IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 December 2024, accepted 31 December 2024, date of publication 6 January 2025, date of current version 22 January 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3525726

RESEARCH ARTICLE

Anomaly Usage Behavior Detection Based on Multi-Source Water and Electricity Consumption Information

WENQING ZHOU[1], CHAOQIANG CHEN[2], QIN YAN[1], (Member, IEEE), BIN LI[1], (Student Member, IEEE), KANG LIU[3], YINGJUN ZHENG[1,4], HONGMING YANG[1], (Member, IEEE), HUI XIAO[1], AND SHENG SU[1], (Senior Member, IEEE)

[1]School of Electrical and Information Engineering, Changsha University of Science & Technology, Changsha 410114, China
[2]State Grid of China, Changsha Electric Power Company Ltd., Changsha 410014, China
[3]College of Electrical and Information Engineering, Hunan University, Changsha 410082, China
[4]State Grid of China, Jinhua Electric Power Company Ltd., Jinhua 321200, China

Corresponding authors: Sheng Su (eessheng@163.com) and Qin Yan (qin.yan@csust.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 51777015.

ABSTRACT The construction of smart cities contributes to promoting residents' life convenience and sustainable energy development. Despite these advancements, the challenge of fully analyzing and understanding residents' energy usage behaviors leads to inefficient energy use and potential economic losses. Current resident anomaly detection technologies rely on single-source energy data, lacking detailed behavior pattern analysis. Hence, this paper proposes a method to detect abnormal residential water and electricity usage by incorporating multi-source information. Specifically, the correlation between water and electricity usage of residential customers is analyzed based on real metering data and the use of the Copula distribution function, followed by the integration of two innovative data mining techniques to form an anomaly detection framework. The distance correlation coefficient algorithm is used to measure the relevance of users' water and electricity usage data. Then, the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm is utilized to cluster the distance correlation coefficient for users and detect abnormal users whose distance correlation coefficient curves deviate from the normal user clusters. This multi-source approach avoids single-source bias by improving the data accuracy over one-dimensional methods. Experiments are implemented in a real low-voltage transformer area to prove the validity of the proposed method.

INDEX TERMS Advanced metering infrastructure, behavior analysis, distance correlation coefficient, multi-source information, smart cities.

I. INTRODUCTION

Smart cities improve the construction of urban infrastructure by applying information and communication technology to create an efficient, convenient, and eco-friendly management system. It aims to fulfill residents' economic, cultural, and social needs, leading to sustainable development in social, economic, and ecological aspects [1], [2], [3]. The smart grid acts as the foundation and support for smart city development and provides energy security for smart cities. As the new trend of power grid development in today's world, smart grid not only has superior resource allocation capabilities, but also can realize the interaction of multi-source information and multi-business systems, which holds considerable importance for the establishment and growth of smart cities [4].

In the context of smart cities, abnormal detection of residential water and electricity consumption has become an important task. Public utility companies suffer from huge economic losses and safety hazards due to energy theft, metering errors, abnormal energy use, etc. For example,

The associate editor coordinating the review of this manuscript and approving it for publication was Hao Wang.

according to the statistical data from the Northeast Group, the losses for energy companies covering 125 countries around the world have reached 96 billion dollars [5]. Furthermore, research in the field indicates a projected rise in worldwide consumption of water and energy by 55% and 80% respectively by the mid-21st century [6]. Therefore, abnormal detection of water and electricity is needed to help ensure the efficient use of resources and prevent resource waste or over-consumption. In this context, Advanced Metering Infrastructure (AMI), as a pivotal part of smart cities, gathers, relays, and displays data on the consumption of electricity, water, gas, and heat. This is achieved through the widespread deployment of smart meters, along with advanced wireless communication and data management technologies. AMI not only revolutionizes the mode of energy services but also offers extensive data support for a comprehensive analysis of consumer behavior. The implementation of AMI enhances resource management efficiency and lays the groundwork for promoting a more sustainable urban lifestyle.

Since water and electricity are the two most frequently consumed resources in residents' daily lives, abnormal detection of their data is crucial to ensuring the sustainable operation of the city and the quality of residents' lives. Most existing research therefore focuses on the detection of abnormal data in water or electricity using artificial intelligence techniques [7]. The methods can be mainly divided into machine learning and deep learning. To address the challenge of high data labeling costs, Kou et al. developed an anomaly detection method based on a contrastive learning network, which effectively utilizes unlabeled data to detect abnormal power consumption [8]. Wang et al. proposed an improved Canopy-K Means unsupervised algorithm to tackle the difficulty of classifying similar power consumption patterns among users. By combining the Canopy-K Means algorithm with the isolation forest algorithm, they jointly constructed an abnormal power consumption detection model that leverages multi-layer fused feature data analysis [9]. In response to issues related to dimensionality and low data resolution, Ghamkhar et al. introduced a method for detecting abnormal water meter data by integrating DBSCAN with Lempel-Ziv complexity features [10]. Moghaddass and Wang design a model based on smart meter data to detect user-level abnormal events [11]. This model categorizes abnormal events into different levels, thereby better-assisting utility companies in planning and maintenance. Taking the fact that users may adjust the voltage coils to realize electricity theft into consideration, Leite and Mantovani use the voltage information recorded by smart meters to identify abnormal electricity usage based on voltage anomalies [12]. Zhu et al. introduce a method for network detection that leverages a hybrid-order approach to representation learning [5], while Yang et al. establish an innovative method for detecting anomalous electricity usage patterns, utilizing normalized covariant measures to assess the linkage between non-technical loss and electricity consumption [13]. In addition, Shin et al. combined the XG Boost and Light GBM algorithms based on machine learning to develop a model for predicting indoor water leakage in urban areas, which provides a decision-making basis for water leakage problems [14].

