Study of Electricity Data Processing Based on KNN Interpolation and Wavelet Soft-Threshold Denoising Methods

Kexin Xie, Jun Wang, Chengzhi Liu, and Yan Wang

***Abstract*—This paper addresses the challenges of missing data and noise in three-phase power system signals. A KNN-based imputation method that combines horizontal and vertical interpolation accurately reconstructs missing values. Meanwhile, wavelet soft-thresholding effectively suppresses noise, resulting in a high SNR and low RMSE. Correlation analysis reveals strong linear relationships among the three-phase currents, confirming the consistency and reliability of the pre-processed data. These results provide a solid foundation for improving power system monitoring and signal analysis. Future work will focus on optimizing and extending these techniques.**

***Index Terms*—soft-threshold wavelet denoising, correlation analysis, KNN-based imputation method, horizontal and vertical interpolation.**

I. Introduction

POWER systems are fundamental to modern society, and their stable operation is essential for sustaining socio-economic activities and ensuring public welfare. The operation and maintenance of these systems depend on accurate and complete data, which is vital for stability assessment, fault diagnosis, and operational decision-making. However, in practical data acquisition, electricity data signals frequently suffer from missing data and noise contamination due to sensor malfunctions, communication failures, and environmental disturbances. These data quality issues severely compromise analytical accuracy and the reliability of predictive models.

To address these challenges, advanced data preprocessing is necessary, with particular attention to two key aspects: missing data imputation and noise reduction. Missing values, often caused by unpredictable factors, disrupt the continuity and temporal coherence of time-series data. Various imputation techniques have been proposed, from traditional statistical interpolation (e.g., mean, median, mode imputation [1]) to more sophisticated models such as multiple imputation [2], chained equations [3], and autoregressive moving average models for multivariate time series [4]. For time-series datasets exhibiting periodicity, researchers have explored mixture factor analysis and multiscale CNN-based frameworks to better capture temporal dependencies [5], [6]. Deep learning methods, including transformer-based models, have shown strong performance in reconstructing complex missing patterns [7]. Additionally, techniques like full information multiple imputation for linear regression models with missing response variables offer a statistically rigorous approach to handling uncertainty [8]. These methods increasingly emphasize optimization and adaptability to specific application scenarios [5], [9]. Traditional approaches, such as simple deletion or mean imputation, are inadequate, as they fail to preserve the inherent structure and temporal dynamics of power system data.

Similarly, noise in signals introduces fluctuations that obscure meaningful patterns and hinder subsequent analysis. Researchers have developed various denoising methods, including Fourier and wavelet transforms, adaptive filtering, and machine learning-based approaches. For example, Fourier clustering with adversarial mechanisms [10], fractional Fourier transforms [11], and adaptive Fourier decomposition based on energy distribution [12] have been shown to be effective in various domains. Wavelet transform techniques have proven highly flexible for denoising, with soft and hard thresholding methods, featuring improved threshold functions, widely applied in signal and image denoising to increase signal-to-noise ratio (SNR) and reduce root mean square error (RMSE) [13], [14]. Modified thresholding strategies have addressed limitations such as discontinuities and fixed deviations, yielding superior results in applications ranging from signal processing to underwater target detection [15], [16].

In machine learning-based denoising, random forests and their variants have shown robustness in noisy data classification tasks [17], [18]. Similarly, adaptive Kalman filtering has been applied to denoising, where iterative parameter updates enable adaptation to non-stationary noise [19]. Hybrid frameworks that integrate wavelet thresholding with adaptive mechanisms have improved power system stability under noisy conditions by identifying low-frequency oscillations [20], and wavelet packet decomposition combined with random forests has enhanced fault diagnosis in rotating machinery [21].

These developments highlight the importance of optimization and domain-specific adaptation in mitigating the impact of noise [22], [23]. In this paper, we adopt the wavelet soft-thresholding method, which enhances denoising performance

Manuscript received December 9, 2024; revised July 5, 2025.

This work was supported in part by the Natural Science Foundation of Hunan Province under Grant 2023JJ50080 and 2025JJ70357, and the Scientific Research Funds of Hunan Provincial Education Department under Grant 24A0637.

K. X. Xie is an undergraduate student of School of Mathematics and Finance, Hunan University of Humanities, Science and Technology, Loudi 417000, P. R. China (e-mail: 3469018119@qq.com).

J. Wang is an undergraduate student of School of Mathematics and Finance, Hunan University of Humanities, Science and Technology, Loudi 417000, P. R. China (e-mail: 1918838015@qq.com).

