Construction Site Monitoring Data Processing Based on Detecting Anomalies and Improved Variational Mode Decomposition

Yixiao Shao[a], Tengfei An[b], Yafei Qi[c], Wenli Liu[*]

School of Civil and Hydraulic Engineering, Huazhong University of Science and Technology, Wuhan, Hubei, China

[a]shaoyx@hust.edu.cn, [b]tf_an@hust.edu.cn
[c]feifeifeibiu333@163.com, [*]liu_wenli@hust.edu.cn

ABSTRACT. Anomalies and noise are prevalent in the time series data extracted from sensors at construction sites, which can hinder the assessment of safety levels and risks. This study aims to detect anomalies and denoise real-time monitoring data from sensors, thereby facilitating early risk warning and enhancing accuracy of real-time status. To achieve this objective, we propose a framework that integrates Extended Isolation Forest, Whale Optimization Algorithm, and Variational Mode Decomposition models. The effectiveness of the framework is validated using a dataset obtained from sensors deployed during the construction of a deep pit foundation. The proposed approach successfully denoises the dataset without anomalies with a root mean square error of 0.0389 and signal-to-noise ratio of 24.09. Consequently, our approach effectively preprocesses data to enable improved decision-making and enhance security risk management capabilities.

Keywords: Deep pit foundations; Variational mode decomposition; De-noise; Anomaly

1 INTRODUCTION

Technological advancements in sensing and data processing have facilitated the efficient monitoring of engineering data throughout the construction process of deep pit foundations, enabling a more comprehensive analysis of geotechnical safety concerns. The extraction and transmission of data from the sensors is not flawless; anomalies and noise seriously affect the quality and completeness of the data, and random disturbances can exacerbate the problem [11]

Anomalies can be detected using both supervised and unsupervised techniques. The Support Vector Machine[3] and Random Forest are two commonly used supervised algorithms that can detect anomalies with high performance (i.e. low false alarm rate). The datasets used to apply supervised techniques need to have high quality labelling in complicated engineering situations. However, the datasets are often incomplete and

Z. Ahmad et al. (eds.), *Proceedings of the 2023 5th International Conference on Structural Seismic and Civil Engineering Research (ICSSCER 2023)*, Atlantis Highlights in Engineering 24,
https://doi.org/10.2991/978-94-6463-312-2_27

require human labelling [12], which takes time. Unsupervised techniques, on the other hand, do not require labelling and have been promoted for use in detecting anomalies in sensor data. However, false positive rates are often significant and detection rates are consistently poor. The Wavelet Transform, Fourier Transform and Empirical Mode Decomposition (EMD) are common methods for processing signals, although each has drawbacks. Examples include the inability of the Fourier transform to handle non-stationary, non-linear signals that change frequency with time, and the modal aliasing that can occur with the components of EMD.

Against this backdrop, this research endeavors to address the following inquiry: How can real-time monitoring data be effectively utilized to detect anomalies, eliminate noise, and evaluate geotechnical safety hazards? In order to respond to this question, we propose a hybrid intelligent data strategy that amalgamates EIF with enhanced Variational Mode Decomposition (VMD) models. This approach aims at efficiently identifying anomalies and denoising monitoring data in order to enhance its quality for safety risk assessment. The unsupervised anomaly detection algorithm Extended Isolation Forest (EIF) exhibits comparable performance to supervised algorithms.

The EIF shares a similar underlying principle with the Isolation Forest (IF), but it overcomes the limitations associated with biased tree branching in IF[7]. EMD is widely employed for signal analysis, enabling the identification and decomposition of signals into their primary "modes" across various time-frequency applications. Huang et al [8] state that EMD can effectively handle nonlinear and non-stationary processes while remaining adaptable. In geotechnical monitoring, Variational Mode Decomposition (VMD) serves as a noise-robust, variational, non-recursive method for multi-resolution decomposition. VMD outperforms EMD in terms of noise robustness when decomposing vibration signals, and it also mitigates issues like modal aliasing and endpoint effects more effectively than EMD does.

