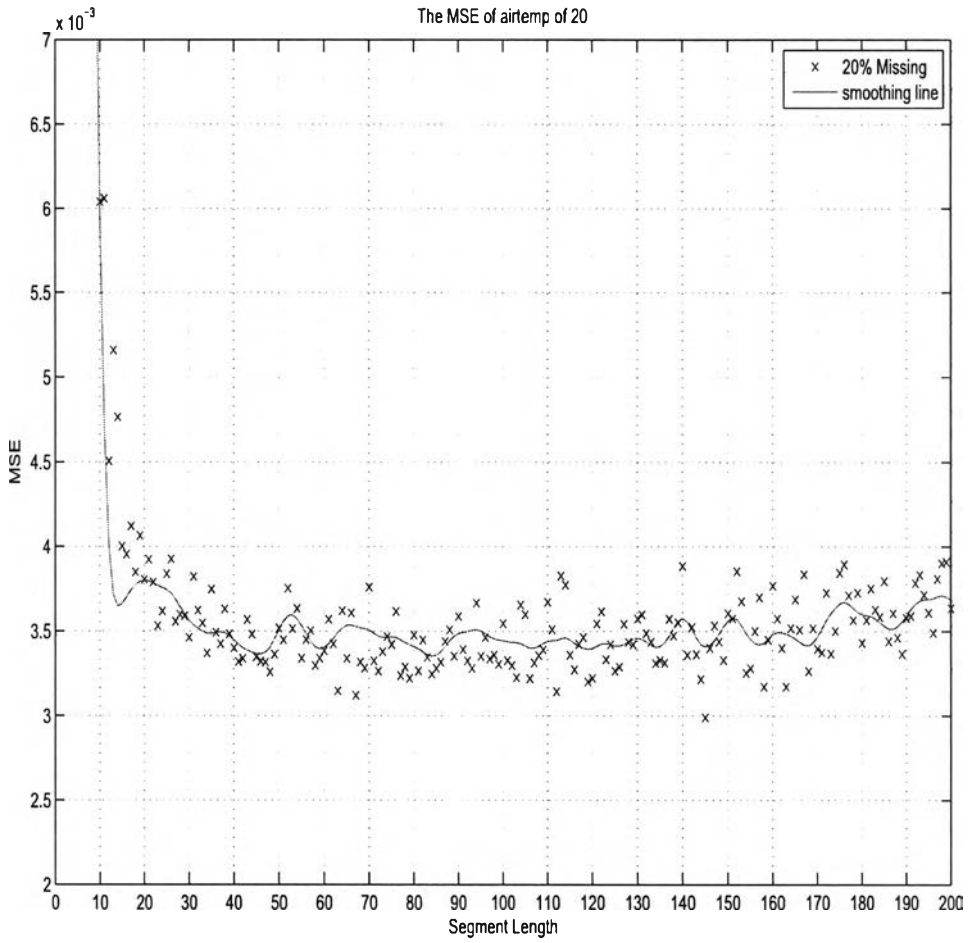


Figure 4.39: Scatter plots of the reference MSE values versus segment length at 20% missing of the air temperature data set.

