A Statistical Approach to Estimate Imbalance-Induced Energy Losses for Data-Scarce Low Voltage Networks

Lurui Fang, *Student Member, IEEE*, Kang Ma, *Member, IEEE*, Ran Li, *Member, IEEE*, and Zhaoyu Wang, *Member, IEEE*, Heng Shi

***Abstract*—Phase imbalance in the UK and European low voltage (415V, LV) distribution networks causes additional energy losses. A key barrier against understanding the imbalance-induced energy losses is the absence of high-resolution time-series data for LV networks. It remains a challenge to estimate imbalance-induced energy losses in LV networks that only have the yearly average currents of the three phases. To address this insufficient data challenge, this paper proposes a new customized statistical approach, named as the CCRE (Clustering, Classification, and Range Estimation) approach. It finds a match between the network with only the yearly average phase currents (the data-scarce network) and a cluster of networks with full-time series of phase current data (data-rich networks). Then CCRE performs a range estimation of the imbalance-induced energy loss for the cluster of data-rich networks that resemble a data-scarce network. The Chebyshev's inequality is applied to narrow down this range, which also represents the confidence interval of the imbalance-induced energy loss for the data-scarce network. Case studies reveal that, given such few data from the data-scarce networks, more than 80% of these networks are classified to the correct clusters and the confidence of the imbalance-induced energy loss estimation is 89%.**

***Index Terms*— energy loss, low voltage, phase imbalance, power distribution, three-phase power**

I. INTRODUCTION

IMBALANCE-induced energy losses in the UK and European low voltage (415, LV) distribution networks account for up to 35% of the energy losses on distribution wires [1]. This is mainly due to the significant phase imbalance in the UK's LV networks [2], [3], [4]. Data from Western Power Distribution (a UK distribution network operator) show that over 50% of their LV networks have the peak current of the "heaviest" phase exceeding that of the "lightest" phase by more than 50%, e.g. it is common to have a peak current of 300 A on one phase and 150 A on another phase, causing the phase residual current to be comparable to or even larger than phase currents [5]. The phase residual current then causes an imbalance-induced energy loss. Imbalance-induced energy losses are also widespread in distribution networks in other countries [6], [7]. Therefore, understanding imbalance-induced energy losses are important for distribution network operators (DNOs) to evaluate the total cost of phase imbalance and the potential benefit of phase balancing [8], [9].

There exist a number of references that focus on imbalance-induced energy losses. Reference [10] calculates the energy loss on the neutral wire of overhead lines in the distribution network, using Carson's equations to model the lines. Reference [11] calculates neutral energy losses, based on the ratio between the equivalent neutral line resistance and line resistance of a transposed three-phase line. Reference [12] calculates the neutral energy loss caused by non-linear three-phase loads. Reference [13] calculates the neutral energy loss in medium-voltage distribution networks due to load imbalance. Reference [14], [15] calculates the energy losses in distribution networks, including energy losses on both the phases and the neutral wire.

The above references all require networks to have high-resolution time series data (e.g., data collected every 15 minutes or of a comparable resolution) or load curves. However, only a small portion of LV networks, the data-rich networks, have high-resolution time-series data, whereas the majority of LV networks only have data collected once a year, i.e., they are data-scarce networks. Therefore, a major challenge to understanding imbalance-induced energy losses is the lack of data. Existing imbalance-induced energy loss estimation methods are not applicable to data-scarce networks.

This paper makes an original contribution by addressing this gap. To do this, we propose a new customized statistical approach named as CCRE, which consists of three stages: Clustering, Classification, and Range Estimation. This approach overcomes the insufficient data challenge by finding a cluster of data-rich networks whose features match the data-scarce network through clustering and classification, using only the yearly average currents of the three phases as the feature. Then this approach performs a range estimation of the imbalance-induced energy loss for the cluster of data-rich networks that resemble the data-scarce network. This range is narrowed down by applying the Chebyshev's inequality formula to counter the impact of outliers. This is the confidence interval of the imbalance-induced energy loss for the data-

L. Fang, K. Ma, R. Li, and H, Shi are with the University of Bath, Bath, U. K. (correspondence author: K. Ma. E-mail: K.Ma@ bath.ac.uk).

Z. Wang is with Iowa State University, IA, 50010, USA. (Email: wzy@iastate.edu)

scarce network.

Because the yearly average phase currents are widely available data in LV networks, this research enables the DNO to estimate imbalance-induced energy losses on a mass scale across its business area, without the need to deploy high-resolution monitoring devices. This is economically appealing in terms of significant cost savings. According to [16], if all UK's 900,000 LV networks were to be made data-rich, the total cost of deploying and maintaining pervasive monitoring systems would be approximately two billion British pounds, which can be saved. The proposed method enables the DNO to evaluate a key cost of phase imbalance for the majority of the LV networks that are data-scarce, because imbalance-induced energy losses constitute a cost, which occurs year by year until the three phases are rebalanced. This cost is a key input for the cost-benefit analysis of phase balancing solutions.

The rest of this paper is organized as follows: Section II presents the clustering and classification methodology. Section III presents the range estimation of the imbalance-induced energy loss. Section IV performs case studies. Section V concludes this paper.

II. METHODOLOGY

To calculate the imbalance-induced energy loss, two variables, phase residual currents and the impedance data, are required as inputs. However, these two variables are not normally available in LV networks, which have the yearly average current for each phase [17]. On the other hand, we have time series phase current data collected from 800 data-rich LV networks throughout a year and these networks cover a wide range of regions (urban, suburban, and rural areas). Therefore, this paper proposes a CCRE approach to estimate the phase residual currents for any data-scarce LV network, based on the available data from the 800 networks. The CCRE approach consists of three stages: clustering, classification, and range estimation. The reason for having the clustering stage is to extract representative characteristics of the phase residual currents (expressed in the form of cumulative density functions) from the 800 data-rich networks, thus narrowing down the 800 data-rich networks into a few representative classes. Then, the purpose of the classification stage is to find the best match between the data-scarce network and one of the representative classes. Finally, the reason for applying the range estimation is to account for the uncertainty in the imbalance-induced energy loss estimation. Multiple scenarios on the impedance are considered. The overall flowchart of the CCRE approach is presented in Fig. 1. It should be noted that all input current data are magnitudes only.

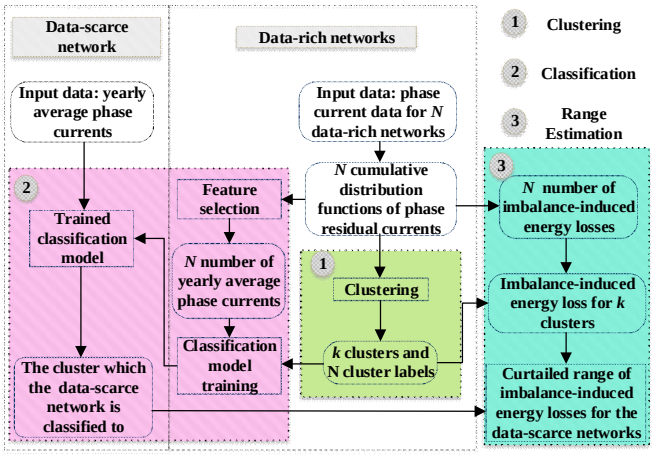

Fig. 1 Overview of the CCRE approach

A. Data pre-processing

On one hand, we have time-series phase current data collected from N (in this case, N = 800 but the methodology supports a generic dataset) data-rich LV substations throughout a year at an interval of 15 minutes. These substations, within Western Power Distribution (a UK DNO)'s business area, cover a good mix of geographical areas (urban, suburban, and rural) and customer composition (domestic, commercial, and industrial). For example, Cardiff city center is selected as an urban area with a large number of commercial customers; Monmouthshire is selected as a representative rural area [5]. These data are the deliverables of the project "Low Voltage Network Templates". Reference [5] presents a detailed description of these data and this project. On the other hand, the majority of the UK's LV networks are data-scarce, where time-series data are unavailable. For data-scarce networks, the protection systems (e.g. Schneider Sepam series 20) in the substations record the yearly average currents of the three phases [17] – these are the very few available data for data-scarce networks.

