World Journal of Engineering Research
Print ISSN: 2959-9865
Online ISSN: 2959-9873
DOI: https://doi.org/10.61784/wjer3121

SLOPE DEFORMATION DATA CALIBRATION BASED ON LINEAR REGRESSION AND SEGMENTED IDENTIFICATION ALGORITHMS

GuiTao Li, ShiJie Zhang*
School of Transportation and Vehicle Engineering, Shandong University of Technology, Zibo 255000, Shandong, China.
**Corresponding Author: ShiJie Zhang*

Abstract: This study develops an algorithmic framework for slope deformation monitoring data calibration and staged deformation identification. The selected modeling section focuses on correcting systematic deviation in fiber-optic displacement measurements and identifying transition nodes in surface displacement sequences. A univariate linear regression model is established between fiber-optic displacement data and vibrating-wire displacement reference data, and the calibration parameters are solved by minimizing the residual sum of squares. Five-fold cross-validation is used to evaluate stability, while MSE, RMSE, MAE, and R2 measure calibration accuracy. The full-sample fitting result reaches MSE = 7.195, RMSE = 2.682, MAE = 1.308, and R2 = 0.999, indicating that the calibrated displacement closely follows the reference sequence. For staged deformation identification, the method combines Savitzky-Golay smoothing, Hampel outlier detection, sliding Welch t-test screening, physical duration and monotonicity constraints, and segmented fitting. The final transition nodes are identified at indices 5261 and 8861, dividing the sequence into slow uniform deformation, accelerated deformation, and rapid deformation. The average velocities of the three stages are 0.1753, 0.6115, and 4.1211 mm/h, and the fitted R2 values are 0.947975, 0.983713, and 0.998814. The results show that the proposed framework can reduce sensor bias, identify deformation-stage evolution, and support monitoring-based warning analysis.
Keywords: Linear regression; Sensor data calibration; Savitzky-Golay smoothing; Hampel detection; Segmented deformation identification

1 INTRODUCTION

Slope deformation monitoring increasingly depends on continuous multi-source sensor data. Fiber-optic displacement meters, vibrating-wire displacement meters, surface displacement observations, and time-indexed monitoring sequences provide complementary descriptions of deformation evolution. In practical monitoring, however, raw data are affected by sensor zero drift, installation deviation, missing values, short-term disturbances, and local jumps. If these errors are not corrected, subsequent deformation interpretation may overestimate or underestimate the true displacement trend. A reliable analytical procedure should therefore begin with data calibration and should make the monitored sequence closer to a stable reference measurement before extracting deformation-stage information [1].

The central research issue is how to build a compact algorithmic framework that can both correct systematic measurement deviation and identify staged deformation transitions. The calibration part needs to map fiber-optic displacement observations to vibrating-wire reference data with minimum residual error, while retaining the original temporal trend. The stage-identification part needs to distinguish real transition nodes from short-lived engineering disturbances or sensor anomalies. This distinction is important because a temporary velocity spike is not necessarily a structural stage transition. A useful model must therefore combine statistical testing, signal smoothing, anomaly screening, physical duration constraints, and segmented fitting accuracy [2].

The research scheme contains two connected modules. In the calibration module, valid paired observations are extracted after missing-value removal. A univariate linear regression relation is established between the measured fiber-optic sequence and the reference sequence. The objective function minimizes the squared residuals between calibrated values and reference values. Least squares estimation is used to solve the proportional and offset parameters, and five-fold cross-validation evaluates model stability through MSE, RMSE, MAE, and R2 [3]. The calibrated values are then compared with the original and reference data to verify whether the systematic deviation is reduced.

