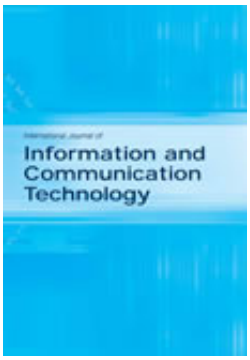


International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

A data-driven full hierarchy topology identification method for low-voltage distribution area

Yutong Chen, Zhihong Zheng, Pengyu Zhang, Wei Wang, Chu Pei, Yinghua Li

DOI: 10.1504/IJICT.2025.10071098

Article History:

Received:	11 March 2025
Last revised:	26 March 2025
Accepted:	26 March 2025
Published online:	27 May 2025

A data-driven full hierarchy topology identification method for low-voltage distribution area

Yutong Chen, Zhihong Zheng*, Pengyu Zhang, Wei Wang, Chu Pei and Yinghua Li

State Grid Shanxi Electric Power Company electric power Science, Research Institute,
Shanxi, 030000, China
Email: weibin279@qq.com
Email: weibin279@sina.cn
Email: zhangpy145@sina.com
Email: wangw134@163.com
Email: peichu279@163.com
Email: liyh778@qq.com
*Corresponding author

Abstract: The inaccurate topology of low-voltage distribution station area leads to inefficient fault disposal and affects user experience. In view of the above situation, this paper proposes a data-driven full hierarchical topology identification method for low-voltage distribution station areas. The density peak K-means (DPK-means) algorithm is employed in this method to discern the user's phase based on the similarity analysis of 'distribution transformer-branch box-metre box-user'. Furthermore, the Kendall correlation coefficients between the voltage curves at each hierarchy in the low-voltage distribution station area are calculated and normalised to identify subordinate relations. The proposed method enables the recognition of user phases and hierarchical subordinate relations within the low-voltage distribution station area. Finally, the effectiveness of the proposed method is analysed and verified in the actual distribution station area.

Keywords: low-voltage distribution station area; full hierarchy topology identification; the similarity of voltage curves; density peak K-means; DPK-means algorithm; Kendall correlation coefficient.

Reference to this paper should be made as follows: Chen, Y., Zheng, Z., Zhang, P., Wang, W., Pei, C. and Li, Y. (2025) 'A data-driven full hierarchy topology identification method for low-voltage distribution area', *Int. J. Information and Communication Technology*, Vol. 26, No. 15, pp.78–95.

Biographical notes: Yutong Chen is a Professor-level Senior Engineer, with a Master's degree, and graduated from North China Electric Power University in 2008. He worked in State Grid Shanxi Electric Power Research Institute. His research interests include distribution automation technology.

Zhihong Zheng is a Senior Engineer, with a Master's degree, and graduated from North China Electric Power University in 2016. He worked in State Grid Shanxi Electric Power Research Institute. His research interests include distribution automation technology.

Pengyu Zhang is an Engineer, with a Master's degree and graduated from Taiyuan University of Technology in 2020. He worked in State Grid Qingxu County Power Supply Company. His research interests include distribution automation technology.

Wei Wang is a Senior Engineer, with a Master's degree, and graduated from Taiyuan University of Technology in 2015.He worked in State Grid Shanxi Electric Power Research Institute. His research interests include distribution automation technology.

Chu Pei is a Senior Engineer, with a Master's degree, and graduated from Huazhong University of Science and Technology in 2016. She worked in State Grid Shanxi Electric Power Research Institute. Her research interests include distribution automation technology.

Yinghua Li is a Senior Engineer, with a Master's degree, and graduated from Taiyuan University of Technology in 2007. He worked in State Grid lvliang Power Supply Company. His research interests include distribution automation.

1 Introduction

The low voltage distribution network is situated at the terminal of the power grid and directly caters to end-users. However, due to the complexity of wiring in the low-voltage distribution station area, there are inevitable errors in the record and archive, which leads to inefficient fault disposal and affects user experience. Therefore, it is very meaningful to identify the topology of low-voltage distribution station area (Martins et al., 2002; Zelenskii et al., 2022; Mittal and Verma, 2023).

The primary objective of low-voltage distribution station areas topology identification is to organise the physical connection relation between the 'distribution transformer-branch-metre box-user' and determine the user phase. At present, there are many references focus on the topology identification of low-voltage distribution station area. Angelopoulos et al. (2020) based on the multiple mapping relationship between the switching state of distribution network and the measured voltage data. The improved conformal prediction method is used to integrate and output the possible operating topologies, and various possible alternative topologies and their probabilities are given. A distribution network topology identification method based on CNN and improved conformal prediction is constructed. Srinivas and Wu (2022) proposed a probabilistic method based on un-scented Kalman filtering and Newton Raphson (NR) iterative accurate identification of network topology. The influence of SMs and μ PMU on measurement noise is analysed, and the study of the acceptable noise level is quantified. The influence of identification algorithms on network state estimation. Jafarian et al (2020) use Monte Carlo simulations, he explored different DER generation and load values. Two hidden layer feedforward deep neural networks are used to classify different topologies. Anderson and Yu (2021) proposes an algorithm based on graph signal processing to detect changes in the topology and to identify the new topology after reconfiguration using smart metre voltage magnitude measurements by comparing the smoothness of the signal on the possible topologies. Pappu et al. (2017) used principal

component analysis and its graph theory interpretation to identify load phase connectivity from energy measurements of time series. The method deduces the steady-state network topology from the energy measurement of a smart metre. A new data-driven method is proposed to identify the underlying network topology of low-voltage distribution networks. Padullaparti et al. (2022) proposed a robust phase recognition algorithm based on supervised machine learning. The algorithm can accurately identify the phase connectivity of AMI metres with significant PV generation. Ma et al. (2022) line parameters and nodal voltage phase angles are estimated using the Newton-Raphson method based on nodal measurements of real and reactive power injections, as well as voltage magnitudes. Poudel et al. (2023) proposed a state estimation model selection method based on graph theory. The first stage is to reduce the sample space of possible exchange combinations by enumerating the candidate topologies based on graphs. Provide candidate topologies for the second phase. The second stage is model selection based on state estimation. The aim is to solve state estimates for each candidate topology to obtain system variables that are consistent with system measurements. Fabbiani et al. (2021) proposed an online learning program to estimate a network access matrix that captures topological information and line parameters. A recursive recognition algorithm is provided by using the phasor measurement of voltage and current. Liu et al. (2021) proposed a user variation relation identification method based on derivative dynamic time bending (DDTW) algorithm and density-based noisy spatial clustering application (DBSCAN) algorithm. Pengwah et al. (2021) proposes an enhanced graph learning algorithm with backtracking to generate graph clusters and select the best fit candidates from them using a set of optimisation criteria. Smart metres are used to measure the load of the grid to estimate the voltage sensitivity factor related to changes in the load current. These coefficients are then used to reconstruct the network topology. Li et al. (2021) K-means clustering algorithm and Pearson voltage correlation coefficient analysis principle were used to complete the topology model of low-voltage distribution network.

