Research Article

Processing Method of Missing Data in Dam Safety Monitoring

Wei Wei^{1,2,3}, Chongshi Gu^{1,2,3}, and Xiao Fu^{1,4}

¹College of Water Conservancy and Hydropower Engineering, Hohai University, Nanjing 210098, China

²State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Hohai University, Nanjing 210098, China

³National Engineering Research Center of Water Resources Efficient Utilization and Engineering Safety, Hohai University, Nanjing 210098, China

⁴Nanjing Hydraulic Research Institute, Nanjing 210029, China

Correspondence should be addressed to Chongshi Gu; csgu@hhu.edu.cn

Received 23 March 2021; Accepted 23 June 2021; Published 3 July 2021

Academic Editor: Tianwei Zhang

Copyright © 2021 Wei Wei et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

A large amount of data obtained by dam safety monitoring provides the basis to evaluate the dam operation state. Due to the interference caused by equipment failure and human error, it is common or even inevitable to suffer the loss of measurement data. Most of the traditional data processing methods for dam monitoring ignore the actual correlation between different measurement points, which brings difficulties to the objective diagnosis of dam safety and even leads to misdiagnosis. Therefore, it is necessary to conduct further study on how to process the missing data in dam safety monitoring. In this study, a data processing method based on partial distance combining fuzzy C-means with long short-term memory (PDS-FCM-LSTM) was proposed to deal with the data missing from dam monitoring. Based on the fuzzy clustering performed for the measurement points of the same category deployed on the dam, the membership degree of each measurement point to cluster center was described by using the fuzzy C-means clustering algorithm based on partial distance (PDS-FCM), so as to determine the clustering results and preprocess the missing data of corresponding measurement points. Then, the bidirectional long short-term memory (LSTM) network was applied to explore the pattern of changes of measurement values under identical clustering conditions, thus processing the data missing from monitoring effectively.

1. Introduction

As an indicator of the safety and work performance of concrete dam, the original measurement data always attract much attention for its integrity and accuracy [1–3]. However, due to the impact of such influencing factors, such as complicated and ever-changing operating conditions and human error, it is often difficult to prevent the actual measurement data from suffering a loss to varying extents [4]. The missing data will have an influence on the thorough assessment of dam for its serviceability, which can even lead to misjudgment in some cases. As for the traditional means to process the missing data, they include mean value processing, regression processing, expectation maximization processing, and so on [5, 6]. Nevertheless, these methods commonly suffer such problems due to the inability to reflect the correlation between samples or attributes, low processing accuracy, and limited applicability [7].

In recent years, as the study on machine learning and other emerging disciplines deepens [8–11], some new approaches for data processing have been proposed [12, 13]. By improving the k-nearest neighbors imputation method, Shichao Zhang proposed a gray processing method (GKNN) to replace the traditional Euclidean distance with gray distance [14]. Similarly, Tutz G put forward a distance-based weighted nearest neighbors imputation method [15]. However, the process for finding the adjacent samples for each incomplete sample as part of the method requires global traversal, which makes computation overly complex at the time of dealing with a huge amount of data. Chen applied LS-SVM [16] to fill the data, based on which the global optimal solution was worked out and small-scale data were made applicable. However, most of the SVM-based methods involve artificial judgment, which makes them heavily reliant on the personal experience gained by the

operator. Proposed by Takagi and Sugeno, the TS nonlinear regression model [17–19] is capable of describing nonlinear problems linearly by constructing local linear models and associating these linear models with membership function. However, this model ignores the correlation between input and output, which leads to the low accuracy in processing sequence information. Through the combination between autocoding and the genetic algorithm, Abdella et al. put forward a processing method which takes the missing data as the independent variable for cost function, and the genetic algorithm was applied to optimize the cost function of the missing data to be solved [20, 21]. Based on the aforementioned method, Nelwamondo et al. introduced the dynamic programming theory to build multiple self-encoders and then selected the optimal model for each incomplete sample to process the missing data [22]. However, such methods ignore the correlation between different sample attributes, which causes the accuracy of missing data processing to be low. The newly proposed autoregressive integrated moving average model (ARIMA) hybrid method shows the advantages of both the autoregressive integrated moving average and artificial neural network, which improves its generality. Due to the need for a large data sample size, the practicability of it is not as high as that required [23].