In the staged deformation module, the surface displacement series is first smoothed by Savitzky-Golay filtering, and velocity and acceleration sequences are calculated. Hampel filtering detects instantaneous velocity jumps, while a sliding Welch t-test evaluates whether the velocity distributions before and after a candidate position differ significantly [4]. Search ranges and physical constraints are introduced to avoid unreasonable early or terminal nodes. The final stage boundaries are selected by minimizing segmented fitting residuals under monotonic velocity-increase constraints. Linear, quadratic polynomial, and exponential models are then used to represent slow uniform deformation, accelerated deformation, and rapid deformation. The model is finally assessed through R2, RMSE, F-test information, and average stage velocity, forming a complete data correction and deformation-stage recognition workflow [5]. In addition, the workflow keeps each computational step transparent: calibration parameters, residual indicators, transition-node criteria, stage velocities, and fitting scores are all reported together, so the corrected monitoring sequence and the segmented deformation pattern can be checked consistently in later warning-oriented analysis for operational use.

2 DATA CALIBRATION AND STAGED DEFORMATION IDENTIFICATION MODEL

2.1 Linear Calibration Model

2.1.1 Specific analysis

This section designs a data calibration model based on univariate linear regression. The model uses the least-squares regression method to correct the deviation of new fiber-optic displacement-meter data, so that the corrected data are as close as possible to the reference data from the traditional vibrating-wire displacement meter [6].

The two displacement time-series datasets from the same slope and monitoring point are obtained. The data are read from Excel files using pandas tools, and fiber-optic displacement-meter data A and vibrating-wire displacement-meter data B are extracted. Data A is used as the data to be corrected, while data B is used as the reference data. Missing samples in either data sequence are removed, and complete valid data pairs are retained to construct the model training samples. Considering that the error of data A mainly comes from sensor zero drift and installation deviation, which appear as overall offset and proportional bias, a univariate linear regression model is used to describe the mapping relationship between data A and data B. Five-fold cross-validation is then adopted to test model performance. MSE, RMSE, MAE, and R2 are calculated to reflect calibration error and fitting stability.

After obtaining the result, the univariate linear regression algorithm is trained on all valid samples, and the calibration relationship from data A to data B is obtained. The corresponding corrected results are then calculated.

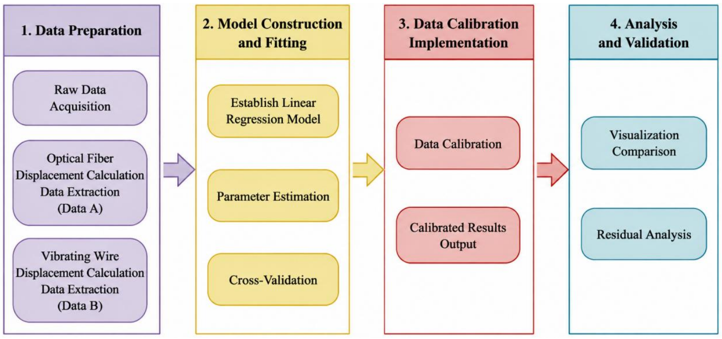

Figure 1 Analysis Flow of Displacement Data Calibration

Figure 1 describes the calibration workflow from data extraction and preprocessing to linear model construction, parameter fitting, correction-result output, and performance analysis. The flow clarifies how the sensor-bias correction is implemented as a regression-based data processing procedure.

2.1.2 Data preprocessing

During data preprocessing, data A and data B are first extracted from the corresponding dataset through Python pandas tools. The column names and data dimensions are checked to ensure correct data reading. Missing values are then processed by removing samples for which either A or B is empty, and complete valid data pairs are retained. Through these steps, the original displacement data are cleaned and organized into valid samples for modeling.

2.1.3 Model establishment

In the model construction stage, a calibration relationship model is built according to the systematic bias between fiber-optic displacement-meter data A and vibrating-wire displacement-meter data B. The objective function describes the error between the corrected data and the reference data, thereby forming a minimum-error calibration model based on univariate linear regression.