At present, the methods for topology identification of low-voltage distribution network can be divided into signal injection method and data analysis method. The signal injection method is dependent on hardware devices in practical applications. The use of this method is faced with the problem of high cost. Furthermore, the signal injection method is highly susceptible to electromagnetic interference, thereby resulting in erroneous recognition outcomes. The low-voltage distribution network, situated at the terminal of the power grid and in direct proximity to end-users, represents a pivotal element within the construction of a smart power grid. With the continuous improvement of advanced metering infrastructure (AMI) in the distribution network configuration, a large number of user data resources are generated. Data analysis method refers to mining the intrinsic information of the data by analysing the characteristics of the electricity volume in the metre and using relevant algorithms (Niu et al., 2022).Therefore, data analysis method has broad application prospects. However, due to the higher degree of similarity in the voltage fluctuation curve within the low-voltage distribution station areas, there is a propensity for identification errors when determining the user's phase. Simultaneously, previous studies have rarely distinguished the hierarchical subordination relation within low-voltage distribution station areas, hindering comprehensive topological identification across all hierarchy.

The existing exploration methods can be divided into signal injection method and data analysis method. The signal injection method uses a direct physical connection, so that the results are reliable. However, the signal injection method relies on specialised

hardware and is costly. Not suitable for large-scale deployment. Data analysis requires no hardware modification, so the cost is low. At the same time, data analysis can be used to deal with massive metre data. However, the data analysis method has some disadvantages, such as high dependence on preset parameters and weak ability to distinguish high similarity voltage curves.

To sum up, the biggest difficulties associated with topology identification are the following. On the one hand, users in the same area are highly similar because of the close electrical distance and the degree of voltage fluctuations. The traditional clustering method is difficult to distinguish the phase attribution, leading to misjudgment. On the other hand, the existing methods mainly focus on the direct correlation of the house-hold-variable relationship, and ignore the multi-level membership relationship of 'branch box-table box-user', It is difficult to build a complete topology.

In order to solve the above problems, this paper proposes a data-driven full-hierarchy topology identification method for low-voltage distribution station areas based on. The multi-dimensional features of the user's voltage fluctuation are extracted and utilised, followed by employing the DPK-means algorithm to cluster the voltage fluctuation curve for precise identification of each user's phase. K-means clustering algorithm requires predefined cluster number. The random selection of initial cluster centre may lead to local optimal results. The proposed method uses DPK-means algorithm to improve phase recognition. By calculating the sample density, the cluster centre and number are automatically determined, eliminating the dependence on preset parameters. The distance calculation is optimised with Gaussian kernel function to avoid the clustering fluctuation caused by random initialisation of traditional K-means. The Kendall correlation coefficients are computed and normalised to establish the hierarchical relation between 'branch box-metre box-user' within the distribution substation area, thereby enabling comprehensive topological identification of the low-voltage distribution station areas. The normalised Kendall coefficient is used to construct hierarchical relationships: the voltage correlation difference of equipment at different levels is amplified, and the fuzzy membership relationship problem under high similarity is solved. At the same time, the hierarchical topology is reconstructed to realise the complete topology of 'transformer – branch box – table box – user'. Finally, the topology recognition results are verified on the experimental bench.

The main contributions of this paper are summarised as follows:

1 This paper proposed a method based on DPK-means clustering to identify the user phase. DPK-means clustering method overcomes the disadvantages that K-means needs to set the number of clusters in advance and randomly selection of initial cluster centres may lead to the results falling into local optimum. DPK-means clustering method can improve the stability of clustering results.

2 The hierarchical subordinate relation in the low-voltage distribution station area is obtained by calculating the voltage Kendall coefficients between branch boxes, metre boxes and users in pairs. The normalisation of Kendall coefficients can enhance the visibility of the results. The proposed method has significant advantages in the case of high voltage similarity.

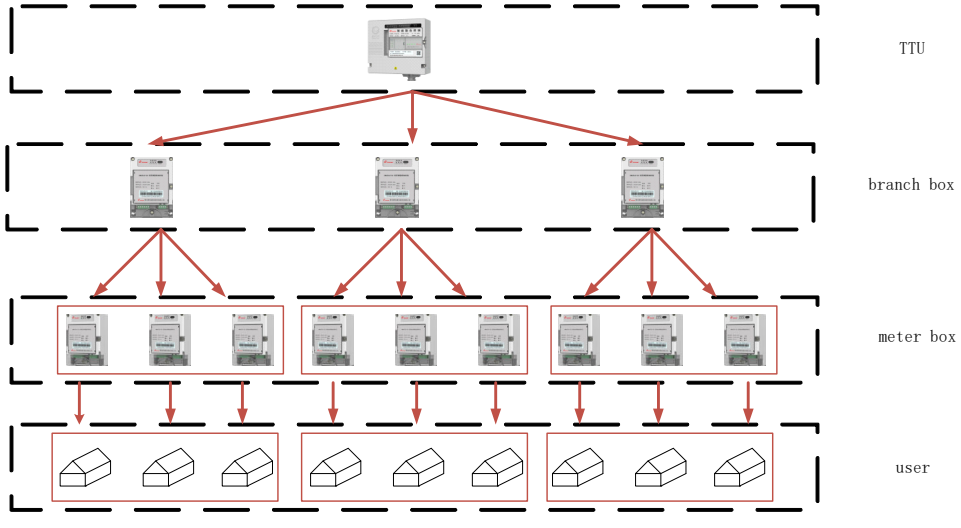
The rest of this paper is organised as follows: Section 2 introduce the structure of low-voltage distribution area. Section 3 a method of obtaining user phase by using DPK-means and a method of obtaining hierarchical subordinate relation of low-voltage

distribution station area by calculating Kendall correlation coefficient are given. Section 4, the reliability of the method is verified in the actual low voltage distribution station area. Finally, Section 5 concludes this article.

