Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/74/20240481

Prediction of building settlement amount based on the ARIMA-LSTM-GXBoost combined model

Qiaojun Chen[1,3,*], Jiale Qin[2,4], Tianyu Zhang[1,5], Jiahe Guang[2,6], Jiahao Gao[1,7]

[1]School of Safety Science and Engineering, Liaoning Technical University, Xingcheng, Huludao City, Liaoning Province, China
[2]School of Electronics and Information Engineering, Liaoning Technical University, Xingcheng, Huludao City, Liaoning Province, China

[3]cqj0925@163.com
[4]qinjiale1212@163.com
[5]1159192693@qq.com
[6]guangjiahe@qq.com
[7]1977134323@qq.com
*corresponding author

Abstract. Accurate prediction of building settlement amounts is crucial for ensuring the safety of building structures and human lives. Addressing issues such as the scarcity of building settlement measurement data and the lack of weighting in combined prediction model results, which have not yet led to further improvements in prediction accuracy, this study first selected building CJ06 in Yizheng City, Yangzhou, as the research subject. Polynomial fitting was applied to the obtained non-equidistant time series data, and the fitting data with a preferable polynomial order was chosen to replace the original numerical data. Then, using the first 140 days of data as the training set and the subsequent 63 days of data as the test set, an ARIMA (AutoRegressive Integrated Moving Average) model was initially established for the building settlement, to address the linear features in the fitted sequence. This was followed by using LSTM (Long Short-Term Memory) to correct the ARIMA model's prediction residuals, and combining the results with the XGBoost (Extreme Gradient Boosting) prediction model. Weights were determined using the reciprocal error method for combined prediction, and the results of the combined prediction model were compared with those of the ARIMA-LSTM, ARIMA, LSTM, and GXBoost models. The comparison periods were chosen as follows: 143d, 158d, 173d, 188d, and 203d; ultimately, the maximum absolute relative error and mean square error (MSE) were used as the evaluation metrics for model prediction accuracy. The results showed that the ARIMA-LSTM-GXBoost prediction model's results were more closely aligned with actual values, with a maximum absolute relative error and mean square error of 1.68% and 0.03, respectively. The prediction accuracy was higher than the other four comparison models, achieving high-precision prediction of building settlement amounts.

Keywords: ARIMA, LSTM, GXBoost, Weighting, Prediction Accuracy.

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/74/20240481

1. Introduction

As cities continue to develop rapidly, the growth in the number and scale of constructions in the building industry also changes dramatically, making the safety issues of building structures critical for ensuring the safety of human lives [1]. During the construction and operation phases of buildings, the structures often undergo settlement and deformation due to the effects of natural factors such as building structures, load actions, and the geology, hydrology, soil, and atmospheric temperatures of foundation engineering. When these settlements and deformations exceed certain limits, they can jeopardize the safety of the buildings. Therefore, adopting scientific and effective methods to predict and analyze buildings has become a focal point for engineers. Thus, accurate prediction of building settlement amounts is of great importance [2].

In light of this, this paper addresses the issues of sparse original data and non-equidistant time series by proposing the use of polynomial interpolation fitting. By adjusting the degree of the polynomial and considering the impact of factors such as goodness of fit, overfitting, and interpretability, a more optimal fitting function is chosen to replace the original data with the fitted continuous data, thereby resolving the issues associated with sparse original data and non-equidistant time series. In this research process, an ARIMA (AutoRegressive Integrated Moving Average) model is first established based on the aforementioned processed data to extract linear features from the fitted sequence. Subsequently, an LSTM (Long Short-Term Memory) model is used to correct the prediction residuals of the ARIMA model. Finally, the corrected ARIMA model predictions are merged with the predictions from the XGBoost (Extreme Gradient Boosting) model. To enhance prediction accuracy, the reciprocal error method is used to determine the weights for the weighted combined prediction model, and a comparative analysis is conducted against the results of single models and unweighted combined models.

2. Introduction to Model Principles

2.1. Introduction to ARIMA Model Principles

ARIMA [3] (AutoRegressive Integrated Moving Average) is a classic method of time series forecasting introduced by scholars BOX and JENKINS in 1976. The core idea of the ARIMA model is to transform non-stationary time series into stationary time series through differencing, which has proven effective for processing building settlement data. Then, regression analysis and moving average models [4] are applied to the transformed stationary series to facilitate prediction.