The objective function considers the overall deviation between corrected values and reference values and follows the principle of minimizing residual sum of squares:

$$f(a,b)=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{1}$$

where xi denotes the data A collected by the fiber-optic displacement meter at the ith moment, yi denotes the reference data B collected by the vibrating-wire displacement meter at the same moment, y-hat-i denotes the corrected result of data A, a and b denote the proportional calibration coefficient and offset calibration coefficient, and n denotes the number of valid samples. A smaller function value indicates a better calibration effect.

Since the error of data A mainly comes from sensor zero drift and installation deviation, the deviation between data A and reference data B can be approximated as the superposition of proportional deviation and fixed offset. Therefore, the calibration relationship is established as

$$\hat{y}_i=ax_i+b \tag{2}$$

where a corrects the proportional difference between the fiber-optic displacement meter and the reference device, and b corrects the sensor zero drift and installation deviation. Through this linear relationship, original data A can be mapped into corrected values close to reference data B.

After establishing the calibration relationship, the regression expression is substituted into the objective function, and least squares is used to solve the parameters a and b that minimize the error. The LinearRegression algorithm completes parameter fitting by taking data A as the input variable and data B as the output variable. Five-fold cross-validation is adopted to evaluate model stability and generalization. The valid samples are divided into five subsets, four of which are used for training and one for validation in each loop. Finally, five given pre-correction values are substituted into the trained model to obtain corresponding corrected results.

2.1.4 Model solution

The univariate linear regression method is used to correct the new fiber-optic displacement-meter data A. Sensor zero drift, installation deviation, and systematic offset between data A and reference data B are considered together. A linear mapping relation from data A to data B is established, and five-fold cross-validation evaluates the model.

The five-fold cross-validation result shows that RMSE and MAE are small, indicating low calibration error. R2 is close to 1, indicating that the model can describe the variation pattern of data B well. The small standard deviations of the indicators show good stability and generalization. In the full-sample fitting result, MSE is 7.195, RMSE is 2.682, MAE is 1.308, and R2 is 0.999. This is close to the cross-validation result, indicating that no obvious overfitting occurs. The five validation values are then predicted, and the pre-correction and corrected data are shown in Table 1.

Table 1 Data Calibration Results

Pre-correction data x	Corrected data y
7.132	5.911
18.526	15.946
84.337	73.907
123.554	108.446
167.667	147.296

Table 1 shows that as the pre-correction data x increases, the corrected data y also increases, indicating that the model maintains the changing trend of the original displacement data. Since the model slope is less than 1 and the intercept is negative, all corrected values are smaller than the corresponding pre-correction values. This agrees with the model-parameter pattern, showing that data A is generally larger than reference data B and needs proportional reduction and offset correction.

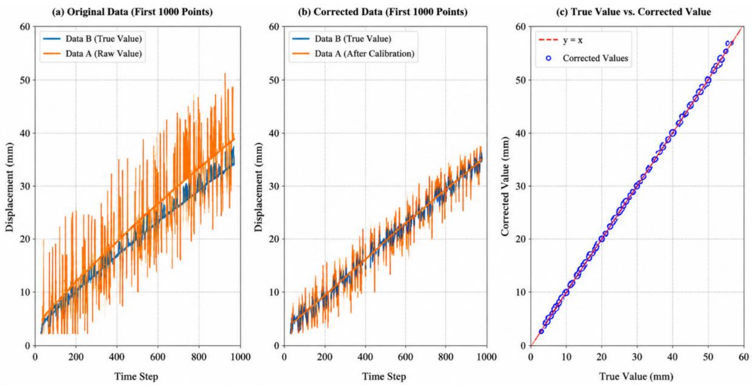

Figure 2 Comparison of Original, Corrected, and Reference Data

Figure 2 compares data A before and after calibration with reference data B. The corrected sequence overlaps the reference sequence more closely, and the scatter plot is concentrated near the y = x line, indicating that systematic deviation is reduced.