C. Z. Liu is an associate professor of School of Mathematics and Finance, Hunan University of Humanities, Science and Technology, Loudi 417000, P. R. China (corresponding author to provide e-mail: it-rocket@163.com).

Y. Wang is a master's student of School of Information, Hunan University of Humanities, Science and Technology, Loudi 417000, P. R. China (e-mail: 2256678418@qq.com).

through improved threshold functions, optimized wavelet basis selection, and adaptive thresholding strategies. This approach effectively suppresses noise while preserving critical signal features.

While traditional preprocessing methods such as simple mean interpolation and Fourier-based denoising offer basic handling of missing and noisy data, they often fail to address the nonlinear and non-stationary characteristics of real-world power system signals. To overcome these limitations, in this paper we propose a hybrid interpolation framework that integrates horizontal (within-day) and vertical (cross-day) interpolation to capture both short-term and periodic temporal patterns. A weighted K-nearest neighbors (KNN) algorithm dynamically adjusts the weights of neighboring points to better reflect the true data distribution. The final imputation values are optimized by combining horizontal and vertical estimates through weighted averaging, enhancing continuity and structural integrity. For denoising, the proposed method employs wavelet soft-thresholding with adaptive refinements to suppress interference while retaining key signal characteristics. This integrated approach avoids information loss common in deletion-based or crude imputation strategies, enhances signal clarity and accuracy, and provides a reliable foundation for predictive modeling and power system optimization. Its modular and generalizable design makes it broadly applicable to diverse time-series domains.

II. Missing Data Imputation

To ensure data continuity and modeling accuracy, addressing missing values is a critical step in power system data preprocessing. We adopt a multidimensional interpolation strategy that leverages both horizontal and vertical temporal structures. Specifically, we integrate the KNN algorithm [24] with a weighted averaging scheme to improve imputation precision while preserving the underlying dynamics of the data.

A. Correlation-driven Justification

To support the proposed interpolation strategy, we performed a correlation analysis on the three-phase current signals. The Pearson correlation coefficients among I_a, I_b, and I_c are summarized in Table I. All coefficients exceed 0.98, indicating strong positive linear relationships. This interdependence confirms the structural consistency of the signals and supports the use of cross-phase information in the imputation process.

TABLE I
Pearson correlation coefficients between three-phase currents.

	I_a	I_b	I_c
I_a	1.0000	0.9920	0.9825
I_b	0.9920	1.0000	0.9825
I_c	0.9825	0.9825	1.0000

These results also suggest that missing values in one phase may be reliably estimated using the data from the other two, thus providing a solid basis for future expansion into cross-phase imputation schemes.

B. Multidimensional KNN-Based Interpolation

Building on the aforementioned correlation analysis, we design a multidimensional KNN-based interpolation strategy that leverages both short-term and long-term temporal structures to reconstruct missing values in three-phase current signals. Specifically, this method combines horizontal interpolation and vertical interpolation to ensure accurate estimation.

The use of KNN enables flexible adaptation to local data characteristics, while distance-based weighting ensures that closer observations exert greater influence on the imputation result. The overall interpolation process consists of the following steps:

Step 1 Identification of missing values: Analyze the time series dataset to detect and localize missing values.

Step 2 Nearest neighbor selection: For each missing data point, compute the Manhattan distances to its $K = 5$ nearest neighbors. The Manhattan distance between t_i and t_j is given by

$$d_{ij} = |t_i - t_j|.$$

Step 3 Weighted mean estimation: Estimate the missing value $\hat{y}_i$ using the values of its K nearest neighbors $y_j (j = 1, 2, \cdots, K)$, weighted by their distances d_{ij}, i.e.,

$$\hat{y}_i = \frac{\sum_{i=1}^{K} (y_i / d_{ij})}{\sum_{i=1}^{K} (1/d_{ij})}.$$

This weighting scheme assigns higher influence to closer neighbors, ensuring that the imputed value is more representative of the surrounding data points.