In this study, we put forward a comprehensive framework comprising of anomaly identification and data cleansing techniques to acquire an enhanced dataset that can enhance the capability for evaluating safety hazards. Anomalous data can provide incorrect information for decision making and security risk assessment. Anomaly detection is important for risk alerts and for building high quality datasets. The data set without anomalies can achieve better performance in the denoising process. The WOA-VMD is useful to obtain the optimisation parameters, which can avoid the uncertainty arising from the selection of parameters according to manual experience. The feasibility and effectiveness of our proposed approach are presented with an engineering case.

2 Detecting Anomalies

In the existing literature, there are numerous methods that have been created and developed for anomaly detection [4]. The three types of existing sensor measurement methods are rule-based, supervised learning-based and unsupervised learning-based[8].

In particular, the application of the above methods to anomaly detection presents a number of difficulties. For example, rule-based techniques are unable to detect harmful events for which no rules have been established. In fact, rule-based systems can only

detect events where there are rules. In supervised learning, the training data must be labelled, otherwise the algorithms cannot be used [1]. However, unsupervised learning approaches can train on unlabelled data.

Hariri et al [7] originally proposed the EIF model. Isolation Forest (iForest), a model-free anomaly detection technique, is extended. The EIF creates an extended collection of Isolation Forest trees by extracting features from each monitoring dataset (e.g., steel shotcrete wave force and building settlement). An anomaly score is generated as each new piece of monitoring data is mapped to one of these IFrees. It is considered normal if its anomaly score falls below a predetermined threshold. If it does not, the monitoring data is considered abnormal.

An anomaly score is generated as each new piece of monitoring data is mapped to one of these IFrees. It is considered normal if its anomaly score falls below a predetermined threshold. If it does not, the monitoring data is considered abnormal. Conversely, anomalous data are outliers and can be separated after a small number of cycles of random partitioning. A binary tree can serve as a visual representation of the partitioning procedure employed during the isolation phase. The earlier a point undergoes partitioning, the higher its chances of being classified as an atypical point. **Figure 1** depicts an illustration of this partitioning process, where the 'red' leaf node is more likely to be identified as an anomaly.

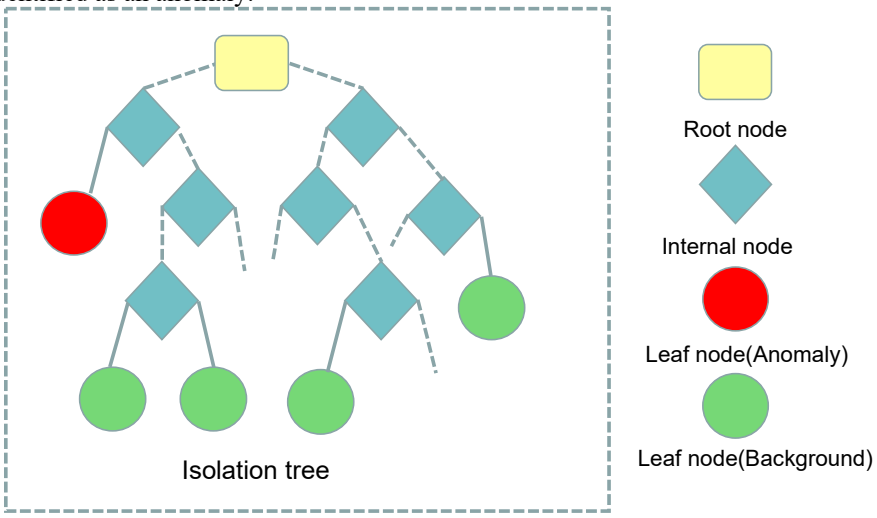


Fig. 1. Example of the structure of an *i*Tree

The node of the isolation tree (T) can have two types of nodes: external nodes without children or internal nodes with one test and two daughter nodes (Tl, Tr). The number of external nodes is denoted as n, the number of internal nodes is n−1, and the total number of nodes in an iTree is 2n−1 [9]. A test consists of an attribute q and a split value p. To create an iTree from a database X = {x1..., xn} containing n instances from a d-variate distribution, we recursively partition X by randomly choosing an attribute q and a partition value p, and we repeat this process until one of the following conditions

is met:(1) the tree reaches a height limit, (2) |X| = 1 or all data in X have the same values [9].