It is worth noting that most current methods only rely on a single electricity or water consumption information, which limits the comprehensiveness and reliability of anomaly detection. For example, residential users exhibit significant differences in their electricity usage behavior due to factors such as family size, occupation, and lifestyle habits. In addition, the electricity usage patterns of normal users can resemble those of certain abnormal users [12]. Therefore, the information contained in the electricity usage data is limited, and it may be difficult to accurately distinguish abnormal electricity consumption behavior from normal electricity consumption fluctuations based on meter data alone [15]. Similarly, a single water consumption data may also be insufficient to differentiate between seasonal changes in normal water use and water leakage events. In addition, these methods usually ignore the possible correlation between water and electricity consumption behaviors, thus limiting a deeper understanding and research on the energy consumption behavior of residential users. Therefore, user behavior analysis and anomaly detection research that integrates water and electricity data can help to more comprehensively understand and reveal residents' energy consumption patterns, provide a more reliable basis for resource management in smart cities, and help achieve a more efficient and sustainable operation mode for cities.

Thus, this paper proposes a method for anomaly detection using water and electricity metering data collected by AMI. Firstly, a qualitative analysis of users' usage behavior is conducted using daily and hourly scale water and electricity metering data. Then, the bivariate joint distribution of daily water and electricity usage data is quantitatively analyzed using the Copula distribution function to further determine the correlation of daily water and electricity metering data. Based on the correlation analysis, a method for detecting abnormal water and electricity usage behavior of residential users is proposed. The method mainly combines two data mining methods, the distance relationship coefficient algorithm and DBSCAN. The distance correlation coefficient curve is plotted by calculating the distance correlation coefficient between electricity usage and water usage in the transformer service area (TSA), while DBSCAN is used to cluster the user distance correlation coefficients and detect abnormal behavior users. Finally, a practical example within the low-voltage TSA domain substantiates the applicability of the suggested approach. The proposed method combines information from multiple sources to avoid the biases of relying on a single data source, thereby increasing accuracy and better detecting abnormal behaviors of users. The method not only reduces energy waste, but also provides a more precise decision-making basis for urban resource allocation and services to support more efficient and eco-friendly management of resources within smart cities.

The rest of this paper is organized as follows. In Section II, the correlation analysis of water and electricity usage behaviors is presented. Section III presents the detection framework and methodology for abnormal behavior users. Section IV presents the results of numerical simulations. In Section V, the conclusions and future works are pointed out.

II. CORRELATION ANALYSIS OF WATER AND ELECTRICITY USAGE

Compared to relying solely on a single data source, integrating multi-source measurement data including water and electricity allows for a more comprehensive capture of usage details and their underlying correlations, thereby depicting user behavior patterns more accurately. To conduct a more detailed analysis of the correlation, quantitative and qualitative analyses are performed using real TSA data collected by smart meters.

A. QUALITATIVE ANALYSIS OF WATER AND ELECTRICITY CORRELATION

The data for this study is gathered by State Grid Corporation of China (SGCC) smart meters and originates from the integrated demonstration area of multiple meters in China [16]. Accumulated minute-level water and electricity usage data of a residential user in the demonstration area from January 1, 2021, to March 31, 2021, are used to obtain hourly and daily water and electricity usage data, as shown in FIGURE 1.

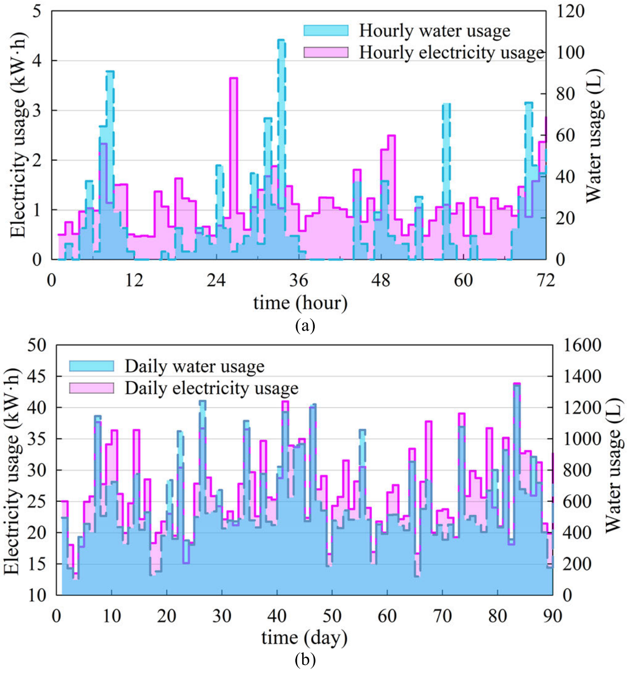


FIGURE 1. Users' water and electricity usage at different time scales. (a) Hourly usage. (b) Daily usage.

In FIGURE 1(a), the hourly electricity usage primarily lies in the range of 0.5 kW·h to 3 kW·h, while the hourly water usage primarily lies in the range of 0 L to 60 L. It is evident that the water usage curve exhibits a clear period of zero usage during nighttime hours, contrasting with the electricity usage curve which lacks an equivalent zero-usage period, attributable to factors such as the standby operation of electrical appliances. Furthermore, the patterns of fluctuation in electricity and water usage exhibit notable discrepancies, with the amplitude of variation significantly diverging. Residential electricity usage is relatively stable, whereas water consumption can increase dramatically due to residents' water usage behaviors being concentrated in specific periods. Therefore, the correlation between water and electricity usage behavior on the hourly time scale is not obvious.

In FIGURE 1(b), the daily electricity usage primarily lies in the range of 10 kW·h to 40 kW·h, while the daily water usage primarily lies in the range of 100 L to 1200 L. The daily water and electricity usage exhibit a certain periodicity, with the line chart showing noticeable synchronization and a roughly proportional fluctuation amplitude. It's clear that a distinct relationship exists between daily water and electricity usage among low-voltage residential users. This suggests the potential of utilizing this correlation to detect abnormal water and electricity usage. The results show that the correlation between water and electricity usage among residential users varies at different time scales. The correlation between hourly water and electricity consumption is weak, while the correlation becomes stronger at the daily scale. Although the users' consumption patterns remain unchanged, the observed time scale can lead to different results in the correlation between water and electricity usage. To further validate the correlation between daily water and electricity consumption, a quantitative analysis will be conducted using distribution functions.