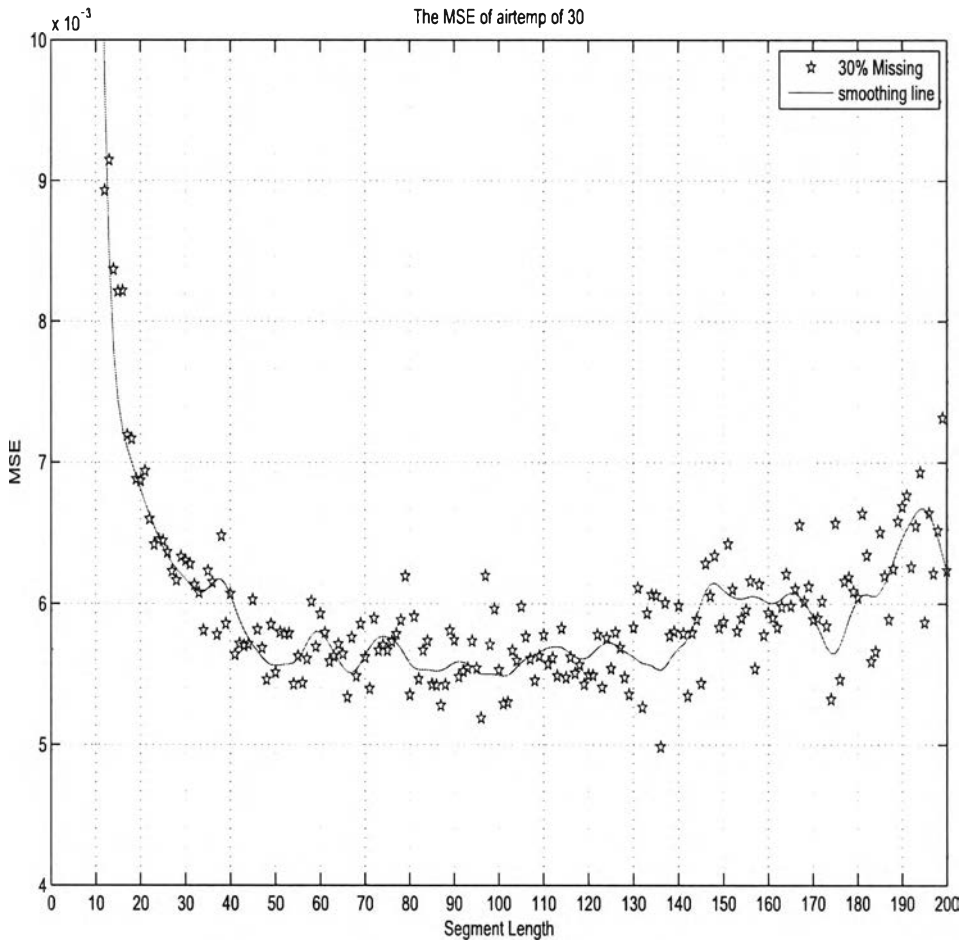


Figure 4.40: Scatter plots of the reference MSE values versus segment length at 30% missing of the air temperature data set.

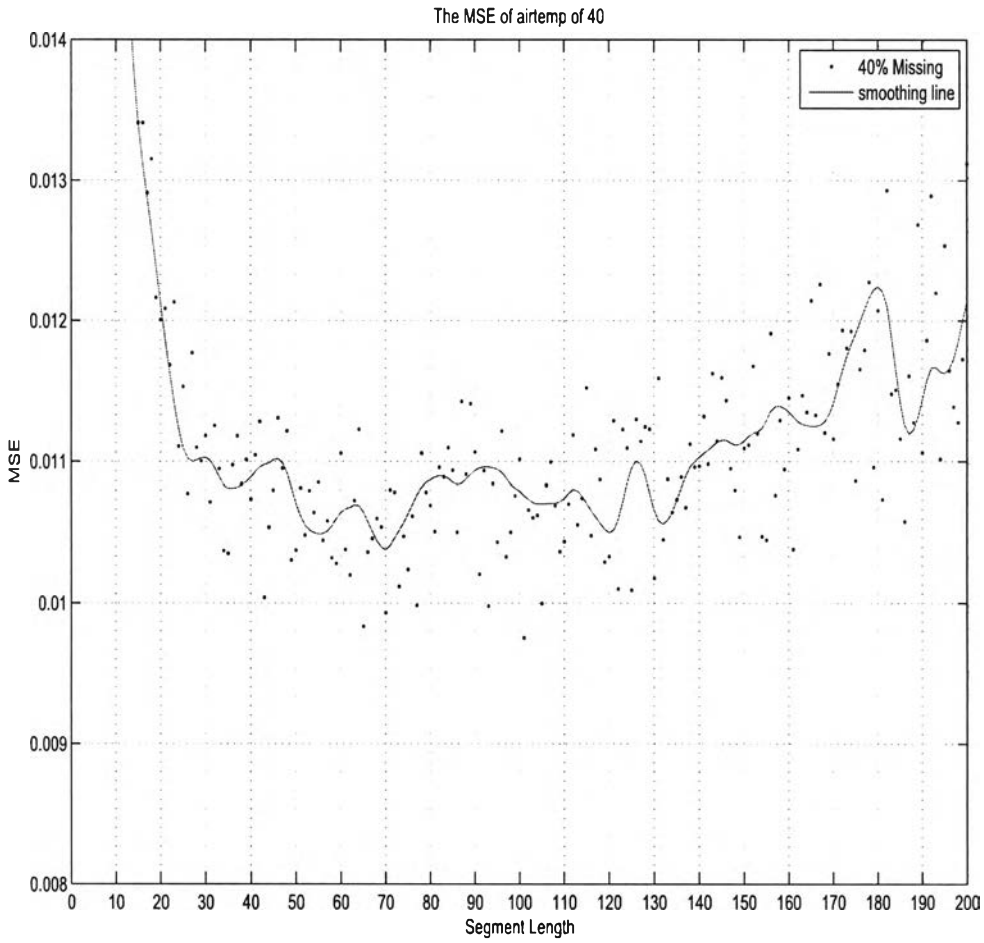


Figure 4.41: Scatter plots of the reference MSE values versus segment length at 40% missing of the air temperature data set.