The length of path (h(x)) is determined by the number of edges that x traverses from the root node to an external node in an iTree. To estimate the average path length (E(h(x))) of iTree, we utilize analysis techniques borrowed from Binary Search Tree (BST). The instance x's anomaly score (s) is defined as:

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}} \quad (1)$$

$$c(n) = 2H(n-1) - \left(\frac{2(n-1)}{n}\right) \quad (2)$$

Specific details of the assessment process can be found in Liu et al [9].

3 De-noising Data

Raw monitoring data will always contain noise due to the spatio-temporal ambiguity and complexity of working conditions in deep foundations. In this case, the noise hinders accurate data processing and decision making. Traditional signal denoising techniques include low-pass filtering, Wiener filtering [2], adaptive learning and Kalman filtering. Despite their effectiveness, these techniques have drawbacks as they eliminate or reduce valuable features.

Multi-mode noise is often combined with monitoring data. A new approach for analyzing signals is needed to decompose a signal with multiple components into distinct intrinsic mode functions that are limited to specific frequency bands (BLIMFs). The frequency domain signal segmentation and component separation are efficiently determined by the VMD. Furthermore, VMD has demonstrated the ability to effectively separate signals, enhance resistance against noise interference, and optimize computational efficiency. Consequently, we will employ VMD as a data denoising technique in this investigation, as elaborated upon in the subsequent section.

3.1 Variational Mode Decomposition

A non-recursive decomposition technique, called VMD, was proposed by Dragomiretskiy and Zosso [5] and is used for adaptive and quasi-orthogonal signal decomposition. A multicomponent seismic trace can be simultaneously decomposed into a limited number of band-limited intrinsic mode functions (IMFs). The traditional Wiener filter is generalised by the VMD into numerous adaptive bands. Wiener filtering is one of the most widely used techniques in signal processing, particularly for source separation and signal denoising. The short-time Fourier transform (STFT) is commonly used in the time-frequency domain when applied to audio [10]. Compared to EMD-based adaptive decomposition techniques, the VMD algorithm is more noise-resistant [5]. The following is an introduction to the theories and concepts surrounding VMD.

Definition 1: (Intrinsic Mode Function)

Intrinsic Mode Functions are amplitude-modulated-frequency-modulated (AM-FM) signals, which is different from the definition of EMD.

$$\mu_k(t) = A_k(t)\cos(\emptyset_k(t))$$

Where the phase $A_k(t)$ is envelope of $\mu_k(t)$ and $\emptyset_k(t)$ is a non-decreasing function. The equation of phase $\emptyset_k(t)$ and instantaneous frequency $\omega_k(t)$ is as follow:

$$\omega_k(t) = \frac{d\emptyset_k(t)}{dt} \geq 0$$

Definition 2: (Total Practical IMF Bandwidth)

The total practical bandwidth of an IMF is estimated as Eq.(). Depending on the actual IMF, either of these terms may be dominant.

$$BW_{AM-FM} = 2(\Delta f + f_{FM} + f_{AM})$$

It is necessary to consider the decomposition layers k and the penalty factor in order to impose constraints on the VMD that will affect the performance of the algorithm. The centre frequency method is currently popular. Without a reliable basis, this method primarily calculates the value of k by observing the centre frequency at different values of k. It can only calculate the number of modes, k, but not the penalty parameter, which ultimately results in suboptimal noise reduction. To achieve a better noise reduction result, the whale optimisation technique is used to adaptively determine the two parameters mentioned above.