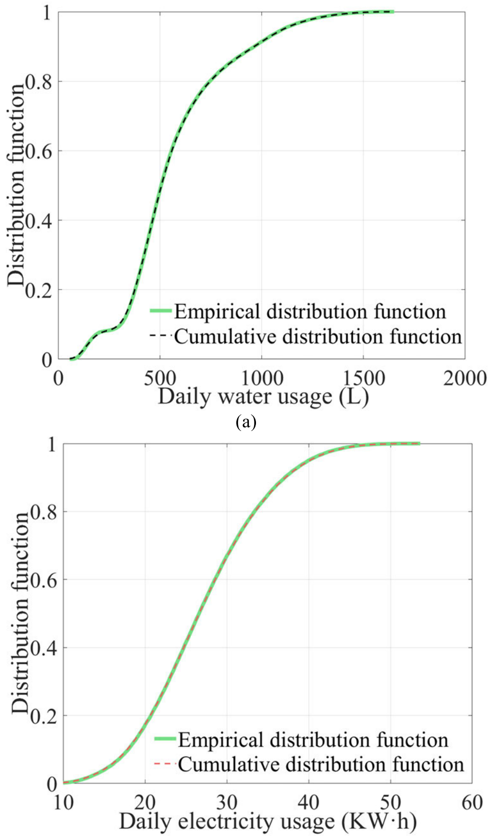
B. QUANTITATIVE ANALYSIS OF WATER AND ELECTRICITY CORRELATION

To delve deeper into the relationship between users' water and electricity consumption patterns, this section utilizes a distribution function for a quantitative analysis of the joint bivariate distribution concerning both water and electricity usage data. The data utilized in this research is obtained through SGCC smart meters in China's multi-meter integration demonstration zone, including data on daily electricity and water consumption for 50 standard household customers, spanning from October 1, 2020, through September 30, 2021 [16].

The copula function outlines the interdependence characteristics between variables and has been effectively applied in the fields of statistics and finance in recent years [17], [18]. The bivariate Copula function $C(u, v)$, which corresponds to the joint distribution of two random variables, is defined in the work by reference [19]. The copula distribution function is employed to analyze the relationship between the daily consumption of water X and electricity Y.

1) DETERMINATION OF MARGINAL PROBABILITY DISTRIBUTION FUNCTIONS FOR WATER AND ELECTRICITY

To determine the distribution types of daily water usage X and daily electricity usage Y, a non-parametric approach is employed to estimate the cumulative distribution function of daily electricity and water usage [20]. The empirical distribution function can effectively approximate the actual distribution function, the related results are shown in FIGURE 2. As can be seen, the difference between the results of the non-parametric estimation and the empirical distribution function is very slight.



(a)
(b)

FIGURE 2. The distribution functions of X and Y. (a) Daily water usage X. (b) Daily electricity usage Y.

2) COPULA FUNCTION SELECTION

After determining the empirical distribution and kernel distribution estimation of X and Y, the appropriate Copula function structure can be selected according to the shape of the bivariate frequency histogram$(U_i, V_i)(i = 1, 2, \cdot\cdot\cdot, n)$ shown in FIGURE 3. The frequency histogram can be used as an estimate of the (U, V) joint density function (i.e., Copula density function). Copula functions can be classified into five distinct categories [21]. Since (U, V) joint density function (i.e., Copula density function) has a symmetric tail, either the bivariate normal Copula or the t-Copula can be chosen to describe the correlation structure of daily water usage X and electricity daily usage Y [18].

3) PARAMETER ESTIMATION AND MODEL EVALUATION

Based on the historical data of daily water usage X and electricity usage Y, the maximum likelihood method is utilized for estimating the parameters of bivariate normal Copula and bivariate t-Copula. Table 1 shows the parameters obtained by model estimation. ρ is the linear correlation coefficient between the variables, and k is the degree of freedom.

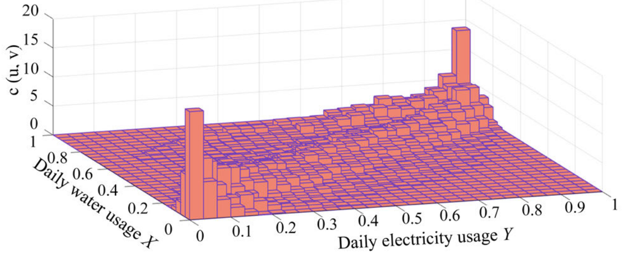


FIGURE 3. Frequency histogram of X and Y.

TABLE 1. Parameters by maximum likelihood estimation.

Function type	Spearman	Kendall	Fitting parameters
normal Copula	$\begin{bmatrix}1 & 0.8624\\ 0.8624 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.8753\\ 0.8753 & 1\end{bmatrix}$	$\rho = 0.8727$
t-Copula	$\begin{bmatrix}1 & 0.8649\\ 0.8649 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.8801\\ 0.8801 & 1\end{bmatrix}$	$\rho = 0.9325$ $k = 4$

To ascertain the most fitting Copula function for the joint distribution of water and electricity usage, the concept of the empirical Copula function and the square Euclidean distance index for goodness-of-fit detection are introduced [22].

Let $(x_i, y_i)(i = 1, 2, \cdots, n)$ be the sample selected from the two-dimensional sequence (X, Y). Also, denote the empirical distribution functions of X and Y as $F_n(x)$ and $G_n(y)$, and define the sample empirical function Copula as follows:

$$\overline{C_n}(u, v) = \frac{1}{n}\sum_{i=1}^{n} I_{[F_n(x_i)\leq u]} I_{[G_n(y_i)\leq v]}, \quad u, v \in [0, 1] \tag{1}$$

where $I_{[\cdot]}$ is an indicative function, when $F_n(x_i) \leq u$, $I_{[F_n(x_i)\leq u]} = 1$, otherwise $I_{[F_n(x_i)\leq u]} = 0$.

Euclidean distance of bivariate normal Copula(C_{Ga}) and bivariate t-Copula(C_{t}) with empirical Copula(C_{n}) using the following equation:

$$\begin{cases} d_{\mathrm{Ga}}^2 = \sum_{i=1}^{n} \left|\overline{C}_n(u_i, v_i) - C_{\mathrm{Ga}}(u_i, v_i)\right|^2 \\ d_{\mathrm{t}}^2 = \sum_{i=1}^{n} \left|\overline{C}_n(u_i, v_i) - C_{\mathrm{t}}(u_i, v_i)\right|^2 \end{cases} \tag{2}$$

The square of Euclidian distance between the bivariate normal Copula and the empirical Copula is 4.0722, and the square of Euclidian distance between the bivariate t-Copula and the empirical Copula is 3.4272. Therefore, it can be judged that the bivariate t-Copula and empirical Copula model can better fit the correlation between daily water usage X and daily electricity usage Y.