Step 4 Multidimensional interpolation: Perform interpolation in two dimensions: horizontally (within the same day) to obtain $\hat{y}_i$, and vertically (at the same time across different days) to obtain $\bar{y}_i$. The final imputed value is computed as the average, so we have

$$y_{\text{final}} = \frac{\hat{y}_i + \bar{y}_i}{2}.$$

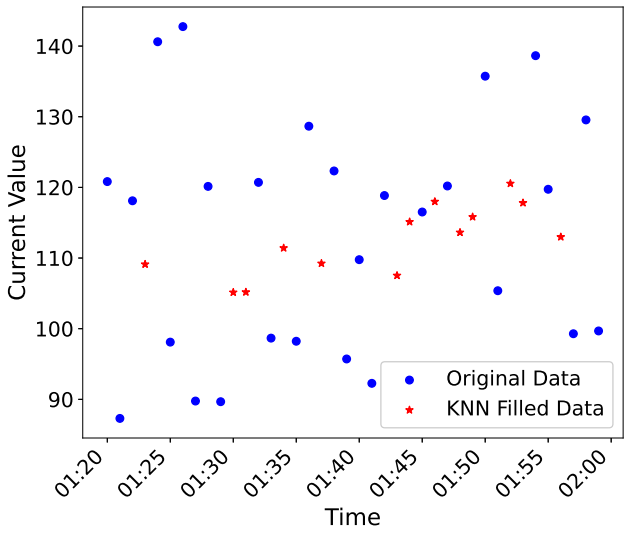


Fig. 1. Example of missing data imputation.

This multidimensional interpolation framework effectively maintains the temporal continuity of the data while leveraging both intra-day and inter-day correlations. As illustrated

in Fig. 1, the comparison between original and imputed data confirms the superior performance of the proposed method. By employing this strategy, the accuracy and robustness of missing data estimation are significantly enhanced, providing a reliable foundation for downstream model training and predictive analytics.

III. Noise Suppression via Wavelet Soft-Thresholding

To address noise interference and enhance signal fidelity, we adopt a soft-threshold wavelet denoising algorithm as our primary preprocessing tool. This method is well-suited for power signals, as it selectively attenuates noise while preserving critical waveform features essential for downstream analysis.

A. Soft-threshold wavelet denoising

The soft-threshold wavelet denoising method is particularly effective for preprocessing power data, as it suppresses small-amplitude noise while retaining significant signal components embedded in the wavelet coefficients. The denoising procedure involves the following steps:

Step 1 Wavelet decomposition: By selecting the wavelet basis and setting the decomposition level to 3, we apply the wavelet transform to obtain the wavelet coefficients w_{jk}, where j denotes the decomposition level and k denotes the coefficient index.

Step 2 Soft-threshold processing: A soft-threshold function is applied to the wavelet coefficients to suppress noise. The thresholded coefficients $\hat{w}_{jk}$ are computed as

$$\hat{w}_{j,k} = \begin{cases} \operatorname{sgn}(w_{jk})(|w_{jk}| - \lambda), & |w_{jk}| \geq \lambda, \\ 0, & |w_{j,k}| < \lambda, \end{cases}$$

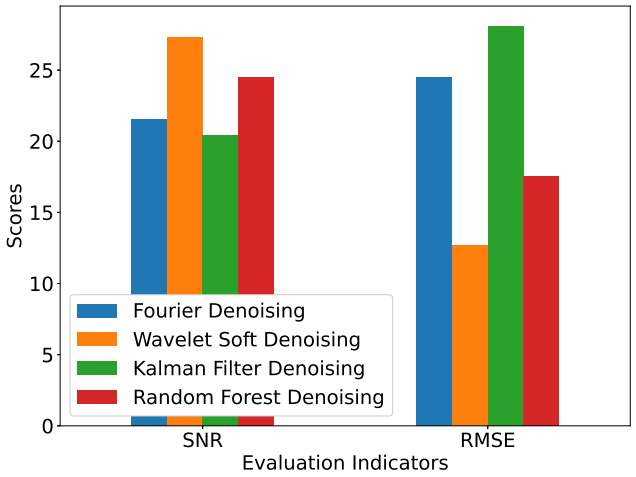
where $\operatorname{sgn}(w_{jk})$ is the sign function, and λ is a predefined threshold that determines the level of denoising.

Step 3 Wavelet reconstruction: The processed wavelet coefficients $\hat{w}_{jk}$ are subjected to an inverse wavelet transform to reconstruct the denoised power data. The SNR and the RMSE are used to evaluate the performance of the denoising algorithm.