The phase residual current $I_{prc}(t)$ is a key variable. For the 800 data-rich LV networks with time series phase current data, the time series phase residual current is given by

$$I_{prc}(t) = [I_a^2(t) + I_b^2(t) + I_c^2(t) - I_a(t)I_b(t) - I_b(t)I_c(t) - I_a(t)I_c(t)]^{1/2} \quad (1)$$

where $I_a(t), I_b(t), I_c(t)$ denote the currents on phase a, b, and c at time t, respectively.

In reality, the time series of phase residual currents for different LV networks have different lengths because there are minor missing data, e.g. the time series for Network 1 has 35,040-time points, whereas the time series for Network 2 only has 35,028 time points. This paper resolves this problem by transforming each time series of phase residual currents into a cumulative distribution function (CDF). This is suitable because this paper is only concerned about the imbalance-induced energy loss over a year (this is the basis for calculating the annual cost of the imbalance-induced energy loss), rather than the power loss at any specific time point.

For each data-rich network, the time series of phase residual currents are transformed into a probability density function of

the phase residual currents through kernel density estimation (KDE) [18], as given by (2).

$$f(I_n) = \frac{1}{n \cdot h}\sum_{t=1}^{n} K\left(\frac{I_n - I_n(t)}{h}\right) \quad (2)$$

where I_n denotes the phase residual current; $I_n(t)$ is the phase residual current at time t; n denotes the sample size; h denotes the kernel bandwidth. In this paper, the kernel function K is chosen to be the Gaussian kernel [19], given by

$$K\left(\frac{I_n - I_n(t)}{h}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{I_n - I_n(t)}{h}\right)^2} \quad (3)$$

where the bandwidth $h = 1.06 \cdot \sigma \cdot n^{-\frac{1}{5}}$ [18]; σ denotes the standard deviation of the sample data; n denotes the sample size.

For each data-rich network, its probability density function of the phase residual currents is transformed into a CDF. Therefore, there are a total of 800 phase residual current CDFs for the 800 data-rich LV networks.

B. *Clustering*

Agglomerative hierarchical clustering and k-means clustering are applied to cluster these 800 phase residual current CDFs into k clusters. The reason why we use the agglomerative hierarchical clustering and k-means clustering is because they are commonly used classic clustering methods [20], [21]. The agglomerative hierarchical clustering method starts by taking each CDF as its own cluster; then it generates higher-level clusters by merging clusters with the least dissimilarity between each other until eventually achieving only one cluster [20]. This subsection presents three detailed aspects: 1) distance metrics; 2) the selection of the number of clusters, and 3) the evaluation of clustering results.

Both Euclidean distance (ED) [16] and Jensen-Shannon distance (JSD) [22] are applied to calculate the dissimilarity between any two CDFs.

1) *Determine the number of clusters*

In this paper, the number of clusters k is determined by a bi-objective optimization model. The optimization model aims to minimize the weighed sum of: 1) an overlap ratio; and 2) the relative within-cluster sum of squared distances. The optimization model is given by

$$\min_k \; C \cdot r(k) + s(k) \quad \text{subject to} \;\; 2 \le k \le k_{up};\ k \text{ is an integer} \quad (4)$$

where C is a weighting factor ($C > 0$); $r(k)$ is the overlap ratio defined in (5); $s(k)$, defined in (6), is the relative within-cluster sum of squared distances as a function of k; $k_{up} = \text{argmax}\, r(k)$. $0 \le r(k) < 1$ and $0 \le s(k) < 1$.

Now this paper defines the overlap ratio $r(k)$. Because this paper estimates the annual imbalance-induced energy loss which is proportional to the sum of data-rich network's squared phase residual currents over a year, the clustering results are considered "good" if different clusters are distinguishable from each other in terms of their distributions of the sums of squared phase residual currents over a year. In other words, each cluster shall have a distinct distribution of the sum of squared phase residual currents as compared to other clusters. To quantify such a distinctiveness, the overlap ratio is defined in (5).

$$r(k) = n_o(k)/N \quad (5)$$

where k denotes the number of clusters; $r(k)$ is the overlap ratio as a function of k; n_o is the number of data-rich networks that have the same sum of squared phase residual currents across different clusters (the shadow area as illustrated in Fig. 2). N denotes the total number of data-rich networks. An illustration of the overlap ratio is given in Fig. 2.

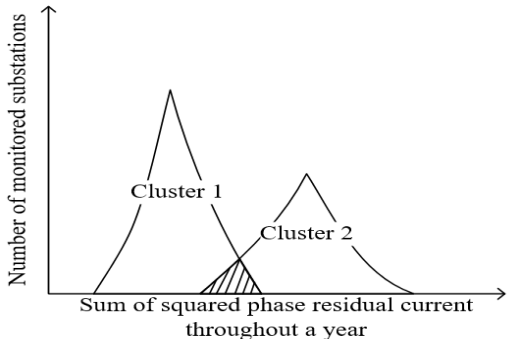

Fig. 2 The objective overlap area

The shadow area in Fig. 2, i.e., the overlap of the two clusters 1 and 2, represents n_o in (4) – this can be easily extended to k clusters. The overlap ratio $r(k)$ is the shadow area divided by the total area of all clusters. When k increases from 2 to the maximum number of clusters (800 in this case), $r(k)$ first increases then decreases to zero. Denote k_{up} as the k value when $r(k)$ reaches the maximum.

Now this paper defines the relative within-cluster sum of squared distances $s(k)$, as given by

$$s(k) = \frac{\sum_{j=1}^{k}\sum_{i\,(x_i \in \text{cluster } j)}(x_i - \bar{x}_j)^2}{\sum_i (x_i - \bar{x}_t)^2} \quad (6)$$

where x_i denotes the ith element in cluster j; $\bar{x}_j$ is the prototype of cluster j; $\bar{x}_t$ is the medoid of all elements.

2) *Evaluate clustering results*

After determining the number of clusters k, the agglomerative hierarchical clustering process is straightforward. The results show that the agglomerative hierarchical clustering with Euclidean distance yields the least overlap ratio, as compared to k-means with Euclidean distance, k-means with Jensen-Shannon distance, and agglomerative hierarchical clustering with Jensen-Shannon distance. The numerical results and detailed discussions are presented in section IV (case studies). Therefore, the agglomerative hierarchical clustering with Euclidean distance is chosen as the method for clustering the 800 phase residual current CDFs. The clustering output is a cluster label for each data-rich network, indicating which cluster this network belongs to. The medoid of each cluster is selected to be the prototype of this cluster [20].

C. *Classification*

Given the clustering outputs, the classification process consists of the following steps: 1) feature vectors (input data for classification) are determined for both the data-scarce and data-rich networks; 2) the feature vectors and cluster labels for the 800 data-rich networks are used to train the classification model by applying multiclass support vector machine (MSVM) and

kAdaBoost; MSVM and kAdaBoost then classify the data-scarce network to an existing cluster of data-rich networks. The classification results are validated by 10-fold cross-validation.

1) Determine feature vector

Data-scarce networks do not have time series data and they account for the majority of the UK's LV networks. They only have data collected once a year. According to [17], this paper suggests that the yearly average currents for three phases ($\bar{I}_a, \bar{I}_b$ *and* $\bar{I}_c$) be chosen as the known data for data-scarce networks: 1) DNOs can obtain them directly with current d evice in a low-cost fashion for millions of networks and these data do not require any deployment of high-resolution monitoring devices; 2) the features derived from these data allow for a relatively high classification accuracy.

Given the yearly average phase currents ($\bar{I}_a, \bar{I}_b$ *and* $\bar{I}_c$), this paper proposes a feature vector consisting of two features: the virtual average phase residual current value $\overline{I_{vprc}}$ and virtual average balanced current value $\overline{I_{vbc}}$. They can be readily calculated from the yearly average phase currents:

$$\overline{I_{vprc}} = \sqrt{\bar{I}_a^{\,2} + \bar{I}_b^{\,2} + \bar{I}_b^{\,2} - \bar{I}_a\bar{I}_b - \bar{I}_a\bar{I}_c - \bar{I}_b\bar{I}_c} \tag{7}$$

$$\overline{I_{vbc}} = ave(\bar{I}_a, \bar{I}_b, \bar{I}_c) \tag{8}$$

where $\bar{I}_a, \bar{I}_b$ *and* $\bar{I}_c$ denote the yearly average phase currents. Therefore, the feature vector $\mathbf{x}_i = [\overline{I_{vprc}}, \overline{I_{vbc}}]$ is available for the data-scarce network.