According to the bivariate t-Copula distribution function and its distribution parameters, the fitting probability density function C(u, v) of the joint distribution of water and electricity usage for residential users can be drawn, as shown in FIGURE 4. In the bivariate t-Copula probability density distribution diagram, coordinates u and v are the cumulative distribution function of daily water usage X and daily electricity usage Y respectively. Also, coordinates $C(u, v)$ is the probability density under daily water usage X and daily electricity usage Y. In order to facilitate identification, the combination of electricity and water with a higher probability density is marked as fuchsia, while the combination with a lower probability density is marked as blue.

In FIGURE 4(a), most observation points fall on both sides of the main diagonal of daily electricity usage and water usage, indicating a strong positive correlation between electricity usage and water usage sequence. The left and right ends show an obvious warping trend. When the residential users leave home, the corresponding performance is a synchronous reduction of water and electricity usage on the left end. When many people are resting at home, it is easy to show synchronous expansion of water and electricity at the right end. Because the number of such days is relatively concentrated, it is easy to show the bulge on the bivariate probability density map.

In FIGURE 4(b), apart from the diagonal bulge, the left and right sides drop rapidly, indicating a strong correlation in the daily water and electricity usage among normal users. Furthermore, the probability of two combinations with high electricity and low water usage, or high water and low electricity usage, is very low. This observation further supports the positive correlation between electricity usage and water usage among normal users.

Based on the above analysis of the strong correlation and synchronization between the daily water and electricity usage behavior of residential users, the strong correlation between the two can be used to detect whether there is abnormal water and electricity usage behavior.

III. DESIGN OF WATER AND ELECTRICITY ABNORMAL USAGE BEHAVIOR DETECTION SCHEME

In this paper, the distance correlation coefficient and density clustering are used to detect the abnormal usage behavior of water and electricity. According to the analysis results in Section I, the analysis involves computing the distance correlation coefficients between individual electricity and water usage for all users within the high-loss TSA, afterwards the distance correlation coefficient curve is drawn. Thereafter, DBSCAN is used for clustering, identifying those individuals as anomalous users whose distance correlation coefficients significantly diverge from the cluster of regular users, indicating a weak correlation in their water and electricity consumption patterns.

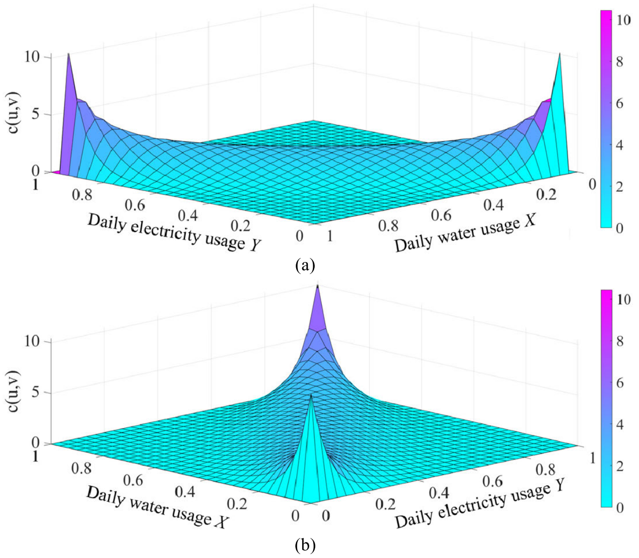


FIGURE 4. Bivariate t-Copula probability density distribution. (a) Front view. (b) Side view.

A. DISTANCE CORRELATION COEFFICIENT

The distance correlation coefficient is enhanced by the Pearson correlation coefficient, which can measure the nonlinear correlation between variables [23], [24], [25]. To assess the correlation between water usage X and electricity usage Y, the distribution function $F(Y)$ of Y can be compared with the conditional distribution function $F(Y|X)$ of Y under X. The higher the similarity between the two, the less electricity usage is influenced by water usage and the weaker the correlation between water usage and electricity usage. To facilitate easier calculation, the following characteristic function is used to replace the distribution function:

$$f_{XY}(s,t) = E\exp\left[\mathrm{i}\langle s, X\rangle + \mathrm{i}\langle t, Y\rangle\right] \tag{3}$$

$$f_X(s) = f_{XY}(s,0) = E\exp\left[\mathrm{i}\langle s, X\rangle\right] \tag{4}$$

$$f_Y(t) = f_{XY}(0,t) = E\exp\left[\mathrm{i}\langle t, Y\rangle\right] \tag{5}$$

where E is the mathematical expectation; i is an imaginary number unit; s and t are real vectors; $\langle\rangle$ is the dot product. If and only if $f_{XY}(s,t) - f_X(s)f_Y(t) = 0$, X and Y are not correlated, and vice versa.

The value range of the distance correlation coefficient is [0, 1]. As the distance correlation coefficient reflecting the relationship between water and electricity usage approaches 1, it indicates an increasingly strong linkage between the two. On the contrary, the closer the coefficient is to 0, the weaker the relationship between water and electricity consumption. Considering the strong correlation between water and electricity usage in normal users, anomalous behaviors in water and electricity consumption will reduce the coupling strength between the two, leading to a decrease in their correlation. Consequently, the distance correlation coefficient for users exhibiting abnormal behaviors tends to be lower than that for normal users.

Given a sufficiently large sample size, the distance correlation coefficient can describe the association between variables and capture a broad spectrum of relationships without constraining the relationship to a specific form. The distance correlation coefficient allows for the independent assessment of the direct coupling strength between a user's electricity and water usage when identifying abnormal users, without taking into account the impact from other users. This methodology has extensive applicability in the field of engineering [26]. It is worth mentioning that there are some drawbacks to identifying abnormal hydro users only through the distance correlation coefficient. Variations in abnormal time periods of water and electricity usage, coupled with the duration of creep, can cause fluctuations in the distance correlation coefficient of users. Consequently, it is challenging to simply set a specific threshold for distinguishing abnormal users. Besides, when a large user base is connected to the platform, a large number of distance correlation coefficient curves overlap with each other, making it difficult to identify abnormal users through curves. Therefore, incorporating clustering becomes necessary to differentiate and filter abnormal users.

