

Research on Bridge Monitoring Data Fusion Algorithm based on Temperature Correlation

Yujie Yang¹, Jiaye Wu^{2,3}, Changrui Tan³, Chunlin He¹

¹ Sichuan University of Science & Engineering, Zigong, Sichuan, 643000, China

² Southwest Petroleum University, Chengdu 610500, China

³ Sichuan Central Inspection Technology Inc., Zigong 643030, China

Abstract

Firstly, the bridge monitoring data is preprocessed, and then the correlation coefficient between deflection and temperature is obtained by Pearson correlation coefficient, indicating that there is a strong correlation between temperature and deflection. In machine learning, random forest has certain randomness, which can avoid overfitting, and has improved generalization ability than a single tree model. XGBoost uses Boost learning theory to strengthen weak links in learning, which is beneficial to better learning on large-scale complex data sets, and the effect is more obvious. Therefore, this paper proposes a fusion model of random forest and XGBoost to predict bridge monitoring data, and introduces temperature correlation as training data to input into the fusion model to improve prediction accuracy.

Keywords

Bridge Health Monitoring; Pearson Coefficient; Random Forest and XGBoost Prediction; Temperature Correlation Prediction.

1. Introduction

Weather changes, environmental erosion, natural disasters and increasing traffic loads will constantly change the performance of Bridges, and even lead to the degradation of Bridges in long-term use [1]. The application of bridge health monitoring has been recognized as an attractive tool that can improve bridge health and safety and provide early warning of structural damage [2-4]. A typical BHM system typically provides a variety of useful real-time information, such as temperature, cracks, deflection, and strain. These information can be used to judge the health of the bridge according to the corresponding diagnostic methods. Therefore, bridge health monitoring and safety assessment are the inevitable requirements for the sustainable development of bridge engineering.

The detection efficiency of sensors installed in bridge construction gradually decreases during long-term operation, which affects the reliability of monitoring data. In addition, poor environmental conditions also have a serious impact on the quality of data collected [2]. In addition, the types of data collected are different and the amount of data is huge. Therefore, the difficulty in the current research and application of bridge health monitoring system is often not the collection of monitoring data, but how to extract useful information from the long-term accumulation of massive data [5].

Machine learning is an effective method to predict bridge monitoring data. Linear regression, support vector machine, neural network and so on have been used to predict bridge monitoring data. In literature [6], support vector machine and particle swarm optimization were combined to predict the bridge earthquake damage. XGBoost model in literature [7] was used to predict the travel time of expressway using the data of detecting vehicles, and XGBoost model had considerable advantages in forecasting accuracy and efficiency. Literature [8] uses BP neural network algorithm to create a long-







span bridge structural response prediction model. Literature [9] proposes a bridge prediction data analysis method based on LSTM neural network. The results show that compared with traditional BP neural network, LSTM model has a higher performance.

2. Data Set Description and Modeling Techniques

2.1 Data Set Specification

The monitoring object is the bridge of Beijing Bailiwei Logistics Park. The bridge logistics park consists of 4-lane road surface and unloading platform. The main structure of the bridge logistics park is beam and plate structure, with a length of about 527m and a width of 16m. Due to the large increase of large tonnage logistics transport vehicles with the increase of business volume, it is easy to cause the bearing capacity of bridge pavement is too large, which greatly increases the load of bridge pavement, resulting in abnormal deformation of bridge, resulting in concrete cracks, spalling and other phenomena. The main monitoring items of the system include: bridge structural stress and strain, vibration characteristics, bridge deflection, structural temperature, and environmental monitoring. The main hardware of the monitoring system is shown in Table 1. In bridge structures, the characteristics of beam deflection (according to which the bridge can make a timely response to the state of the bridge under the combined action of load and external environment, because it is not easy to produce noise and is sensitive to structural damage) can truly reflect the health of the bridge. Therefore, we used a random forest and XGBoost fusion model based on temperature correlation to predict deflection data in bridge health monitoring.