To address the problem mentioned above, this study proposes a missing data processing method based on the partial distance fuzzy C-means (PDS-FCM) model and long short-term memory (LSTM) network. Considering the correlation between the measurement points, it relies on the membership degree of each measurement point to cluster center to carry out cluster analysis and preprocess missing data. On this basis, the bidirectional LSTM network is introduced to further process the long-sequence missing data, thus ensuring the accurate processing of missing data. The basic principle is detailed as follows.

2. Processing Method of Dam Missing Data

First, the core of the missing data processing method proposed in this study is to use PDS [24] instead of Euclidean distance strategy to represent the correlation between the measurement points of a certain type of monitoring quantity. Then, the FCM clustering algorithm is adopted to construct the membership relationship between each measurement point and the cluster center, thus achieving fuzzy clustering for the measurement points and the preprocessing of missing data. Finally, the bidirectional LSTM network is introduced to construct the training model of missing data and to process the long series missing measurement values for the measurement points under identical clustering conditions, thus achieving the effective processing of missing data.

2.1. Preprocessing Method of Missing Data

2.1.1. Characterization of Correlation between Measurement Points. When the measurement points of a certain type of monitoring quantity encounter the missing measurement

value, it is difficult to measure the correlation between measurement points using those traditional distance measurement indicators. To solve this problem, this study applies PDS to characterize the correlation between different measurement points.

Suppose $X = \{x_i, x_i \in \mathbb{R}^n, i = 1, 2, \dots, n\}$ is an incomplete dataset with the number of measurement points n and the sequence of measurement value s . $x_i = [x_{i1}, x_{i2}, \dots, x_{is}]^T$, $i = 1, 2, \dots, n$, indicates the measurement value of the i^{th} measurement point. $I = [I_{il}] \in \mathbb{R}^{n \times s}$ represents a function describing the missing data of the measurement point, which is expressed as in the following equation:

$$I_{il} = \begin{cases} 0, & l = t^*, \\ 1, & l = t, \end{cases} \quad (1)$$

where I_{il} suggests that the measurement value of the i^{th} measurement point is complete at time t and missing at time t^* .

Based on equation (1), the correlation between x_i and x_k in the incomplete dataset X can be expressed as follows:

$$d_{\text{part}}(x_i, x_k) = \sqrt{\frac{s \sum_{l=1}^s I_{il} |x_{il} - x_{kl}|^2}{\sum_{l=1}^s I_{il}}}, \quad (2)$$

where d_{part} refers to the partial distance.

2.1.2. Fuzzy Clustering of Measurement Points and Preprocessing of Measurement Values. Based on equation (2) and FCM clustering algorithm, fuzzy clustering is performed in this section for the measurement points of a certain type of monitoring quantity, and the measurement value of the measurement point is preprocessed. Similar to most clustering algorithms, the FCM algorithm splits the dataset into a number of subsets according to the similarity between samples. Each subset represents a cluster, and the center of sample distribution in the cluster is taken as the cluster center. The difference is that this method achieves the transformation of membership from 0 or 1 to $[0, 1]$ by means of fuzzy processing. Equation (2) is applied to construct the PDS-FCM clustering model of the following equation:

$$\begin{cases} \min J(U, V) = \sum_{k=1}^K \sum_{i=1}^n u_{ik}^z d_{\text{part}}(x_i, v_k)^2, \\ u_{ik} \in [0, 1], \quad i = 1, 2, \dots, n, k = 1, 2, \dots, K, \\ \text{s.t.} \quad \sum_{k=1}^K u_{ik} = 1, \quad i = 1, 2, \dots, n, \\ 0 < \sum_{i=1}^n u_{ik} < n, \quad k = 1, 2, \dots, K, \end{cases} \quad (3)$$

where u_{ik} represents the membership degree, which indicates the degree of the measurement point x_i falling into the k^{th} cluster, denoted as $U = [u_{ik}] \in \mathbb{R}^{n \times K}$. v_k denotes the cluster center of the k^{th} cluster, and the corresponding center matrix is expressed as $V = [v_k] = [v_1, v_2, \dots, v_K]^T \in \mathbb{R}^{K \times s}$.