A wide array of clustering algorithms exists, such as systematic, partition, hierarchical, and density-based clustering. Notably, the DBSCAN algorithm (Density-Based Spatial Clustering of Applications with Noise) stands out among density clustering methods. Its unique feature is the ability to determine the cluster count without a preset number, adapting automatically based on sample density in the given space, making it suitable for data sets of any shape [27]. For detailed definitions of sample points within DBSCAN, one may consult [28].

B. COMBINED DETECTING FRAMEWORK

Taking the water and electricity usage data in TSA as an example, the identification process of abnormal users with water and electricity in combination with FIGURE 5 is described as follows:

(1) Gather data on water and electricity usage from users via smart meters, and pre-process the water and electricity usage data of access users. The processed electricity usage data sequence is defined as Y_i and the water usage data sequence is defined as X_i, where i is the i-th user.

(2) Establish a distance correlation coefficient model for the water and electricity usage data of all users in TSA, so as to be able to calculate the distance correlation coefficient between electricity and water usage in different time periods.

(3) Utilizing the DBSCAN algorithm, users are clustered based on distance correlation coefficients to subsequently detect potential anomalous users with weak correlations in their water and electricity usage patterns.

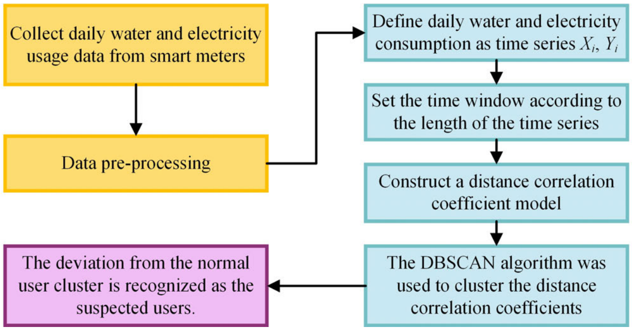

Collect daily water and electricity usage data from smart meters
Data pre-processing
Define daily water and electricity consumption as time series X_i, Y_i
Set the time window according to the length of the time series
Construct a distance correlation coefficient model
The DBSCAN algorithm was used to cluster the distance correlation coefficients
The deviation from the normal user cluster is recognized as the suspected users.

FIGURE 5. **Detection framework of the combined method.**

IV. ANALYSIS OF NUMERICAL EXPERIMENTS

Within this segment, a real usage dataset of TSA smart meters is used and the data is preprocessed. The data is collected by SGCC from a comprehensive multimeter demonstration area in China. The TSA connects to 42 residential users with a power loss of approximately 8.9%. Daily water and electricity usage data is collected for 95 days from January 1, 2021, to April 5, 2021 [16].

A. CONSTRUCTION OF DISTANCE CORRELATION COEFFICIENT MODEL

For users numbered from 1 to 42, daily consumption of electricity and water is characterized by the time series Y_1 to Y_{42} and X_1 to X_{42}, respectively. The distance correlation coefficient model is constructed for the water and electricity usage of the users. To better capture the information of user data across various time intervals, the distance correlation coefficient model is obtained from different time lengths. Time series lengths range from 1 to 30 days, 1 to 35 days, 1 to 40 days, and so forth. In the form of the time window of 14 time periods on the 1st to 95th day. This model is then used to calculate the distance correlation coefficient between the user's electricity usage Y_1-Y_{42} and water usage X_1-X_{42}, and draw the result of the distance correlation coefficient as shown in FIGURE 6.

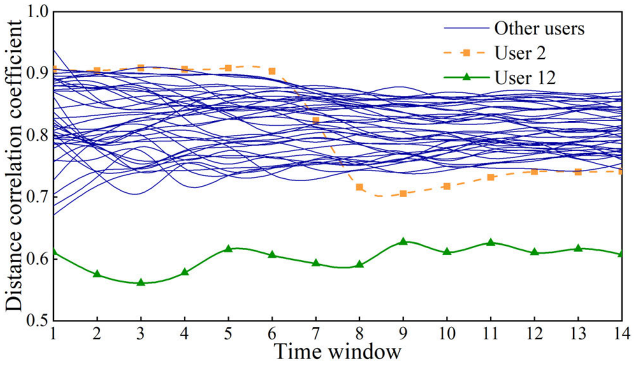


FIGURE 6. **Correlation coefficient curve of water and electricity distance.**

The consumer electricity usage Y_1-Y_{42} and water usage X_1-X_{42} fluctuate with the passage of time, and the corresponding distance correlation coefficient also fluctuates. Also, the calculation results tend to be stable with the increase in the calculation interval. In general, some users in the TSA will reduce the water usage due to the occasional dining out and other factors, resulting in a drop of the distance correlation coefficient. Nonetheless, such incidental elements are insufficient to notably lower the entire distance correlation coefficient, thereby categorizing the individual as an outlier.

The curve of user 12 deviates from that of most users. The range of distance correlation coefficient of most users is [0.7, 0.9], while the distance correlation coefficient of user 12 fluctuates around 0.6. Hence, it can be preliminarily judged that user 12 has the possibility of abnormal electricity usage. Meanwhile, the curve of user 2 has a high distance correlation coefficient in the first six periods, while there is a sudden drop from the 6th period and stabilizes at 0.7 or below after the 8th period. This may be caused by normal electricity usage in the first period and abnormal electricity usage in the later period.

If the user did not exhibit abnormal behavior, its distance correlation coefficient does not gravely deviate from the normal users, artificial according to such easy to confuse the user experience based on the curve from 2 users. In addition, different time periods and durations of abnormal power use will cause a difference in the distance correlation coefficient of abnormal users, and it is difficult to set a definite threshold to distinguish abnormal users. Hence, integrating a clustering approach is essential to differentiate and filter out abnormal users.