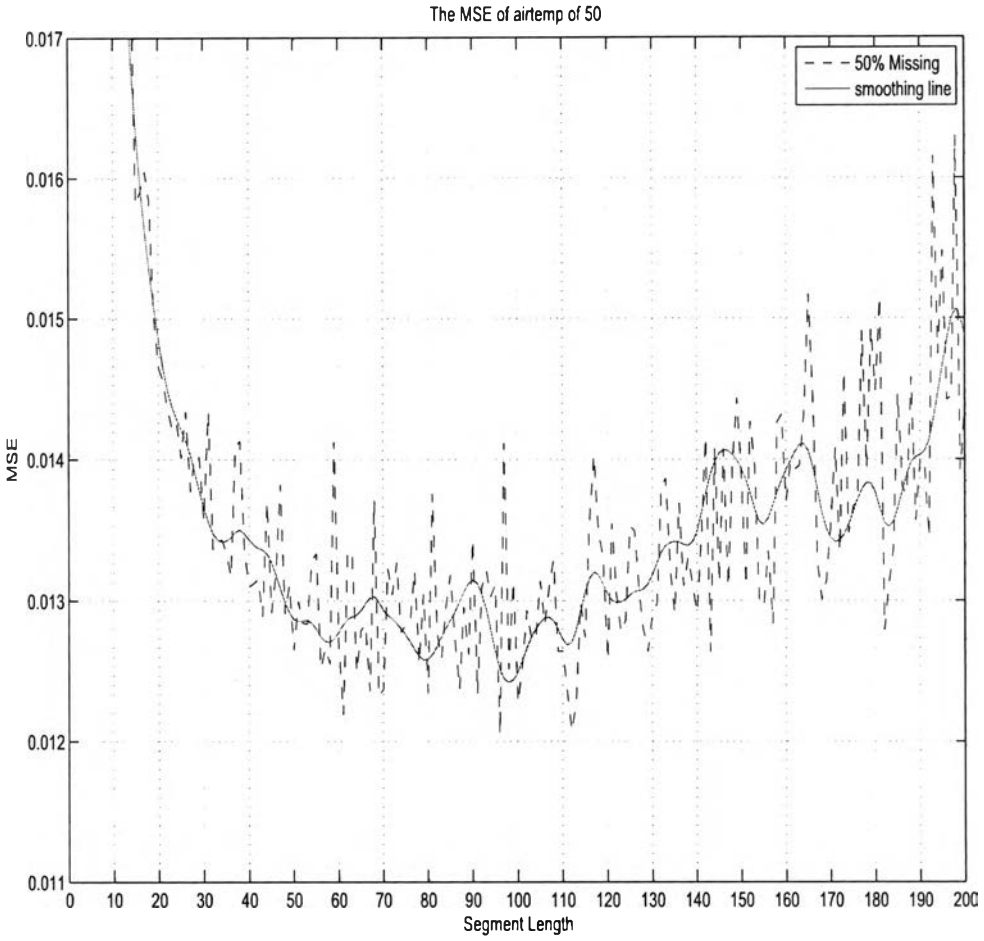


Figure 4.42: Scatter plots of the reference MSE values versus segment length at 50% missing of the air temperature data set.