$z \in (1, \infty)$ is a fuzzy parameter, which indicates the degree to which these measurement points belong to each cluster.

Considering the equality constraints on membership degree in equation (3), the Lagrange multiplier method is used in this section to solve equation (3). The augmented Lagrange function applied to this process is as follows:

$$J_\lambda(U, V) = \sum_{k=1}^K \sum_{i=1}^n u_{ik}^z d_{\text{Part}}(x_i, v_k)^2 + \sum_{i=1}^n \lambda_i \left(\sum_{k=1}^K u_{ik} - 1 \right), \quad (4)$$

$$v_{kl} = \frac{\sum_{i=1}^n u_{ik}^z x_{il}}{\sum_{i=1}^n u_{ik}^z}, \quad l = 1, 2, \dots, s; k = 1, 2, \dots, K, \quad (5)$$

$$u_{ik} = \left[\sum_{t=1}^K \left(\frac{d_{\text{Part}}(x_i, v_k)^2}{d_{\text{Part}}(x_i, v_t)^2} \right)^{(1/(z-1))} \right]^{-1}, \quad i = 1, 2, \dots, n; k = 1, 2, \dots, K. \quad (6)$$

With the application of equations (5) and (6), the membership matrix U and center matrix V that meet the requirements of the accuracy of dam safety monitoring are obtained through iterative solution. The maximum membership of each measurement point to each cluster center is treated as the final clustering result.

According to the membership relationship between the missing data measurement points and the central points, the missing positions of the measurement values are preprocessed as follows:

$$\begin{cases} X_{il}^* = U_{ik} V_{kl}, & l = t^*, \\ X_{il}^* = X_{il}, & l = t, \end{cases} \quad (7)$$

where X_{il}^* denotes the measurement value of the i^{th} measurement point at the time l after preprocessing, and V_{kl} indicates the measurement value of the i^{th} center at the time l .

The ultimate clustering result is determined by equation (6), and the missing data are preprocessed by equation (7) based on which the data missing from dam safety monitoring can be effectively processed.

2.2. Implementation Technology of the PDS-FCM-LSTM Processing Method for the Data Missing from Dam Safety Monitoring. On the basis of the abovementioned research, the bidirectional LSTM network is introduced in this section to establish the mapping relationship between the measurement values of measurement points under identical clustering conditions, based on which the PDS-FCM-LSTM processing method is proposed.

2.2.1. Construction of Mapping Relationship between Measurement Values of the Measurement Points under Identical Clustering Conditions. The advantage of the LSTM network

where $\lambda = [\lambda_1, \lambda_2, \dots, \lambda_n]^T$ represents the Lagrange multiplier.

Based on equation (2) and the equality constraint of u_{ik} in equation (3), the necessary condition for the clustering objective function J_λ in equation (4) to reach the minimal is expressed as follows:

is that it is capable of controlling the information transfer of time series data by introducing input gate, forget gate, and output gate to the unit [9, 25–28]. It is assumed that the data of dam safety measurement point i are missing from time j to time $j + p$, which require reprocessing. For the unit at time l , the data x_{il}^* ($l < j$) of measurement point i are defined as output variables, while the data x_{ml}^* ($m \neq i, l < j$) of other measurement points in the same cluster are treated as input variables. On this basis, the mapping relationship between the measurement values of the forward LSTM network is established, as shown in Figure 1.