B. DBSCAN CLUSTER ANALYSIS

According to the description in Section II, the distance correlation coefficients between electricity usage Y_1-Y_{42} and water usage X_1-X_{42} are taken as data sets for DBSCAN clustering. The minimum number of samples required for cluster formation, MinPts, is set as 3. The k-dist curve is used to obtain the neighborhood radius Eps. The distance of the nearest k neighbors is selected and sorted from largest to smallest, and the k-dist curve is obtained as shown in FIGURE 7. The distance of 0.1 corresponds to the inflection point of the curve is set as E_{ps} The distance correlation coefficient calculated between electricity usage Y_1-Y_{42} and water usage X_1-X_{42} over various time periods is utilized as the feature. Moreover, Kernel principal components analysis is adopted to reduce it to 2 dimensions, and after dimensional reduction to a dimensionless feature space, the horizontal and vertical coordinates represent the transformed characteristic quantities. The clustering visualization results obtained are shown in FIGURE 8.

Cluster 1 in FIGURE 8 identifies normal users, while noise points identify suspected abnormal users. User 2 and User 12 are identified as users with abnormal water and electricity usage, which is consistent with the on-site inspection results. Compared to FIGURE 6, the analysis presented in FIGURE 8 offers clearer and more explicit results. This enhanced clarity facilitates the accurate identification of abnormal users, particularly in scenarios with a high volume of users or challenging identification conditions.

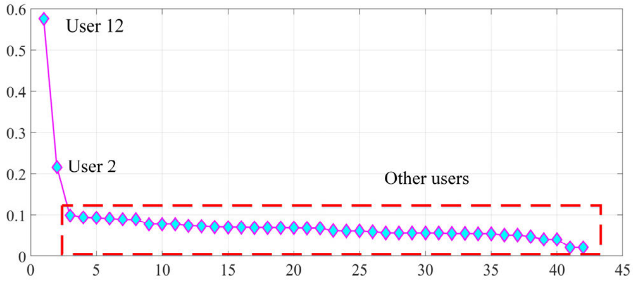


FIGURE 7. Distance correlation coefficient k-distance curve diagram of each user.

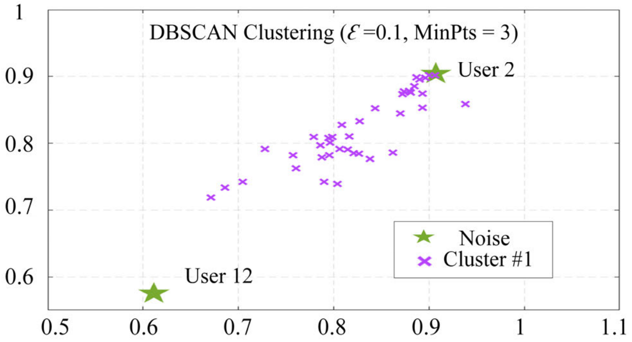


FIGURE 8. The results of the DBSCAN clustering results of each user's distance correlation coefficient.

C. COMPARATIVE EXPERIMENT

To further validate the advantages and performance of the multi-source information-based anomaly detection method, this section uses the real high-loss TSA from Chapter 4, Sections A and B, as the detection target. Two single-source information methods are selected for comparative experiments. References [12] and [29] present electricity consumption data-based detection methods, which perform anomaly detection by analyzing the potential relationship between power loss and users' electricity consumption data, and have been widely applied in practical engineering.

In [12], the correlation between power loss and electricity consumption is quantified using maximum mutual information. The correlation degree between variables is positively correlated with the maximum mutual information value. The greater the value, the larger this value is, the greater the suspicion of anomaly in the corresponding user. Conversely, the smaller the value, the lesser the suspicion of anomaly. Figure 9 displays the calculated maximum mutual information value between the overall power loss and the electricity data of all users.

By comparing FIGURE 8, FIGURE 9, and the previous analysis, it is evident that when applying the method proposed in [12] to identify abnormal users within the high-loss TSA, the maximum mutual information value of user 32 is the highest, followed by user 31. User 2 has the lowest maximum mutual information value, and user 12 is in the lower middle level, so the correlation appears to be weak. Thus, the method erroneously classified users 32 and 31 as abnormal users.

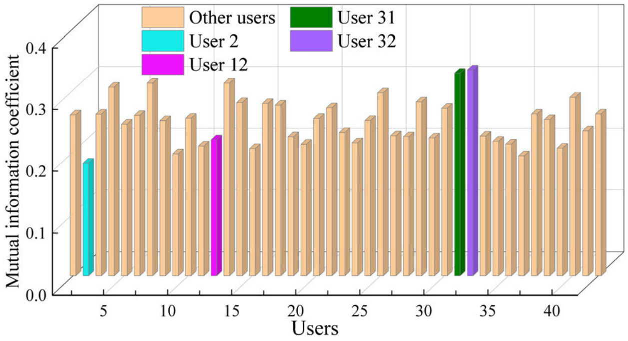


FIGURE 9. Result chart of mutual information coefficient.

TABLE 2. Granger test results of power loss in TSA.

Assumption	Significance level
Power loss is not the Granger cause of user 2	0.6622
User 2 is not the Granger cause of power loss	0.3970
Power loss is not the Granger cause of user 12	0.3979
User 12 is not the Granger cause of power loss	0.5476

Reference [29] used the Granger attribution method to build a model between power loss and electricity usage to identify users with abnormal behavior. The results are shown in Table 2, for the hypothesis that "User 2 is not the Granger cause of power loss and User 12 is not the Granger cause of power loss", none of their significance levels met the condition of less than 0.05, so the change of user 2 and user 12 is not the cause of the change of power loss. The results indicate that this method did not successfully detect two users with abnormal behavior.

In conclusion, compared to the two aforementioned methods that rely solely on electricity usage information for anomaly detection, the multi-information detection method proposed in this paper is effective in identifying abnormal water and electricity usage. It provides a clearer understanding of residents' energy consumption patterns and offers effective support for ensuring the efficient use of resources and preventing resource wastage.