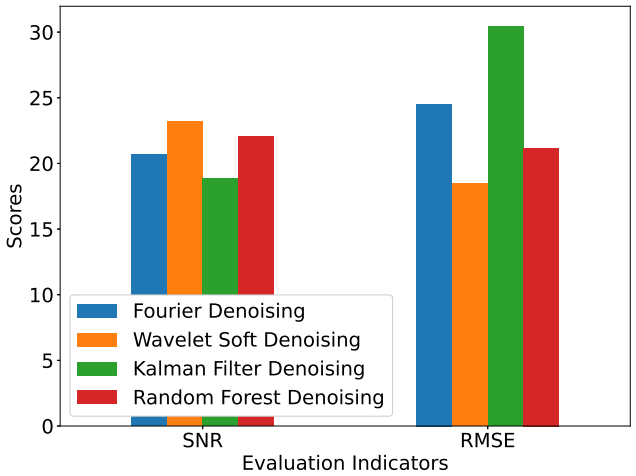
B. Analysis of denoising results

To evaluate the performance of soft-threshold wavelet denoising, we compared it with several other denoising techniques applied to the three-phase current data. These techniques included Fourier denoising [25], Kalman filter denoising [26], and random forest denoising [27]. In the comparative analysis, each method was evaluated based on its ability to increase SNR and reduce RMSE.

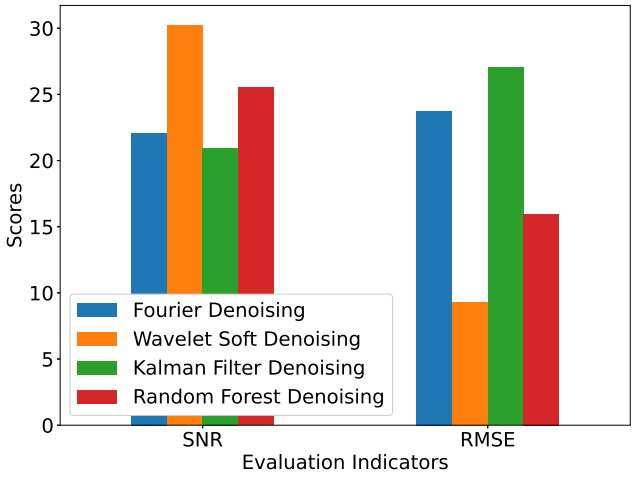
Fig. 2 shows the experimental results of the denoising methods. It can be seen that the wavelet soft-threshold denoising method outperforms the other three techniques (Fourier, Kalman filter, and random forest) for all three phase currents (I_a, I_b, and I_c), achieving the highest SNR and the lowest RMSE. This demonstrates that wavelet soft-thresholding effectively preserves the essential features of the signal while efficiently removing noise. These findings underscore the effectiveness of wavelet soft-thresholding as a robust denoising tool for power system signals, particularly in scenarios with pronounced high-frequency interference.



(a) I_a



(b) I_b



(c) I_c

Fig. 2. Comparison of denoising methods for I_a, I_b, and I_c.

In addition, Fig. 3 presents a visual comparison between the original and denoised waveforms of the three-phase current signals. The blue curves represent the raw signals, which exhibit significant fluctuations and high-frequency noise. In contrast, the orange curves correspond to the denoised signals, displaying smoother and more stable profiles.

This comparison clearly demonstrates the effectiveness of the proposed method in suppressing noise while preserving the essential structural characteristics of the signals.

IV. Conclusion

This paper presented an integrated framework for addressing two critical data quality issues in power system signals: missing data imputation and noise suppression. By employing a KNN-based interpolation strategy that integrates both horizontal (within-day) and vertical (cross-day) temporal correlations, the method effectively reconstructed missing values and preserved the temporal structure of the data. The integration of these two interpolation dimensions significantly enhanced imputation accuracy and continuity, providing a solid foundation for downstream signal analysis.

In terms of noise reduction, the wavelet soft-thresholding method demonstrated superior performance across various evaluation metrics, achieving higher SNR and lower RMSE compared to Fourier, Kalman filter, and random forest-based methods. This validates its effectiveness in retaining key signal features while suppressing high-frequency noise.

Overall, this study verified the effectiveness of combining data-driven interpolation and wavelet-based denoising for enhancing the quality of power system time-series data. Future work may explore hybrid models that integrate multi-phase information for imputation. Pursuing these directions is expected to significantly advance high-fidelity signal reconstruction, real-time monitoring accuracy, and intelligent control strategies in modern power systems.

References

[1] Z. Zhang, "Missing data imputation: focusing on single imputation," *Annals of Translational Medicine*, vol. 4, no. 1, pp. 9, 2016.