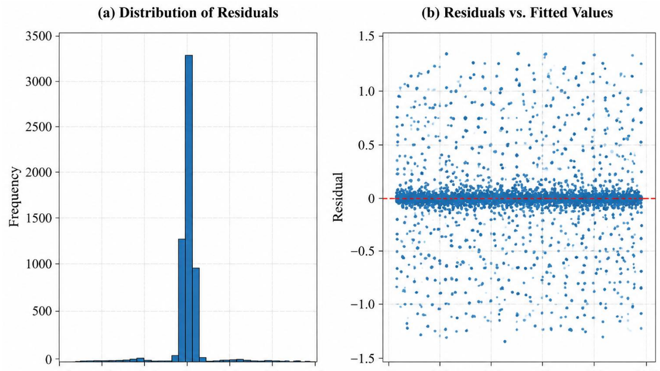

Figure 3 Residual Analysis of Calibration Results

Figure 3 presents the residual distribution and residual time sequence after calibration. The residuals are mainly concentrated around zero and show no obvious time-series drift, which supports the stability and generalization of the linear correction model.

2.2 Staged Deformation Identification Model

2.2.1 Specific analysis

This section designs a three-stage deformation recognition model for stage transition-node identification and segmented modeling. The model performs transition-node recognition, stage division, and segmented modeling for the displacement time series through filtering and denoising, anomaly identification, statistical testing, and segmented fitting [7].

The index and surface-displacement data are extracted, and the index is converted into actual acquisition time according to the sampling start point and sampling frequency. A time-displacement sequence is constructed. Savitzky-Golay filtering is used to smooth the original displacement data, and velocity and acceleration sequences are calculated. Hampel filtering identifies instantaneous abnormal jumps in the velocity sequence. Short-duration abnormal points that recover quickly are treated as noise or engineering disturbance rather than stage transition nodes. A sliding Welch t-test then compares the velocity distributions before and after candidate points to preliminarily screen stage-transition candidates. Finally, three-segment fitting residual minimization and an increasing-stage-velocity constraint determine the final transition nodes, dividing deformation into slow uniform deformation, accelerated deformation, and rapid deformation.

The three stages are modeled by a linear model, a quadratic polynomial model, and an exponential model. The model effects are evaluated, and the average surface displacement velocity of each stage is calculated.

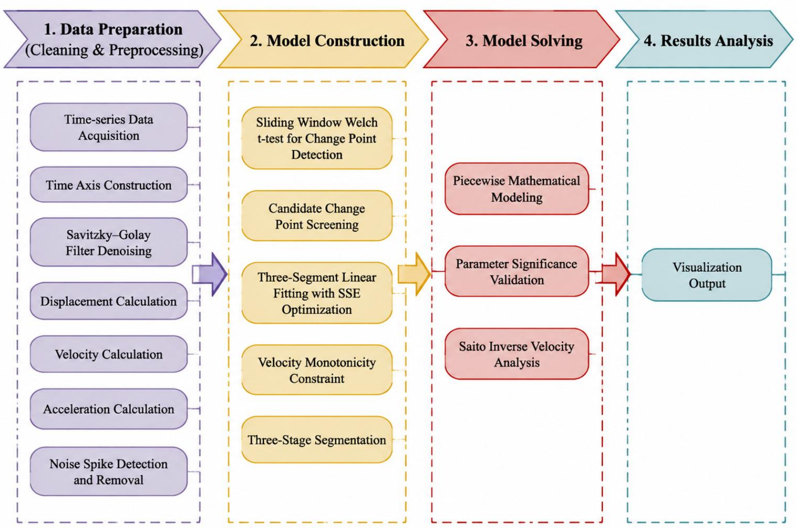

Figure 4 Analysis Flow of Staged Deformation Identification

Figure 4 summarizes the full staged-deformation identification process, including displacement smoothing, anomaly jump screening, Welch t-test candidate selection, segmented fitting, parameter evaluation, and visualization output.