3.2 Whale Optimization Algorithm

For the optimisation of numerical problems, Mirjalili and Lewis developed the Whale Optimisation Algorithm (WOA). The WOA is a swarm intelligence algorithm for continuous optimization problems in meta-heuristic optimization. It is becoming increasingly popular in engineering applications because: (1) it is based on simple ideas and is easy to use; (2) it does not require gradient information; (3) it can avoid local optima; and (4) it can be applied to a wide variety of problems across different disciplines. The WOA has been shown to be equal to or better than some of the currently used computational approaches. Here is the mathematical model: (1) Prey encirclement: Whales are able to locate and encircle prey at any time. Here we assume that the target prey, or a location close to it, is the current best position of the search agent. The other whales (search agents) try to adjust their positions so that they are facing the most effective search agent. The model looks like this:

$$\vec{D} = \left|\mathrm{C}.\overrightarrow{X^*}(t) - \vec{X}(t)\right| \tag{3}$$

$$\vec{X}(\mathrm{t}+1) = \overrightarrow{X^*}(t) - \vec{A}\cdot\vec{D} \tag{4}$$

where t indicates the current iteration, X^* is the position vector of the current best solution obtained through iteration t, $\vec{X}$ is the position vector of each agent, | | is the

absolute value, and "." is an element-by-element multiplication. The coefficient vectors $\vec{A}$ and $\vec{C}$ are calculated as follows:

$$\vec{A} = 2\vec{a} \cdot \vec{r} - \vec{a} \tag{5}$$

$$\vec{C} = 2 \cdot \vec{r} \tag{6}$$

where $\vec{a}$ linearly decreases from 2 to 0 over the course of the iteration and r is a random number[0,1].

4 Case Study

We illustrate and validate our hybrid smart data methodology with an exemplary case study. The site is a major foundation hole for a metro line being built in Wuhan, China, which is shown in Figure 2. The project was chosen because the researchers were working closely with contractors on a number of other studies, and sensors were being used to monitor geotechnical safety issues.

4.1 Case Description

The T-shaped transfer between stations A and B is the chosen metro project. A station with a 13 metre island platform and three levels of double piers is located underground. Shield tunnel reception shafts are located at both ends of the station, which has a total length of 239.2 metres, a total width of 22.5 metres for the standard section, and a height of 22.63 metres to 25.08 metres. With two underground levels and two columns, Subway Station B is an island-style station with a length of 14 metres. The total length of the station is 634.105 metres, while the total width of the standard section is 23.1 metres. The ground elevation of the study area is between 26.0 and 30.7 metres, and the landform is a denudation accumulation ridge area (Grade III terrace).

Fig. 2. Example of deep pit foundation

Examples of sensors installed in the case are presented in Figure 3. After the sensors are installed, the data are transmitted and stored in the web-based monitoring system, as shown in Figure 3.