2.2. Introduction to LSTM Principles

LSTM [5] (Long Short-Term Memory) is a variant of recurrent neural networks designed to address issues such as gradient disappearance and explosion in traditional RNNs (Recurrent Neural Networks) when processing long sequence data. It is a powerful deep learning model, with its structural principle illustrated in Figure 1.

DOI: 10.54254/2755-2721/74/20240481

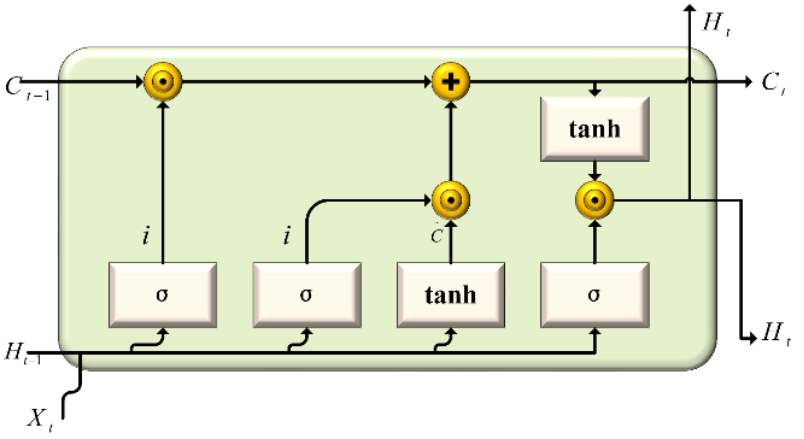

Figure 1. LSTM schematic structure

2.3. Introduction to GXBoost Model Principles

XGBoost (Extreme Gradient Boosting) is a predictive model based on boosting ensemble learning, suitable for predicting building settlement. The fundamental approach involves integrating multiple weak classifiers to establish a strong classifier, optimizing the empirical loss function, continuously fitting the negative gradient of the loss function through iterations, and then generating the optimal weak learner using a line search method [6-7].

3. Model Establishment and Solution

3.1. Data Collection and Analysis

The main research building is the Yizheng City Comprehensive Transportation Hub in Yangzhou, with a land area of 50,841 square meters and a total construction area of 68,202 square meters. The CJ06 monitoring point of the driver and passenger apartment's building settlement process was selected as the research object. Engineering technicians and experts monitored the building settlement at the CJ06 monitoring point every 15 days, as documented in reference [8], and obtained the monitoring data shown in Table 1:

Table 1. CJ06 Actual Settlement Monitoring Values

Time/d	Settlement/mm	Time/d	Settlement/mm
8	1.2	113	5.9
23	1.76	128	6.05
38	2.61	143	6.19
53	2.84	158	6.31
68	4.21	173	6.42
83	5.14	188	6.55
98	5.7	203	6.68

Considering that settlement observation is a slow geological process, and due to the impact of weather, environment, and other factors, the settlement data is a non-equidistant time series, meaning the observation values (or data points) are unevenly distributed over the time axis. To better observe the settlement deformation of the building, this paper chose to use Origin software to perform polynomial fitting on the settlement observation data. Polynomial fitting was chosen, adjusting the polynomial order to 1, 2, 3, 4, 5, 6, 7, 8, 9, with the goodness of fit R^2R^2 and the complexity of the fitting curve as evaluation criteria, selecting the more optimal fitting results to replace the non-continuous time series. Some of the fitting results are shown in Figure 2 below:

DOI: 10.54254/2755-2721/74/20240481

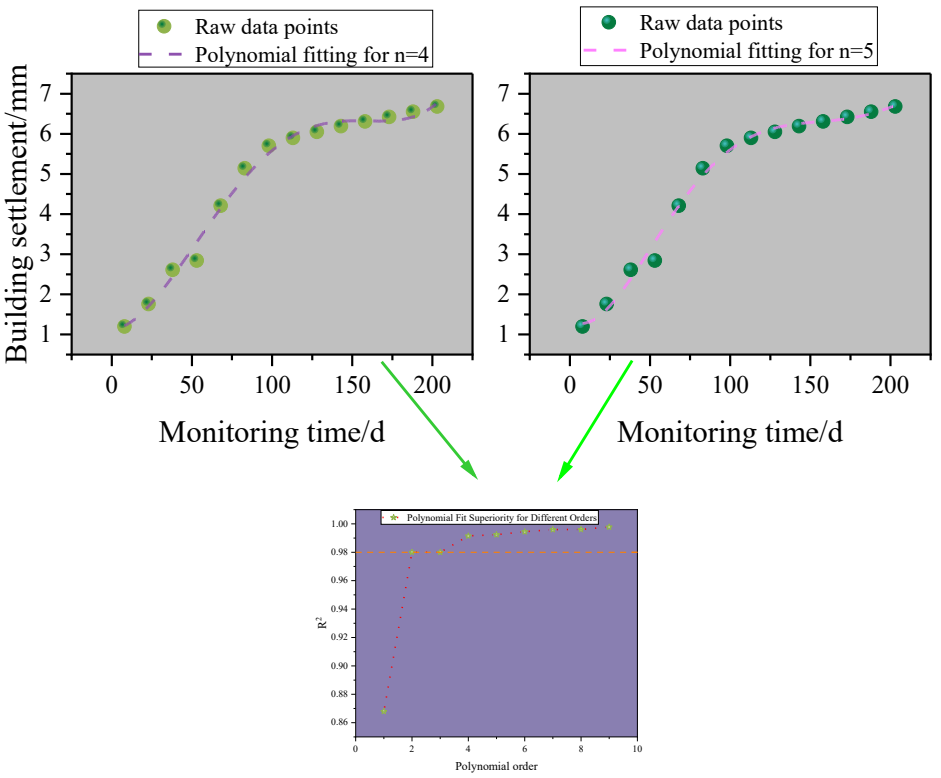

Figure 2. Comparison of partial fitting results and goodness of fit

Based on the analysis of Figure 2, it was found that as the polynomial order n increases, the goodness of fit gradually approaches 1. However, when n=5, the speed at which the goodness of fit approaches 1 significantly decreases. Furthermore, considering that an increase in polynomial order can lead to overfitting phenomena—where the model performs very well on training data but poorly on unknown data—and high-order polynomial models may be difficult to interpret because the model may contain many interacting terms, making it difficult to understand and explain the model. Taking into account the above analysis, the n=5 polynomial fitting result was chosen to replace the original discontinuous data.

3.2. Model Establishment

The application of the ARIMA-LSTM-XGBoost prediction model in the field of building settlement prediction is notably innovative. This combined model integrates a mix of time series analysis, deep learning, and machine learning techniques, providing a powerful tool for predicting building settlement.

Based on the analysis of the principles of the aforementioned three algorithms, to enhance the prediction accuracy of building settlement, a combination of the ARIMA model, LSTM model, and XGBoost model is implemented, employing error correction and a weight distribution-based approach [8-9]. This combination is applied in the field of building settlement prediction, with the combined ARIMA-LSTM-XGBoost prediction model as shown in Figure 3 below:

DOI: 10.54254/2755-2721/74/20240481

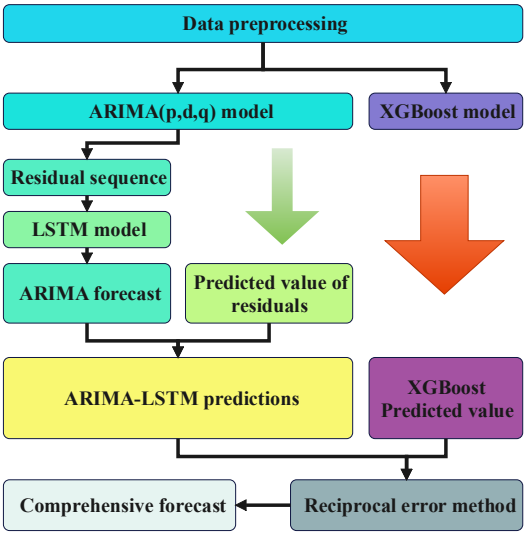

Figure 3. Flowchart of model prediction

3.3. Result Analysis

Based on the above analysis, the ARIMA-LSTM-XGBoost prediction model was constructed using MATLAB software. On the basis of the aforementioned error correction and weight distribution, the actual values of building settlement for the first 140 days were used as the training set, and the actual values for the subsequent 63 days as the test set. The predicted values for the next 43 days of building settlement were obtained and compared with the results of four other models: ARIMA-LSTM, ARIMA, LSTM, and GXBoost. The original data observed in the latter 63 days at 143d, 158d, 173d, 188d, and 203d were chosen as comparison samples, and the results are shown in Figure 4 below.