2 Low-voltage distribution substation area

In the power system, the low-voltage distribution station area refers to the region where a distribution transformer provides low-voltage power supply. The typical low-voltage distribution station area structure is shown in Figure 1. Each low-voltage distribution station area is a distribution transformer, transformer followed by branch box, branch box connecting to the box, box access to the user's structure. The voltage, current. Power data of the main metre in the station area are collected by distribution transformer supervisory terminal unit (TTU) every 15 minutes for continuous monitoring purposes. Summarise and analyse the data uploaded by down-stream equipment. This data can be uploaded to a distribution automation system or an information system for electricity acquisition. The branch box and metre box can collect three-phase voltage, current and power data. On the user side, residential electricity is usually single-phase user. When monitoring the current data of the low-voltage distribution station area, there are instances where certain users' long-term non-usage of electricity renders their current data unsuitable for topology identification. Therefore, it is preferable to gather voltage data from the low-voltage distribution station area. When the low-voltage distribution station area is connected to the load, both the voltage of the station transformer and that of the user undergo changes in response to load fluctuations. Additionally, due to close electrical distances within a given station, there exists a high similarity in voltage fluctuation curves.

Figure 1 Schematic diagram of low-voltage distribution station area structure (see online version for colours)



Low-voltage distribution station area topology identification refers to identifying the full hierarchical network of 'distribution transformer-branch box-metre box-user-phase'. contains the following two parts. The first part identifies the user's phase. Uses DPK-means cluster analysis to obtain the phase classification of all users in the low-voltage distribution station area. The second part establishes the subordinate relation of 'distribution transformer-branch-metre box-user' within the low-voltage distribution station area, normalisation of Kendall coefficient. Derives the hierarchical subordination based on correlation sorting.

3 Topology identification method

3.1 Phase identification based on DPK-means algorithm

Density peaks clustering (DPC) can automatically find cluster centres and achieve clustering algorithms of arbitrary distribution shapes and data. Its core idea is that:

1 compared with other samples around the cluster centre, the samples of the cluster centre have the highest density

2 the distance between the cluster centre sample and another cluster sample with higher density is large (Wahyuningrum et al., 2021).

Therefore, this paper proposes to optimise the K-means algorithm by using the density peak method. Then distinguish the clusters of users in the low-voltage distribution station area, so as to realise the phase differentiation of users.

The process of DPK-means algorithm to realise phase recognition of low-voltage distribution station area is as follows:

1 Input voltage sample data $D = \left\{ U_1^T, U_2^T, \cdots, U_n^T \right\}$ where U_i^T represents the voltage at time T of the i sample.

2 Gaussian Kernel was used to calculate the density ρ_i of the sample points

$$\rho_i = \sum_j \exp\left(-\left(d_{ij} / d_c\right)\right) \tag{1}$$

where d_{ij} represents the distance between sample U_i and sample U_j, and dc represents the truncation distance, usually, the data distance of the top 2% position is selected as the truncation distance.

3 Find other samples that are closest to sample U_i and higher than density ρ_i.

$$\delta_i = \begin{cases} \min\limits_{j:\rho_j > \rho_i} \left(d_{ij}\right), i \geq 1 \\ \max\limits_j \left(d_{ij}\right) \end{cases} \tag{2}$$

where δ_i represents the minimum distance between sample U_i and other samples higher than density ρ_i. But if the sample U_i is the highest sample value, set the

maximum value between the U_i and the rest of the samples to δ_i. j indicates the number of other samples higher than the density ρ_i.

4 Select the cluster centre. The cluster centre is selected by the calculated value of γ:

$$\gamma_i = \rho_i * \delta_i \tag{3}$$

The density peak method takes the sample points with large local density and large distance as the clustering centre. Rath and Srungavarapu (2021) points out that the number of cluster centres K and cluster centres α_1, α_2, … α_K can be determined by the changes in the size of the values.

5 Recalculate the centre of each cluster in terms of $Center = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i$

6 Repartition the cluster and determine the cluster centre.

7 The process of (5) ~ (6) is repeated until the clustering centre no longer changes and the value of the objective function J is minimised.

$$J = \sum_{i=1}^{N} \sum_{j=1}^{K} r_{ij} \left\| x_i - \alpha_k \right\|^2 \tag{4}$$

where $\begin{cases} r_{ij} = 1, & x_i \in C_k \\ r_{ij} = 0, & x_i \notin C_k \end{cases}$, α_K represents the cluster centre.

According to the DPK-means algorithm, the voltage of the metre box and the user voltage data are clustered. The result can distinguish the user's phase. According to the voltage data attribution phase shown in the metre box, determine which phase the user belongs to. Take 12 users as an example, if user 1, user 4, user 7 and user 10 are clustered together with the metre box A, then user 1, user 4, user 7 and user 10 can be divided into A phase users.

3.2 Recognition of hierarchical subordinate based on normalised Kendall coefficient

When the voltage amplitude fluctuations of users in the station area are similar, the introduction of Kendall correlation coefficient to determine the topological connection relation can help intuitively distinguish the differences (Curiac and Micea, 2022). The Kendall voltage correlation coefficient is a quantitative measure that assesses the hierarchy of association between two datasets.

$$S = \sum_{1 \le i < j}^{n} \left[sign\left(p_i - p_j\right) sign\left(q_i - q_j\right) \right] \tag{5}$$

$$D = \sum_{1 \le i < j}^{n} \left[sign\left(p_i - p_j\right) sign\left(q_j - q_i\right) \right] \tag{6}$$

$$\tau = \frac{S - D}{\sqrt{\left(Y_0 - Y_1\right)\left(Y_0 - Y_2\right)}} \tag{7}$$

In the formula $Y_0 = \frac{n(n-1)}{2}$, $Y_1 = \sum_{m=1}^{z} \frac{t_m\left(t_m - 1\right)}{2}$, $Y_2 = \sum_{m=1}^{z} \frac{h_m\left(h_m - 1\right)}{2}$.

$$R_{normalisation} = 10 \times \frac{\tau - \tau_{\min}}{\tau_{\max} - \tau_{\min}} \tag{8}$$

where, X and Y represent two groups of data to be compared, p_i and q_i are elements in X and Y. τ represents the Kendall correlation coefficient. t_m or h_m the number of elements contained in the MTH small set of X or Y. τ_{min} is the minimum value of the feature, τ_{max} is the maximum value of the feature, and the larger the value of $|R|$, the higher the degree of correlation between the two groups of data (Zhuravlyova and Vedishchev, 2021).