Table 1. Summary of main hardware of monitoring system

Monitored Parameters	Device Name	Quantity	Layout instructions	Equipment picture
structural stress/strain	Vibrating wire extensometer	23	Middle beam bottom of monitored beam span	
Deflection/Settlement	Hydrostatic leveling	16	The monitored beam is measured in the middle span	
Humiture	Temperature and humidity sensor	1	Midspan side of monitored beam	
vibration	accelerated speed	4	Midspan side of monitored beam	
Data collection	Multi-function data acquisition instrument	3	Two 1# cabinets and one 2# cabinet	
	Industry cabinet	2	Cabinet 1# is placed on the inside of the column where axis 5-2 intersects with axis J, and cabinet 2# is placed on the inside of the column where axis 6-3 intersects with axis H	

2.2 Data Pre-processing

Due to the massive bridge monitoring data affected by environmental temperature, system error, bridge load and other factors, it will produce complex interference. Therefore, in the face of massive data, preliminary processing is required. The commonly used methods are to correct abrupt values and delete outliers. These methods are used to reduce the errors generated in the analysis process, and try to make the collected signal the same as the real signal to the greatest extent, so as to prevent excessive noise signals in the collected signal from being able to truly express the bridge structure [14]. The analysis error of the final result is too large.

Tens of thousands of data of temperature, vertical displacement, strain and crack of an interconnecting prestressed continuous bridge were collected from 0:00 on January 1, 2023 to 23:00 on June 8, 2023. In this paper, temperature and vertical displacement were taken as an example and processed as follows:

1) When the monitoring value has abnormal value or abrupt value, the abnormal value is deleted and the abrupt value is corrected.

2) When there is missing data, the missing value is processed by means of mean insertion, Lagrange interpolation, time series interpolation and neural network interpolation.

After the outliers are processed, the data is aggregated. The original data is recorded in minutes, for example, 2023/2/8 20:19:00. There are 60 records per hour, for example, 2023/2/8 21:00:00 to 2023/2/8 21:59:00. The time data is segmented to extract "year-month-day-hour", such as "2023-2-8-21", and then the data is aggregated by hour to obtain 24 records in a day, such as "23-2-8-00" to "23-2-8-23", as shown in the figure. This research focuses on the analysis and modeling of hour-level data. Some of the data are shown in Table 1.

Table 2. Partially preprocessed data

Time	Mean Value	Minimum Value	Maximum Value
2023-01-01-21	-1.2874385	-1.28015	-1.29295
2023-01-01-22	-1.277507667	-1.25761	-1.28674
2023-01-01-23	-1.247061333	-1.23543	-1.26227
2023-01-02-00	-1.252282667	-1.24742	-1.25874
2023-01-02-01	-1.250803667	-1.22591	-1.25911
2023-01-02-02	-1.222922667	-1.211	-1.23267
2023-01-02-03	-1.2312455	-1.21686	-1.23572
2023-01-02-04	-1.163424167	-1.13533	-1.21809
2023-01-02-05	-1.162798667	-1.14336	-1.17799
2023-01-02-06	-1.172277667	-1.15878	-1.18784
2023-01-02-07	-1.191281667	-1.17433	-1.2003
2023-01-02-08	-1.165232667	-1.15953	-1.18068
2023-01-02-09	-1.164970167	-1.15453	-1.17683
2023-01-02-10	-1.186745333	-1.17198	-1.20015
2023-01-02-11	-1.226236167	-1.19928	-1.24927
2023-01-02-12	-1.2696645	-1.24733	-1.28787
2023-01-02-13	-1.321911667	-1.28571	-1.35509
2023-01-02-14	-1.3572055	-1.34926	-1.36713
2023-01-02-15	-1.3510245	-1.33818	-1.35797
2023-01-02-16	-1.326253333	-1.30405	-1.34113
2023-01-02-17	-1.295575833	-1.2812	-1.30841

As can be seen from the above table, the average value, minimum value and maximum value of deflection were obtained after data processing, and the average value was taken as data verification, and a total of 3647 sets of data were obtained.

