Railway Sciences

Research on risk identification of railway subgrade deformation based on Bayesian and ICA theories

711

Yi Liu, Fengyan Yang and Hu Wang
China Academy of Railway Sciences Corporation Limited, Institute of Computing Technologies, Beijing, China
Xuanqi Wang
College of Civil Engineering, Huaqiao University, Xiamen, China, and
Chengwen Wu and Hongsheng Yu
China Academy of Railway Sciences Corporation Limited, Institute of Computing Technologies, Beijing, China

Received 2 September 2025
Revised 25 September 2025
Accepted 28 September 2025

Abstract

Purpose – This paper conducts a joint analysis of monitoring data in the hidden danger areas of railway subgrade deformation using a data-driven method, thereby realizing the systematic risk identification of regional hidden dangers.
Design/methodology/approach – The paper proposes a regional systematic risk identification method based on Bayesian and independent component analysis (ICA) theories. Firstly, the Gray Wolf Optimization (GWO) algorithm is used to partition each group of monitoring data in the hidden danger area, so that the data distribution characteristics within each sub-block are similar. Then, a distributed ICA early warning model is constructed to obtain prior knowledge such as control limits and statistics of the area under normal conditions. For the online evaluation process, the input data is partitioned following the above-mentioned procedure and the ICA statistics of each sub-block are calculated. The Bayesian method is applied to fuse online parameters with offline parameters, yielding statistics under a specific confidence interval. These statistics are then compared with the control limits – specifically, checking whether they exceed the pre-set confidence parameters – thus realizing the systematic risk identification of the hidden danger area.
Findings – Through simulation experiments, the proposed method can integrate prior knowledge such as control limits and statistics to effectively determine the overall stability status of the area, thereby realizing the systematic risk identification of the hidden danger area.
Originality/value – The proposed method leverages Bayesian theory to fuse online process parameters with offline parameters and further compares them with confidence parameters, thereby effectively enhancing the utilization efficiency of monitoring data and the robustness of the analytical model.

Keywords Bayesian theory, Grey Wolf Algorithm, Independent component analysis, Railway subgrade, Deformation analysis
Paper type Research article

1. Introduction

Railways play an extremely pivotal role in boosting social development and serving economic construction. With the rapid advancement of railways and the ever-increasing demand for operational quality, higher requirements have been put forward for maintaining the full-life-cycle performance of railway infrastructure. Currently, Chinese railway authorities have adopted a wide range of inspection, detection, and monitoring measures to grasp the

Funding: This work was supported by Science and Technology Research and Development Program Project of China State Railway Group Co., Ltd. (award number: K2024X010).

Railway Sciences
Vol. 4 No. 6, 2025
pp. 711-728
Emerald Publishing Limited
e-ISSN: 2755-0915
p-ISSN: 2755-0907
DOI 10.1108/RS-09-2025-0033

operational status of railway permanent way infrastructure (Niu, Liu, & Yang, 2024; Tian, You, & Wang, 2024; Guo, Liu, & Tao, 2021; Liu, Li, & Feng, 2024). The purpose of these measures is to control track equipment failures and construction-related hazards that affect train operation safety, and to achieve early identification and diagnosis of infrastructure faults. At present, a comprehensive intelligent inspection and monitoring system covering status acquisition, comprehensive assessment, safety early warning, and guidance for maintenance and repair has been established, which has yielded positive results.

At present, the BeiDou Navigation Satellite System (BDS) has emerged as a pivotal new-type infrastructure in the domain of spatiotemporal information and positioning-navigation services (Xie, Zhuang, & Kang, 2021). By developing auxiliary facilities such as reference stations, it has achieved positioning accuracy at the millimeter level—capable of playing a vital role in enhancing the detection and monitoring capabilities of railway infrastructure. Currently, railway infrastructure deformation monitoring systems based on BeiDou technology have been widely put into application. Qin Jian proposed a method for site selection and construction of the railway BeiDou Ground-Based Augmentation System based on the technical characteristics of the BeiDou GBAS, providing fundamental support for deformation monitoring of infrastructure along railways (Qin, Pan, & Tao, 2018). Qiang Xiaojun applied high-precision BeiDou positioning technology to the settlement deformation monitoring of high-speed railway bridges (Qiang, 2020).

The aforementioned research methods primarily involve deploying high-precision BeiDou positioning terminals in risk-prone zones along railways to monitor the safety status of railway permanent way infrastructure. However, they are confronted with multiple challenges: the over-reliance on a single deformation assessment method, insufficient deformation predictability, and the absence of comprehensive quantitative deformation evaluation tools based on multi-source data. These limitations hinder their ability to fully meet the safety protection requirements for high-quality railway operations. Furthermore, researchers have proposed a data-driven approach, which can significantly enhance the deformation early warning capability for infrastructure. Zhu constructed a nonlinear mapping model between GNSS data and precise leveling data using a Back Propagation (BP) neural network, corrected the GNSS data, and effectively eliminated the influence caused by errors in the satellite signal acquisition process (Zhu, Shuang, & Sun, 2023). He Kaitao integrated BeiDou satellite technology with remote sensing satellite technology to develop a geological survey service and management system, capable of providing services such as real-time communication and location tracking for geological survey personnel (He, Li, & Wang, 2012). Lu employs the Long Short-Term Memory (LSTM) model to model and predict the deformation monitoring data of the railway infrastructure monitoring system based on the Global Navigation Satellite System (GNSS), thereby enabling the early warning of railway infrastructure disasters (LU, Pan, & Bai, 2022).

The core idea of the aforementioned methods lies in integrating machine learning and statistical theories, which endows them with significant advantages over traditional approaches in identifying the safety status of infrastructure. Given that risk-prone zones along railways span distances from several hundred meters to over a thousand meters, multiple BeiDou terminals are typically deployed on-site to collectively monitor external risks and hazards along railway corridors. Consequently, a critical and pressing issue demanding immediate resolution arises: how to characterize the overall safety status of these risk-prone zones. Against this backdrop, the paper proposes a method for identifying regional systemic risks by fusing Bayesian theory with independent component analysis, thereby enabling the accurate identification of systemic risks in hazardous zones.

2. Fundamental theories

2.1 GWO algorithm

The Grey Wolf Optimization (GWO) algorithm is developed based on the biological phenomenon of the pyramidal hierarchical mechanism exhibited by wolf packs during prey-

hunting processes (Liu, Huang, & Sun, 2020; Long, Ai, & Zhou, 2018). Its core design concept lies in simulating the optimization behavior of gray wolf packs. By simulating the cooperation, competition, and chasing behaviors of wolf packs, the algorithm searches for the optimal solution.

The social hierarchy mechanism in the GWO algorithm simulates the social structure of gray wolf packs, mainly consisting of four ranks: α-wolves, β-wolves, δ-wolves, and ω-wolves. These four ranks correspond to the leader, sub-leader, sub-sub-leader, and ordinary wolves in a gray wolf pack, respectively. In each iteration of the algorithm, each gray wolf determines its position in the pack based on its fitness value.