For data-rich networks, the above feature vector can be readily derived from the time series phase residual current data throughout a year. Therefore, each data-rich network has a cluster ID (this is an output from the clustering stage) as its label and a feature vector $\mathbf{x}_i = [\overline{I_{vprc}}, \overline{I_{vbc}}]$. Then, the feature vectors and cluster ID for all data-rich networks and the feature vector for the data-scarce network are used as the input data for the classification stage.

2) Classification

The classification is performed by applying two methods, kAdaBoost and MSVM. The reason for choosing MSVM (which uses the support vector machine as the base classifier) is because, by finding the largest margin to separate different classes, the performance of the support vector machine is widely recognized [23], [24]. kAdaBoost is chosen as a candidate because: 1) it reduces the bias of weak leaners by combining the weak learners into a strong learner and it is shown to be resistant against overfitting [25]; and 2) the Gaussian kernel transformation further improve the classification accuracy.

The kAdaBoost, i.e. the kernel-based Adaptive Boost model, is a combination of the kernel transformation and Adaptive Boost [25]. It consists of the following steps:

Firstly, a Gaussian kernel transformation is applied to transform the original feature vectors $\mathbf{x}_i$ for all networks i (both data-rich and data-scarce) into a high-dimensional feature space. Such a transformation improves the classification accuracy by up to 2%. The Gaussian kernel is given by [26]

$$K(\boldsymbol{x}_{ij}, \boldsymbol{x}_{ik}) = \exp\left(-\frac{\|\boldsymbol{x}_{ij} - \boldsymbol{x}_{ik}\|^2}{2\sigma^2}\right) \tag{9}$$

where x_{ij} and x_{ik} denote the j_{th} and k_{th} elements of network i's feature vector $\mathbf{x}_i$, respectively; σ^2 is the variance.

Secondly, the Adaboost.M2 model takes the transformed feature $K(\boldsymbol{x}_{ij}, \boldsymbol{x}_{ik})$ as the input. For Adaboost.M2, it is essentially a "boosting" method that combines a number of weak classification models ("weak models") into a strong classification model ("strong model") [27] . In each iteration of Adaboost.M2, a weak model performs classification with a relatively high error. Given the error from this weak model, the weight parameters of each training sample is updated, i.e. lower weights are assigned to the correctly classified training samples and higher weights are given to the wrongly classified training samples. When the iteration finishes, the strong model is built up as the combination of weak models and it yields an overall low classification error. The strong model is given by [22]:

$$H(x) = \operatorname{argmax} \sum_{t=1}^{T} h_t(x, y) \log \frac{1}{a_t} \tag{10}$$

where h_t is the weak model; a_t denotes the weight parameter. The algorithm and pseudocodes of AdaBoost.M2 are detailed in [25].

An illustration of Adaboost.M2 is given as follows (decision stumps are chosen as the weak models and there are 10 iterations) [28] :

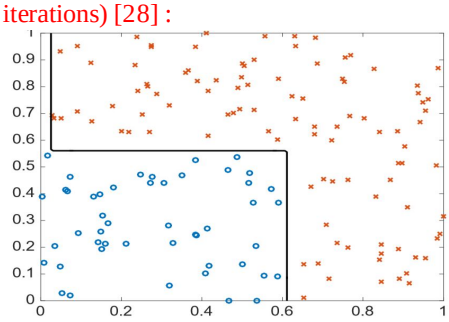

Fig. 3 Illustration of Adaboost.M2

In Fig. 3, the blue and red points represent two classes of data; the black line is built by 10 iterations of Adaboost. The results show that the two classes are successfully separated by the Adaboost model.

The MSVM is the multiclass support vector machine [29], [26]. The MSVM is essentially a one-versus-one framework that extends the support vector machine (a binary classifier) into a multiclass classifier [29]. The one-versus-one framework breaks the original multiclass classification problem down to $\frac{k(k-1)}{2}$ binary classification subproblems. For each binary classification subproblem, the support vector machine aims to find a separating hyperplane in the high-dimensional feature space (as a result of the Gaussian kernel transformation of the feature vectors) to separate the two classes with the maximum margin [24]. This is an optimization problem, as given by [30].

$$\begin{aligned} &\min_{\omega, b} \frac{1}{2}\|\boldsymbol{\omega}\|^2 + C\sum_{i=1}^{N_t} e_i \\ &\text{subject to } y_i(\boldsymbol{\omega}^T \cdot \varphi(\mathbf{x}_i) + b) \geq 1 - e_i \\ &e_i \geq 0 \end{aligned} \tag{11}$$

where $\boldsymbol{\omega}$ and b are the coefficient vector and the interception term, respectively; y_i is the label for training example i ($y_i \in \{-1, 1\}$); $\varphi(\mathbf{x}_i)$ is the transformed feature vector in the high-dimensional space for training example i; $C\sum_{i=1}^{N_t} e_i$ is the regularization term that reduces the generalization error, where C denotes the penalty coefficient; N_t denotes the total number of training examples; e_i represents the infringement an outlier causes.

The above binary support vector machine is extended into MSVM by using the one-versus-one framework [29]: each cluster is compared to each other cluster, where support vector machine (the binary classifier) is used to discriminate one cluster from another. This trains a total of $\frac{k(k-1)}{2}$ binary support vector machines. When given a data-scarce network, a voting is performed among the binary support vector machines and the cluster with the most number of votes wins.

The algorithm and pseudocodes of MSVM are presented in [23]. An illustration of the SVM classification is given in Fig. 4:

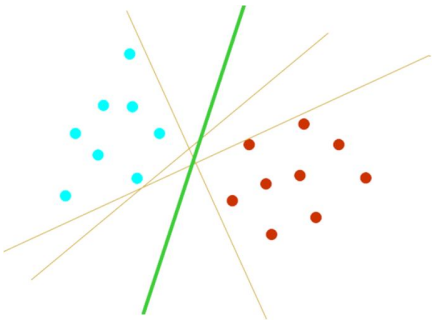

Fig. 4 Illustration of SVM

In Fig. 4, the blue and red points represent two classes. The results show that the two classes are successfully separated by the green line, i.e. the hyperplane derived by SVM with the largest separating margin, as compared with other separating hyperplanes.

The classification process is validated by 10-fold cross-validation, which is a very popular validation method [31] [32]. The 800 data-rich networks are divided into 10 groups of equal size. One of the ten groups of data-rich networks are retained as the validation samples (the "ground truth") to validate the classification; the other 9 groups are used as the training samples to train the classification model. The validation samples are treated as if their cluster IDs were unknown and are fed into the trained classification model. The model then outputs the cluster IDs for the validation samples. These clustering IDs are compared with the true known cluster IDs of the validation samples for validation. This process repeats until every group has served as the validation samples once. This process produces 10 classification accuracy results. The final classification accuracy is their average.

The classification results from the two methods are compared with each other in the case studies. Given the clustering and classification model trained and the data-scarce network, the output of the classification stage is the cluster which this network is classified to.

III. IMBALANCE-INDUCED ENERGY LOSS RANGE ESTIMATION

The classification stage in Section II – C classifies the data-scarce network into an established cluster derived in Section II – B. The maximum range of the imbalance-induced energy loss for this cluster is then derived. This range is then narrowed down to a confidence range by applying the Chebyshev's inequality formula. This confidence range is where the imbalance-induced energy loss of the data-scarce network falls at a predefined confidence level, as cross-validated in Section IV. Detailed steps are given below.