(1) Forget Gate. The output variable of the hidden layer h_{l-1} at time $l - 1$ and the input variable x_{ml}^* at time l are taken. Equation (8) is applied to construct the output variable of the forget gate f_l at time l , which determines whether to maintain the state of the hidden unit c_{l-1} at the previous layer.

$$f_l = \sigma(W_f \cdot [h_{l-1}, x_{ml}^*] + b_f), \quad (8)$$

where f_l represents the output variable after the processing by the forget gate at time l , h_{l-1} denotes the output variable of the hidden layer at time $l - 1$, x_{ml}^* indicates the input variable of each measurement point at time l after preprocessing, b_f refers to the bias term of the forget gate, W_f stands for the weight matrix of the forget gate, and σ means the activation function of sigmoid, which is expressed as follows:

$$\sigma(z) = \frac{1}{1 + \exp(-z)}. \quad (9)$$

(2) Input Gate. The information collected by the input gate is the same as that collected by the forget gate. Based on equations (9) and (10), the output variable of the input gate R_l at time l and the unit state c_l at time l are determined.

FIGURE 1: Schematic diagram of the LSTM unit.

$$R_l = \sigma(W_R \cdot [h_{l-1}, x_{ml}^*] + b_R), \quad (10)$$

$$c_l = f_l * c_{l-1} + R_l * \tanh(W_c \cdot [h_{l-1}, x_{ml}^*] + b_c), \quad (11)$$

where R_l represents the output value after processing by the input gate at time l , W_R indicates the input gate weight matrix, W_c denotes the unit state weight matrix, c_l means the unit state at time l , b_R refers to the bias term of input gate, and b_c stands for the bias term of unit state.

(3) *Output Gate.* The output gate is jointly determined by the updated unit state c_l , the output variable of the hidden layer h_{l-1} at time $l-1$, and the input variable x_{ml}^* at time l . The output variable of the output gate o_l at time l and the output variable of the hidden layer h_l at time l can be determined as follows:

$$\begin{aligned} o_l &= \sigma(W_o \cdot [h_{l-1}, x_{ml}^*] + b_o), \\ h_l &= o_l * \tanh(c_l), \end{aligned} \quad (12)$$

where o_l represents the output variable processed by the output gate at time l , W_o denotes the output gate weight matrix, and b_o indicates the bias term of the output gate.

To process the missing data of measurement value from time j to time $j+p$ more reasonably, a new bidirectional LSTM network method is proposed to process the data from time j to time $j+p$. As shown in Figure 2, model training and validation are carried out on the measurement points before and after time j to time $j+p$, respectively, according to the aforementioned principles based on which the forward and reverse LSTM model parameters are obtained. Then, the remaining measurement point data of the

FIGURE 2: Bidirectional LSTM processing method and idea (the gray box represents the preprocessed value of discrete data and the red box represents the long-sequence missing data).

complete information from time j to time $j + p$ are input into the model. Finally, the calculation value of the measurement point with missing data corresponding to the period from time j to time $j + p$ is obtained, so as to process the missing data bidirectionally.

2.2.2. Implementation Method of the PDS-FCM-BILSTM Processing Method for Dam Missing Data. According to the PDS-FCM-LSTM method for dam missing data processing as proposed above, the implementation process is shown in Figure 3.

Step 1: input the missing dataset of the dam measurement point

Step 2: enter the PDS-FCM layer to preprocess the missing data, set fuzzy parameter z , cluster number K , and threshold ε ($\varepsilon > 0$), and then randomly initialize the partition matrix $U^{(0)}$

Step 3: when the number of iterations is n ($n \geq 1$), equation (5) is used to update the prototype matrix $V^{(n)}$ based on the partition matrix $U^{(n-1)}$

Step 4: update the partition matrix $U^{(n)}$ based on equation (6) and the prototype matrix $V^{(n)}$

Step 5: if the conditions $\forall k, i: \max |u_{ik}^{(n)} - u_{ik}^{(n-1)}| < \varepsilon$ are met, where $k = 1, 2, \dots, K$, $i = 1, 2, \dots, n$, the algorithm stops to output partitioning matrix U and prototype matrix V . Otherwise, if $n \leftarrow n + 1$, return to Step 3.

Step 6: based on the membership relationship between the measurement points with missing data and each cluster center (U, V) , the existing values in the cluster center are weighted and averaged to obtain the preprocessed data X^*

Step 7: establish the bidirectional LSTM processing model for the missing data of the dam. The forward LSTM network model and reverse LSTM network model are constructed according to the time required to process the measurement values.