2.3 Correlation Analysis based on Pearson Coefficient

In a bridge health monitoring system, there is a certain relationship between different monitoring data, for example, when the temperature changes, the strain, deflection and cracks will change because of the temperature difference. The Pearson correlation coefficient is a statistical measure used to measure the degree of linear correlation between two continuous variables. Its value ranges from -1 to 1, indicating the strength and direction of the variables, and the greater the absolute value, the stronger the correlation. Pearson correlation coefficient is calculated as follows:

$$\gamma_{\alpha,\beta} = \frac{m \sum_{i=1}^m \alpha_i \sum_{i=1}^m \beta_i - \sum_{i=1}^m \alpha_i \sum_{i=1}^m \beta_i}{\sqrt{m \sum_{i=1}^m \alpha_i^2 - (\sum_{i=1}^m \alpha_i)^2} \sqrt{m \sum_{i=1}^m \beta_i^2 - (\sum_{i=1}^m \beta_i)^2}} \quad (1)$$

Where: Data samples of two types of bridge monitoring indicators; Indicates the sample data quantity. In this paper, the deflection is taken as an example. Through correlation analysis, it is obtained that the Pearson correlation coefficient between temperature and deflection is -0.89, as shown in Figure 1. It can be seen that there is a strong negative correlation between temperature and deflection.

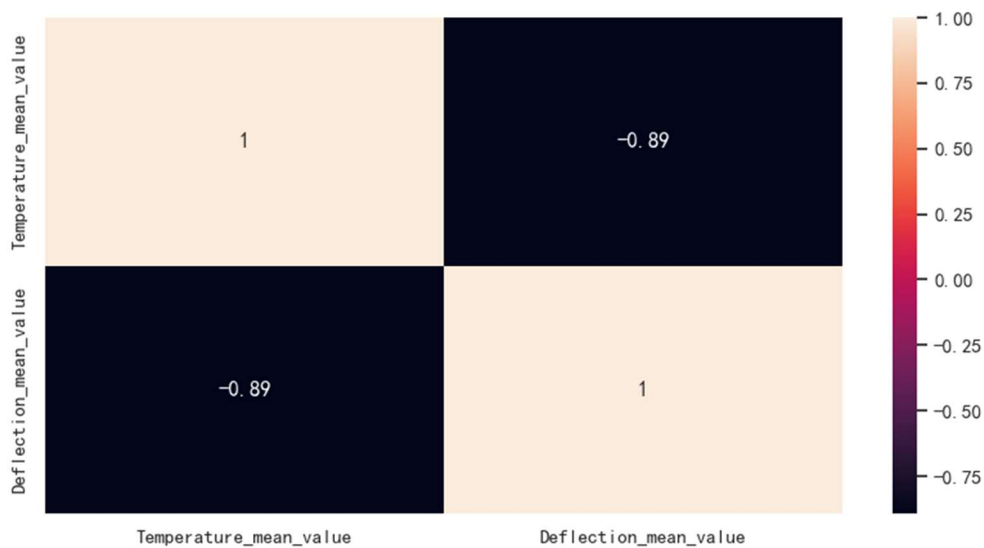


Figure 1. Plot of Pearson's correlation coefficient between temperature and deflection

2.4 Prediction based on Random Forest Regression

As a relatively new machine learning algorithm, one of the advantages of random forest is that it reduces the amount of computation. By classifying or regression the data by repeatedly dichotomizing, compared with the classical neural network machine learning model, the classical algorithm has high prediction accuracy but large amount of computation, while the random forest algorithm takes less time to run the program calculation. The random forest algorithm originated from the classification tree algorithm. In the subsequent development of the algorithm, multiple classification trees were gradually combined to form a random forest, the data of various variables were randomized, converted into many classification decision trees for processing, and finally the results were summarized [15]. In the case of this random application, a forest is established, and the classification

decision trees in the random forest have no direct correlation with each other. After the random forest is generated, when new data samples enter the forest, each classification decision tree will judge and select these new data samples to determine the categories that these data samples belong to respectively. The decision difference between classification decision trees, their own decision classification ability and the macro correlation between each tree are set up. Pass calculate and classify each data node in a random way, compare the classification errors, and then divide the categories to make classification or prediction. The overall process is to use multiple classification decision trees in the forest to train data samples, and then make prediction or classification [16]. The training process of using random forest algorithm for prediction is as follows.