Firstly, the imbalance-induced energy losses for these data-rich networks are calculated for two different earthing systems, TN-C and TN-S. The TN-C earthing system is demonstrated in Fig. 5 [33] :

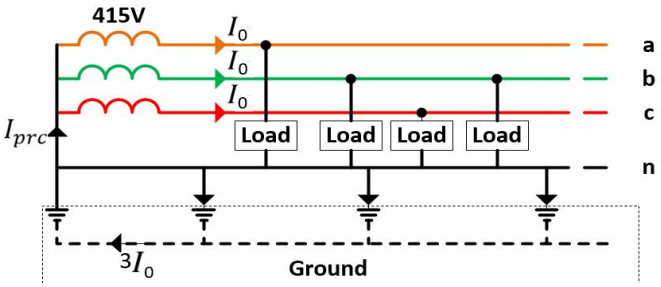

Fig. 5 The TN-C earthing system

For the TN-C earthing system, I_{prc} is the phase residual current that flows into the transformer neutral point from the ground [33]. The imbalance-induced energy loss is given by

$$E_{loss} = I_{prc}(t)^2 * R_g \quad (12)$$

where I_{prc} denotes the phase residual current; R_g is the equivalent ground resistance, which is 0.0953 (Ω/km) · Length (km).

The TN-S earthing system is shown in Fig. 6 [33]:

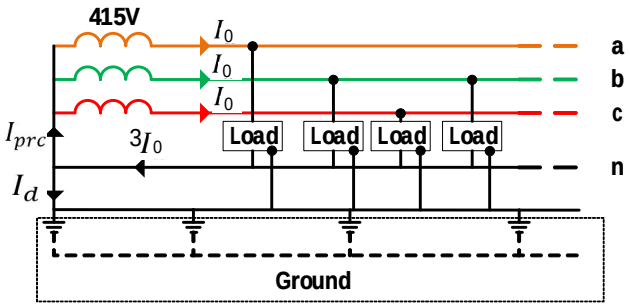

Fig. 6 The TN-S earthing system

For the TN-S earthing system, the protective wire and the neutral wire are separate conductors. When there is phase imbalance, the phase residual current, I_{prc}, flows into the transformer neutral point through the neutral conductor. Therefore, the imbalance-induced energy loss is given by

$$E_{loss} = I_{prc}(t)^2 * R_n \quad (13)$$

where I_{prc} denotes the phase residual current; R_n denotes the neutral wire resistance.

Secondly, given that the clustering stage in Section II – B has already clustered the 800 data-rich networks into N clusters, the

maximum range $[E_{lossmin}, E_{lossmax}]$ of the imbalance-induced energy loss for each cluster is derived, where $E_{lossmin}$ and $E_{lossmax}$ denote the minimum imbalance-induced energy loss and the maximum imbalance-induced energy loss, respectively.

The above maximum range is sensitive to outliers. To counter the impact of outliers, the maximum range is narrowed down to a confidence range. In industry, a common practice is to remove 1 – 2% of the observed data close to the range boundaries [34], assuming that the data follow a Gaussian distribution. The reason why we choose the Chebyshev's inequality formula for the range estimation is that, unlike other methods, it does not require that the data follow any particular classic distribution (e.g. Gaussian distribution). In this paper, the imbalance-induced energy loss results for any cluster of data-rich networks are not assumed to follow any particular classic distribution. Therefore, the Chebyshev's inequality formula is suitable in this case. This paper applies the Chebyshev's inequality formula [35] [36] to narrow down the range of the imbalance-induced energy loss. The Chebyshev's inequality formula states that the probability of a random variable falling beyond $k\sigma$ from its mean is less than $1/k^2$.

$$\text{Prob}(|x-\mu| \geq k\sigma) \leq 1/k^2 \quad (14)$$

where Prob($(|x-\mu| \geq k\sigma)$ denotes the probability that $|x-\mu| \geq k\sigma$; x is the random value of the imbalance-induced energy loss; μ denotes the expectation of the imbalance-induced energy loss; σ is the standard deviation of the imbalance-induced energy loss; k is the coefficient. Reference [37] suggests that the coefficient k be set as 3 to remove outliers, which means that the values falling in the interval $[\mu - 3\sigma, \mu + 3\sigma]$ has a confidence level of 89%.

The confidence range corresponds to removing 11% of data from the original cluster by the Chebyshev's inequality method. An illustration of the "tail cutting" by the Chebyshev's inequality method is shown in Fig. 7.

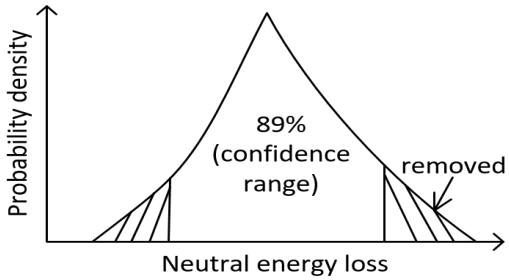

Fig. 7 The distribution of example imbalance-induced energy loss for cluster *i*

In cluster *i*, the distance between the imbalance-induced energy loss of each data-rich network and the mean imbalance-induced energy loss of the cluster is calculated. Then, 11% of the data-rich networks in cluster *i* with larger distances than the rest are removed. The resulting range of the imbalance-induced energy loss is the 89% confidence range of imbalance-induced energy loss for cluster *i*. This effectively counters the impact of outliers and it is applicable to generic probability distributions.

The choice of the 89% confidence level for the range estimation is validated by applying a 10-fold cross-validation. For each cluster of n data-rich networks, n number of imbalance-induced energy loss values are randomly divided into 10 groups of equal size. One of the ten groups of data-rich networks is retained as the validation group, the other 9 groups form a large training group to build a distribution of the imbalance-induced energy loss values. This distribution is narrowed down to the 89% confidence range by applying the Chebyshev's inequality formula. Then, the percentage of the validation samples (the imbalance-induced energy loss values within the validation group) that fall within the distribution is calculated. This process repeats until every group has served as the validation group once. This process outputs 10 values, i.e. the percentages of the validation samples falling within the distribution. These 10 values are averaged and it is found that the average value is close to 89%. In this way, the choice of the 89% confidence level is validated.

The resulting estimation error of the imbalance-induced energy loss is given by

$$error = \frac{abs(AL - EML)}{AL} \quad (15)$$

where AL denotes the actual imbalance-induced energy loss (IIBL) of the LV networks; EML is the mean value of the estimated range of the imbalance-induced energy loss.

IV. CASE STUDY

This section presents the numerical results from applying the methodology in Section II and III. The clustering and classification results are given in Sections IV – A and B, respectively. The imbalance-induced energy losses are calculated in Section IV – C. A discussion is presented in Section IV – D.

A. *Clustering*

The first step of clustering is to determine the number of clusters by solving the bi-objective optimization problem in (5). TABLE I presents the overlap ratio $r(k)$ for different numbers of clusters k.

TABLE I
OBJECTIVE OVERLAP RATIO COMPARISON

Number of clusters	$r(k)$ under the ED metric	$r(k)$ under the JSD metric
6	3.2%	9.8%
7	3.2%	9.8%
8	3.45%	10.1%

In TABLE I, $r(8) > r(7) = r(6)$. $k = 7$ is preferred over $k = 6$ because the former corresponds to a lower sum of within-cluster errors. Therefore, the number of clusters k is chosen to be 7 for both JSD and ED metric.

Given the number of clusters $k = 7$, the second step is to perform the clustering process using both k-means and hierarchical clustering methods, based on JSD and ED distance metrics. The results are presented in TABLE II for comparison.

TABLE II
CLUSTERING METHOD COMPARISON

		$r(k)$	Hierarchical Cophenet
Hierarchical	JSD	9.8%	0.7733

clustering	ED	3.2%	0.7845
K-means	JSD	22.%	
clustering	ED	10.3%	

In TABLE II, the Hierarchical cophenet denotes the cophenet correlation coefficient for the Hierarchical cluster tree, indicating how faithfully the tree represents the dissimilarities among observations (the larger the better). Hierarchical clustering with the ED distance metric yields the lowest overlap ratio and a higher cophenet – this combination is therefore chosen for clustering.

Fig. 8 and Fig. 9 visualize how distinguishable the seven clusters are under: 1) hierarchical clustering with ED metric; 2) hierarchical clustering with JSD metric; 3) k-means with ED metric; and 4) k-means with JSD metric.