[2] S. Sinharay, H. S. Stern and D. Russell, "The use of multiple imputation for the analysis of missing data," *Psychological Methods*, vol. 6, no. 4, pp. 317–29, 2001.

[3] S. I. Khan and A. S. M. L. Hoque, "SICE: an improved missing data imputation technique," *Journal of Big Data*, vol. 7, no. 1, pp. 37, 2020.

[4] I. Sumertajaya, E. Rohaeti, A. Wigena, et al., "Vector autoregressive-moving average imputation algorithm for handling missing data in multivariate time series," *IAENG International Journal of Computer Science*, vol. 50, no. 2, pp. 727–735, 2024.

[5] D. Jeong, C. Park and Y. M. Ko, "Missing data imputation using mixture factor analysis for building electric load data," *Applied Energy*, vol. 304, pp. 117655, 2021.

[6] L. Zhang, Q. Jiang, M. Gao, et al., "Time Series Missing Imputation Framework to Combine Multi-scale CNN and Weighted KNN Based on Information Entropy," *Proceedings of the 2024 Guangdong-Hong Kong-Macao Greater Bay Area International Conference on Digital Economy and Artificial Intelligence*, pp. 571–579, 2024.

[7] A. Lotfipoor, S. Patidar and D. P. Jenkins, "Transformer network for data imputation in electricity demand data," *Energy and Buildings*, vol. 300, pp. 113675, 2023.

[8] L. Song and G. Guo, "Full Information Multiple Imputation for Linear Regression Model with Missing Response Variable," *IAENG International Journal of Applied Mathematics*, vol. 54, no. 1, pp. 77–81, 2024.

[9] M. C. Wang, C. F. Tsai and W. C. Lin, "Towards missing electric power data imputation for energy management systems," *Expert Systems with Applications*, vol. 174, pp. 114743, 2021.

[10] Y. Li, I. N. R. Ugli, Y. I. H. Ugli, et al., "Optimizing Models and Data Denoising Algorithms for Power Load Forecasting," *Energies*, vol. 17, no. 21, pp. 5513, 2024.

[11] M. Y. Zhai, "Seismic data denoising based on the fractional Fourier transformation," *Journal of Applied Geophysics*, vol. 109, pp. 62–70, 2014.

[12] Z. Wang, F. Wan, C. M. Wong, et al., "Adaptive Fourier decomposition based ECG denoising," *Computers in Biology and Medicine*, vol. 77, pp. 195–205, 2016.

[13] N. Zhang, P. Lin and L. Xu, "Application of weak signal denoising based on improved wavelet threshold," *IOP Conference Series: Materials Science and Engineering*, vol. 751, no. 1, pp. 012073, 2020.

[14] J. Lu, L. Hong, Y. Dong, et al., "A new wavelet threshold function and denoising application," *Mathematical Problems in Engineering*, vol. 2016, no. 1, pp. 3195492, 2016.

[15] H. Cui, R. Zhao and Y. Hou, "Improved threshold denoising method based on wavelet transform," *Physics Procedia*, vol. 33, pp. 1354–1359, 2012.

[16] Y. Zhang, W. Ding, Z. Pan, et al., "Improved wavelet threshold for image de-noising," *Frontiers in Neuroscience*, vol. 13, pp. 39, 2019.

[17] I. Reis, D. Baron and S. Shahaf, "Probabilistic random forest: A machine learning algorithm for noisy data sets," *The Astronomical Journal*, vol. 157, no. 1, pp. 16, 2018.

[18] N. H. Agjee, O. Mutanga, K. Peerbhay, et al., "The impact of simulated spectral noise on random forest and oblique random forest classification performance," *Journal of Spectroscopy*, vol. 2018, no. 1, pp. 8316918, 2018.

[19] H. D. Hesar and M. Mohebbi, "An adaptive Kalman filter bank for ECG denoising," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 1, pp. 13–21, 2020.

[20] J. Chen, X. Li, M. A. Mohamed, et al., "An adaptive matrix pencil algorithm based-wavelet soft-threshold denoising for analysis of low frequency oscillation in power systems," *IEEE Access*, vol. 8, pp. 7244–7255, 2020.

[21] Z. Wang, Q. Zhang, J. Xiong, et al., "Fault diagnosis of a rolling bearing using wavelet packet denoising and random forests," *IEEE Sensors Journal*, vol. 17, no. 17, pp. 5581–5588, 2017.