The variation in voltages within the same station is smaller compared to that across different stations, resulting in closely clustered values during Kendall correlation coefficient calculations and potentially leading to judgment errors. To address this issue, equation (8) can be employed for normalising the Kendall correlation coefficient, thereby enhancing data visualisation effects.

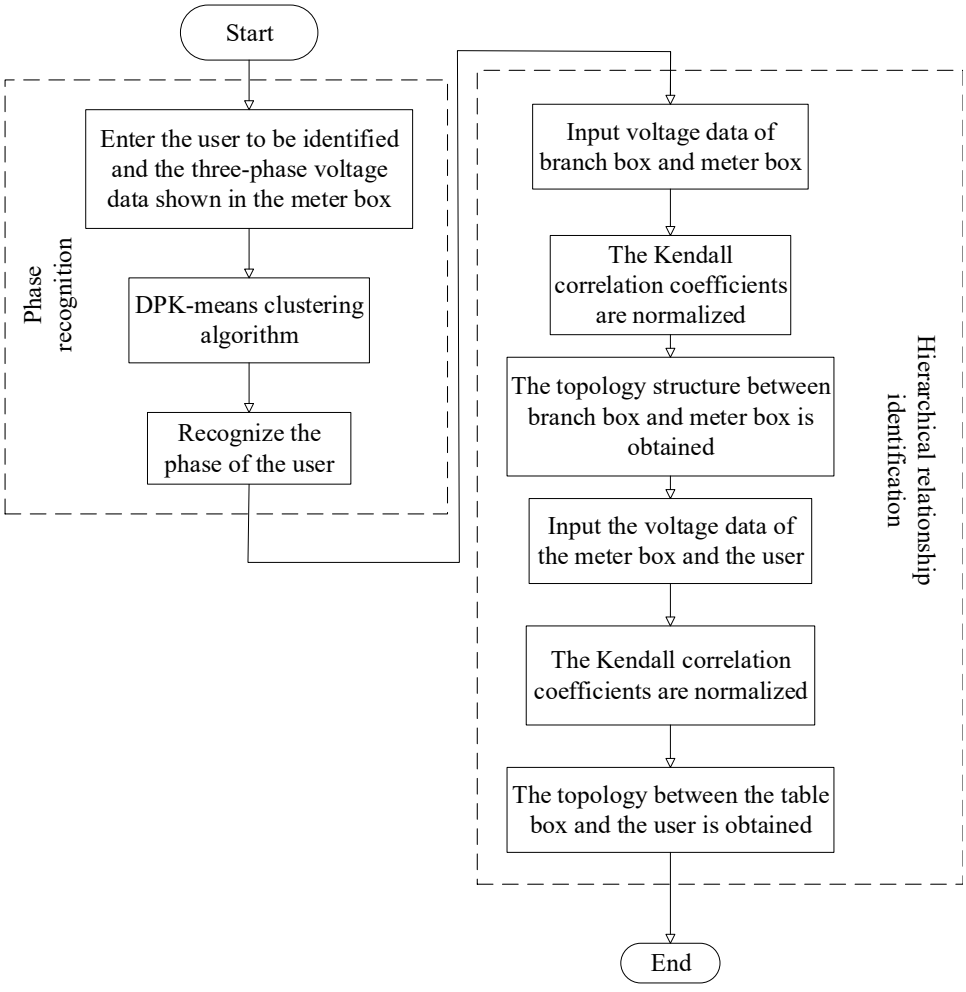
3.3 The overall technical route is identified by the topology of the low-voltage distribution area

The DPK-means algorithm is initially employed to cluster the input user and three-phase voltage data of any metre box, enabling the determination of the user's corresponding phase. Consequently, accurate phase recognition for each user can be achieved. The Kendall correlation coefficients between the branch box, the boxes, and the users in the low-voltage distribution station area are subsequently normalised. Based on these correlations, we can identify the connections between the branch box and the box, as well as between the box and the user within this station area. This identification process enables us to effectively discern hierarchical relation within the station area. The final identification algorithm flow of phase relation and hierarchical relation in low-voltage distribution station area is shown in Figure 2.

4 Case study

The voltage data from a specific day in a distribution station area located in Shanxi Province, China were selected for analysis. The smart metre had a sampling interval of 15 minutes, allowing for the collection of 96 voltage readings per day. According to the collected information, the low-voltage distribution station can be classified into four hierarchical hierarchy: a distribution transformer, the 4 branch boxes are respectively numbered as X_1, X_2, X_3, X_4. The 20 boxes are respectively numbered as Y_1, Y_2, …Y_{20}, 120 users are respectively numbered as Z_1, Z_2, …Z_{120}.

Figure 2 Topology identification process for low-voltage distribution area



4.1 Low-voltage power distribution area user phase identification

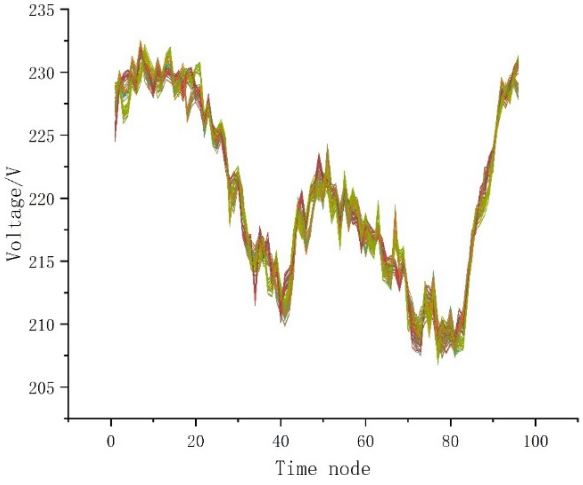
The data in this paper come from the intelligent fusion terminal of a subsidiary company of the state grid in Shanxi Province, China, and the specific cooperation unit is the power supply Bureau in Shanxi Province. Data collection followed 30 consecutive days of voltage data, the collection frequency was collected every 15 minutes, and 96 voltage sampling points were generated in a single day. The intelligent fusion terminal can cover voltage data including transformers, branch boxes, metre boxes, and users to ensure full coverage of the distribution station area.

The voltage data from the metre box's three-phase and all users' daily voltage trend curves are combined to form a voltage sequence, draw the voltage trend curve as shown in Figure 3.