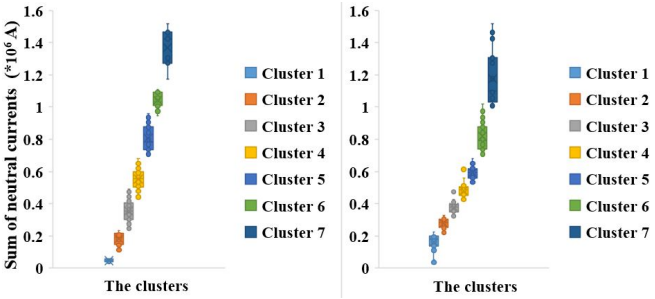

Fig. 8 Hierarchical (left) and K-means (right) clustering results with ED metric

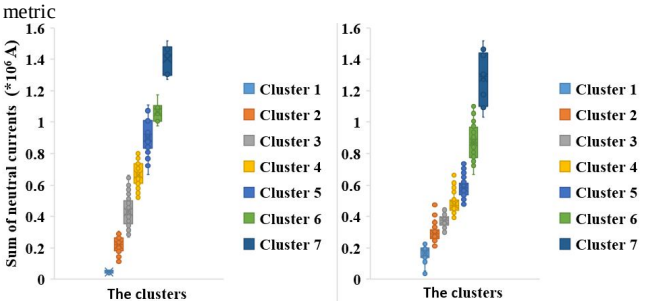

Fig. 9 Hierarchical (left) and K-means (right) results with JSD metric

In these diagrams, each cluster is resembled as a bar. Fig. 8 and Fig. 9 show that hierarchical clustering with the ED distance metric yields the most distinguishable seven clusters as compared to other methods.

The phase residual current CDFs of the data-rich networks within each cluster are plotted as a heat map in Fig. 10.

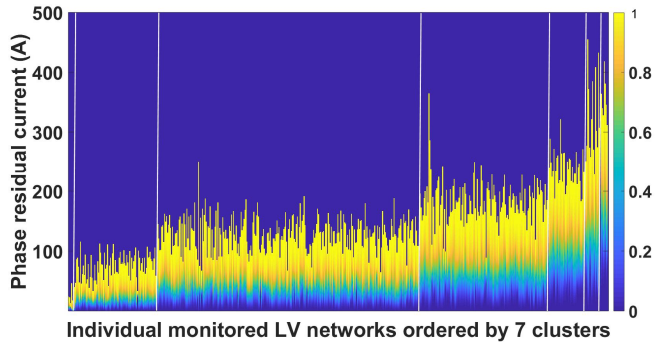

Fig. 10 The heat map of the squared phase residual current CDFs of the data-rich networks within each cluster

In Fig. 10, the diagram is separated into seven intervals by six vertical white lines, where each interval corresponds to a cluster (from Cluster 1 in the left to Cluster 7 in the right). Each blue-yellow vertical line represents the phase residual current CDF of a data-rich network belonging to the cluster. Each red vertical line represents each cluster's prototype. This figure demonstrates that each cluster has its own phase residual current CDF tendency, which is distinctive from other clusters. In addition, Cluster 1 accounts for 1.09% of the data-rich networks in this study; Clusters 2 – 7 account for 15.25%, 49%, 23.96%, 6.72%, 2.72%, and 1.27% of the data-rich networks, respectively.

B. *Classification*

According to Section II – C, the virtual average balanced current and virtual average phase residual current are the features used for classification in this sub-section. This feature is derived from yearly average currents of three phases ($\bar{I}_a, \bar{I}_b$ *and* $\bar{I}_c$), recorded once a year by a relay protection metering function. The distribution of the features for each cluster is plotted in Fig. 11.

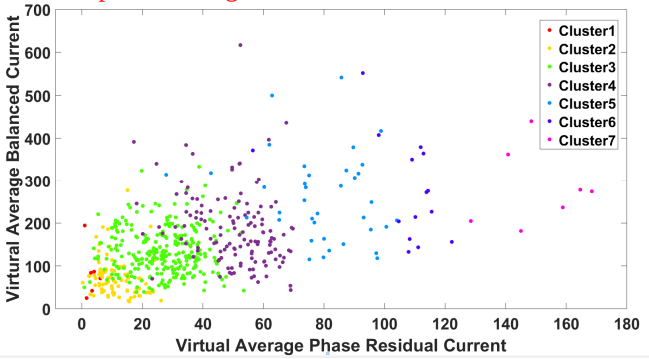

Fig. 11 Data-rich networks' feature distribution

Fig. 11 shows that the features for different clusters overlap to a large extent. This overlap reflects the data scarcity, i.e., the available feature is rather limited.

From case studies, we find that the Gaussian-kernel-based MSVM and kAdaBoost achieve higher classification accuracies than alternative classification methods such as k-Nearest Neighbours (KNN) and decision tree. The comparison of the classification accuracies is presented in Fig. 12.

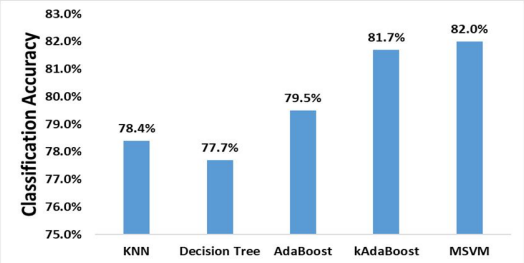

Fig. 12 The classification results comparison of different methods

From Fig. 12, the MSVM achieves the highest classification accuracy of 82%, followed by kAdaBoost which achieves a classification accuracy of 81.7% and adaptive boost (AdaBoost) which achieves 79.5% accuracy. KNN and decision tree achieve 78.4% and 77.7% accuracies, respectively. In comparison, a blind guess would give an accuracy of only 14.29%.

The confusion matrices for the classification results by MSVM and kAdaBoost are presented in Fig. 13.

Fig. 13 Confusion matrices for the MSVM (left) and kAdaBoost (right) methods

The confusion matrices in Fig. 13 demonstrate the classification accuracies in details. For instance, for the MSVM classification, column two shows that the data-scarce network which should be classified into Cluster 2 has 5% probability of being misclassified into Cluster 1, 22% probability of being misclassified into Cluster 3.

Both classification methods require only virtual average balanced current and virtual average phase residual current, derived from the yearly average currents of three phases ($\bar{I}_a, \bar{I}_b$ *and* $\bar{I}_c$), as the feature from data-scarce LV networks. This means it can be implemented in a cost-effective manner using existing devices only.

For example, a data-scarce network has the yearly average phase currents $[\bar{I}_a, \bar{I}_b, \bar{I}_c] = [219.1\text{A}, 182.4\text{A}, 224.1\text{A}]$. These data are transformed into a feature vector $\mathbf{x}_i = [\overline{I_{vbc}}, \overline{I_{vprc}}] = [208.5\text{A}, 39.4\text{A}]$. Given this feature vector, this data-scarce network is classified into Cluster 4 by applying either MSVM or kAdaBoost.

C. *Imbalance-induced energy losses estimation*

The resistance of the path on which the phase residual current flows is affected by many factors, including the length of the path, the resistivity of the cables and the ground, ambient condition, and the topology, etc. To account for the complicated nature, this paper considers multiple scenarios on the resistance and estimates the imbalance-induced energy losses for these scenarios. According to [38], the length of the UK's LV networks normally ranges from 0.9 km to 2.1 km; the resistivity of the ground is 0.0953 Ω/km; the resistivity of the neutral conductor ranges from 0.168 Ω /km to 0.320 Ω /km. Therefore, for TN-C earthing system, the ground resistance R_g varies from 0.0858 Ω to 0.2001 Ω; for TN-S earthing system, the neutral conductor resistance R_n varies from 0.1512 Ω to 0.6720 Ω;

For the TN-C earthing system, the confidence range of the imbalance-induced energy losses for each cluster is plotted in Fig. 14:

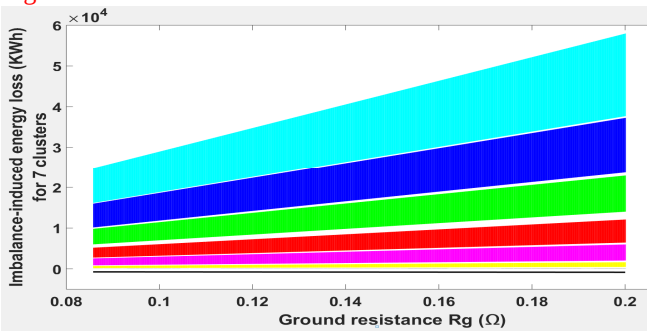

Fig. 14 The confidence range of the imbalance-induced energy losses of TN-C earthing system for the clusters