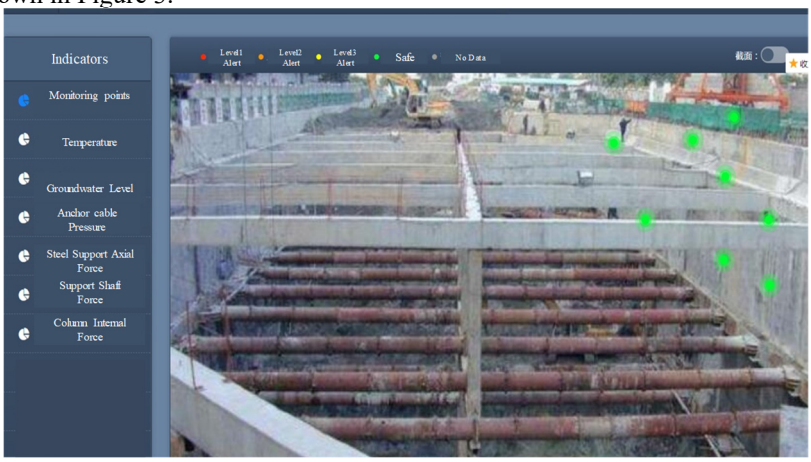


Fig. 3. Web-based monitoring data system

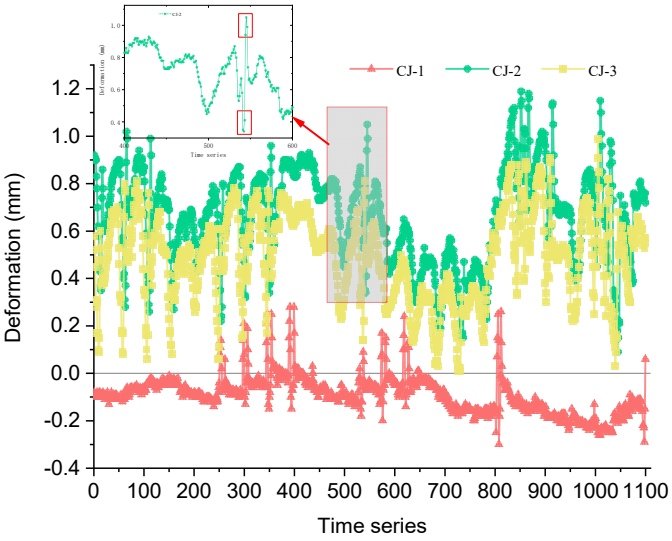
4.2 Anomaly Monitoring Data Detection

The data source and monitoring points consist of four installed sensors, with one serving as a reference point and the other three (CJ1, CJ2, CJ3) serving as analysis points. Figure 4a indicates that there are no apparent anomalies in the billing data. Anomalies are identified through one-dimensional (CJ1), two-dimensional (CJ1 and CJ2), and three-dimensional (CJI, CJ2, and CJ3) analyses. During anomaly detection, each point is assigned an anomaly value based on various training set sizes. The highest anomaly value from each training set is used to determine sensitivity to anomalies. Figure 4b displays the outcomes of dimensional analysis and anomaly detection using different training sets.

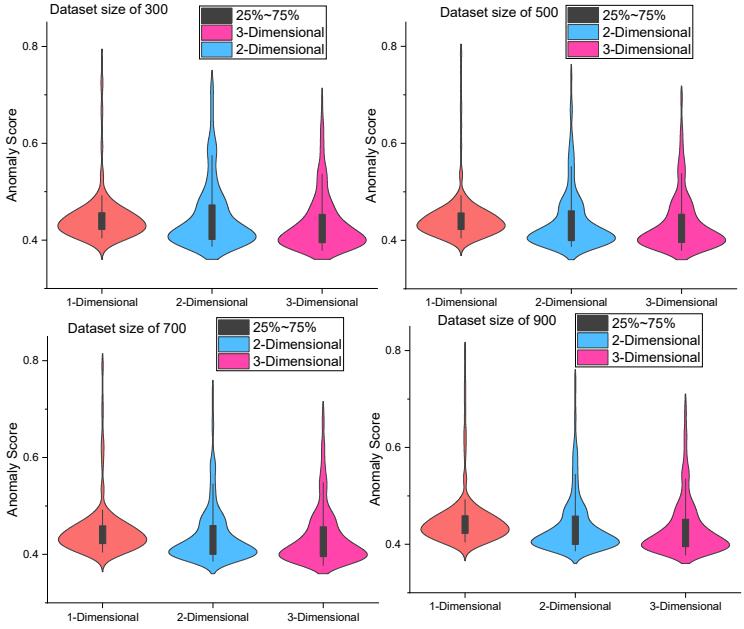
We classify data points with anomaly scores exceeding 0.6 as outliers and analyze the anomaly scores of these outliers. Figure 4b illustrates that one-dimensional data exhibits higher anomaly scores compared to two-dimensional and three-dimensional data. The anomaly scores for one-dimensional data vary with changes in the size of the training set, but they consistently exceed 0.78. Conversely, the anomaly scores for two-dimensional and three-dimensional data remain below 0.74. The highest value observed for the anomaly score of two-dimensional data ranges from 0.71 to 0.74, while for three-dimensional data it falls between 0.69 and 0.71. The outlier anomaly scores in high-dimensional datasets are more concentrated and tend to be lower overall.