Step 8: input the measurement values of each measurement point in the same cluster before and after the time slot T into the forward and reverse LSTM models, so as to obtain the weight W and bias b of each parameter after training and learning

Step 9: input the preprocessed measurement point data as sample into the forward and backward LSTM models for obtaining the forward and backward output values at time slot T , with their mean values taken as the ultimate result of missing data processing

According to the above implementation steps, the implementation concept of PDS-FCM-LSTM is obtained.

2.2.3. Evaluation of the Missing Data Processing Effect. In this study, the average absolute error MAE and the average absolute percentage error MAPE are taken as the evaluation indicators of the data processing effect for the measurement points, which are defined as follows:

$$MAE = \frac{1}{|\hat{X}_m|} \sum_{\hat{x}_{ij} \in \hat{X}_m} |r_{ij} - \hat{x}_{ij}|, \quad (13)$$

$$MAPE = \sqrt{\frac{1}{|\hat{X}_m|} \sum_{\hat{x}_{ij} \in \hat{X}_m} \left| \frac{r_{ij} - \hat{x}_{ij}}{r_{ij}} \right|}, \quad (14)$$

where $\hat{X}_m$ represents the set of all processing values, $\hat{x}_{ij}$ denotes the processing value, r_{ij} indicates the measurement value of the processing value, and $\hat{x}_{ij}$ corresponds to the measurement value of the measurement point with complete information.

Using equations (13) and (14) and the error allowable values listed in the technical specifications for dam safety monitoring, the effect of missing data processing is thoroughly evaluated, and the processed data meeting the

FIGURE 3: PDS-FCM-LSTM concept.

accuracy requirements are taken as the effective data for dam safety monitoring.

3. Numerical Example

3.1. Project Summary. The dam is a concrete parabolic double curvature arch dam [29–32] with the height of 240 m and the crest arc length of 774.69 m. It consists of 39 dam sections from left to right. In order to monitor the horizontal displacement of arch crown, arch shoulder, and 1/4 arch ring, 8 inverted plumb lines and 10 direct plumb lines are deployed in the dam body to form 5 groups of direct and inverted plumb lines, forming a total of 20 measurement

points. The plumb survey points on the left shoulder of No.4 dam section are TCN01~TCN02, TCN03~TCN06 on No.11 dam section, TCN07~TCN14 on No.19 ~23 dam section, TCN015~TCN18 on No.33 dam section, and TCN19~TCN20 on the right shoulder of No.37 dam section. The specific distribution of the measurement points is shown in Figure 4.

3.2. The Data Processing. Based on the time series of complete measurement data on the deformation to plumb lines from January 1, 2014, to December 31, 2019, the experiment is designed. Take TCN11 as an example, and the

FIGURE 4: Layout of measurement points of a concrete arch dam.

FIGURE 5: TCN11 measurement values.

measurement data (31 data in total) from October 27, 2016, to March 29, 2017, are purposely deleted to construct the incomplete dataset (the data of other measurement points are complete). The missing data are shown in Figure 5.

The PDS-FCM-LSTM method proposed in this study is adopted for data processing. First, the PDS-FCM data processing model is constructed for the dam deformation measurement points. According to the measurement values obtained from the dam deformation measurement points between January 1, 2014, and December 31, 2019, the partial distance between each measurement point and the membership matrix U is calculated when the minimum value is taken for the objective function J_λ . The results can be seen from Figure 6.

Then, according to the clustering principle proposed in the second section, all measurement points are clustered. The clustering results are given in Tables 1 and 2. It can be seen that the measurement points are divided into multiple clusters; TCN11 belongs to the second cluster, and TCN04, 05, and 16 are also one of them. According to equation (7), the preprocessing result of the measurement point TCN11 is shown in Figure 7.

Based on the above processing results, the data of measurement points TCN04, 05, 11, and 16 can be taken as the basic data to be used for secondary processing.