For example, when the ground resistance is 0.143 Ω (a length of 1.5 km, which is the average length of the UK's LV networks), the imbalance-induced energy loss for each cluster is given in Fig. 15:

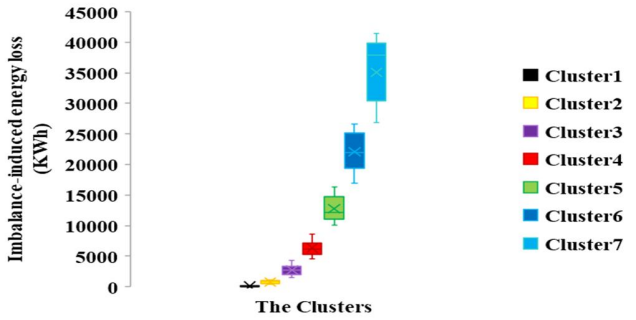

Fig. 15 The example of confidence range of the imbalance-induced energy losses of TN-C earthing system for the 7 clusters

For Cluster 1, the confidence range of the imbalance-induced energy losses is [53.5 kWh, 76.4 kWh] per year. The confidence ranges of the imbalance-induced energy losses for Clusters 2 – 7 are [327.7 kWh, 1162.5 kWh], [1456.9 kWh, 4270.8 kWh], [4601.2 kWh, 8638 kWh], [10005 kWh, 16345 kWh], [16904 kWh, 26615 kWh], and [26914 kWh, 41405 kWh] per year, respectively.

Given an estimation of 900,000 networks throughout the UK and an average electricity price of £ 0.18/kWh, the phase imbalance situation causes 3.01×10^6 to 6.02×10^6 MWh of imbalance-induced energy losses each year, worth £451.2m to £903.0m per annum.

For TN-S earthing system, the neutral conductor resistance R_n varies from 0.1512 Ω to 0.6720 Ω. The confidence range of the imbalance-induced energy losses for each cluster is plotted in Fig. 16:

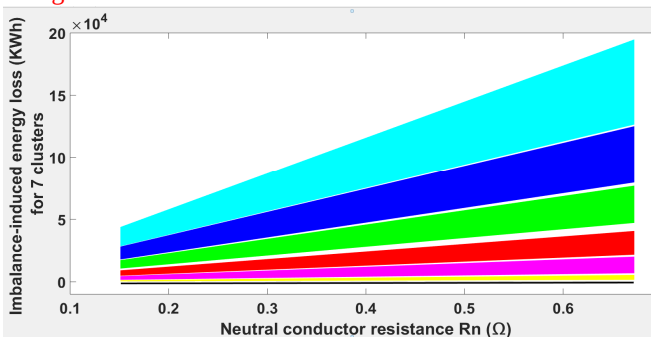

Fig. 16 The confidence range of the imbalance-induced energy losses of TN-S earthing system for the clusters

If the neutral conductor resistance is 0.252 Ω (with a length of 1.5 km and a resistivity of 0.163 Ω/km), the imbalance-induced energy loss for each cluster is presented in Fig. 17.

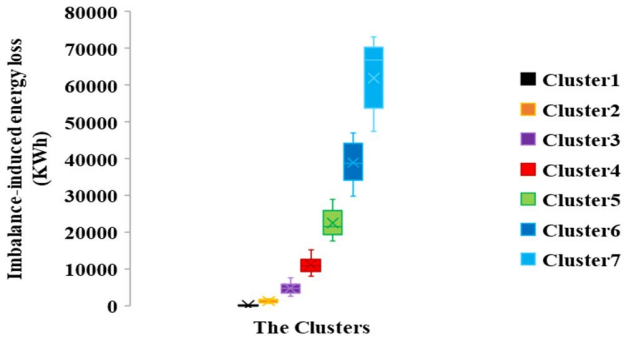

Fig. 17 The example of confidence range of the imbalance-induced energy losses of TN-S earthing system for the 7 clusters

For Cluster 1, the confidence range of the imbalance-induced energy losses is [94.3 kWh, 134.8 kWh] per year. The confidence ranges of the imbalance-induced energy losses for Clusters 2 – 7 are [577.8 kWh, 2050.1 kWh], [2569 kWh, 7531 kWh], [8114 kWh, 15233 kWh], [17644 kWh, 28824 kWh], [29809 kWh, 46934 kWh], and [47461 kWh, 73016 kWh] per year, respectively.

Given an estimation of 900,000 networks throughout the UK and an average electricity price of £ 0.18/kWh, the phase imbalance situation causes 5.3×10^6 to 1.06×10^7 MWh of imbalance-induced energy losses each year, worth £795.3m to £1,592m per annum.

This paper applies a 10-fold cross-validation to validate the confidence range of the annual imbalance-induced energy loss. The results are shown in Fig. 18.

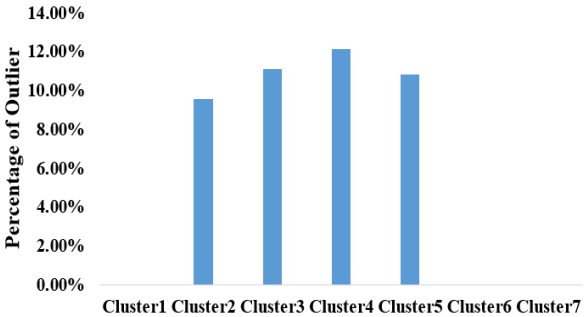

Fig. 18 The percentage of the outlier given by 10 folds cross-validation for the 7 clusters

From Fig. 18, cross-validation results show that 9% of the data-rich networks that belong to Cluster 2 fall beyond the confidence range of Cluster 2; 11%, 12%, and 11% of the data-rich networks that belong to Clusters 3, 4, and 5 falls beyond the respective confidence range of the cluster. Clusters 1, 6, and 7 have 5, 15, and 6 data-rich networks, respectively – too few networks that it is not suitable to remove any data from them. Therefore, the confidence ranges of Clusters 1, 6, and 7 are the maximum range of these clusters.

The same example in Section IV – B is used. Its estimated imbalance-induced energy loss is within a confidence range of [1,074.2 kWh, 2,131.4 kWh] per year, with a confidence level of 89%.

TABLE III presents a few examples showing the estimation errors:

TABLE III
EXAMPLF OF THE CCRE ESTIMATION ERROR

	I_{vprc} (A)	I_{vbc} (A)	Correct cluster	Actual IIBL (kWh)	Classified cluster	Estimated range of IIBL (kWh)	Estimating error
1	87.3	324	Cluster 5	24,520	Cluster6	29,809 – 46,934	61.15%
2	19.7	336	Cluster 3	3,096	Cluster 2	577 –2,050	55.91%
3	98.0	407	Cluster 6	38,350	Cluster 5	17,644—28,824	40.92%
4	17.8	38.1	Cluster 2	1,692	Cluster 2	577 –2,050	19.33%
5	59.9	177	Cluster 4	9,580	Cluster 4	8,114 – 15,233	18.29%
6	145	181	Cluster 7	54,386	Cluster 7	47,461 – 73,016	13.79%

In TABLE III, the first three examples are classified into the wrong clusters, resulting in substantial errors of more than 40%. The last three examples are classified to the correct clusters, resulting in errors of less than 20%.

D. *Discussion*

To estimate the imbalance-induced energy loss, the proposed CCRE approach only requires the yearly average phase currents as the feature from data-scarce networks. This feature can be easily obtained from existing LV networks. This renders the CCRE approach applicability to the majority of the UK's LV networks that are data-scarce, without the need for high-resolution monitoring devices on neutral wires.

In this paper, the 800 CDFs of the phase residual current I_{prc} are used as the input data for clustering. The energy loss is proportional to the square of the phase residual current, i.e. I_{prc}^2. The reason why the CDFs of I_{prc} are used as the input data instead of the CDFs of I_{prc}^2 is because the latter would increase the data dispersion from 0 – 300 to 0 – 90,000. This expands the range of the CDFs to a level too wide for clustering. Furthermore, the clustering results show that the former results in an overlap ratio as low as 3.2%, whereas the latter results in an overlap ratio of more than 20%. Therefore, the former is much better than the latter as the input data for clustering.