2.2.2 Model establishment

The stage-identification model is constructed according to displacement trend, velocity difference, abnormal jump characteristics, and staged evolution. The objective function describes the segmented fitting error under different transition-node combinations and forms a stage-division method combining statistical testing, physical constraints, and minimum residual criteria.

Let the dataset contain N displacement observations, and let the smoothed displacement sequence be

$$\tilde{s}_1,\tilde{s}_2,\ldots,\tilde{s}_N \tag{3}$$

Assume that the two stage-transition nodes are b1 and b2. Here b1 denotes the transition node from slow uniform deformation to accelerated deformation, and b2 denotes the transition node from accelerated deformation to rapid deformation. The whole sequence can be divided into three intervals:

$$\begin{cases} I_1=[1,b_1] \\ I_2=[b_1,b_2] \\ I_3=[b_2,N] \end{cases} \tag{4}$$

Each segment is fitted separately, and the residual sum of squares is calculated. The objective function is defined as

$$F(b_1,b_2)=SSE_1+SSE_2+SSE_3 \tag{5}$$

For the jth stage, the residual sum of squares is

$$\mathrm{SSE}_j=\sum_{i\in I_j}(\tilde{s}_i-\hat{s}_i)^2 \tag{6}$$

where F(b1,b2) is the total segmented fitting error under node combination (b1,b2), SSEj is the residual sum of squares of the jth stage, Ij denotes the index set of the jth stage, s-tilde-i denotes the smoothed displacement value, and s-hat-i denotes the fitted displacement value.

The total residual sum of squares is minimized as

$$\left(b_1^*,b_2^*\right)=\underset{b_1,b_2}{\arg\min}F(b_1,b_2) \tag{7}$$

This yields the optimal transition-node combination. To avoid high computation and prevent local noise from being misidentified as transition nodes, a sliding Welch t-test is used to screen candidate nodes. For a candidate position k, velocity windows with length w are taken before and after it:

$$V_1=\{v_{k-w},v_{k-w+1},\ldots,v_{k-1}\} \tag{8}$$

$$V_2=\{v_k,v_{k+1},\ldots,v_{k+w-1}\} \tag{9}$$

Then Welch t-test is conducted on V1 and V2. If the mean velocities before and after the candidate position differ significantly, the position may be a stage transition node. The screening conditions are

$$\begin{cases}p<0.001\\ |t|>5\end{cases} \tag{10}$$

where p is the significance probability of Welch t-test, and t is the test statistic.

To avoid misjudging instantaneous jumps caused by engineering disturbance or sensor anomaly as real transition points, the Hampel method is introduced to detect abnormal points in the original velocity sequence. For a velocity point vi, a local window is taken near it, and the median velocity and median absolute deviation are calculated. If

$$|v_i-\mathrm{median}(v)|>\lambda\cdot 1.4826\cdot \mathrm{MAD} \tag{11}$$

the point is judged as a local abnormal jump point. Here lambda is the anomaly threshold, median(v) is the local-window velocity median, and MAD is the median absolute deviation in the local window.

To ensure that recognition conforms to the physical law of staged slope deformation, a stage-velocity increasing constraint is added. Let the average velocities of the three stages be v-bar1, v-bar2, and v-bar3, which should satisfy

$$\overline{v_1}<\overline{v_2}<\overline{v_3} \tag{12}$$

This constraint indicates that the displacement sequence should evolve from slow uniform deformation to accelerated deformation and then to rapid deformation. A real transition node must simultaneously satisfy statistical significance, a higher new-stage velocity level, certain duration, and physically reasonable monotonic evolution [8].