Finally, based on the integrity of the measured values before October 27, 2016, and after March 29, 2017, after preprocessing, the bidirectional LSTM network is used to process the long-sequence missing data of the measurement point TCN11. To validate the missing data processing method proposed in this study, the data missing from measurement point TCN11 between October 27, 2016, and March 29, 2017, are exemplified. It is processed by using the PDS-FCM-LSTM processing method, the LSTM network of single measurement point, and PDS-FCM processing method, respectively. The results are shown in Figure 8 and Tables 3 and 4.

It can be seen from Figure 8 and Tables 3 and 4 that the PDS-FCM-LSTM processing method proposed in this study achieves the highest processing accuracy because it gives full consideration to the correlation between measurement points and the pattern of changes in measurement values. PDS-FCM takes into account only the correlation between the measurement points, which makes the outcome of

FIGURE 6: Measurement points of each center cluster.

treatment less satisfactory. The single-point LSTM performs the worst in terms of the processing effect because it considers only the pattern of changes in measurement values while ignoring the correlation between the measurement points. In addition, it can be seen from Table 3 that the

absolute errors of the outcome of PDS-FCM-LSTM treatment proposed in this study are all less significant than that specified for the water project (29), which demonstrates the effectiveness of the missing data processing method proposed in this study.

TABLE 1: Membership degree of each deformation measurement point to clustering center.

Measurement point	Membership of center 1	Membership of center 2	Membership of center 3	Membership of center 4	Membership of center 5
TCN1	0.93	0.00	0.01	0.00	0.06
TCN2	0.16	0.01	0.08	0.00	0.76
TCN3	0.97	0.00	0.00	0.00	0.02
TCN4	0.01	0.93	0.02	0.03	0.01
TCN5	0.06	0.53	0.26	0.06	0.09
TCN6	0.07	0.00	0.03	0.00	0.90
TCN7	0.04	0.00	0.01	0.00	0.95
TCN8	0.05	0.02	0.74	0.01	0.18
TCN9	0.01	0.03	0.01	0.94	0.01
TCN10	0.01	0.10	0.02	0.86	0.01
TCN11	0.00	0.96	0.01	0.01	0.01
TCN12	0.05	0.06	0.78	0.01	0.10
TCN13	0.00	0.00	0.98	0.00	0.01
TCN14	0.05	0.01	0.75	0.00	0.19
TCN15	0.68	0.01	0.04	0.00	0.28
TCN16	0.01	0.94	0.01	0.03	0.01
TCN17	0.12	0.02	0.45	0.01	0.40
TCN18	0.97	0.00	0.00	0.00	0.02
TCN19	0.96	0.00	0.01	0.00	0.03
TCN20	0.85	0.00	0.02	0.00	0.13

TABLE 2: Clustering results of deformation measurement points.

Cluster	Measurement points
Cluster 1	TCN01, TCN03, TCN15, TCN18, TCN19, TCN20
Cluster 2	TCN04, TCN05, TCN11, TCN16
Cluster 3	TCN08, TCN12, TCN13, TCN14, TCN17
Cluster 4	TCN09, TCN10
Cluster 5	TCN02, TCN06, TCN07

FIGURE 7: Preprocessing results of measurement values at measurement point TCN11.

FIGURE 8: Comparison of different processing methods for TCN11 measurement values.

TABLE 3: Results of missing data processing for TCN11 by different methods.

Date	True value (mm)	PDS-FCM (mm)	LSTM (mm)	PDS-FCM-LSTM (mm)
2016/10/27	26.36	29.19	29.26	26.47
2016/11/3	26.53	29.30	29.90	26.51
2016/11/11	26.45	29.43	29.48	26.43
2016/11/17	26.20	28.92	28.51	26.40
2016/11/24	25.79	28.43	28.29	26.14
2016/12/1	25.40	27.81	27.96	25.93
2016/12/8	25.19	27.53	27.67	25.63
2016/12/15	25.18	27.62	27.60	25.31
2016/12/22	25.16	27.61	27.41	25.13
2016/12/29	25.28	27.74	27.00	24.99
2017/1/5	25.40	28.19	26.65	24.81
2017/1/13	25.40	28.26	26.47	24.72
2017/1/19	25.28	27.80	26.35	24.65
2017/1/23	25.14	27.61	26.45	24.56
2017/2/3	23.55	25.18	26.50	24.26
2017/2/9	23.01	24.33	25.63	23.88
2017/2/16	22.64	23.74	25.20	23.12
2017/2/23	21.93	22.89	24.85	22.51
2017/3/1	21.75	22.28	24.28	22.53
2017/3/9	21.48	22.13	23.51	22.67
2017/3/16	21.26	21.87	22.65	22.40
2017/3/23	20.74	21.05	23.03	22.32
2017/3/29	20.43	20.50	21.75	21.99

TABLE 4: Evaluation of data processing results for TCN11.