The algorithm initializes the positions of grey wolves which can be regarded as candidate solutions to the optimization problem. Then, it evaluates the fitness of each grey wolf's position and simulates the social hierarchy within the wolf pack based on the fitness. The mathematical model of this process can be expressed as follows:

$$D = |C \cdot X_P(t) - X(t)| \tag{1}$$

$$X(t+1) = X_P(t) - A \cdot D \tag{2}$$

In the above formula, D represents the distance of individual grey wolf, A and C denote the synergetic coefficient vectors, t indicates the current moment, X_P stands for the position of the alpha wolf, and X represents the position of an ordinary grey wolf. The synergetic coefficient vectors A and C are given as follows:

$$A = 2a \cdot r_1 - a \tag{3}$$

$$C = 2 \cdot r_2 \tag{4}$$

In the above formula, a denotes the convergence factor during the iteration process, which linearly decreases from 2 to 0 as the number of iterations increases. r_1 and r_2 represent random numbers within the interval [0, 1].

The search process of the Grey Wolf Optimization algorithm relies on the α, β, and δ wolves.. During the iteration, the distances between the wolf pack and the prey are determined based on the positions of the α, β, and δ wolves, thereby calculating the distances the wolves need to move. The mathematical model of this process can be expressed as follows:

$$D_\alpha = |C_1 \cdot X_\alpha - X| \tag{5}$$

$$D_\beta = \left|C_2 \cdot X_\beta - X\right| \tag{6}$$

$$D_\delta = |C_3 \cdot X_\delta - X| \tag{7}$$

$$X_1 = X_\alpha - A_1 \cdot D_\alpha \tag{8}$$

$$X_2 = X_\beta - A_2 \cdot D_\beta \tag{9}$$

$$X_3 = X_\delta - A_3 \cdot D_\delta \tag{10}$$

$$X(t+1) = \frac{X_1 + X_1 + X_1}{3} \tag{11}$$

In the above formula, X_α, X_β and X_δ denote the current positions of the alpha, beta, and delta wolves, respectively. C_1, C_2 and C_3 represent random numbers between 0 and 2. D_α, D_β and D_δ

are the distances between the wolves and the prey, respectively. While $A_1, A_2\ A_3$ denote the distance coefficient vectors. X_1, X_2, X_3 represent the updated positions of the wolves. $X(t+1)$ defines the final position of the ω wolf after the $(t+1)^{th}$ iteration.

2.2 Bayesian theory

Bayesian theory is the process of revising the prior probability to the posterior probability based on the information obtained after an event occurs. According to the definition of conditional probability: the probability that one event occurs given that another event has occurred (Deng & Xu, 2018; Guo & Qi, 2014; Zhang, Yuan, & Chen, 2024; Pan, Hu, & Lan, 2019). Therefore, the posterior probability can also be expressed as the conditional probability given new information. The core concept of Bayesian theory is as follows: Initially, the true state of the target event $\widetilde{\theta}$ is unknown, but it is known to follow a probability distribution $P(\widetilde{\theta})$, which is called the prior probability. We can calculate the posterior probability $P(\widetilde{\theta}|E)$ according to the formula, after obtaining new sample information E .If E is a specific event and $\widetilde{\theta} = \theta_j$ is a certain hypothesis, then the conditional probability of $P(\theta_j|E)$ occurring given that E has occurred can be expressed as:

$$P(\theta_j|E) = \frac{P(\theta_j \cap E)}{P(E)} \tag{12}$$

In the above formula, $P(E)$ represents the probability of E event occurring, and $P(\theta_j \cap E)$ represents the probability of both the hypothesis θ_j and E event occurring simultaneously. Its probability can also be expressed as:

$$P(\theta_j \cap E) = P(\theta_j|E) \cdot P(E) = P(E|\theta_j) \cdot P(\theta_j) \tag{13}$$

If $\theta_1, \theta_2, \ldots, \theta_m$ form a partition of the sample space S for hypothesis $\widetilde{\theta}$, and $E \subset S$, $P(\theta_j) \neq 0$, $j = 1, 2, \ldots, m$. $P(E)$ can be defined as:

$$P(E) = \sum_{j=1}^{m} P(E|\theta_j) \cdot P(\theta_j) \tag{14}$$

Given the information event E, Bayes' Theorem can revise the prior probability of the hypothesis $P(\theta_j)$ probability of hypothesis $\widetilde{\theta} = \theta_j$ to the posterior probability $P(\theta_j|E)$.

$$P(\theta_j|E) = \frac{P(\theta_j \cap E)}{P(E)} = \frac{P(E|\theta_j) \cdot P(\theta_j)}{\sum_{j=1}^{m} P(E|\theta_j) \cdot P(\theta_j)} \tag{15}$$

3. Method design

Beidou high-precision positioning terminals can effectively monitor the safety status of railway permanent way infrastructure when deployed in risk-prone areas along railways. In fact, these risk-prone areas along railways mainly include geological disaster-prone spots and key infrastructure along the lines, with spans ranging from several hundred meters to over one thousand meters. Therefore, multiple Beidou terminals are often deployed on-site to jointly monitor external risks and hidden dangers along the railways.

The traditional deformation monitoring methods for infrastructure mainly include two types: one is an active early warning method that provides Beidou deformation monitoring data of a specific monitoring point, and the other directly sets a system threshold and triggers

an alarm when the deformation exceeds the threshold. Although these methods can effectively determine the safety status of infrastructure, the analysis results only reflect the risk and hidden danger conditions of local areas, and fail to describe the overall safety status of risk-prone areas. Therefore, there is an urgent need to conduct joint analysis on various sets of monitoring data in the hidden danger areas along railways, so as to realize the identification of systematic risks in the hidden danger areas.

Accordingly, the paper proposes a methodological framework for railway deformation risk identification, which commences with the partitioning of historical deformation data from monitoring points. Specifically, the Grey Wolf Optimizer (GWO) is employed to segment the historical data into sub-blocks, with the partitioning criterion grounded in data distribution similarity—a strategy that ensures the homogeneity of data characteristics within each sub-block and lays the foundation for subsequent model construction.

Subsequently, PCA model is independently developed for each segmented sub-block to extract a priori knowledge. This knowledge acquisition process yields three core outputs: (1) the prior unmixing matrix corresponding to each prior sub-block, which captures the intrinsic correlation structure of deformation data within the sub-block; (2) the prior statistical metrics that quantify the baseline variation patterns of historical deformation; and (3) the control limits associated with these prior statistics, which serve as critical thresholds for subsequent risk assessment.

For the real-time processing of newly generated deformation data from each monitoring point, the data are first grouped in strict accordance with the prior partitioning scheme derived from historical data—this step ensures consistency between real-time analysis and offline modeling. Leveraging the prior unmixing matrix of the corresponding sub-block, the posterior statistical metrics of the real-time data are computed efficiently. Local risk monitoring for potential hazard zones is then implemented by comparing these posterior statistics against the pre-determined control limits from the a priori knowledge.