V. CONCLUSION

To cope with the challenge of limited information in residents' energy usage detection in the context of smart cities, an abnormal usage behavior detection method of residential water and electricity usage by incorporating multi-source information is proposed, which lays a solid foundation for in-depth analysis of user behavior to further improve energy usage efficiency. The results showed that there were obvious synchronous fluctuations and correlation characteristics in the daily-scale water and electricity usage data. By utilizing the Copula distribution function, a strong correlation and synchronization of residential water and electricity usage behaviors are further demonstrated. This result not only elucidates the intrinsic connections in residential energy usage patterns but also holds significant importance for understanding and optimizing residential energy consumption. Based on this characteristic, a combined method for detecting abnormal behavior of resident users using both the water and electricity metering data is proposed. This method employs the distance correlation coefficient to characterize the strength of the correlation between water and electricity consumption. Clusters the distance correlation coefficients of users through DBSACN, which can effectively identify abnormal users who deviate from the normal group. A case study based on real data further verified the effectiveness of the method. Compared with anomaly detection methods that only use a single metering data, the proposed method can detect abnormal user behavior more accurately, thus reducing the risk of false negatives and false positives.

Although the proposed method for detecting abnormal behavior by integrating multi-source information shows significant results, it still has certain limitations. In some areas, residential properties are used for non-residential purposes, such as offices, storage, or educational training. These situations may cause changes in water and electricity usage patterns and their correlations, thus affecting the accuracy of the model. Future work can consider further enriching the multi-source information by incorporating natural gas energy data and analyzing the relationships between water, electricity, and gas. Additionally, more correlation patterns between water and electricity should be explored to further improve detection accuracy and expand application scenarios.

REFERENCES

[1] M. Talebkhah, A. Sali, M. Gordan, S. J. Hashim, and F. Z. Rokhani, "Comprehensive review on development of smart cities using industry 4.0 technologies," *IEEE Access*, vol. 11, pp. 91981–92030, 2023.

[2] V. Javidroozi, H. Shah, and G. Feldman, "FABS: A framework for addressing the business process change challenges for smart city development," *IEEE Access*, vol. 11, pp. 64850–64885, 2023.

[3] K. Liu, X. Liu, P. Zhang, Y. Xue, B. Li, and S. Su, "Load peak feature-based LSTM autoencoder for electricity theft identification," *Autom. Electr. Power Syst.*, vol. 47, no. 2, pp. 96–104, 2023.

[4] W. Hurst, C. A. C. Montañez, N. Shone, and D. Al-Jumeily, "An ensemble detection model using multinomial classification of stochastic gas smart meter data to improve wellbeing monitoring in smart cities," *IEEE Access*, vol. 8, pp. 7877–7898, 2020.

[5] Y. Zhu, Y. Zhang, L. Liu, Y. Liu, G. Li, M. Mao, and L. Lin, "Hybrid-order representation learning for electricity theft detection," *IEEE Trans. Ind. Informat.*, vol. 19, no. 2, pp. 1248–1259, Feb. 2023.

[6] H. Liu, X. Zhang, L. Gong, Z. Guo, Y. Zhao, J. Xu, and J. Xia, "Multi-scenario simulation and risk analysis of a water-energy coupled system: A case study of Wuhan City, China," *Sustain. Cities Soc.*, vol. 93, Jun. 2023, Art. no. 104518.

[7] Y. Himeur, K. Ghanem, A. Alsalemi, F. Bensaali, and A. Amira, "Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives," *Appl. Energy*, vol. 287, Apr. 2021, Art. no. 116601.

[8] L. Kou, L. Chen, M. Wang, J. Zhang, and Y. Lin, "A contrastive-learning-based abnormal electricity load detection method," *IEEE Internet Things J.*, vol. 11, no. 22, pp. 36619–36632, Nov. 2024.

[9] J. Wang and X. Li, "Abnormal electricity detection of users based on improved canopy-Kmeans and isolation forest algorithms," *IEEE Access*, vol. 12, pp. 99110–99121, 2024.

[10] H. Ghamkhar, M. J. Ghazizadeh, S. H. Mohajeri, I. Moslehi, and E. Yousefi-Khoshqalb, "An unsupervised method to exploit low-resolution water meter data for detecting end-users with abnormal consumption: Employing the DBSCAN and time series complexity," *Sustain. Cities Soc.*, vol. 94, Jul. 2023, Art. no. 104516.

[11] R. Moghaddass and J. Wang, "A hierarchical framework for smart grid anomaly detection using large-scale smart meter data," *IEEE Trans. Smart Grid*, vol. 9, no. 6, pp. 5820–5830, Nov. 2018.

[12] J. B. Leite and J. R. S. Mantovani, "Detecting and locating non-technical losses in modern distribution networks," *IEEE Trans. Smart Grid*, vol. 9, no. 2, pp. 1023–1032, Mar. 2018.

[13] Y. Yang, R. Song, Y. Xue, P. Zhang, Y. Xu, J. Kang, and H. Zhao, "A detection method for group fixed ratio electricity thieves based on correlation analysis of non-technical loss," *IEEE Access*, vol. 10, pp. 5608–5619, 2022.

[14] J. Shin, S. Son, and Y. Cha, "Spatial distribution modeling of customer complaints using machine learning for indoor water leakage management," *Sustain. Cities Soc.*, vol. 87, Dec. 2022, Art. no. 104255.

[15] Y. Peng, Y. Yang, Y. Xu, Y. Xue, R. Song, J. Kang, and H. Zhao, "Electricity theft detection in AMI based on clustering and local outlier factor," *IEEE Access*, vol. 9, pp. 107250–107259, 2021.

[16] S. Su, "Anomaly usage behavior detection based on multi-source water and electricity consumption information," IEEE Dataport, Jan. 2024.

[17] S. Jin, S. Su, Y. Xue, Y. Yang, S. Liu, and Y. Cao, "A review of data-driven electricity theft detection methods and a research outlook for low false positive rate," *Autom. Electr. Power Syst.*, vol. 46, no. 1, pp. 3–14, Jan. 2022.

[18] L. Dong, Y. Liu, H. Tang, and Y. Du, "Bearing data model of correlation probability box based on new G-Copula function," *IEEE Access*, vol. 8, pp. 224565–224577, 2020.

[19] R. B. Nelsen, *An Introduction to Copulas*. Berlin, Germany: Springer, 2006.

[20] A. Majdara and S. Nooshabadi, "Nonparametric density estimation using copula transform, Bayesian sequential partitioning, and diffusion-based kernel estimator," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 4, pp. 821–826, Apr. 2020.