Measurement point	Processing method	MAE	MAPE
TCN11 [#]	BILSTM	2.21	0.08
	PDS_FCM	1.91	0.07
	PDS-FCM-BILSTM	0.58	0.03

4. Conclusion

- (1) To solve the problem that the traditional distance measurement indicators are ineffective in measuring the correlation between the measurement points when the dam suffers incomplete monitoring information, the method for characterizing the correlation between survey points is studied and proposed on the basis of partial distance. Then, the construction technology of the PDS-FCM clustering model is studied, which achieves the preprocessing of data missing from dam safety monitoring and lays a foundation for the effective processing of missing data.
- (2) Considering the correlation and trend of the changes in effect size for monitoring, a PDS-FCM-LSTM processing method intended for the data missing from dam safety monitoring is proposed. It combines the advantages of PDS-FCM and LSTM. The effectiveness of the proposed method is verified in practical engineering application. As the proposed method is universally applicable, it is suited for missing data processing in similar projects.

Data Availability

The data used to support the findings of this study are from a large water conservancy project and not suitable to upload to the network. The data are included within the Supplementary Information files.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (51779086, 52079046, and 51609074) and the Fundamental Research Funds for the Central Universities (B210202017).

Supplementary Materials

Experimental data.pdf provides all measured values of all measuring points of a concrete dam from January 2, 2014, to December 31, 2019. It should be noted that TCN1-20 refers to measuring points 1–20 and the green box represents the data deleted purposely. . (Supplementary Materials)

References

- [1] T. Gebler, "Messdatenauswertung von gewichtsstauauern zur identifikation von schäden," *Wasserwirtschaft*, vol. 110, no. 5, pp. 16–21, 2020.
- [2] Z. X. Xing, L. Ni, and R. J. Ma, "Temperature measurement and data analysis of the roller compacted concrete dam," *Applied Mechanics and Materials*, vol. 212–213, pp. 942–946, 2012.
- [3] H. Gu, T. Wang, Y. Zhu, C. Wang, D. Yang, and L. Huang, "A completion method for missing concrete dam deformation monitoring data pieces," *Applied Sciences*, vol. 11, no. 1, p. 463, 2021.
- [4] W. Ge, X. Wang, Z. Li et al., "Interval analysis of the loss of life caused by dam failure," *Journal of Water Resources Planning and Management*, vol. 147, no. 1, Article ID 04020098, 2021.
- [5] S. Meeyai, "Logistic regression with missing data: a comparison of handling methods, and effects of percent missing values," *Journal of Traffic and Logistics Engineering*, vol. 4, no. 2, 2016.
- [6] A. P. Dempster, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of Royal Statistical Society*, vol. 39, pp. 1–22, 1977.
- [7] R. Deb, A. Wee-Chung Liew, and E. Oh, "A correlation based imputation method for incomplete traffic accident data," in *Proceedings of the The 13th Pacific Rim International Conference on Artificial Intelligence*, pp. 905–912, Springer, Gold Coast, Australia, December 2014.
- [8] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.
- [9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [10] I. Goodfellow, Y. Bengio, and A. Courville, "Machine learning basics," *Deep Learning*, vol. 1, pp. 98–164, 2016.
- [11] L. Zhang, J. Tan, D. Han, and H. Zhu, "From machine learning to deep learning: progress in machine intelligence for rational drug discovery," *Drug Discovery Today*, vol. 22, no. 11, pp. 1680–1685, 2017.
- [12] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, John Wiley & Sons, Hoboken, NJ, USA, 2019.
- [13] I. G. Dragset, "Analysis of longitudinal data with missing values," MS Thesis, Norwegian University of Science and Technology, Trondheim, Norway, 2009.
- [14] S. Zhang, "Nearest neighbor selection for iteratively kNN imputation," *Journal of Systems and Software*, vol. 85, no. 11, pp. 2541–2552, 2012.
- [15] G. Tutz and S. Ramzan, "Improved methods for the imputation of missing data by nearest neighbor methods," *Computational Statistics & Data Analysis*, vol. 90, pp. 84–99, 2015.
- [16] T.-T. Chen and S.-J. Lee, "A weighted LS-SVM based learning system for time series forecasting," *Information Sciences*, vol. 299, pp. 99–116, 2015.
- [17] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE Transactions on Systems, Man, and Cybernetics*, pp. 116–132, 1985.
- [18] J. A. Roubos, S. Molloy, R. Babuška, and H. B. Verbruggen, "Fuzzy model-based predictive control using Takagi-Sugeno models," *International Journal of Approximate Reasoning*, vol. 22, no. 1–2, pp. 3–30, 1999.
- [19] C. Fantuzzi and R. Rovatti, "On the approximation capabilities of the homogeneous Takagi-Sugeno model," in *Proceedings of the IEEE 5th International Fuzzy Systems*, pp. 1067–1072, IEEE, New Orleans, LA, USA, September 1996.
- [20] M. Marseguerra and A. Zoia, "The autoassociative neural network in signal analysis: II. application to on-line monitoring of a simulated BWR component," *Annals of Nuclear Energy*, vol. 32, no. 11, pp. 1207–1223, 2005.
- [21] M. Abdella and T. Marwala, "The use of genetic algorithms and neural networks to approximate missing data in database," in *Proceedings of the IEEE 3rd International Conference on Computational Cybernetics*, 2005. ICC3 2005, pp. 207–212, IEEE, Port Louis, Mauritius, April 2005.