(1) To achieve systematic risk identification at the macro scale, Bayesian comprehensive statistics are further calculated based on the posterior metrics of all sub-blocks. By benchmarking these comprehensive statistics against the preset significance level, the global deformation risk along the entire railway line is evaluated. This integrated approach enables the simultaneous realization of local risk screening for specific hazard areas and global risk assessment, thereby forming a systematic deformation risk identification mechanism. The technical implementation of this method is divided into two interdependent phases: the offline modeling phase and the online early warning phase. The detailed procedural steps are outlined as follows. Offline Modeling Process

- Training Data Preparation: Collect foundational training data for model development, serving as the raw input for subsequent algorithmic processing.
- Data Partitioning: Utilize the GWO method to divide the training data into N sub-blocks (i.e. Sub-block 1, Sub-block 2, ..., Sub-block N). This intelligent optimization algorithm enables rational data segmentation, laying the groundwork for targeted modeling in subsequent steps.
- ICA Modeling: Establish an ICA model for each sub-block (denoted as ICA Model 1, ICA Model 2, ..., ICA Model N). The ICA algorithm is leveraged to extract the intrinsic features and latent patterns embedded within each sub-block of data.
- Distributed Model Integration: Integrate all sub-block ICA models to construct a "Distributed ICA Early-Warning Model", thus completing the offline model development process and providing a foundational framework for online early-warning applications.

(2) Online Warning Process

- Online Data Input: Collect real-time online data, which serves as the target object for early-warning analysis.
- Data Partitioning: Apply the Grey Wolf Optimization (GWO) algorithm to re-partition the online data. This process strictly aligns with the partitioning logic adopted in offline modeling, thereby ensuring consistency in data processing across the offline-online workflow.
- Statistic Computation: Calculate monitoring statistics for the online data within each re-partitioned sub-block, with the aim of capturing characteristic features of abnormal data.
- BIC Calculation: Leverage Bayesian inference to compute the BIC, which is designated as the core metric for risk discrimination.
- Risk Discrimination: Judge whether the computed BIC value is greater than or equal to the preset threshold (β). If the condition "BIC≥β" is satisfied, initiate "systematic risk identification" to pinpoint existing risks within the system; otherwise, continue with routine data monitoring.

Figure 1 shows Overall Design Diagram.

4. Key technologies

4.1 Variable partitioning design

Since hazard-prone areas can be treated as an integrated whole, there exist inherent correlations among monitoring variables. In this paper, multiple sets of monitoring data within the areas are processed in a combined manner, which further enhances the utilization efficiency of deformation monitoring data and the robustness of the early-warning model. After obtaining multiple groups of feature data, the data are divided into several sub-blocks according to their distribution characteristics. Subsequently, an Independent Component Analysis (ICA) early-warning model is established for each sub-block individually.

Therefore, how to partition the reconstructed dataset constitutes the primary task in constructing the distributed ICA early-warning model. The GWO algorithm is adaptable to various types of optimization problems, requiring minimal prior knowledge about the problem itself. It also exhibits advantages such as excellent convergence performance and a low

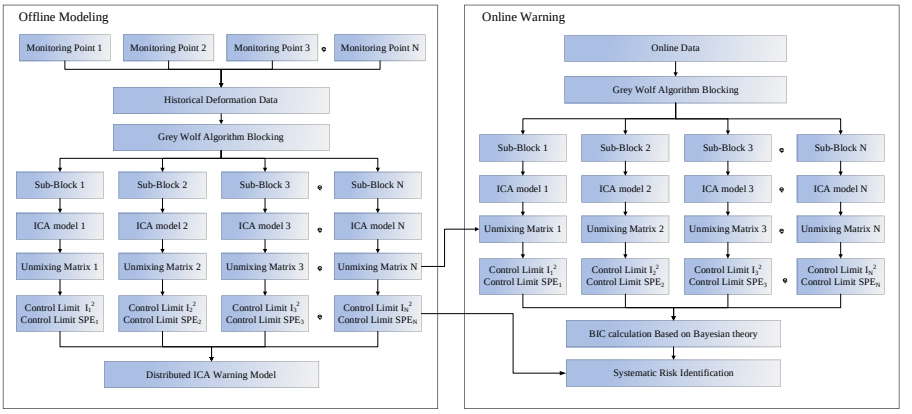


Figure 1. Overall design diagram. Source(s): Authors' own work

tendency to fall into local optima. For these reasons, this paper employs the GWO algorithm to partition the monitoring data, ensuring that the data distribution of variables within each sub-block remains similar.

When partitioning data with the GWO algorithm, the grey wolf pack follows a hierarchical pyramid structure. Specifically, the grouping result determined by the α-wolf represents the optimal partitioning of monitoring variables; the result from the β-wolf denotes the suboptimal partitioning; the outcome of the δ-wolf corresponds to the sub-suboptimal partitioning; and the grouping from the ω-wolf serves as the candidate partitioning.

In the grey wolf optimization process, the target prey corresponds to the optimization function, which is defined as the sum of KL divergences among variables within each group. By selecting KL divergence as the optimization metric, we can ensure that data distributions within individual sub-blocks remain consistent. This consistency enables Independent Component Analysis (ICA) early-warning models for different sub-blocks to effectively extract deformation features specific to their respective sub-blocks. In turn, this ensures that the integrated ICA model can efficiently capture multi-dimensional deformation characteristics across varying distributions. The definition of KL divergence is as follows:

$$KL(P|Q) = \sum P(x)\log\frac{P(x)}{Q(x)} \tag{16}$$

In the above formula, $KL(P|Q)$ denotes the KL divergence, where $P(x)$ and $Q(x)$ represent two different probability distributions of the random variable X. When the two distributions are similar, KL divergence approaches 0; conversely, the more dissimilar they are, the larger the KL divergence becomes.

When the GWO algorithm is applied for partitioning, the first step involves determining the initial grouping schemes generated by considering all permutations and combinations of every two variables, where the total number of grouping schemes is set to match the size of the grey wolf pack; following the determination of initial groupings, iterative calculations are performed using the predefined optimization function, with the core objective of identifying the grouping index corresponding to the minimum KL divergence, which marks the completion of the first iteration cycle; subsequently, the two variables included in the optimal grouping identified in the previous cycle are excluded from the candidate variable pool, and the remaining variables are reorganized and recombined to form new candidate groupings; this entire process, which includes generating new groupings and selecting the optimal one based on KL divergence, is repeated iteratively until all monitoring variables are assigned to respective groups, and the detailed partitioning process of monitoring variables based on the GWO algorithm is described as follows:

(1) Initialize the grey wolf pack. Generate a wolf pack through pairwise variable combinations, where each wolf represents a potential solution in the solution space and contains multiple matrices—each matrix corresponding to a variable partitioning scheme. Simultaneously, initialize the synergy coefficients A and C, along with the linearly decreasing parameter A.

(2) Compute the KL divergence between variables within each matrix of every wolf in the pack, with the sum of these divergences serving as the fitness function.

(3) Determine the social hierarchy of each wolf in the pack based on their respective KL divergence values.

(4) Update the positions of the α-wolf, β-wolf, δ-wolf, and ω-wolf (corresponding to updates of the optimal, suboptimal, and sub-suboptimal solutions in the grouping schemes) using formulas (5) to (11), with reference to the current social ranks of the pack.

(5) Update the synergy coefficients A and C, as well as the linearly decreasing parameter a, in accordance with formulas (3) to (4).

(6) Recalculate the KL divergence for each wolf based on their updated positions, and revise the pack's social hierarchy according to the new fitness values.

(7) Check for satisfaction of the stopping criteria. If the maximum number of iterations is reached or the KL divergence converges to a preset threshold, terminate the algorithm and output the optimal or near-optimal grouping scheme. If the criteria are not met, return to Step 4 to continue iterations until the specified conditions are fulfilled.