Collect three-phase voltage data from any metre box and the data of 120 users for DPK-means clustering analysis. Figure 4(b) in Figure 4 is the descending order γ graph

calculated by equation (3). The three points in Figure 4(b) exhibiting significant changes in γ value correspond to the points depicted in Figure 4(a). Consequently, it can be inferred that these three selected clustering centres represent distinct clusters to which the user belongs. The phases of the remaining 120 users corresponded respectively to the three clustering centres. The clustering results obtained are shown in Figure 4(c), where different clusters represent different phases. Based on a comparison between the clustering results and the original topology findings, a 100% success rate in clustering was achieved.

Figure 3 Voltage fluctuation curve between the watch box and the user on a certain day (see online version for colours)



According to the classification results obtained from the DPK-means clustering algorithm, the corresponding association between users and three phases is determined. Specifically, there are 40 users in Phase A, 40 users in Phase B, and 40 users in Phase C. The voltage profiles of these three categories of power consumers are depicted in Figure 5.

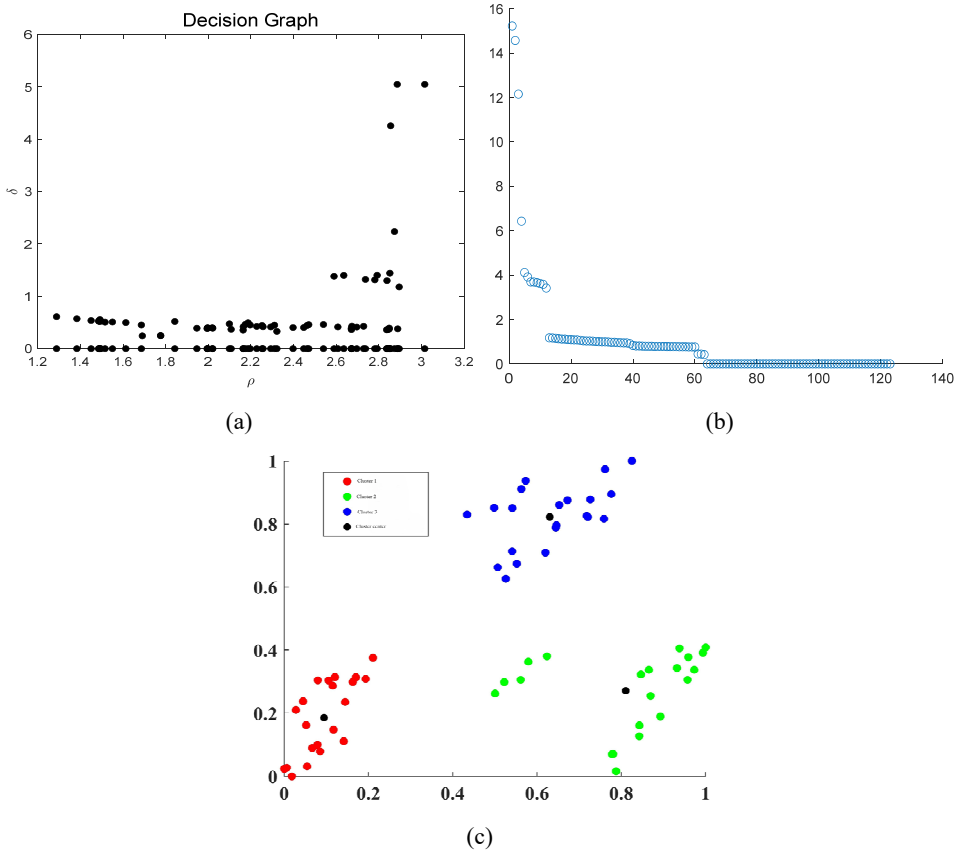
The group of phase A users is{Z3, Z9, Z11, Z13, Z14, Z15, Z17, Z20, Z22, Z28, Z32, Z40, Z41, Z42, Z44, Z49, Z56, Z60, Z65, Z66, Z67, Z68, Z70, Z75, Z79, Z81, Z82, Z83, Z84, Z88, Z90, Z98, Z99, Z102, Z104, Z112, Z114, Z116, Z119, Z120}.

The group of phase B users is{Z1, Z2, Z5, Z7, Z10, Z16, Z23, Z24, Z25, Z27, Z29, Z30, Z34, Z35, Z36, Z37, Z45, Z47, Z50, Z51, Z53, Z54, Z59, Z61, Z62, Z69, Z71, Z76, Z80, Z85, Z86, Z89, Z91, Z92, Z93, Z95, Z103, Z105, Z106, Z107}.

The group of phase C users is {Z4, Z6, Z8, Z12, Z18, Z19, Z21, Z26, Z31, Z33, Z38, Z39, Z43, Z46, Z48, Z52, Z55, Z57, Z57, Z58, Z63, Z64, Z72, Z73, Z74, Z77, Z78, Z87, Z94, Z96, Z97, Z100, Z101, Z108, Z109, Z110, Z111, Z113, Z115, Z117, Z118}.

In order to further prove the accuracy of the proposed method in phase recognition, the proposed method is compared with K-means algorithm, and the results are shown in Table 1. According to the voltage data provided by the intelligent fusion terminal and input into the method model in this pa-per, the full-level topology recognition results are obtained. The recognition accuracy can be obtained by comparing the result with the original construction topology.

Figure 4 Clustering results of phase relation in low-voltage distribution area based on DPK-means (see online version for colours)



The clustering results of the K-means algorithm are unstable, and the 5-day voltage data collected are grouped to obtain different identification errors in the clustering results of the K-means algorithm in Day2, Day3 and Day5.

Table 1 Comparison of accuracy between K-means algorithm and DPK-means algorithm

Method	*Phase recognition accuracy*				
	Day1	*Day2*	*Day3*	*Day4*	*Day5*
K-means	100%	49.7%	95.7%	100%	80%
DPK-means	100%	100%	100%	100%	100%

The K-means algorithm iterates to assign data points to the nearest cluster centre and updates the centre point until convergence. However, the K-means algorithm is sensitive to the initial centre and cannot handle non-spherical clusters. DPC algorithm automatically identifies cluster centres based on local density and relative distance, which is suitable for discovering clusters with arbitrary shapes. But the parameter DPC algorithm has high computational complexity. Poor effect on high density overlapping areas. The method used in this paper is DPK-means algorithm to identify the phase of

users in the platform area, and this method is used to eliminate the dependence on cluster number and initial centre, and improve the stability. After clustering, the complexity of subsequent hierarchical analysis can be reduced.