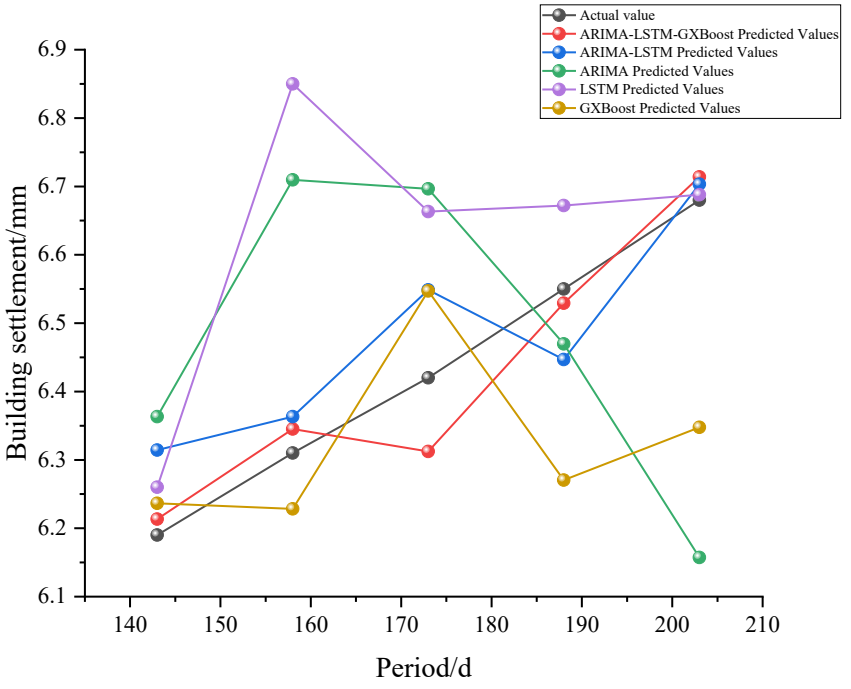

Figure 4. Comparison of prediction results of different models

DOI: 10.54254/2755-2721/74/20240481

From the analysis of Figure 4, it can be seen that the prediction result curve of the ARIMA-LSTM-GXBoost prediction model fits more closely with the actual value curve, and its trend of change is almost consistent with it. In vertical comparison with the other four models, it was found that its prediction results are superior to those four models, which have a greater trend of variation. In summary, it was found that the prediction results of ARIMA-LSTM-GXBoost are the best.

Considering the comparison of prediction accuracy among different models, two evaluation metrics were introduced: Absolute Percentage Error (APE) and Mean Squared Error (MSE). APE is a measure used to quantify the difference between an estimated value and the actual value. It combines the concepts of absolute error and relative error to express the size of the error in percentage terms. MSE is a common metric used to assess the difference between model predictions and actual observations. It calculates the average of the squares of the model's prediction errors and then takes the square root. The formulas are as follows:

$$|\text{Absolute relative error}| = \left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100\%$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Where y_i represents the actual value of the settlement, and $\hat{y}_i$ represents the actual predicted value of the combined model.

Using the above formulas, the APE and MSE of different prediction models were calculated, and the results are shown in Table 2 below:

Table 2. Comparison of absolute relative error as well as MSE results

Actual Value	Absolute Percentage Error/%				
	ARIMA-LSTM-GXBoost Predicted Value	ARIMA-LSTM Predicted Value	ARIMA Predicted Value	LSTM Predicted Value	GXBoost Predicted Value
6.6338	3.29	4.80	6.39	6.22	6.70
6.6487	3.31	4.69	6.36	6.20	6.32
6.6640	3.32	4.73	5.78	5.65	6.41
6.6797	3.14	4.87	5.88	5.70	5.70
6.6956	3.07	4.73	5.92	5.93	5.75
6.7119	3.05	4.04	5.98	5.83	5.76
MSE	0.046	0.096	0.1632	0.1564	0.1665
Maximum Relative Error/%	3.32	4.87	6.39	6.22	6.70