4.2 ICA model design

Assume that the deformation monitoring data set $X = [x_1, x_2, \ldots, x_n]^T \in R^{n\times m}$ reconstructed by partitioning as described above contains m variables with n samples. The data set is decomposed into p sub-blocks, and each sub-block can be denoted as $X_i (i = 1, 2, \ldots, p)$.Then the variables within the i^{th} sub-block can be expressed as $X_i = [X_{1,i}, X_{2,i}, \ldots, X_{l,i}]$ where $X_{l,i}$ denotes the j^{th} measurement variable in the i^{th} sub-block. By applying ICA to decompose the i^{th} sub-block, X_i can be represented as a linear combination of j unknown independent components $S_i = [S_{1,i}, S_{2,i}, \ldots, S_{j,i}]$, that is:

$$X_i = A_i S_i + E \tag{17}$$

In the above formula, A_i denotes the mixing matrix, S_i represents the independent component matrix, E denotes the residual matrix, and X_i stands for the i^{th} sub-block, which has been normalized and whitened. In this formula, if the residual matrix is negligible, the ICA model can be expressed as:

$$X_i = A_i S_i \tag{18}$$

The core of ICA lies in calculating the independent component matrix S_i and the mixing matrix A_i solely from the observed data X_i. The relationship between the estimated independent component matrix S_i and the data X_i can be expressed as:

$$\widehat{S}_i = W_i X_i \tag{19}$$

In the above formula, $\widehat{S}_i$ denotes the independent component matrix estimated from the original deformation monitoring data, and W_i represents the unmixing matrix. For the data X_i, the independent component matrix can be estimated using the FastICA algorithm. After obtaining the estimated independent component matrix $\widehat{S}_i$ and the unmixing matrix W_i, for the test data of the sub-block i, the statistical monitoring metrics I_i^2 and SPE_i are constructed, namely:

$$I_i^2 = (W_i x_{new,i})^T (W_i x_{new,i}) = S_{new,i}^T S_{new,i} \tag{20}$$

$$SPE_i = (x_{new,i} - \widehat{x}_{new,i})^T (x_{new,i} - \widehat{x}_{new,i}) \tag{21}$$

In the above formula, $x_{new,i}$ represents the newly input sample data of the i^{th} sub-block. I_i^2 and SPE_i reflect the changes in multivariate variables. It is necessary to determine the confidence interval for the normal state of the hidden danger area before conducting online monitoring. The statistical limits of I_i^2 and SPE_i can be obtained using the univariate kernel

density estimation method. Assuming that $y_1, y_1, \ldots, y_n$ the samples are independent and identically distributed, the mathematical description of kernel density estimation is as follows:

$$f_m(x) = \frac{1}{mh_m} \sum_{i=1}^{m} k\left(\frac{y - y_i}{h_m}\right) \tag{22}$$

In the above formula, f_m is the kernel density estimate for the unknown probability density function f, where m denotes the sample size, y_i represents the process data of the sub-block, $k(\cdot)$ denotes the kernel function, and h_m denotes the smoothing parameter or bandwidth. For each sub-block, all statistics and their corresponding control limits are established. If a statistic falls below its corresponding statistical threshold, it suggests that the area represented by the sub-block is generally stable; otherwise, it suggests the presence of potential risks.

$$I_i^2 \le I_{i,th}^2 \tag{23}$$

$$SPE_i \le SPE_{i,th} \tag{24}$$

4.3 Decision-making system design

Through the integration of the GWO algorithm and ICA models, the optimal ICA model and corresponding control limits for each sub-block are derived. Systematic risk identification of hazard-prone areas is achieved by fusing sub-block statistics via Bayesian comprehensive inference. Specifically, the statistical information of I_i^2 and SPE_i within each sub-block is converted into conditional probabilities, with the conversion method detailed as follows:

$$P_{I^2}(x_i|N) = \exp\left\{-\frac{I_{i,new}^2}{I_{i,th}^2}\right\} \tag{25}$$

$$P_{I^2}(x_i|F) = \exp\left\{-\frac{I_{i,th}^2}{I_{i,new}^2}\right\} \tag{26}$$

$$P_{SPE}(x_i|N) = \exp\left\{-\frac{SPE_{i,new}}{SPE_{i,th}}\right\} \tag{27}$$

$$P_{SPE}(x_i|F) = \exp\left\{-\frac{SPE_{i,th}}{SPE_{i,new}}\right\} \tag{28}$$

In the above formula, x_i denotes the sample data of the block, F represents the fault state, and N represents the normal state. $SPE_{i,th}$ and $I_{i,th}^2$ respectively denote the confidence limits of SPE and I^2 for the i^{th} sub-block, while $SPE_{i,new}$ and $I_{i,new}^2$ respectively represent the SPE *and* I^2 statistical values corresponding to the new sample x_i of the i^{th} sub-block. Then, the anomaly probability of x_i can be expressed as:

$$P_{I^2}(F|x_i) = \frac{P_{I^2}(x_i|N)P(F)}{P(x_i)} \tag{29}$$

$$P_{I^2}(x_i) = P_{I^2}(x_i|N)P(N) + P_{I^2}(x_i|F)P(F) \tag{30}$$

When the significance level is β, $P(N)$ can be defined as $1 - \beta$, and $P(F)$ is defined as β. Then, the final Bayesian comprehensive statistic can be expressed as:

$$BIC = \sum_{i=1}^{B} \left\{ \frac{P_{I^2}(x_i|F)P_{I^2}(F|x_i) + P_{SPE}(x_i|F)P_{SPE}(F|x_i)}{\sum_{i=1}^{B} P_{I^2}(x_i|F) + P_{SPE}(x_i|F)} \right\} \tag{31}$$

In the process of active early warning for regional risks, the BIC statistic is used to determine whether the region is stable. The control limit of BIC is set as the significance level β. When the calculated BIC value exceeds the significance level β, it indicates that there are potential risks in the region; otherwise, when the calculated BIC value is lower than the significance level β, it indicates that the region is in a normal state.

5. Simulation experimental analysis

To verify that the proposed method can effectively enable active early warning for railway permanent way infrastructure, this paper selects railway infrastructure in a typical section for field measurement data analysis. The railway safety management department has deployed 1 reference station and 10 monitoring stations in this hazard-prone area to monitor the displacement of the permanent way infrastructure. Among them, the maximum distance between the reference station and any monitoring station does not exceed 300 m, and the monitoring stations are evenly distributed on both sides along the railway line. The monitoring data covers a time span from September 1, 2022, to October 10, 2022, totaling 40 days. The data acquisition frequency is 1 epoch every 10 seconds, with a solution cycle of 10 minutes, resulting in a total of 5,760 data points.

Figure 2(a)–(j) show the time series plots of raw data from each monitoring point in the risk-prone area. It can be seen from the figures that the monitoring data can reflect the deformation trend of the railway permanent way infrastructure within 40 days. Meanwhile, relatively severe spike features caused by external environmental factors can also be observed. Most monitoring points show relatively stable changes within 40 days; however, monitoring points 5 and 9 still exhibit a distinct settlement trend.

First, a statistical characteristic analysis was conducted on each group of monitoring data in the hazard-prone area. Figure 3(a)–(j) present the probability density fitting plots of each monitoring point, where the blue histograms represent the probability density distribution of the original data, the red curves denote the fitting results of the traditional Gaussian probability density distribution, and the green curves stand for the fitting results of the α-stable distribution. Table 1 lists the fitting parameters of the α-stable distribution for each monitoring point.