- 1) Input the data sample, divide the data into training set S , test set T , and select the feature dimension F . Start to determine the random forest parameters, including the number of initial classification trees t , the number of features f of each node, the minimum data sample s of each node and the minimum value limit of information gain m .
- 2) The training set S of data samples is randomly extracted and replaced. After randomization to $S(i)$, it is used as the sample of tree root nodes of classification decision making in the forest, and the algorithm will start training from the root node.
- 3) Determine the information gain of the current node data sample. If the training has reached the termination condition at this time, this node is defined as a leaf node, and the predicted output of the leaf node is taken as the average value of each sample in the current node data sample, and then the decision training of other nodes is continued. If the algorithm training does not reach the termination condition at this time, the feature dimension F of the data set is randomized by the same method to form f -dimensional features. According to these features, the data is classified and searched to find the best-matched one-dimensional feature k and its threshold t_h , bounded by the threshold t_h . Determine whether it belongs to the left ($< t_h$) or the right ($\geq t_h$) until h reaches the leaf node that no longer splits, and output the predicted value [17].
- 4) Repeat the algorithm process to ensure that all nodes have made decision judgments or are recorded as leaf nodes.
- 5) Repeat all algorithm training steps until every classification decision tree in the forest has been trained and output the predicted value.

2.5 Predictions based on the XGBoost Model

XGBoost is an acronym for "Extreme gradient boosting" proposed by Chen and Guestrin [18]. The XGBoost algorithm is an effective machine learning method for classification, regression and prediction. The algorithm utilizes multiple decision trees generated by the Classification Regression Tree (CART) algorithm for sequential training, and the sum of the results from each tree is used as the final prediction. Due to the Gini branching property of the CART algorithm, the algorithm branches by selecting the best features and is therefore suitable for research subjects where the input features are coupled. The XGBoost algorithm also improves the efficiency of the operation by parallelizing the selection of features for each tree. The samples are randomly sampled to prevent overfitting of the model to some extent. The principle of GBDT algorithm is to train a tree on a training set and sample labels, and then use this tree to predict the training set to get the predicted value of each sample, subtract the predicted value of the samples from the labels to get the residuals, and then fit the residuals of each sample when training the second tree. The residuals for each sample are obtained after training, and then the n th tree is trained analogously. XGBoost is more efficient than the GBDT algorithm and achieves integrated learning of multiple CART sub-modules through gradient boosting [18,19]. The XGBoost algorithm performs a second-order Taylor expansion of the loss function and adds a regularization term to the function. As a result, the variance of the model is reduced, leading to more efficient optimal solutions and avoiding overfitting. The loss function of the XGBoost algorithm is defined as:

$$L^{(t)} = \sum_{i=1}^n l(\alpha_i, \beta_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \quad (2)$$

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T \omega_j^2 \quad (3)$$

where $l(\alpha_i, \beta_i^{(t-1)} + f_t(x_i))$ is the number of samples from the i th The residuals of the predictions from the first sample to that the predicted residuals of the second iteration, and x_i is the predicted residual from the first i sample, and $\Omega(f_t)$ is the regular term, and ω_j^2 is the fraction of leaf nodes, and T is the number of leaf nodes, and γ is the coefficient, and λ is the coefficient of the sum of all leaf nodes' L_2 coefficients of the sum of regularization weights of the regular term. The GBDT algorithm is used when the regularization term is 0. The second order Taylor expansion of the loss function of the XGBoost algorithm is as follows:

$$L^{(t)} = \sum_{i=1}^n \left[l(\alpha_i, \beta_i^{(t-1)} + g_i f_t(x_i)) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \quad (4)$$

where g_i is the first order derivative term of the loss function, and h_i is the second order derivative term of the loss function.