After determining the optimal transition nodes, the displacement time series is divided into three stages and mathematical models are built separately. For the first stage, displacement changes slowly and velocity is approximately stable, so a linear model is used:

$$S_1(t)=v_1t+s_0 \tag{13}$$

For the second stage, velocity gradually increases and the displacement curve shows accelerated growth, so a quadratic polynomial model is used:

$$S_2(t)=a_2t^2+b_2t+c_2 \tag{14}$$

For the third stage, displacement growth velocity increases significantly, and deformation enters a rapid development stage, so an exponential model is used:

$$S_3(t)=A(e^{\alpha t}-1)+B \tag{15}$$

After model parameters are solved, R2, RMSE, and F-test information are used to evaluate model fitting effect.

$$R^2=1-\frac{\sum_{i=1}^{n}\left(S_i-\widehat{S}_i\right)^2}{\sum_{i=1}^{n}(S_i-\overline{S})^2} \tag{16}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(S_i-\widehat{S}_i\right)^2} \tag{17}$$

Finally, the average surface displacement velocity of each stage is calculated. Let the initial and final displacements of the jth stage be sj,start and sj,end, and the initial and final times be tj,start and tj,end. The average velocity is

$$\overline{v_j}=\frac{s_{j,\mathrm{end}}-s_{j,\mathrm{start}}}{t_{j,\mathrm{end}}-t_{j,\mathrm{start}}} \tag{18}$$

2.2.3 Model solution

Savitzky-Golay smoothing, Hampel anomaly detection, sliding Welch t-test, and segmented fitting are applied to the surface displacement sequence. Based on displacement trend, significant velocity change, abnormal disturbance, and fitting accuracy, two stage-transition nodes are finally determined, and the displacement sequence is divided into slow uniform deformation, accelerated deformation, and rapid deformation.

The recognition criteria include duration, amplitude, statistical testing, monotonicity, and physical rationality. A velocity deviation must last no less than 3 h, corresponding to no fewer than 18 sampling points. The new-stage average velocity should be more than 1.5 times the old-stage average velocity. The Welch t-test should satisfy $p < 0.001$. The stage velocities should satisfy the monotonic relation v-bar1 < v-bar2 < v-bar3. The stage direction should conform to the irreversible evolution from slow uniform deformation to accelerated deformation and rapid deformation [9].

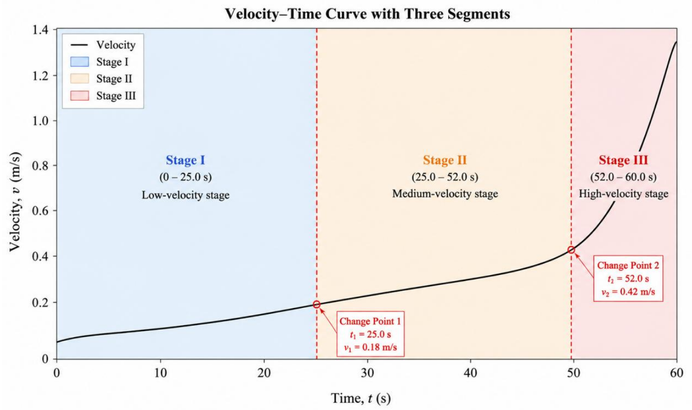

Figure 5 Surface Displacement Sequence and Three-Stage Division

Figure 5 shows the complete surface displacement sequence and the three-stage division result. The blue region denotes slow uniform deformation, the orange region denotes accelerated deformation, and the red region denotes rapid deformation. The two red dashed lines correspond to indices 5261 and 8861, which visually verifies the rationality of the stage division.

Table 2 Model Test Indicators and Average Surface Displacement Velocity by Stage

Stage	Index range	Average velocity/(mm/h)	Model	R2	RMSE/mm
Slow uniform	1-5261	0.1753	Linear model	0.947975	10.3724
Accelerated deformation	5261-8861	0.6115	Quadratic polynomial	0.983713	11.6345
Rapid deformation	8861-10000	4.1211	Cubic polynomial	0.998814	7.6402

Table 2 shows that the displacement sequence is divided into three stages, and the two transition nodes correspond to indices 5261 and 8861. The average velocities of the three stages are 0.1753, 0.6115, and 4.1211 mm/h, showing an obvious increasing trend. The first stage adopts a linear model with R2 = 0.947975, the second stage adopts a quadratic polynomial model with R2 = 0.983713, and the third stage adopts a cubic polynomial model with R2 = 0.998814 and the smallest RMSE. These results indicate that segmented modeling can describe the staged deformation process effectively [10].