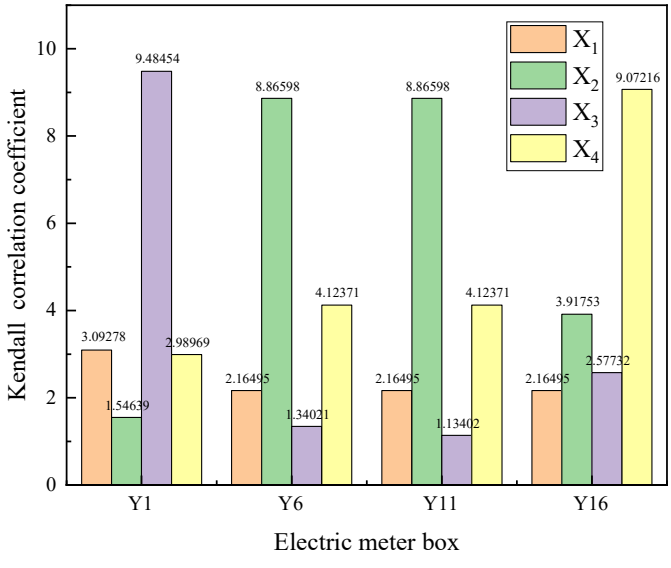
4.2 Identification of low-voltage distribution station area hierarchy membership relation

When assessing the hierarchical relation between individual areas, Kendall voltage correlation coefficients are computed using equation (8) and subsequently arranged in descending order based on these coefficients. The maximum similarity value indicates the highest hierarchy of association between them.

Taking the analysis of the relation between the metre box and the branch box as an example, due to the large size of the table, only the normalised Kendall coefficients between some metre boxes and branch boxes are listed in the calculation example. In order to see the correlation characteristics between them intuitively and clearly, a bar chart is now used to characterise their correlation, as shown in Figure 5. The Kendall coefficient of Y1 and X3 is 9.4854. The correlation between Y1 and X3 is the highest, and it can be seen that table box Y1 belongs to branch box X3. In the same way, we can determine that Y6 belongs to X2, Y16 belongs to X4, and Y11 belongs to X2.

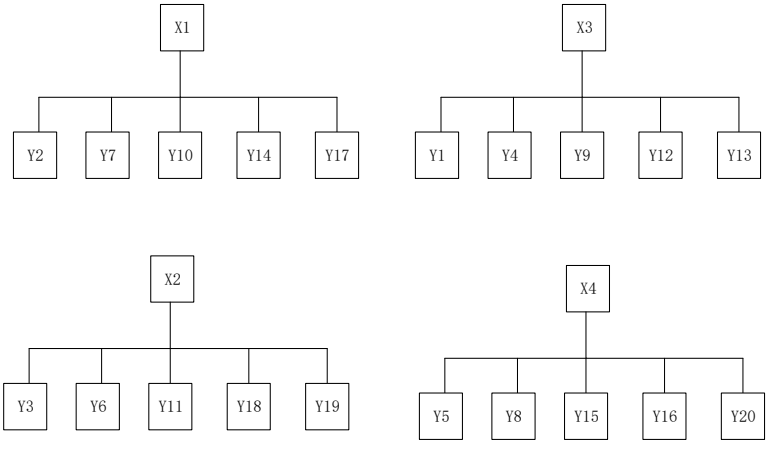
According to the subordinate relation between boxes Y1, Y6, Y11, and Y16 and branches X1, X2, X3, and X4 respectively, it can be inferred that box Y1 is affiliated with branch box X3 with a similarity coefficient of 9.91. Similarly, it can be deduced that Y6 belongs to X2, Y11 belongs to X2, and Y16 belongs to X4. The same methodology can be applied to determine the affiliation of boxes Y1–Y20 with branches X1–X4. The results are shown in Figure 6.

Figure 5 Correlation representation diagram of part branch box and table box (see online version for colours)



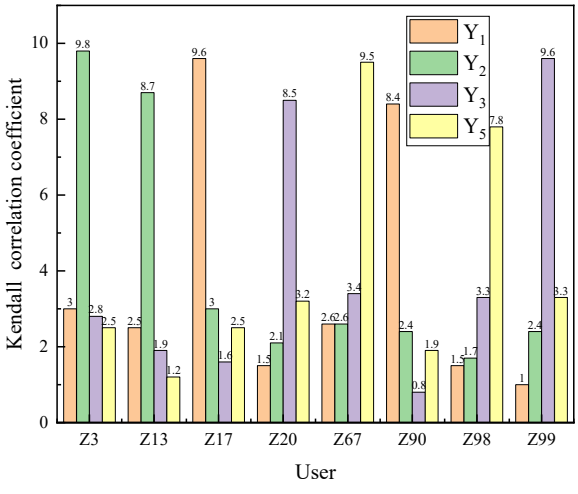
When using the same method to determine the subordinate between the metre box and the user. Under the same phase, even if users do not belong to their subordinate metre boxes, the similarity between them is still high. However, in different phases, even if users belong to the metre box, the similarity coefficient between them is very low. Therefore, based on the user phases identified in Section 4. A, we calculate the normalised Kendall correlation coefficient between users in phase A and telephone boxes to determine their subordinate relation.

Figure 6 Subordinate relation between branch box and metre box



The calculation example in Figure 7 presents the normalised Kendall coefficients between various metre boxes and users.

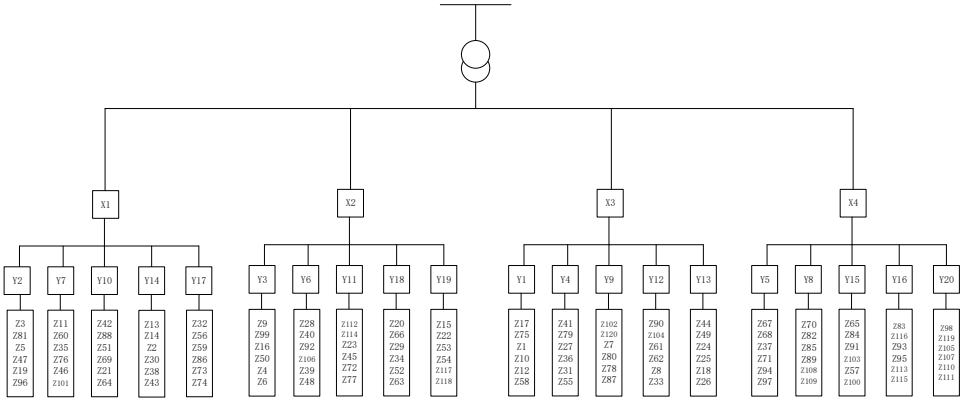
Figure 7 Part of the table box and user correlation representation diagram (see online version for colours)