[22] S. Postalcioglu, K. Erkan and E. D. Bolat, "Comparison of Kalman filter and wavelet filter for denoising," *International Conference on Neural Networks and Brain*, vol. 2, pp. 951–954, 2005.

[23] B. R. Manju and M. R. Sneha, "ECG denoising using wiener filter and kalman filter," *Procedia Computer Science*, vol. 171, pp. 273–281, 2020.

[24] S. Daberdaku, E. Tavazzi and B. Di Camillo, "A combined interpolation and weighted K-nearest neighbours approach for the imputation of longitudinal ICU laboratory data," *Journal of Healthcare Informatics Research*, vol. 4, no. 2, pp. 174–188, 2020.

[25] X. Zhang and S. Jiang, "Application of fourier transform and butterworth filter in signal denoising," *2021 6th International Conference on Intelligent Computing and Signal Processing*, pp. 1277–1281, 2021.

[26] S. Park, M. S. Gil, H. Im, et al., "Measurement noise recommendation for efficient Kalman filtering over a large amount of sensor data," *Sensors*, vol. 19, no. 5, pp. 1168, 2019.

[27] M. Hibino, A. Kimura, T. Yamashita, et al., "Denoising random forests," *arXiv preprint* arXiv:1710.11004, 2017.

[28] J. Adler and I. Parmryd, "Quantifying colocalization by correlation: the Pearson correlation coefficient is superior to the Mander's overlap coefficient," *Cytometry Part A*, vol. 77, no. 8, pp. 733–742, 2010.

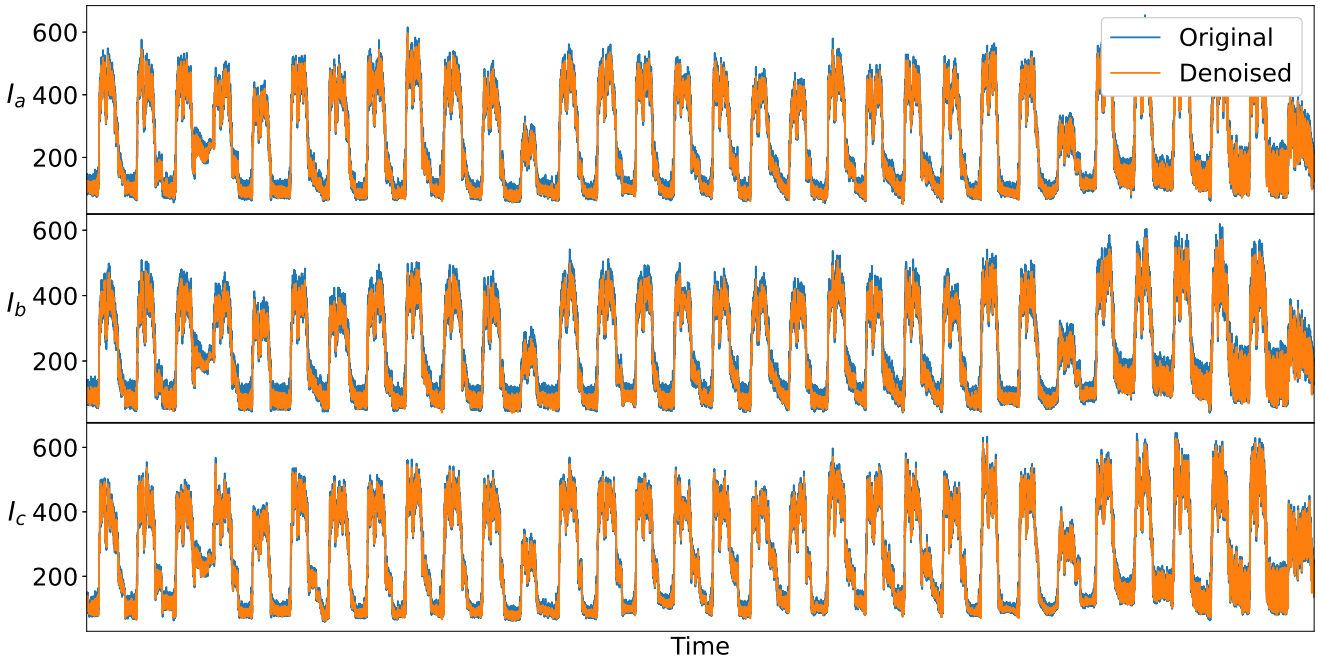


Fig. 3. Comparison between original (blue) and denoised (orange) three-phase current signals using wavelet soft-thresholding.