3 CONCLUSION

This study builds an integrated algorithmic framework for slope deformation monitoring data calibration and staged deformation identification. The calibration model uses univariate linear regression and least-squares parameter estimation to correct systematic bias in fiber-optic displacement data. Five-fold cross-validation and full-sample fitting show strong performance, with MSE = 7.195, RMSE = 2.682, MAE = 1.308, and R2 = 0.999. The corrected values retain the variation trend of the original data while reducing the systematic deviation relative to the vibrating-wire reference sequence.

The staged deformation model combines Savitzky-Golay smoothing, Hampel anomaly detection, sliding Welch t-test, duration and amplitude criteria, monotonic velocity constraints, and segmented fitting. The final transition nodes are identified at indices 5261 and 8861, dividing the surface displacement sequence into slow uniform deformation, accelerated deformation, and rapid deformation. The three average velocities are 0.1753, 0.6115, and 4.1211 mm/h, and the fitted R2 values reach 0.947975, 0.983713, and 0.998814. The limitation is that the model is verified on the given monitoring sequences, and broader multi-sensor validation would improve robustness. Future work can connect the calibrated sequence and staged deformation indicators with real-time warning thresholds and adaptive online updating.

Future research should be developed in three directions. First, the calibration model can be extended from a single linear mapping to a multi-sensor fusion model that incorporates temperature, rainfall, groundwater level, excavation disturbance, and other environmental variables, so that nonstationary field conditions can be described more fully. Second, the staged-identification module can be connected with rolling-window analysis and online-learning mechanisms, allowing transition thresholds and stage-boundary decisions to update automatically as new monitoring data accumulate. Third, field validation should be expanded to slopes with different geological structures and deformation modes, and the calibrated displacement indicators can be linked with graded early-warning thresholds to support real-time risk assessment and maintenance decisions.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Waskom M L. Seaborn: statistical data visualization. Journal of Open Source Software, 2021, 6(60): 3021.

[2] Montiel J, Halford M, Mastelini S M, et al. River: machine learning for streaming data in Python. Journal of Machine Learning Research, 2021, 22(110): 1-8.

[3] Herzen J, Lässig F, Piazzetta S G, et al. Darts: user-friendly modern machine learning for time series. Journal of Machine Learning Research, 2022, 23(124): 1-6.

[4] Middlehurst M, Ismail-Fawaz A, Guillaume A, et al. aeon: a Python toolkit for learning from time series. Journal of Machine Learning Research, 2024, 25(289): 1-10.

[5] Makridakis S, Spiliotis E, Assimakopoulos V. M5 accuracy competition: results, findings, and conclusions. International Journal of Forecasting, 2022, 38(4): 1346-1364.

[6] Makridakis S, Spiliotis E, Assimakopoulos V. The M5 uncertainty competition: results, findings and conclusions. International Journal of Forecasting, 2022, 38(4): 1365-1385.

[7] Torres J F, Hadjout D, Sebaa A, et al. Deep learning for time series forecasting: a survey. Big Data, 2021, 9(1): 3-21.

[8] Lim B, Zohren S. Time-series forecasting with deep learning: a survey. Philosophical Transactions of the Royal Society A, 2021, 379(2194): 20200209.

[9] Januschowski T, Gasthaus J, Wang Y, et al. Forecasting with trees. International Journal of Forecasting, 2022, 38(4): 1473-1481.

[10] Hewamalage H, Bergmeir C, Bandara K. Recurrent neural networks for time series forecasting: current status and future directions. International Journal of Forecasting, 2021, 37(1): 388-427.