According to the analysis in Figure 7, Z3 has the highest correlation coefficient with Y2, 9.8. In the normalised Kendall coefficient table between 20 boxes and 120 users, Z3 also

has the highest correlation coefficient with Y2. Therefore, it can be determined that Z3 belongs to the branch under the box of Y2. Similarly, Z99 belongs to Y3, Z17 to Y1, and Z67 to Y5. The correlation coefficient between Z13 and Y2 is 9.68, the highest value in Figure 7. The metre box exhibiting the highest correlation coefficient with Z13 in the normalised Kendall coefficient table is Y14. Therefore, Z13 does not belong to Y2, but to Y14. The reason why Z13 has a higher correlation coefficient with Y2 than with other boxes is that Y2 and Y14 belong to the same branch box X1. Similarly, Z20 does not belong to Y3, It has the highest similarity with it in Figure 7, but to Y20, Z90 belongs to Y12, and Z98 belongs to Y20.

Figure 8 Topology of a district in Shanxi Province, China



According to the above steps, the users in Phase B and Phase C divided in Section 4. A are calculated again, and the results of the subordinate relation between the remaining users and the metre box can be obtained.

The hierarchical topological relation between 'branch box-metre box-user' can be accurately determined by performing the aforementioned steps of calculating the normalised Kendall coefficient analysis between the branch box and metre box, as well as between the metre box and user, as illustrated in Figure 8. The topology is compared with the original topology of the original station area to obtain 100% accuracy.

The core steps of the method in this paper include DPK-means clustering and normalised Kendall coefficient calculation, whose time complexity and space complexity are shown in Table 2.

Table 2 The complexity of the method in this paper

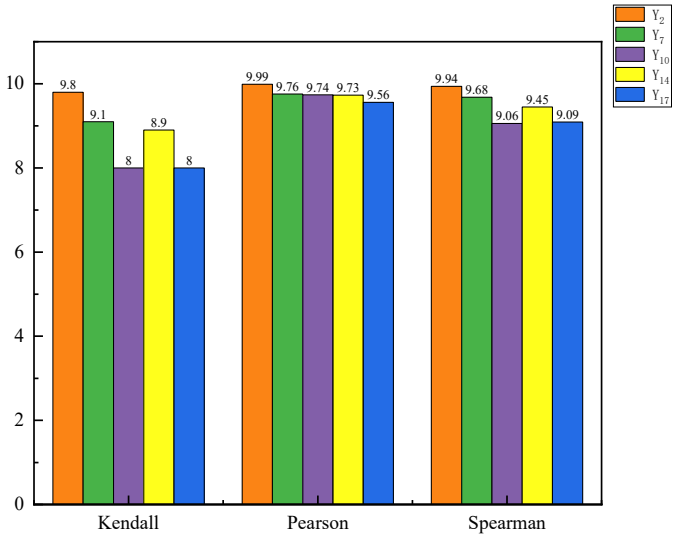
	Time complexity	*Space complexity*
DPK-means clustering	$O(n^2)$	$O(n)$
Normalised Kendall coefficient	$O(m*n*logn)$	$O(m*n)$

Note: n indicates the number of users. m stands for number of tier devices.

In the comparison of the three common correlation analysis methods, Kendall coefficient has a significant advantage in distinguishing the membership relationship between the user and the table box in the distribution area. Because users under the same branch box have highly similar characteristics when they distinguish their own table boxes. Figure 9

shows the performance of Z3 and Y2, Y7, Y10, Y14 and Y17 in three correlation analysis methods.

Figure 9 Comparison of three correlation analyses (see online version for colours)



The method presented in this paper shows nearly perfect accuracy in standard scenarios. When the data is missing, linear interpolation can en-sure the recognition accuracy above 98% and maintain high robustness. The main limitations focus on extreme load unbalance, ultra-large scale computing efficiency and dynamic topological adaptability. In the future, load balancing pre-processing, distributed optimisation and incremental learning will further improve the universality of the method.

Compared with the most advanced signal injection topology identification methods, the method proposed in this paper has obvious advantages in accuracy, robustness, cost and scalability, as shown in Table 3.

Table 3 Analysis of this method and signal injection method

	Topology identification method	*Textual method*
Accuracy	High (static scene)	High (dynamic scene)
Robustness	Low (noise sensitive)	High (anti-noise, missing data)
Cost deployment	Extremely high	Very low
Topological adaptability	Need to re-inject the signal, the real-time is poor	Support online update, incremental clustering

Through the comparison, the proposed method solves the inherent defects of the signal injection method at a very low cost while maintaining high accuracy and provides a revolutionary tool for the topology management of low-voltage distribution area.

5 Conclusions

The paper proposes a data-driven full hierarchical topology identification method for low-voltage distribution station areas. The method can simultaneously discern the phase relation and hierarchical structure among users within a single station. This approach exhibits the following characteristics:

1 The issue of uncertain initial clustering centre and cluster number is addressed by optimising the K-means algorithm with the density peak method. The DPK-means algorithm is employed to cluster voltage data from metre boxes and users to identify the user phase.

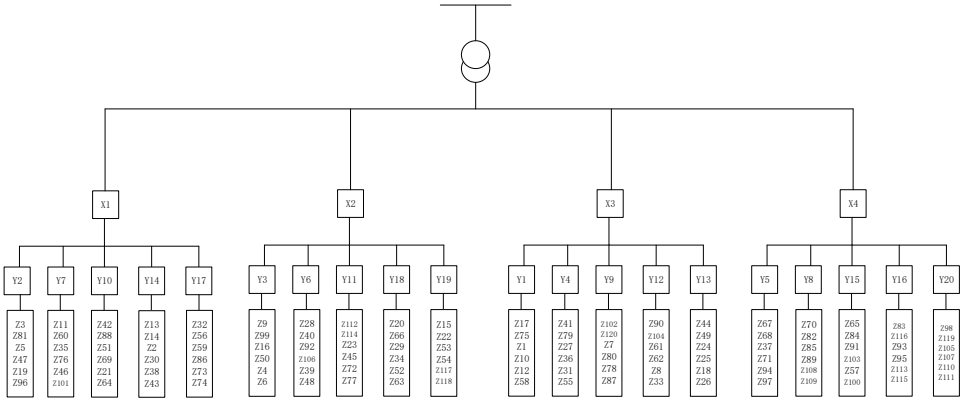
2 The utilisation of normalised Pearson correlation coefficient can effectively address the challenge posed by the high similarity in voltage among users within the low-voltage distribution station area, thereby facilitating their affiliation identification.