As observed in Figure 3, the α-stable distribution fitting can better describe the true probability density distribution of the data. The α-stable distribution curves do not completely overlap with the traditional Gaussian distribution curves, which indicates that all groups of monitoring data in this area do not follow a Gaussian distribution. From the fitting results, none of the data groups exhibit significant heavy-tailed or skewed characteristics, suggesting that no severe deformation has occurred in the hazard-prone area. This finding is consistent with the relatively stable variation trend shown in the time-series plots of each group of monitoring data. Additionally, the α-stable distribution parameters of all monitoring data groups listed in Table 1 are not equal to 2, which further verifies that all groups of monitoring data follow a non-Gaussian distribution. For such non-Gaussian monitoring data, the Independent Component Analysis (ICA) modeling approach can be considered to identify the abnormal status of the hazard-prone area.

Figure 4 illustrates 10 independent components (ICs) decomposed from the 10 groups of monitoring data via the Independent Component Analysis (ICA) algorithm. Asdepicted in the figure, the 10 ICs exhibit a generally stable temporal variation trend; nonetheless, three specific components (i.e. ICA-1, ICA-2, and ICA-3) manifest significant fluctuations on

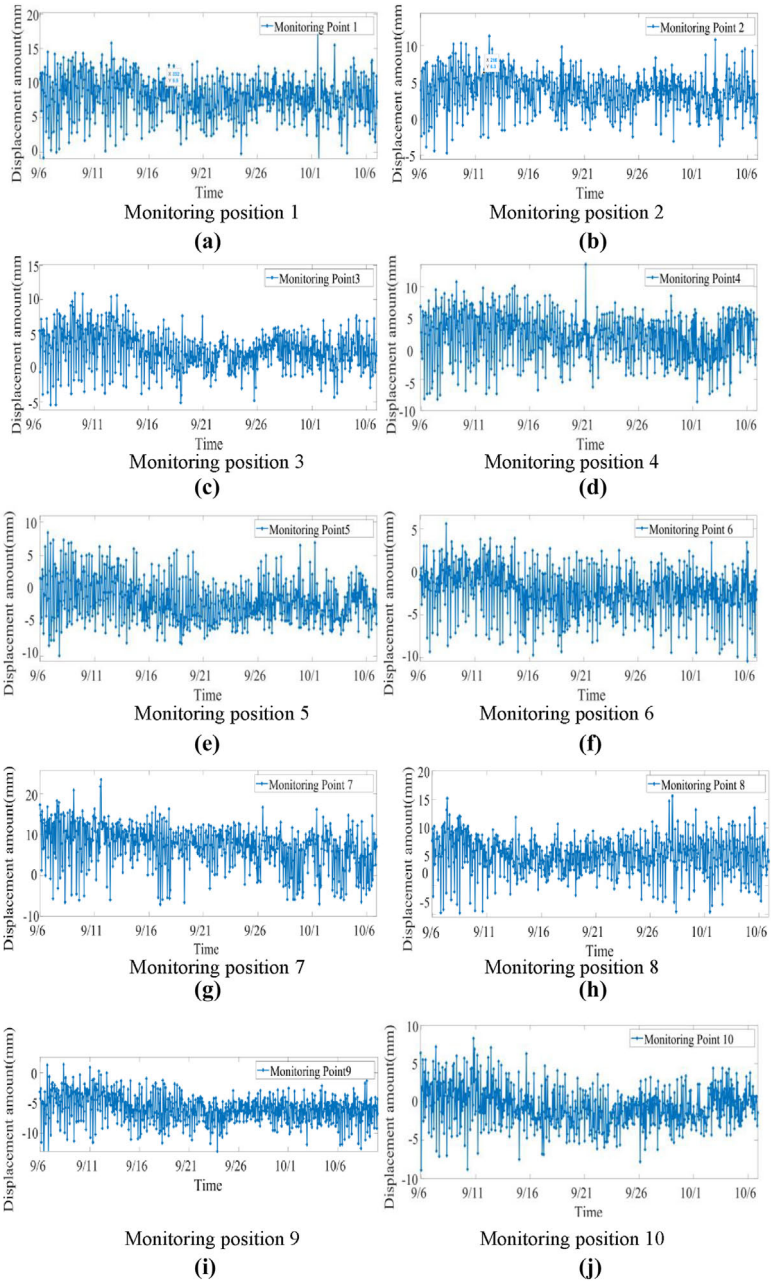


Figure 2. Railway permanent infrastructure deformation monitoring data. Source(s): Authors' own work

September 11th. This phenomenon implies that the spatial regions corresponding to these three ICs experienced notable external disturbances during the aforementioned time window.

To establish the ICA model, the first 30-day segment of the monitoring data was designated as the training dataset. The core objective of this step is to derive the control limits for the I^2