Again, four monitoring locations (ZCL-02-21, ZCL-02-22, ZCL-04-C6, ZCL-04-C7) are selected to analyse the axial forces on the steel columns. Figure 4a of our results demonstrates our ability to conclude that the monitoring data from ZCL-04-C6 is anomalous. The results of our analysis of the detection of anomalous data in different dimensional settings are shown in Figure 4b. As shown in Figure 5, the higher the dimensionality of the monitoring data, the less sensitive iForest is to detecting anomalies. By

examining high dimensional monitoring data, we can distinguish between single and multi-dimensional anomalies. The higher the dimension of the monitoring data, the higher the value of the anomaly and the easier it is to identify the origin of the anomaly.

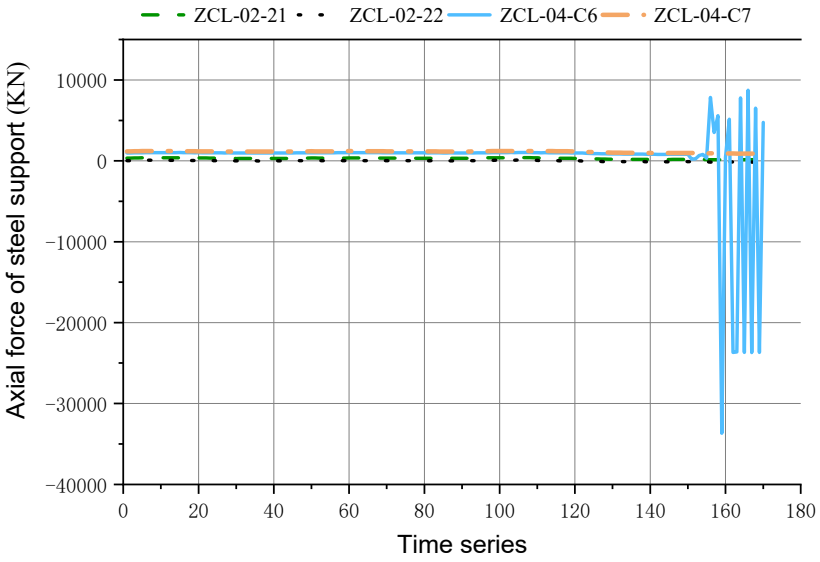


(a) Examplesof monitoring data

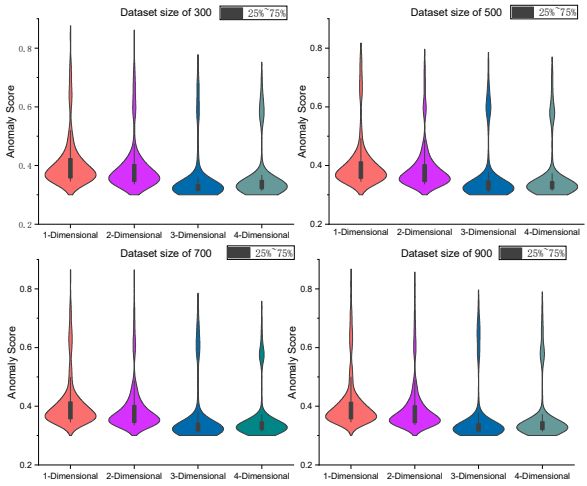


(b) Settlement data: different dimensional and sizes of training set

Fig. 4. Examples of monitoring and settlement data



(a) Examples of monitoring data



(b) Steel support axial force: different dimensional and training set sizes

Fig. 5. Examples of monitoring steel support axial force data

4.3 Monitoring Data De-noising

For the purpose of denoising the sample data, we use a dataset with 1000 monitoring points, depending on the duration and frequency of data collection [6]. According to some studies, multiple "modes" in the original signal may coexist in the same IMF component, or some "modes" may not be properly detected, leading to under-decomposition or leakage decomposition when the number of decomposition modes (K) is too