2.6 Evaluation Indicators

In order to accurately judge the prediction effect of the model, this paper adopts Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Decidability Coefficient R^2 as the evaluation criteria for the prediction accuracy of the algorithm, and the formulas are shown below:

$$MAE = \frac{1}{N} \sum_{i=1}^N |x_i - y_i| \quad (5)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - y_i)^2} \quad (6)$$

$$R^2 = 1 - \frac{\sum_{i=1}^N (x_i - y_i)}{\sum_{i=1}^N (x_i - z_i)} \quad (7)$$

where: x_i is the actual structural displacement monitoring value; y_i is the displacement prediction value; z_i is the average of the true values; N is the number of samples; MAE and $RMSE$ indicates the overall reliability of the predicted data, the smaller the index indicates the smaller the prediction error, the more reliable the predicted results; R^2 Indicates the degree of fit between the predicted value and the actual value, between 0 and 1, the closer to 1, the better the prediction effect, the better the fit between the actual value and the predicted value, that is, the better the prediction performance.

3. Data Modeling and Validation Analysis

3.1 Random Forest, XGBoost Do Comparative Tests

Sampling deflection monitoring data, respectively, using random forest and XGBoost 2 kinds of models for prediction, sampling base (root) environment of jupyter for model building, set up random forest regression tree 100, and compare the prediction accuracy of the curves generated by the 2 methods of modeling, fitting performance by MAE , $RMSE$, and R^2 was evaluated, and the prediction results are shown in Fig. 2 and Fig. 3.