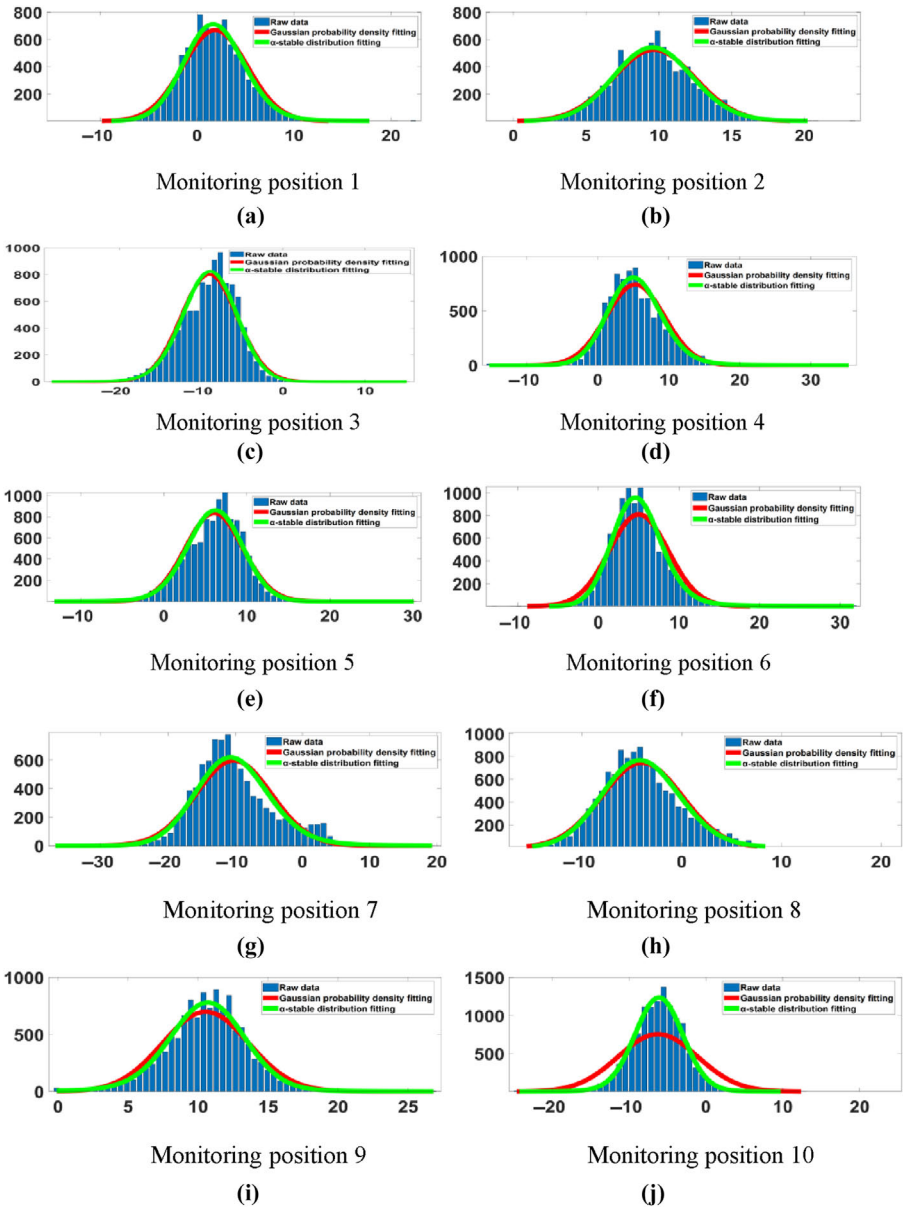


Figure 3. The PDF fitting of deformation monitoring data for each monitoring point. Source(s): Authors' own work

statistic and the SPE statistic, where these limits function as quantitative criteria for evaluating the stability level of the entire monitored area. Subsequently, the remaining 10-day segment of the data was utilized as the test dataset to validate and assess the dynamic stability of the target area. In essence, the operational principle of ICA-based stability monitoring lies in computing the I^2 nd SPE statistics at each discrete time step, followed by a comparative analysis between these computed statistics and the pre-derived control limits. If the value of either statistic

Table 1. The index of PDF fitting

Monitoring location	α	β	γ	δ
Monitoring Point 1	1.9306	1	2.2019	1.9218
Monitoring Point 2	1.9580	1	1.9253	9.7076
Monitoring Point 3	1.9241	0.9146	2.0472	4.9340
Monitoring Point 4	1.9073	1	2.6612	5.4221
Monitoring Point 5	1.9522	−1	2.3638	6.0013
Monitoring Point 6	1.8682	1	2.1881	5.1104
Monitoring Point 7	1.8987	1	3.8263	−9.9856
Monitoring Point 8	1.9355	1	2.7765	−3.8972
Monitoring Point 9	1.8411	−0.3266	1.9491	10.5127
Monitoring Point 10	1.9159	−0.2234	2.2377	−6.1875

Source(s): Authors' own work

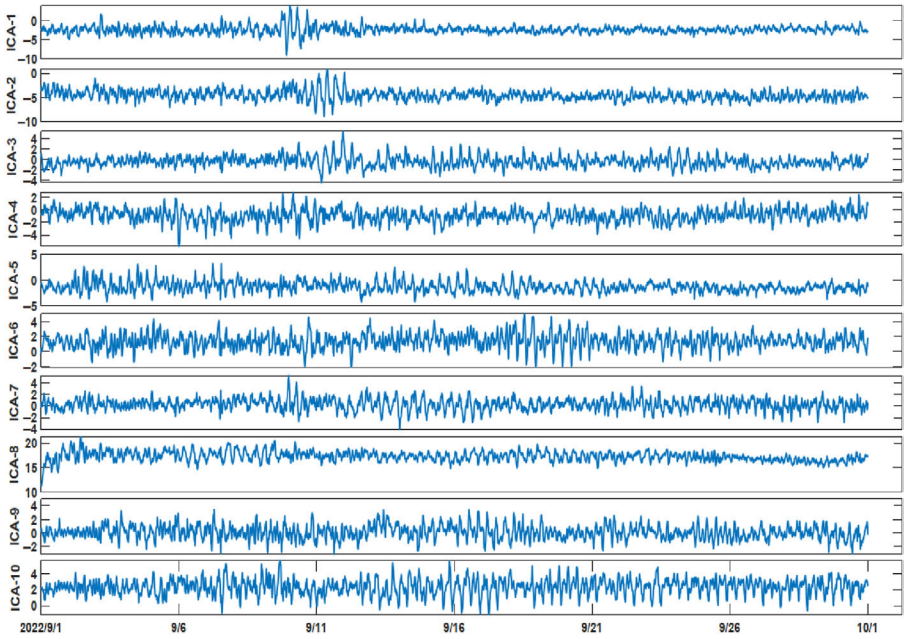


Figure 4. Independent components of ICA model. Source(s): Authors' own work

exceeds its corresponding control limit, this observation indicates the occurrence of an abnormal state within the monitored area.

Figure 5(a) and (b) present the results of ICA analysis conducted on the 10 groups of monitoring data. Specifically, Figure 5(a) illustrates the temporal evolution of the I^2 statistic for the monitoring data at each discrete time point, while Figure 5(b) depicts the temporal variation of the Squared Prediction Error (SPE) statistic across the same time sequence. In both subfigures, the black curves represent the I^2 and SPE values computed via the ICA algorithm at each corresponding time step, whereas the red curves denote the pre-derived control limits, which are calibrated using the training dataset.

As inferred from the graphical results, neither the I^2 statistic nor the SPE statistic exhibits a persistent exceedance of their respective control limits over the entire monitoring period. This

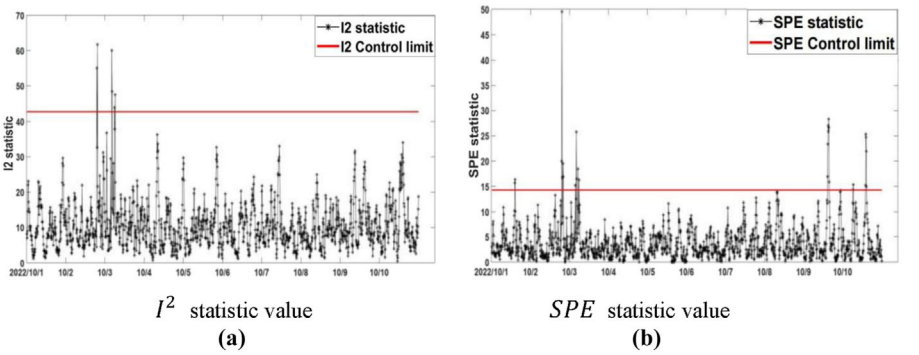


Figure 5. Statistic value of ICA model. Source(s): Authors' own work

observation indicates that the geological system of the monitored area maintained a state of relative stability for the majority of the observation window. Furthermore, a distinct anomaly can be identified: around October 3rd, both the I^2 and SPE statistics simultaneously exceeded their corresponding control limits. This co-occurrence of statistic exceedance suggests that the geological region under investigation experienced a transient increase in activity during this specific time interval. The consistency between the I^2 and SPE anomaly signals provides cross-validation, thereby further verifying the accuracy and reliability of the proposed ICA-based stability monitoring framework.

Additionally, it is noteworthy that the SPE statistic also exhibited a transient exceedance of its control limit around October 10th, a phenomenon not clearly captured by the I^2 statistic. Concurrently, the control limit for the SPE statistic is quantitatively smaller than that for the I^2 statistic. This result demonstrates that, in the context of geological stability monitoring, the SPE statistic possesses higher sensitivity in detecting subtle abnormal signals associated with incipient geological changes, compared to the I^2 statistic.