The CCRE approach is designed to be generic. To apply the CCRE approach to other countries, it would require the following two groups of input data for the country in question: 1) the time-series phase current data monitored throughout a year from at least hundreds of data-rich LV networks (these data are used as the training data); and 2) the yearly average phase currents for the data-scarce network (these limited data are called the feature). In general, the more representative the

training data are, the more accurate the estimated phase residual current for the data-scarce network is.

This paper considers phase residual current profiles and there is a fundamental difference between a load profile and a phase residual current profile. The former depends on the number of customers and types of customers, whereas the latter depends on how evenly (or unevenly) customers are allocated across the three phases. Because urban, suburban, sub-rural, and rural areas have very different customer densities and types of customers, their load profiles are different – the classification of load profiles into these four areas is justified. However, different types of areas may have the same degree of phase imbalance, i.e. customers in these areas are allocated in the same uneven fashion, thus resulting in similar phase residual current profiles. On the other hand, two networks in the same type of areas (e.g. urban) may have very different degrees of phase imbalance, resulting in vastly different phase residual current profiles. To substantiate the above statements, Fig. 19 shows that the yearly average values of the three-phase total load currents (which represent the loading levels) are hardly correlated with the yearly average phase residual currents. This reflects that the load profile and the phase residual current profile have fundamentally different characteristics almost independent from each other. Therefore, it would no longer be justified to classify the phase residual current profiles into urban, suburban, sub-rural, and rural classes. Instead, the clustering and classification are based on the real data set of the phase residual current profiles from 800 data-rich LV networks.

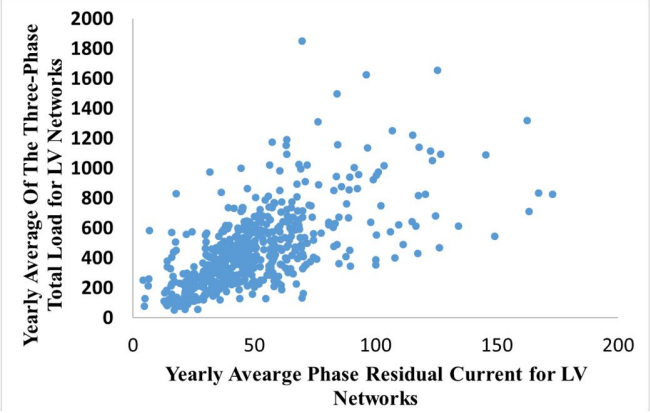

Fig. 19 The non-correlation between the yearly average values of the three-phase total load currents and the yearly average phase residual currents

There can be full current measurements from high-voltage (132 kV / 33 kV) and medium-voltage (33 kV / 11 kV) distribution substations as well as customer billing data. However, these measurements are not normally available from low-voltage (11 kV / 415 V, LV) substations downwards (inclusive), i.e. in LV networks. This is because of the cost-benefit issue: there are millions of LV networks in the UK. To obtain full current measurements (time-series data) for all of them, each LV network should have per-phase monitoring devices and a communication system and there should be a data center with the data processing capability sufficient to handle the vast amount of data. The total investment cost for these would be billions of British Pounds, but the benefit from these measurements does not justify the cost. This is the reason why distribution network operators (DNOs) choose not to monitor the vast majority of the UK's LV networks. Furthermore, even if smart meter data are available for all customers (which is not the case in the UK), which phase each customer is connected to is not normally known [39], [40]. Because of the above field limitations, state estimation cannot be performed for LV networks.

The load loss factor method is popular for calculating energy losses. To calculate the imbalance-induced energy loss by the load loss factor method, it requires the historical average phase residual current and maximum phase residual current as an input, which is not available for data-scarce LV networks. Furthermore, according to [41], the load loss factor is suggested to be updated each month to minimize the error of the estimation. For the data-scarce networks, the cost for engineers to update the load loss factors for 900,000 LV networks every month would be unimaginably high. Therefore, the load loss factor method is not applicable to data-scarce LV networks. In contrast, the proposed CCRE method specifically targets data-scarce networks, utilizing limited existing data without the need to deploy additional monitoring devices.

Increasing the available features will improve the accuracy of the classification. If the sum or average of the phase residual currents over a year is known for data-scarce networks. The CCRE approach then achieves an accuracy of 96.8%. This accuracy is much higher than if only the average phase residual currents are known. However, increasing features will pose more requirements for the monitoring of the LV networks, resulting in more costs. Gaussian-kernel-based MSVM gives a slightly higher classification accuracy (82%) than kAdaBoost (81.7%).

Phase imbalance causes two costs: 1) the imbalance-induced energy loss; and 2) the additional network investment cost. The value of this paper is to find out whether the 1st cost element is significant or not and how significant it is for both highly phase-imbalanced LV networks and not-so-imbalanced LV networks. Furthermore, this paper calculates the 1st cost for one year only. In reality, this cost occurs year by year until the three phases are fully balanced.

A cost-benefit analysis for any phase balancing solution requires that these two benefits be considered: 1) the imbalance-induced energy loss saving; and 2) the network investment cost saving. If the phase balancing completely eliminates the imbalance, then the imbalance-induced energy loss (which this paper estimates) would be eliminated and its cost saved, i.e. the 1st benefit. However, this paper does not cover the 2nd benefit, which is a different challenging topic. The future work will be performing a full cost-benefit analysis for phase balancing solutions considering the above two benefits together, the lack of data in LV networks, and the uncertainty associated with the phase balancing capability.

V. CONCLUSIONS

This paper addresses an unsolved problem faced by utility companies, i.e., estimating imbalance-induced energy losses for

data-scarce low voltage (415V, LV) networks with only the yearly average phase currents data.

The 800 LV data-rich networks with full time-series of phase currents data are clustered into 7 clusters, where each cluster represents networks of similar phase residual current profiles. Then, at the classification stage, cross-validation results show that nearly 82% of the data-scarce networks with only the yearly average phase currents data are classified to the correct clusters. The confidence interval of the imbalance-induced energy loss for the data-scarce network is derived with a confidence level of 89%.

APPENDIX

The phase residual current is the vector sum of the phase currents:

$$\overrightarrow{I_{prc}} = \overrightarrow{I_a} + \overrightarrow{I_b} + \overrightarrow{I_c} \quad (16)$$

In the absence of phasor measurements, it is assumed that the phase currents are 120° apart from each other. Therefore,

$$\begin{aligned}\overrightarrow{I_{prc}} &= I_a\cos 0^\circ + jI_a\sin 0^\circ + I_b\cos -120^\circ \\ &\quad + jI_b\sin -120^\circ + I_c\cos 120^\circ \\ &\quad + jI_c\sin 120^\circ \\ &= (I_a - \frac{1}{2}I_b - \frac{1}{2}I_c) + j(\frac{\sqrt{3}}{2}I_c - \frac{\sqrt{3}}{2}I_b) \quad (17)\end{aligned}$$

$$|I_{prc}| = \sqrt{\left(I_a - \frac{1}{2}I_b - \frac{1}{2}I_c\right)^2 + \left(\frac{\sqrt{3}}{2}I_c - \frac{\sqrt{3}}{2}I_b\right)^2}$$

$$|I_{prc}| = \sqrt{I_a^2 + I_b^2 + I_c^2 - I_aI_b - I_bI_c - I_aI_c}$$

where I_{prc} is the phase residual current; I_a, I_b and I_c denote the magnitudes of the phase currents.

REFERENCES

[1] J. D. Watson, N. R. Watson, and I. Lestas, "Optimized dispatch of energy storage systems in unbalanced distribution networks," *IEEE Transactions on Sustainable Energy*, vol. PP, no. 99, pp. 1-1, 2017.

[2] M. Chindris, A. Cziker, A. Miron, H. Balan, and A. Sudria, "Propagation of unbalance in electric power systems," in 2007 9th International Conference on Electrical Power Quality and Utilisation, 2007, pp. 1-5.