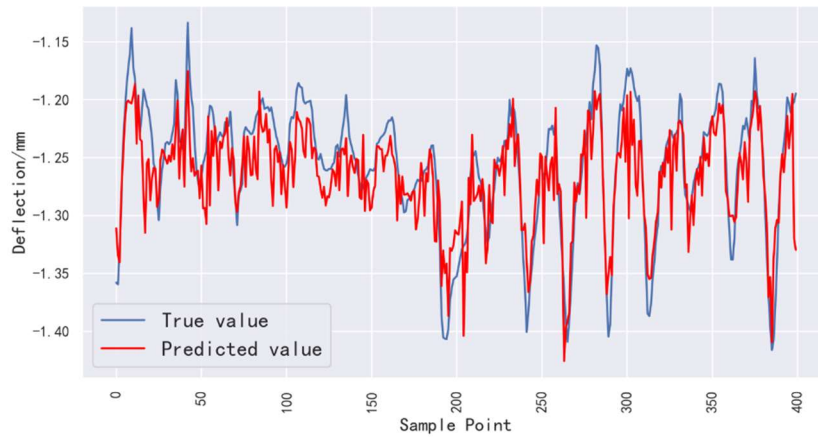


Figure 2. Random forest prediction results

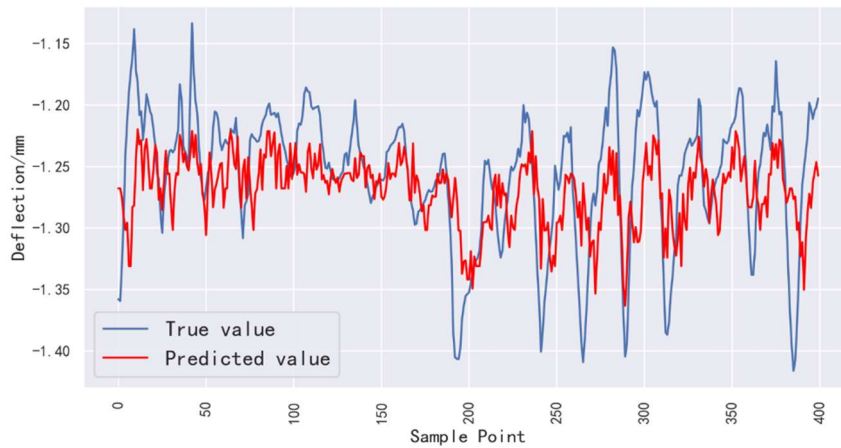


Figure 3. XGBoost prediction results

3.2 Prediction based on Random Forest and XGBoost Fusion Models

In machine learning, Random Forest has a certain degree of randomness, which can avoid overfitting, and there is an improvement in the generalization ability over the tree model alone, and XGBoost uses the Boost learning theory to strengthen the learning of weak links, which is conducive to doing better learning of large-scale complex datasets, and the effect is more obvious. Therefore, combining the two models, the prediction results are shown in Figure 4.

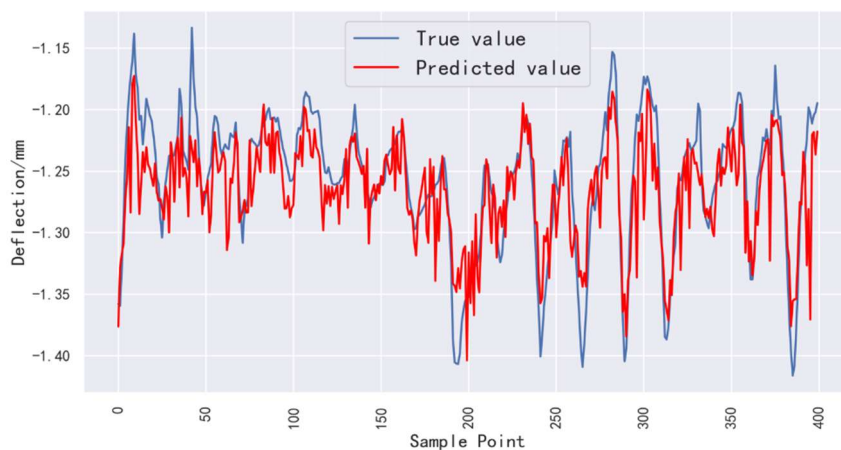


Figure 4. Fusion model prediction results

The prediction errors of the three models, Random Forest, XGBoost and fusion model, are shown in Table 3, from which the prediction accuracies of the two models, Random Forest and XGBoost, can be seen

Table 3. Comparison of prediction errors of random forest, XGBoost and fusion models

mould	<i>MAE</i>	<i>RMSE</i>	R^2
random forest	0.0705	0.0933	0.9113
XGBoost	0.0798	0.0977	0.9064
fusion model	0.0646	0.0732	0.9403

From the table, it can be seen that *MAE*, the *RMSE* and R^2 all outperform the models predicted by Random Forest and XGBoost alone with higher prediction accuracy.

3.3 Random Forest and XGBoost Fusion Model Prediction based on Temperature Correlation

Based on the inference of temperature correlation, the fusion temperature factor is used as training data and input into the fusion model model of Random Forest and XGBoost based on temperature correlation, and the accuracy of the fusion model without considering temperature correlation is compared, and the prediction results are shown in Figure 5.