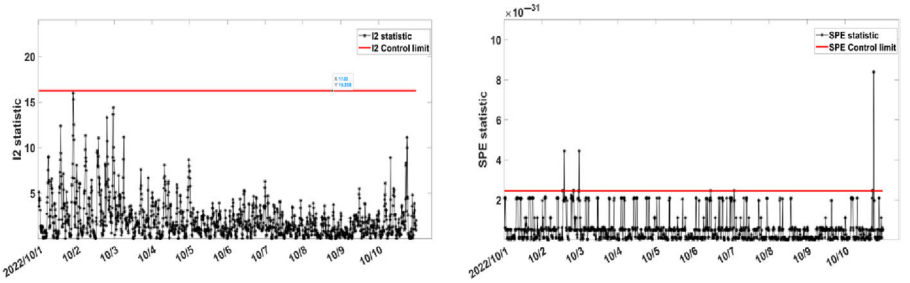
To further enhance the robustness of the ICA model and improve the utilization of monitoring data, as well as to conduct a multi-dimensional assessment of the risk status in the hidden danger area, a distributed ICA model is established by grouping each set of monitoring data. In this paper, the GWO algorithm is employed for optimization. First, the grey wolf pack is set as all pairwise permutations and combinations of each group of monitoring data. The corresponding KL divergence for each combination is calculated, and the group corresponding to the minimum KL divergence is selected as the leader wolf. Then, iterations are performed until all grouping results are determined. The purpose of this process is to ensure that the data distributions within each group are similar. The grouping results are shown in Table 2. A total of 10 sets of monitoring data are evenly divided into 5 sub-blocks, specifically:

Table 2. Block results of each monitoring point in risky area

Sub-block number	Monitoring point
Sub-block 1	Monitoring Point 5, 6
Sub-block 2	Monitoring Point 1, 10
Sub-block 3	Monitoring Point 2, 4
Sub-block 4	Monitoring Point 8, 9
Sub-block 5	Monitoring Point 3, 7

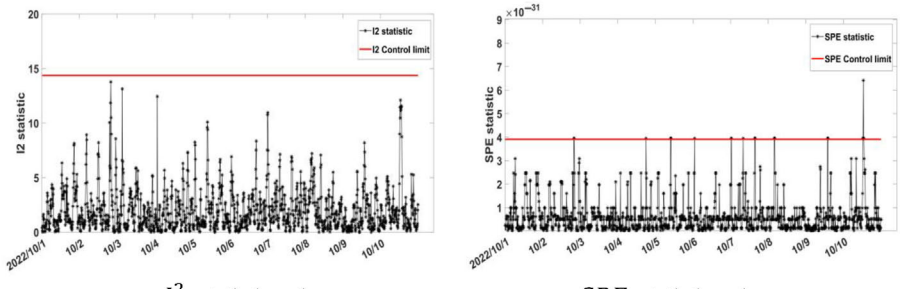
Source(s): Authors' own work

Figure 6–10 show the I^2 and SPE statistics corresponding to the five groups of monitoring data. It can be observed from the figures that the monitoring data in each group do not exceed the control limits in a large range, which is basically consistent with the results of the ICA model for the 10 groups of monitoring data mentioned above. In addition, it can also be observed from the figures that there are partial differences between the statistics of each group



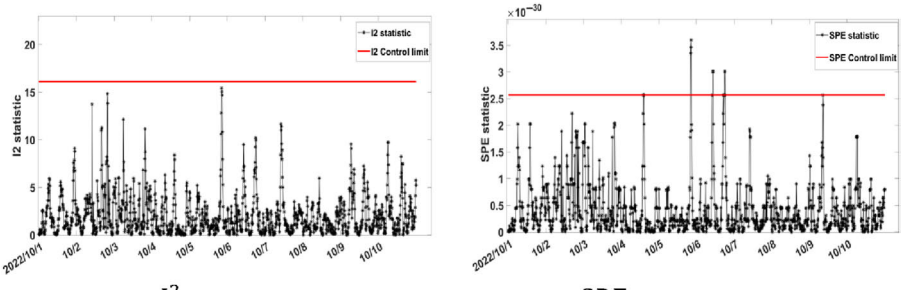
I^2 statistic value (a) SPE statistic value (b)

Figure 6. Statistic value of block1. Source(s): Authors' own work



I^2 statistic value (a) SPE statistic value (b)

Figure 7. Statistic value of block2. Source(s): Authors' own work



I^2 statistic value (a) SPE statistic value (b)

Figure 8. Statistic value of block3. Source(s): Authors' own work

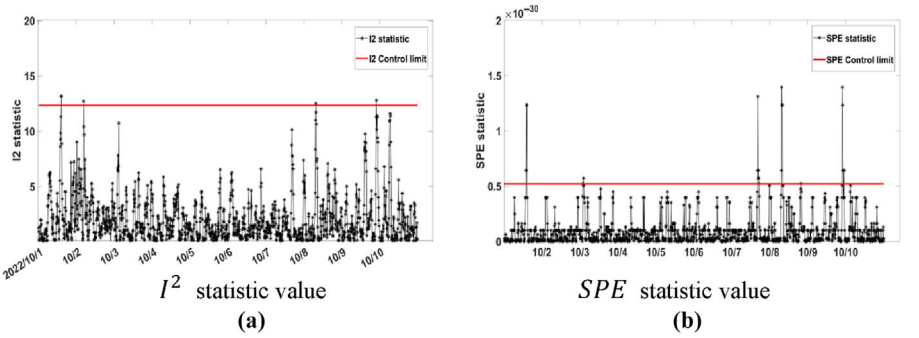


Figure 9. Statistic value of block4. Source(s): Authors' own work

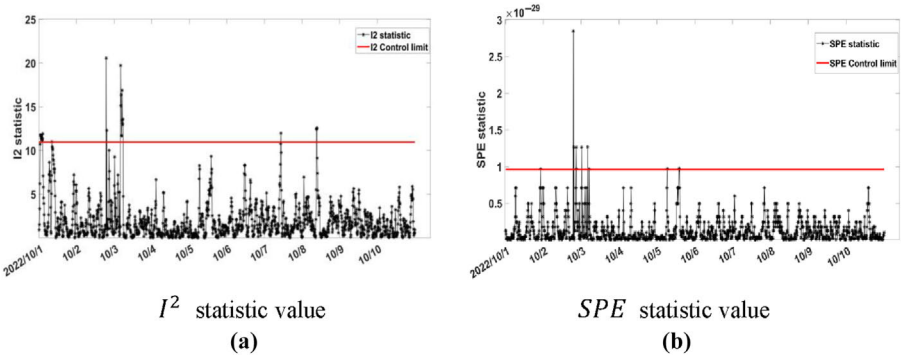


Figure 10. Statistic value of block5. Source(s): Authors' own work

and those of the 10 groups. For example, the SPE statistic of sub-block 4 showed an abnormality around October 8th, while the SPE statistic shown in Figure 5 did not exhibit obvious abnormalities in this period. The reason is that the SPE statistic in Figure 5 is a risk assessment indicator based on the entire hidden danger area, whereas each group is based on two sets of variables, thus being more capable of reflecting specific detailed features. Furthermore, the I^2 and SPE control limits of each sub-block are different from those shown in Figure 5. This is because the statistical limits of each sub-block are derived from the training data of the respective sub-blocks.