- [22] F. V. Nelwamondo, D. Golding, and T. Marwala, "A dynamic programming approach to missing data estimation using neural networks," *Information Sciences*, vol. 237, pp. 49–58, 2013.
- [23] Ü. Ç. Büyüksahin and Ş. Ertekin, "Improving forecasting accuracy of time series data using a new ARIMA-ANN hybrid method and empirical mode decomposition," *Neuro-computing*, vol. 361, pp. 151–163, 2019.
- [24] R. J. Hathaway and J. C. Bezdek, "Fuzzy c-means clustering of incomplete data," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 31, no. 5, pp. 735–744, 2001.
- [25] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: continual prediction with LSTM," in *Proceedings of the 1999 Ninth International Conference on Artificial Neural Networks ICANN 99*, Edinburgh, UK, September 1999.
- [26] O. Vinyals, L. Kaiser, T. Koo et al., "Grammar as a foreign language," 2014, <https://arxiv.org/abs/1412.7449>.
- [27] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6645–6649, IEEE, Vancouver, Canada, May 2013.
- [28] C. Gulcehre, K. Cho, R. Pascanu, and Y. Bengio, "Learned-norm pooling for deep feedforward and recurrent neural networks," in *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 530–546, Springer, Ghent, Belgium, September 2014.
- [29] S. Chen, C. Gu, C. Lin, and M. A. Hariri-Ardebili, "Prediction of arch dam deformation via correlated multi-target stacking," *Applied Mathematical Modelling*, vol. 91, pp. 1175–1193, 2021.
- [30] J. Hu and F. Ma, "Statistical modelling for high arch dam deformation during the initial impoundment period," *Structural Control and Health Monitoring*, vol. 27, Article ID e2638, 2020.
- [31] D. Yuan, B. Wei, B. Xie et al., "Modified dam deformation monitoring model considering periodic component contained in residual sequence," *Structural Control and Health Monitoring*, vol. 27, Article ID e2633, 2020.
- [32] B. Wu, J. Niu, H. Su et al., "An approach for deformation modulus mechanism of super-high arch dams," *Structural Engineering and Mechanics*, vol. 69, pp. 557–566, 2019.