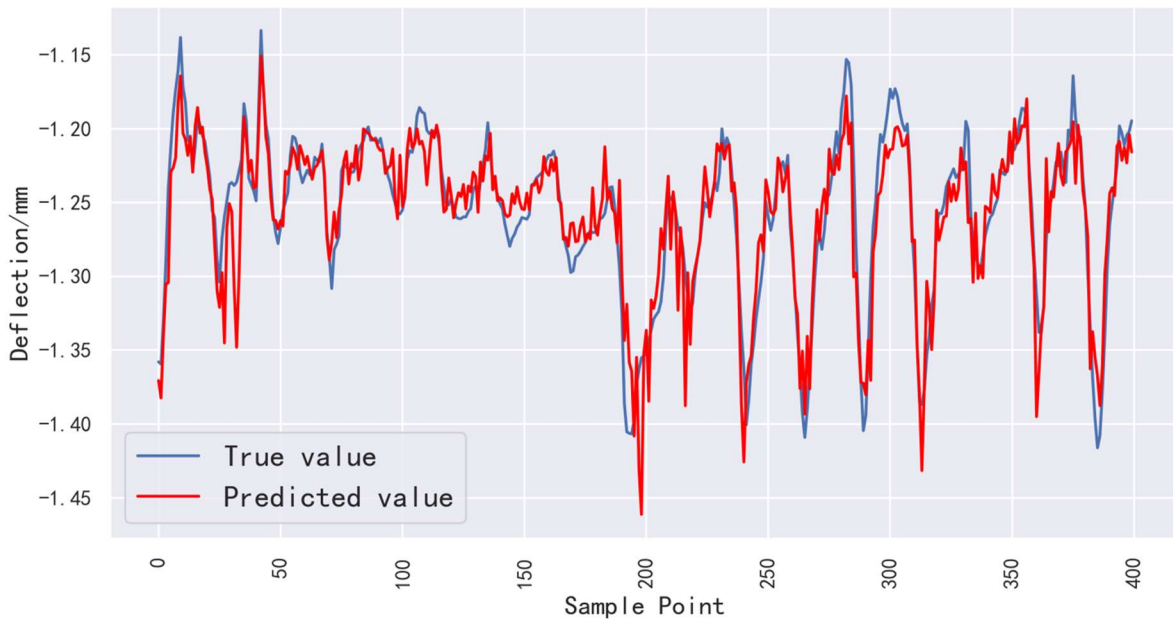


Figure 5. Temperature correlation prediction results

The prediction errors are shown in Table 4, from which it can be seen that the *MAE* and *RMSE* have been reduced by 22.13% and 6.3%, respectively, and R^2 improved by 3.1%. Therefore, the incoming temperature correlation is feasible as a prediction scheme for multi-source data and has better results.

Table 4. Comparison of two kinds of fusion prediction error

mould	<i>MAE</i>	<i>RMSE</i>	R^2
Temperature dependence not taken into account	0.0646	0.0732	0.9403
Consider temperature dependence	0.0503	0.0686	0.9695

4. Conclusion

In this paper, based on the bridge deflection and temperature monitoring data, firstly, the monitoring program of this project and the data sources are introduced, the correlation between temperature and bridge deflection is analyzed using Pearson correlation coefficient, then the prediction based on Random Forest and XGBoost model is established, and then the two models are fused for prediction, and finally the fusion prediction model based on the correlation between the temperature is established in order to MAE, the RMSE and R^2 as evaluation indexes, which proved the accuracy and effectiveness of the established prediction model.

The findings of the study include the following:

- (1) In this paper, taking deflection as an example, the Pearson correlation coefficient between temperature and deflection is obtained as -0.89 through correlation analysis, which indicates a strong negative correlation between temperature and deflection.
- (2) Establishing data prediction based on Random Forest and XGBoost models, both of which have high prediction accuracy and comparable prediction performance.
- (3) Establishing a fusion model data prediction based on Random Forest and XGBoost, comparing the fusion model with Random Forest and XGBoost. MAE, the RMSE and R^2 are all better than the models predicted by Random Forest and XGBoost alone, with higher prediction accuracy.
- (4) Establishing a fusion model bridge monitoring data prediction based on temperature correlation and comparing it with a fusion model that does not take into account temperature correlation, the MAE and RMSE decreased by 22.13% and 6.3%, respectively, and R^2 An improvement of 3.1% was achieved.

Since the temperature change has a large impact on the changes of each part of the bridge, it is feasible to introduce the temperature correlation as a multi-source data prediction scheme and has better results, which provides a strong guarantee for the bridge monitoring data, and is of great significance for the bridge health monitoring and real-time early warning.