Figure 11 shows the distributed ICA model established based on the statistics of the five sub-blocks, where the control limit corresponds to the confidence level. In the figure, the red line represents the control limit corresponding to the confidence level, and the black line represents the statistic at each moment. It can be observed from the figure that the statistics in each group do not exceed the control limit in a large range, which is basically consistent with the results of the ICA model for the 10 groups of monitoring data mentioned above. In addition, it can be observed from the figure that the distributed ICA model exhibits local abnormalities not only on October 3rd and October 10th but also on October 8th. This indicates that the distributed ICA model can reflect the local abnormal information of each group compared with the traditional ICA model. Furthermore, from the I^2 and SPE statistics of each sub-block, it is evident that the geology of the region was indeed relatively active on October 3rd, but the overall situation still did not reach a severe level. In fact, the control limit only serves as an alarm threshold, and the distance between the statistic and the control limit can

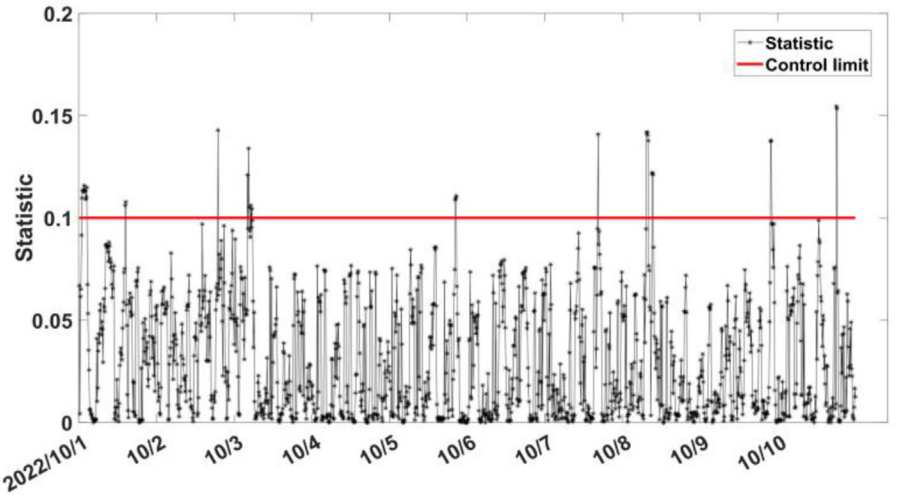


Figure 11. Each component and residual. Source(s): Authors' own work

also be used to determine the overall stability of the hidden danger area. Both the traditional ICA model and the distributed ICA model have cases where individual discrete points exceed the control limit threshold. However, for the overall risk-prone area, it can be stipulated that the area is judged to have entered a relatively severe risk state only when several consecutive statistics exceed the control limit threshold.

6. Conclusion

This paper proposes a regional systematic hazard identification method based on Bayesian theory and Independent Component Analysis (ICA). The method comprises two functional modules, namely offline modeling and online evaluation. In the offline modeling phase, the Grey Wolf Optimizer (GWO) algorithm is utilized for dataset partitioning. This partitioning operation ensures that variables within each sub-block present consistent distribution characteristics, which in turn enhances the utilization efficiency of deformation monitoring data in hazard-prone areas and improves the robustness of the established model. After completing the dataset partitioning, the ICA algorithm is separately applied to each sub-block for model construction, thereby forming a distributed ICA early-warning model dedicated to the hazard-prone area. During this modeling process, key prior knowledge is derived, including the control limits corresponding to the I^2 statistic and the SPE statistic. For the online evaluation phase, the newly updated monitoring data are processed in accordance with the aforementioned partitioning and modeling procedures. This processing step generates the control metrics for each sub-block of the updated data. Subsequently, Bayesian theory is employed to perform fusion analysis on the statistics of all sub-blocks. The fused statistical results are then compared with pre-defined confidence parameters. Based on this comparison, the systematic hazard identification for the target hazard-prone area is ultimately achieved.

References

Deng, X., & Xu, Y. (2018). Multimode non-Gaussian process fault detection based on Bayesian-ICA. *Control Engineering of China, 25*(3), 402–407.

Guo, B., & Qi, P. (2014). Based on the improved ICA fault diagnosis of industrial processes. *Industrial Instrumentation and Automation, 2014*(3), 11–15.

Guo, J., Liu, J., & Tao, K. (2021). Research on comprehensive management technology of railway infrastructure inspection data. *Railway Technical Innovation*, *2021*(6), 110–117.

He, K., Li, Z., & Wang, D. (2012). Overview on the design of the service and management system for field geological survey based on the remote sensing and Beidou satellites. *Journal of Geomechanics*, *18*(3), 203–212.

Liu, Y., Huang, X., & Sun, Q. (2020). Integrated optimization control of performance and jet noise of turbofan engine based on improved grey wolf optimization algorithm. *Journal of Nanjing University of Aeronautics and Astronautics*, *52*(4), 532–539.

Liu, Y., Li, P., & Feng, B. (2024). Analysis and prediction of railway infrastructure deformation monitoring data based on fractional order statistical theory. *IEEE Access*, *2024*(11), 001121203000001.

Long, Z., Ai, X., & Zhou, H. (2018). Network traffic predicting model based on improved grey wolf optimization algorithm. *Application Research of Computers*, *35*(6), 1845–1848.

LU, Z., Pan, P., & Bai, X. (2022). Prediction model for displacement data of railway infrastructure. *Railway Computer Application*, *31*(03), 12–18.

Niu, D., Liu, J., & Yang, F. (2024). Research and practice on big data analysis technology for inspection and monitoring of high speed railway infrastructure. *China Railway*, *2024*(2), 1–11.

Pan, Q., Hu, Y., & Lan, H. (2019). Information fusion progress: Joint optimization based on variational Bayesian theory. *Acta Automatica Sinica*, *2019*(17), 1207–1223.

Qiang, X. (2020). Application of Beidou positioning technology in settlement monitoring of high speed railway. *Railway Engineering*, *60*(7), 81–84.

Qin, J., Pan, P., & Tao, C. (2018). Construction of railway Beidou ground-based AUgmentation system and base station site selection. *Railway Computer Application*, *27*(3), 11–14.

Tian, X., You, M.., & Wang, J. (2024). High-speed railway infrastructure detection data management platform and application. *Railway Computer Application*, *33*(11), 49–55.

Xie, Z., Zhuang, J., & Kang, C. (2021). Internet of Things technology and application based on Beidou system. *Journal of Nanjing University of Aeronautics and Astronautics*, *53*(3), 329–337.

Zhang, L., Yuan, L., & Chen, N. (2024). Bridge weighing-in-motion algorithm theory based on Bayesian posterior estimation and tests. *Journal of Vibration and Shock*, *2024*(8), 20–27.

Zhu, Y., Shuang, M., Sun, D., & Guo, H. (2023). Algorithm and application of foundation displacement monitoring of railway cable bridges based on satellite observation data. *Applied Sciences-Basel*, *13*(5), 2868. doi: 10.3390/app13052868.

Corresponding author
Chengwen Wu can be contacted at: 731736482@qq.com

Chengwen Wu received his Master's degree from China Academy of Railway Sciences in 2024. He is an engineer and his work primarily involves the exploration of key technologies in railway digitalization and communication network optimization.