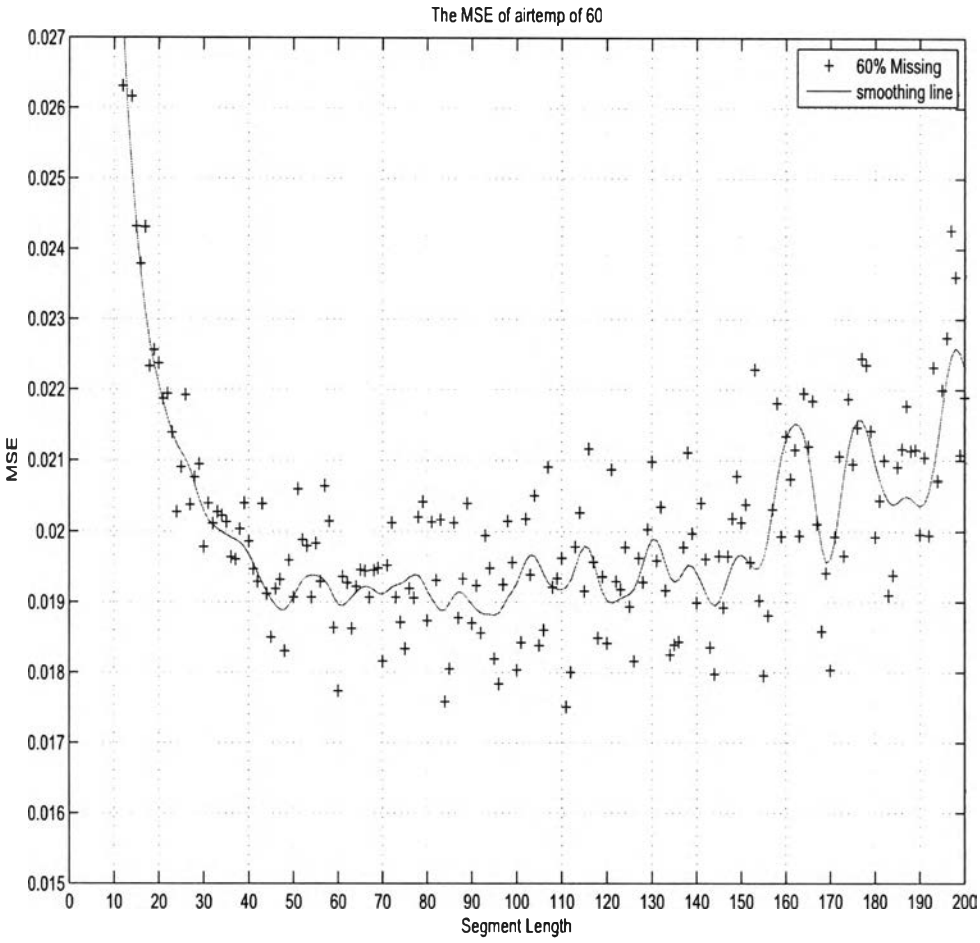


Figure 4.43: Scatter plots of the reference MSE values versus segment length at 60% missing of the air temperature data set.

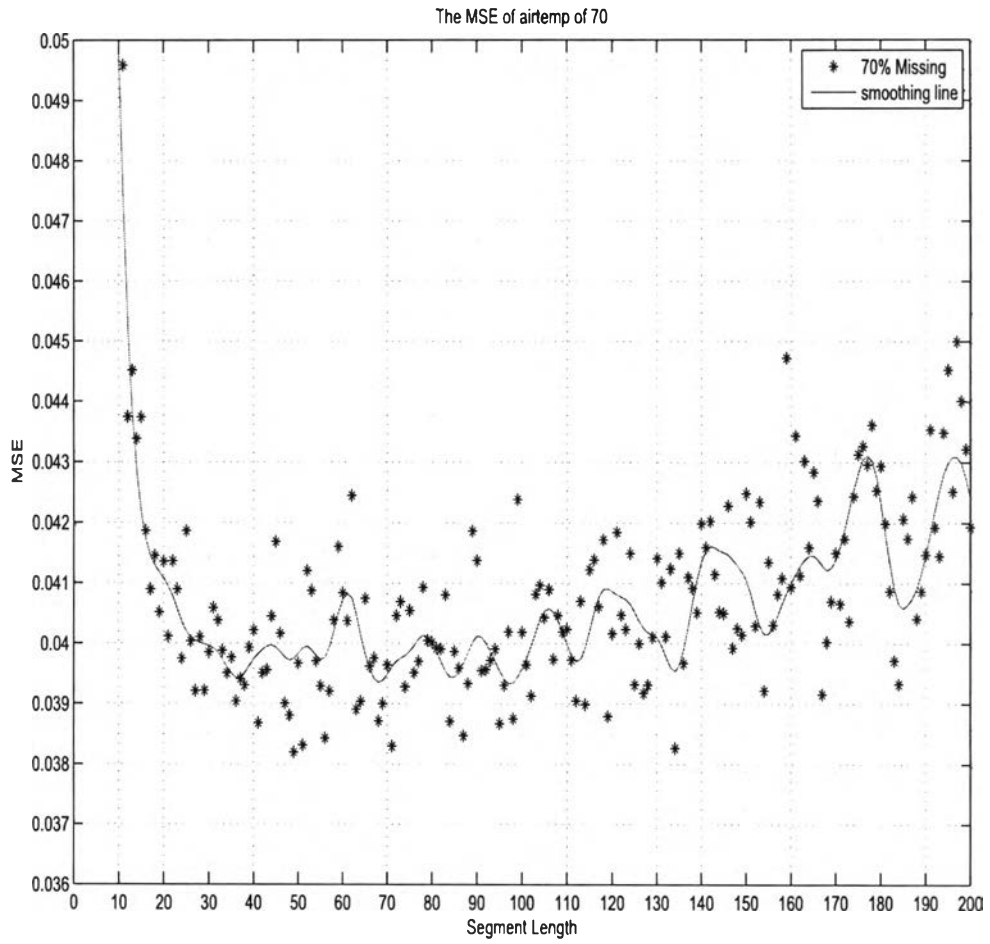


Figure 4.44: Scatter plots of the reference MSE values versus segment length at 70% missing of the air temperature data set.