References

- [1] Li A, Ding Y, Wang H, Guo T (2012) Analysis and assessment of bridge health monitoring mass data-progress in research/development of "structural health monitoring". *Sci China Technol Sci* 55(8):2212-2224.
- [2] Worden K, Cross E (2018) On switching response surface models, with applications to the structural health monitoring of bridges. *Mech Syst Signal Process* 98:139-156.
- [3] Xia Q, Cheng Y, Zhang J, Zhu F (2016) In-service condition assessment of a long-span suspension bridge using temperature-induced strain data. *J Bridge Eng* 22(3):04016124.
- [4] Zhou G, Li A, Li J, Duan M (2018) Structural health monitoring and time-dependent effects analysis of self-anchored suspension bridge with extra-wide concrete girder. *Appl Sci* 8(1):115.
- [5] Forstner E, Wenzel H (2011) The application of data mining in bridge monitoring projects: exploiting time series data of structural health monitoring. In: 2011 22nd International Workshop on Data-base and Expert Systems Applications. IEEE, pp 297-301.
- [6] WANG Ertao, GAO Huiying, SUN Hai et al. Research on group seismic damage prediction model of urban bridges based on PSO-SVM[J]. *Earthquake Defense Technology*, 2017, 12(01): 185-193.
- [7] Chen, Zhen, and Wei Fan. "A freeway travel time prediction method based on an XGBoost model." *Sustainability* 13.15 (2021): 8577.
- [8] TIAN Zhuang, FAN Qiwu, WANG Changjie. Application of deep learning in bridge response prediction and health monitoring[J]. *Journal of Railway Engineering*, 2021, 38(06): 47-52.
- [9] LIU Xi, QIAO Shengwei, CHEN Hang et al. Prediction method of bridge monitoring data based on LSTM neural network[J]. *Guangzhou Architecture*, 2022, 50(04): 33-38.

- [10] ZHAO Y, ZHAO X, YAN LP, et al. Reconstruction of the statistical characteristics of electric fields in enclosures with an aperture based on random forest regression [J] IEEE Transactions on Electromagnetic Compatibility, 2020, 62(4):1151-1159.
- [11] SUN H, IANG L, WANG C, et al. Prediction of the electrical strength and boiling temperature of the substitutes for greenhouse gas SF₆ using neural network and random forest [J] IEEE Access, 2020, 8:124204-124216.
- [12] S.M. Deng, Research on fault diagnosis of bridge monitoring sensors based on data-driven [D]. Chongqing Jiaotong University, 2022. DOI:10.27671/d.cnki.gcjtc.2021.000105.
- [13] XUE Guohua, LI Minghui, HAN Yuxuan et al. Research on multi-source data prediction algorithm based on correlation analysis for bridge structural health monitoring[J]. Railway Construction, 2022, 62(11):73-79.
- [14] Mehrani E, Ayoub A, Ayoub A. Evaluation of fiber optic sensors for remote health monitoring of bridge structures[J]. Materials and Structures, 2009, 42: 183-199.
- [15] Cao Taoyun. Research on the importance of variables based on random forest [J]. Statistics and Decision, 2022, 38(4):60-63. (in Chinese).
- [16] Wang Weitong, Fan Haidong, Liang Chengsi, et al. Research on prediction model of NO_x emission from offset boiler outlet based on Random forest algorithm [J]. Thermal Power Generation, 2019, 51(4):96-104.
- [17] Liu Junyu, Chen Hui, Zhang Fufeng, et al. Construction and implementation of Electric power Enterprise Market competition prediction Model based on improved Random forest algorithm [J]. Mechanical Design and Manufacturing Engineering, 2019, 50(6):85-88.
- [18] J. Dong, Y. Chen, B. Yao, X. Zhang, N. Zeng, A neural network boosting regression model based on XGBoost, Appl. Soft Comput. 125 (2022), 109067.
- [19] K. Song, F. Yan, T. Ding, L. Gao, S. Lu, A steel property optimization model based on the XGBoost algorithm and improved PSO, Comp. Mater. Sci. 174 (2020), 109472.