[21] W. P. J. Philippe, S. Zhang, S. Eftekharnejad, P. K. Ghosh, and P. K. Varshney, "Mixed copula-based uncertainty modeling of hourly wind farm production for power system operational planning studies," *IEEE Access*, vol. 8, pp. 138569–138583, 2020.

[22] Y. Liu and Y. Chen, "Dynamic reliability evaluation of high-speed train gearbox based on copula function," *IEEE Access*, vol. 10, pp. 51792–51803, 2022.

[23] H. Gong, Y. Li, J. Zhang, B. Zhang, and X. Wang, "A new filter feature selection algorithm for classification task by ensembling Pearson correlation coefficient and mutual information," *Eng. Appl. Artif. Intell.*, vol. 131, May 2024, Art. no. 107865.

[24] D. Edelmann, T. F. Móri, and G. J. Székely, "On relationships between the Pearson and the distance correlation coefficients," *Statist. Probab. Lett.*, vol. 169, Feb. 2021, Art. no. 108960.

[25] H. Zhu, X. You, and S. Liu, "Multiple ant colony optimization based on Pearson correlation coefficient," *IEEE Access*, vol. 7, pp. 61628–61638, 2019.

[26] A. Davari, S. Islam, T. Seehaus, A. Hartmann, M. Braun, A. Maier, and V. Christlein, "On Mathews correlation coefficient and improved distance map loss for automatic glacier calving front segmentation in SAR imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5213212.

[27] Y. Aref, K. Cemal, Y. Asef, and S. Amir, "Automatic fuzzy-DBSCAN algorithm for morphological and overlapping datasets," *J. Syst. Eng. Electron.*, vol. 31, no. 6, pp. 1245–1253, Dec. 2020.

[28] K. Liu, B. Li, Y. Xue, Y. Yang, Y. Xu, A. Liu, and S. Su, "User electricity theft detection method based on transfer entropy density clustering," *Proc. CSEE*, vol. 42, no. 20, pp. 7535–7546, 2022.

[29] Y. Zheng, F. Chen, H. Yang, and S. Su, "Edge computing based electricity-theft detection of low-voltage users," *Frontiers Energy Res.*, vol. 10, Jun. 2022, Art. no. 892541.

WENQING ZHOU received the B.S. degree in electrical engineering from Xiangnan University, Chenzhou, China, in 2016, and the M.S. degree in physics from Wuhan University of Technology, Wuhan, China, in 2019. She is currently pursuing the Ph.D. degree in electrical engineering with Changsha University of Science & Technology (CSUST), Changsha, China.

Her research interests include the application of artificial intelligence and big data in power systems.

CHAOQIANG CHEN received the B.S. degree in electrical engineering from Wuhan University of Hydraulic & Electrical Engineering (WHUEE), Wuhan, China, in 1996, and the M.S. degree in electrical engineering from Wuhan University (WHU), Wuhan, in 2001.

He joined Changsha Electric Power Corporation, Changsha, China, in 1996. He is currently the Vice CEO of Changsha Electric Power Corporation. His research interest includes the operation and maintenance of distribution systems.

QIN YAN (Member, IEEE) received the B.S. degree in electrical engineering from WHU, China, in 2010, and the M.Eng. and Ph.D. degrees in electrical engineering from The Texas A&M University, College Station, TX, USA, in 2012 and 2018, respectively.

She is currently a Lecturer with CSUST, Changsha, China. Her research interests include plugin electric vehicles, smart grid, distributed energy resources, and optimization algorithms.

BIN LI (Student Member, IEEE) received the B.S. degree in mathematics and in applied mathematics from Hunan City University, Yiyang, China, in 2016, and the M.S. degree in electrical engineering from CSUST, Changsha, China, in 2019, where he is currently pursuing the Ph.D. degree in electrical engineering.

His research interests include the advanced application of metering data and climatic disasters in power systems.

KANG LIU received the B.S. degree in engineering and the M.S. degree in electrical engineering from CSUST, Changsha, China, in 2020 and 2023, respectively. He is currently pursuing the Ph.D. degree in electrical engineering with HNU, Changsha.

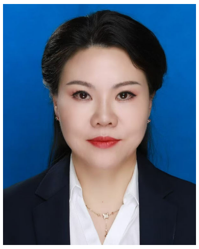
His research interests include the application of artificial intelligence and big data in power systems.

YINGJUN ZHENG received the B.S. and M.S. degrees in electrical engineering from CSUST, Changsha, China, in 2020 and 2023, respectively.

He is currently an Engineer with Jinhua Electric Power Corporation. His research interest includes the advanced application of metering data in power systems.

HONGMING YANG (Member, IEEE) received the M.Sc. degree in electrical engineering from WHU, in 1997, and the Ph.D. degree in electrical engineering from HUST, in 2003.

She was a Research Associate with The Hong Kong Polytechnic University, from 2009 to 2010, a Visiting Scholar with North Carolina State University, from 2010 to 2011, and a Research Fellow with the University of Newcastle, from 2013 to 2014. She joined CSUST, Changsha, China, in 1997, where she is currently a Full Professor. Her research interests include renewable energy in power systems and power markets.



HUI XIAO was born in China, in 1975. She received the M.S. degree in electrical engineering from Hunan University, Changsha, China, in 2002, and the Ph.D. degree in electrical engineering from Wuhan University, China, in 2015.

From July 2017 to August 2018, she was a Visiting Scholar with the School of Electrical and Electronic Engineering, University of Denver, USA. Since November 2005, she has been with the College of Electrical and Information Engineering, Changsha University of Science & Technology, Changsha, where she has been a Professor, since June 2020. Her research interests include new energy grid connection technology, power system operation and control, grid artificial intelligence, and power quality.

SHENG SU (Senior Member, IEEE) received the B.S. degree in electrical engineering from Wuhan University of Hydraulic & Electrical Engineering, Wuhan, China, in 1998, the M.S. degree in electrical engineering from Wuhan University, Wuhan, in 2002, and the Ph.D. degree in electrical engineering from Huazhong University of Science and Technology (HUST), Wuhan, in 2009.

He joined CSUST, Changsha, China, in 2002. From 2004 to 2007, he was a Research Assistant with The Hong Kong Polytechnic University, Hong Kong, China. He is currently a Professor with CSUST. His research interests include advanced application of metering data, climatic disaster defense, and cyber-attack defense of power systems.