small. A particular 'mode' in the signal can be 'pulled' into many IMF components if K is too large, leading to over-decomposition. The settings are optimised using the WOA to avoid information loss or unacceptable decomposition effects. The WOA-VMD population size is set to 10, the maximum number of iterations is set to 30, the range of iterations for K is set to 4-6, and the range for is set to 20-1000. The results are shown in Figure 6. The results of correlation coefficient and the envelope entropy are summarized in Table 1.

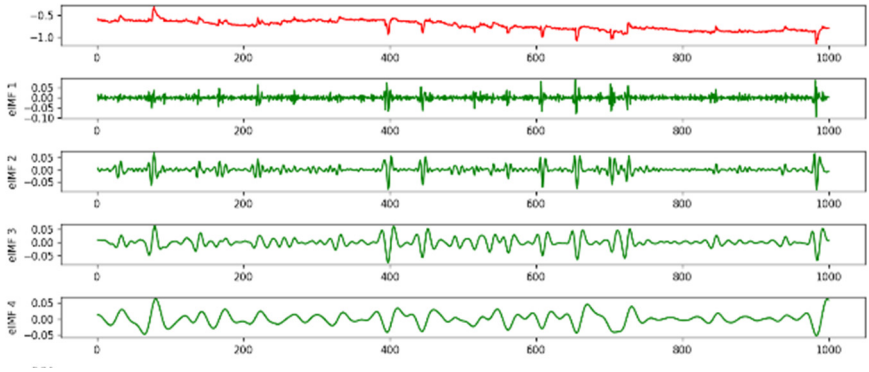


Fig. 6. WOA-VMD results of sample data

Table 1. Results of the optimization of VMD parameters by the WOA.

K	α	Envelope Entropy	Correlation Coefficient			
			IMF1	IMF2	IMF3	IMF4
4	95	6.7963	0.981	0.181	0.115	0.090

The root mean square error (RMSE) and signal-to-noise ratio (SNR) are utilized to calculate the reconstructed signal and original data signal. Figure 7 presents the calculation of these two indicators with different K-values for result validation. The definitions of RMSE and SNR can be found in Eq. (7) and Eq. (8).

$$RMSE = \sqrt{\frac{1}{n}\sum_n (f_0(n) - f_1(n))^2} \quad (7)$$

$$SNR = 10 \times \log_{10}\left(\frac{\frac{1}{n}\sum_n f_0^2(n)}{\frac{1}{n}\sum_n (f_0(n) - f_1(n))^2}\right) \quad (8)$$

Where, f_0 is the original signal data, f_1 is the reconstructed signal data.

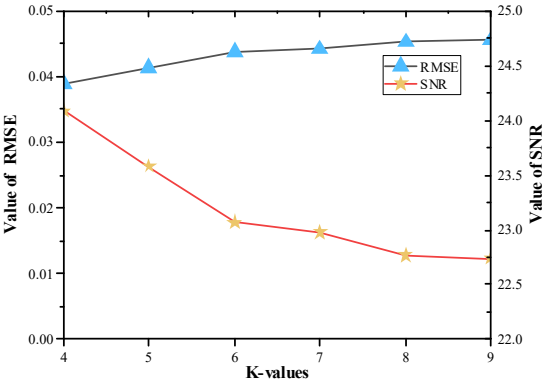


Fig. 7. Two indicators under different *K* values

5 Conclusion

To enhance the accuracy of monitoring data collected from sensors in construction, it is crucial to conduct anomaly detection and noise elimination. Our research focuses on devising an innovative intelligent data methodology that can effectively identify anomalies and eliminate noise in monitoring data, with a specific emphasis on assessing geotechnical safety threats. The proposed framework encompasses:

To detect abnormal points, the Extended Isolation Forest method gathers characteristics from each monitoring dataset. By applying Variational Mode Decomposition, harmonic noise is eliminated to enhance data quality. We illustrate the practicality and effectiveness of our proposed approach using the case study of the Wuhan metro project. The outcomes indicate that by utilizing EIF and enhanced VMD, we can achieve a remarkable level of accuracy in anomaly detection and data denoising. Our findings reveal that our innovative technique exhibits an RMSE value of 0.0389 and an SNR value of 24.09 for identifying anomalies. The capability of EIF and VMD to accurately identify anomalies and improve surveillance data has been successfully demonstrated.

Although our method could not detect all anomalies, it can help site management to better understand the hazards associated with geotechnical safety. Furthermore, we claim that our method can improve the quality of data collected from sensors in deep foundation pits with few errors. As a result, the noise in the monitoring data obtained from sensors can be effectively reduced using the unique smart data approach we have developed.

Reference

1. Ahmed, M., Mahmood, A. N., and Hu, J. (2016). A survey of network anomaly detection techniques. Journal of Network and Computer Applications, 60: 19-31. DOI: 10.1016/j.jnca.2015.11.016

2. Aschero, G., and Gizdulich, P. (2010). De-noising of surface EMG with a modified Wiener filtering approach. Journal of Electromyography and Kinesiology, 20(2): 366-373. DOI: 10.1016/j.jelekin.2009.02.003
3. Bhavsar, Y. B., and Waghmare, K. C. (2013). Intrusion detection system using data mining technique: Support vector machine. International Journal of Emerging Technology and Advanced Engineering, 3(3): 581-586. RUL: Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 3, March 2013)
4. Cha, Y. J., and Wang, Z. (2018). Unsupervised novelty detection–based structural damage localisation using a density peaks-based fast clustering algorithm. Structural Health Monitoring, 17(2): 313-324. DOI: 10.1177/1475921717691260.
5. Dragomiretskiy, K., and Zosso, D. (2013). Variational mode decomposition. IEEE Transactions on Signal Processing, 62(3): 531-544. DOI: 10.1109/TSP.2013.2288675.
6. Han, L., Zhang, R., Wang, X., Bao, A., and Jing, H. (2019). Multi-step wind power forecast based on VMD-LSTM. IET Renewable Power Generation, 13(10): 1690-1700. DOI: 10.1049/iet-rpg.2018.5781.
7. Hariri, S., Kind, M. C., and Brunner, R. J. (2021). Extended isolation forest. IEEE Transactions on Knowledge and Data Engineering. 33(4): 1479-1489. DOI: 10. 1109/TKDE.2019.2947 676.
8. Huang, H. B., Yi, T. H., and Li, H. N. (2020). Anomaly identification of structural health monitoring data using dynamic independent component analysis. ASCE Journal of Computing in Civil Engineering, 34(5), 04020025. DOI: 10.1061/(ASCE)CP.1943-5487.000090.
9. Liu, F. T., Ting, K. M., and Zhou, Z. H. (2012). Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data (TKDD), 6(1): 1-39. DOI: 10.1145/2133360.2133363.
10. Samuel T. Thurman and James R. Fienup (2008). Wiener filtering of aliased imagery. International Society for Optics and Photonics,7076: 179-189. DOI: 10.1117/12.794994.
11. Y., Tang, Z., Li, H., & Zhang, Y. (2019). Computer vision and deep learning-based data anomaly detection method for structural health monitoring. Structural Health Monitoring, 18(2): 401-421. DOI:10.1177/1475921718757405.
12. Guo S. Y., Ding, L.Y., Luo, H., and Jiang, X.Y. (2016). A Big-Data-based platform of workers' behavior: Observations from the field. Accident Analysis and Prevention, 93: 299-309. DOI: 10.1016/j.aap.2015.09.024.