The data-driven full-level topology recognition method proposed in this paper shows high precision and robustness in low-voltage distribution area. But there is still some room for scalability and optimisation of application scenarios. The core directions of future research are as follows:

1 To find the optimal global topological management requirements of ultra-large scale distribution networks suitable for megacities or provincial power grids.

2 The high proportion of distributed energy causes the voltage fluctuation pattern to be complicated. In the future, the topological identification accuracy of distributed energy grid should be improved.

Figure 10 Topology of a district in Shanxi Province, China



Declarations

All authors declare that they have no conflicts of interest. The datasets used and analysed during the current study available from the corresponding author on reasonable request.

This work was supported by Research and Application of Fault Detection and Location Technology in Low voltage Distribution Network Project of Shanxi Electric Power Company electric power Science Research Institute. Under Grant 52053023000U.

References

Anderson, O. and Yu, N. (2021) 'Detect and identify topology change in power distribution systems using graph signal processing', in *2021 IEEE PES Innovative Smart Grid Technologies Europe* (*ISGT Europe*), pp.1–6, IEEE.

Angelopoulos, A., Bates, S., Malik, J. and Jordan, M.I. (2020) *Uncertainty Sets for Image Classifiers Using Conformal Prediction*, arxiv preprint arxiv:2009.14193.

Curiac, C.D. and Micea, M. (2022) 'Evaluating research trends using key term occurrences and multivariate Mann-Kendall test', in *2022 International Symposium on Electronics and Telecommunications* (*ISETC*), pp.1–4, IEEE.

Fabbiani, E., Nahata, P., De Nicolao, G. and Ferrari-Trecate, G. (2021) 'Identification of AC distribution networks with recursive least squares and optimal design of experiment', *IEEE Transactions on Control Systems Technology*, Vol. 30, No. 4, pp.1750–1757.

Jafarian, M., Soroudi, A. and Keane, A. (2020) 'Distribution system topology identification for DER management systems using deep neural networks', in *2020 IEEE Power and Energy Society General Meeting* (*PESGM*), pp.1–5, IEEE.

Li, Y., Zhao, Q., Liang, D. et al. (2021) 'Low voltage distribution network impedance topology model based on measurement information', *Power System and Clean Energy*, Vol. 37, No. 4, pp.15–22+31.

Liu, S., Huang, C., Hou, S. et al. (2021) 'Relation recognition method of household variable based on DDTW distance and DBSYAN algorithm', *Automation of Electric Power Systems*, Vol. 45, No. 18, pp.71–77.

Ma, L., Wu, L., Liu, N. and Pei, W. (2022) 'A two-step approach for multi-topology identification and parameter estimation of power distribution networks', *CSEE Journal of Power and Energy Systems*, November 2024, Vol. 10, No. 6, pp.2446–2456.

Martins, L.S., Martins, J.F., Pires, V.F. and Alegria, C.M. (2002) 'The application of neural networks and Clarke-Concordia transformation in fault location on distribution power systems', in *IEEE/PES Transmission and Distribution Conference and Exhibition*, Vol. 3, pp.2091–2095, IEEE.

Mittal, S. and Verma, A. (2023) 'Topology identification of low voltage distribution networks using smart meter data', in *2023 International Conference on Power, Instrumentation, Control and Computing* (*PICC*), pp.1–6, IEEE.

Niu, W., Wang, Z. and Liu, X. (2022) 'Distribution network topological relationship automatic recognition based on power line wideband carrier', *Electronic Technology and Software Engineering*, Vol. 2022, No. 17, pp.144–147.

Padullaparti, H., Veda, S., Wang, J., Symko-Davies, M. and Bialek, T. (2022) 'Phase identification in real distribution networks with high PV penetration using advanced metering infrastructure data', in *2022 IEEE Power and Energy Society General Meeting* (*PESGM*), pp.1–5, IEEE.

Pappu, S.J., Bhatt, N., Pasumarthy, R. and Rajeswaran, A. (2017) 'Identifying topology of low voltage distribution networks based on smart meter data', *IEEE Transactions on Smart Grid*, Vol. 9, No. 5, pp.5113–5122.

Pengwah, A.B., Fang, L., Razzaghi, R. and Andrew, L.L. (2021) 'Topology identification of radial distribution networks using smart meter data', *IEEE Systems Journal*, Vol. 16, No. 4, pp.5708–5719.

Poudel, S., Ramachandran, T., Veeramany, A., Francis, C. and Reiman, A.P. (2023) 'Topology identification using graph theory informed state estimation-based model selection for power distribution systems', *IEEE Transactions on Industrial Informatics*, Vol. 20, No. 3, pp.3563–3573.

Rath, A. and Srungavarapu, G. (2021) 'New model predictive and algorithm DPC based shunt active power filters (SAPFs)', in 2021 *1st International Conference on Power Electronics and Energy* (*ICPEE*), pp.1–6, IEEE.

Srinivas, V.L. and Wu, J. (2022) 'Topology and parameter identification of distribution network using smart meter and μPMU measurements', *IEEE Transactions on Instrumentation and Measurement*, Vol. 71, No. 2, pp.1–14.

Wahyuningrum, T., Khomsah, S., Suyanto, S., Meliana, S., Yunanto, P.E. and Al Maki, W.F. (2021) 'Improving clustering method performance using K-means, mini batch K-means, BIRCH and spectral', in *2021 4th International Seminar on Research of Information Technology and Intelligent Systems* (*ISRITI*), pp.206–210, IEEE.

Zelenskii, E.G., Tuchina, D.S., Kononov, Y.G. and Kozhevnikov, V.M. (2022) 'Mobile laboratory for identification of radial distribution network topology', in *2022 IEEE International Multi-Conference on Engineering, Computer and Information Sciences* (*SIBIRCON*), pp.2110–2115, IEEE.

Zhuravlyova, M. and Vedishchev, V. (2021) 'Application of Kendall's W coefficient to identify groups of statistically related variables', in *2021 3rd International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency* (*SUMMA*), pp.763–768, IEEE.