For more visualized results, different models' Absolute Percentage Error bar comparison charts were drawn using Origin software, as shown in Figure 5 below:

DOI: 10.54254/2755-2721/74/20240481

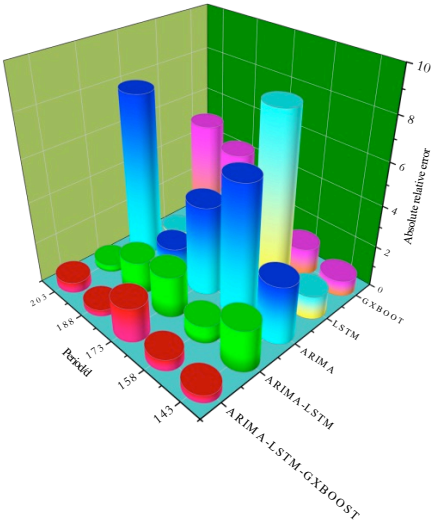

Figure 5. Column comparison of absolute relative errors for different models

From the analysis of Figure 5, it is known that among these five prediction models, the ARIMA-LSTM-GXBoost prediction model has the smallest absolute percentage error, followed by the ARIMA-LSTM prediction model, while the prediction accuracies of the other three single models are inferior to the above two combined prediction models. Since combined models can compensate for the shortcomings of single models, improving the model's generalization ability and prediction accuracy, choosing a combined prediction model for predicting building settlement amounts results in higher prediction accuracy. The prediction accuracy of the ARIMA-LSTM-XGBoost model, which employs error correction and weight distribution, is superior to that of the ARIMA-LSTM model, and its MSE value is also minimized.

4. Conclusion

(1) Choosing a polynomial order of n=5 for fitting the original monitoring data shows better fitting results. Using the fitted 1-203d building settlement data as real data compensates for the deficiencies of the original data and the non-equidistant time series, among other issues. This approach enables the extraction of effective information, providing a certain reference value for subsequent prediction problems.

(2) Compared with ARIMA-LSTM, ARIMA, LSTM, and XGBoost prediction models, the ARIMA-SLTM-XGBoost model offers superior prediction effects and smaller errors. It compensates for the deficiencies of single models and unweighted combined models, improving prediction accuracy to a certain extent and providing a strong reference for relevant departments and engineering technicians.

References

[1] Yan, X. H., & Xu, P. L. (2020). Prediction analysis of building settlement based on GM(1,1) and Verhulst models. Geospatial Information, 18(09), 110-112+130+8.

[2] Sun, C., Zhang, J., Ye, S., et al. (2023). Application of a combined model based on wavelet decomposition in the settlement prediction of high-rise buildings. Surveying and Spatial Geographic Information, 46(10), 221-224.

[3] Yu, J. Q., Nie, J. K., Zhao, A. J., et al. (2022). Short-term power load forecasting based on feature mining of ARIMA-GRU. Journal of Electric Power Systems and Automation, 34(03), 91-99.

[4] Li, W. C., Pan, F. Q., Hu, P., et al. (2022). Short-term passenger flow forecasting for urban buses based on the seasonal differencing integrated moving average autoregressive model. Journal of Jinan University (Natural Science Edition), 36(03), 308-314.

DOI: 10.54254/2755-2721/74/20240481

[5] Kang, J. F., Fu, Y., Fang, L., et al. (2023). Prediction of inbound tourist numbers based on a weighted least squares ARIMA-LSTM model: A case study of Shanghai. Journal of Zhejiang University (Science Edition), 50(04), 508-520.
[6] Shi, J. Q., & Zhang, J. H. (2019). Load forecasting method based on multi-model fusion Stacking ensemble learning. Chinese Journal of Electrical Engineering, 39(14), 4032-4042.
[7] Tan, H. W., Yang, Q. L., Xing, J. C., et al. (2022). Photovoltaic power generation forecasting using an XGBoost-LSTM combined model. Journal of Solar Energy, 43(08), 75-81.
[8] Sha, A. M., Lü, F. R., & Wang, X. D. (2023). Research on building settlement prediction using a GM(1,1)-Logistic combined forecasting model. Popular Science & Technology, 25(08), 17-21.
[9] Fan D, Sun H, Yao J, et al. Well production forecasting based on ARIMA-LSTM model considering manual operations[J]. Energy, 2021, 220: 119708.