[3] S. Weckx, C. Gonzalez, and J. Driesen, "Reducing grid losses and voltage unbalance with PV inverters," in 2014 IEEE PES General Meeting | Conference & Exposition, 2014, pp. 1-5.

[4] J. Zhu, G. Bilbro, and C. Mo-Yuen, "Phase balancing using simulated annealing," *IEEE Transactions on Power Systems*, vol. 14, no. 4, pp. 1508-1513, 1999.

[5] "LV network templates for a low-carbon future," https://www.westernpower.co.uk/docs/Innovation/Closed-projects/Network-Templates/LVNT-Appendix-A-Knowledge-Management.aspx.

[6] T. Routtenberg, Y. Xie, R. M. Willett, and L. Tong, "PMU-Based Detection of Imbalance in Three-Phase Power Systems," *IEEE Transactions on Power Systems*, vol. 30, no. 4, pp. 1966-1976, 2015.

[7] W. Jiye, Z. Nan, and H. Hanyong, "Three-phase imbalance prediction: A hazard-based method," in 2016 IEEE International Conference on Power and Renewable Energy (ICPRE), 2016, pp. 226-231.

[8] K. Ma, R. Li, and F. Li, "Quantification of Additional Asset Reinforcement Cost From 3-Phase Imbalance " *IEEE Transactions on Power Systems*, vol. 31, no. 4, pp. 2885 - 2891 July, 2016.

[9] K. Ma, F. Li, and R. Aggarwal, "Quantification of Additional Reinforcement Cost Driven by Voltage Constraint Under Three-Phase Imbalance," *IEEE Transactions on Power Systems*, vol. 31, no. 6, pp. 5126 - 5134 18 January 2016.

[10] W. H. Kersting, "The computation of neutral and dirt currents and power losses," in IEEE PES Power Systems Conference and Exposition, 2004., 2004, pp. 213-218 vol.1.

[11] S. Pajic, and A. E. Emanuel, "Effect of Neutral Path Power Losses on the Apparent Power Definitions: A Preliminary Study," *IEEE Transactions on Power Delivery*, vol. 24, no. 2, pp. 517-523, 2009.

[12] J. C. Montano, P. Salmeron, and J. P. Thomas, "Analysis of power losses for instantaneous compensation of three-phase four-wire systems," *IEEE Transactions on Power Electronics*, vol. 20, no. 4, pp. 901-907, 2005.

[13] L. Ochoa, R. Ciric, A. Padilha-Feltrin, and G. Harrison, "Evaluation of distribution system losses due to load unbalance," in 15th Power Systems Computation Conference PSCC 2005, pp. 1-4.

[14] A. J. Urquhart, and M. Thomson, "Impacts of Demand Data Time Resolution on Estimates of Distribution System Energy Losses," *IEEE Transactions on Power Systems*, vol. 30, no. 3, pp. 1483-1491, 2015.

[15] J. C. López, J. F. Franco, and M. J. Rider, "Optimisation-based switch allocation to improve energy losses and service restoration in radial electrical distribution systems," *IET Generation, Transmission & Distribution*, vol. 10, no. 11, pp. 2792-2801, 2016.

[16] R. Li, C. Gu, F. Li, G. Shaddick, and M. Dale, "Development of Low Voltage Network Templates Part I: Substation Clustering and Classification," *IEEE Transactions on Power Systems*, vol. 30, no. 6, pp. 3036-3044, 2015.

[17] S. Electric. "Sepam™ Series 20 Protective Relays User's Manual," https://www.schneider-electric.com/resources/sites/SCHNEIDER_ELECTRIC/content/live/FAQS/221000/FA221290/en_US/63230-216-208C1_Sepam_Series_20_User_Manual.pdf.

[18] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*: Chapman and Hall/CRC Bernard. W. Silverman.

[19] S. S. Keerthi, and C.-J. Lin, "Asymptotic Behaviors of Support Vector Machines with Gaussian Kernel," *Neural Computation*, vol. 15, no. 7, pp. 1667-1689, 2003.

[20] S. Aghabozorgi, A. Seyed Shirkhorshidi, and T. Ying Wah, "Time-series clustering – A decade review," *Information Systems*, vol. 53, no. Supplement C, pp. 16-38, 2015.

[21] B. Mirkin, *Clustering for Data Mining A Data Recovery Approach*: Chapman & Hall/CRC, 2005.

[22] S. DeDeo, X. R. Hawkins, S. Klingenstein, and T. Hitchcock, "Bootstrap Methods for the Empirical Study of Decision-Making and Information Flows in Social Systems," *Entropy*, vol. 15, no. 6, 2013.

[23] A. Ng. "CS229 lecture notes: Support Vector Machines," http://cs229.stanford.edu/notes/cs229-notes3.pdf.

[24] J. Huang, Z. Jiang, L. Rylands, and M. Negnevitsky, "SVM-based PQ disturbance recognition system," *IET Generation, Transmission & Distribution*, vol. 12, no. 2, pp. 328-334, 2018.

[25] Y. Freund, and A. R. E. Schapire, "Experiments with a new boosting algorithm," in The Thirteenth International Conference on International Conference on Machine Learning, Bari, Italy, 1996.

[26] S. R. Gunn. "Support Vector Machines for Classification and Regression," http://m.svms.org/tutorials/Gunn1997.pdf.

[27] Y. Freund, and R. E. Schapire, "Experiments with a new boosting algorithm," in Proceedings of the Thirteenth International Conference on International Conference on Machine Learning, Bari, Italy, 1996, pp. 148-156.

[28] J. Duchi. "CS229 lecture note: Boosting algorithms and weak learning," http://cs229.stanford.edu/extra-notes/boosting.pdf.

[29] N. Mehra, and S. Gupta, "Survey on multiclass classification methods," *International Journal of Computer Science and Information Technologies*, vol. 4, pp. 572-576, 2013.

[30] L. Wang, D. Wang, and C. Hao, "Intelligent CFAR Detector Based on Support Vector Machine," *IEEE Access*, vol. 5, pp. 26965-26972, 2017.

[31] Q. Leming, P. S. Routh, and K. Kyungduk, "Wavelet deconvolution in a periodic setting using cross-validation," *IEEE Signal Processing Letters*, vol. 13, no. 4, pp. 232-235, 2006.

[32] B. Gu, V. S. Sheng, K. Y. Tay, W. Romano, and S. Li, "Cross Validation Through Two-Dimensional Solution Surface for Cost-Sensitive SVM," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1103-1121, 2017.

[33] A. A.Sallam, and O.P.Malik, *Electric Distribution System*, p.^pp. 88-95: John Wiley & Sons, Inc, 2011.

[34] S. J. Jeffrey, J. O. Carter, K. B. Moodie, and A. R. Beswick, "Using spatial interpolation to construct a comprehensive archive of Australian climate data," *Environmental Modelling & Software*, vol. 16, no. 4, pp. 309-330, 2001.

[35] J. G. Saw, C. K. Y. Mark, and T. C. Mo, "Chebyshev Inequality with Estimated Mean and Variance," *The American Statistician*, vol. 38, no. 2, pp. 130-132, 1984.
[36] D. Knuth, *The Art of Computer Programming: Fundamental Algorithms*: Reading, MA, USA: Addison-Wesley, 1997.
[37] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median," *Journal of Experimental Social Psychology*, vol. 49, no. 4, pp. 764-766, 2013/07/01/, 2013.
[38] G. Strbac, C. K. Gan, M. Aunedi, V. Stanojevic, P. Djapic, J. Dejvises, P. Mancarella, A. Hawkes, and D. Pudjianto. "Benefits of Advanced Smart Metering for Demand Response based Control of Distribution Networks ";
http://www.energynetworks.org/assets/files/electricity/futures/smart_meters/Smart_Metering_Benerfits_Summary_ENASEDGImperial_100409.pdf.
[39] M. Xu, R. Li, and F. Li, "Phase Identification With Incomplete Data," *IEEE Transactions on Smart Grid*, vol. 9, no. 4, pp. 2777-2785, 2018.
[40] H. Pezeshki, and P. J. Wolfs, "Consumer phase identification in a three phase unbalanced LV distribution network." pp. 1-7.
[41] K. Malmedal, and P. K. Sen, "A Better Understanding of Load and Loss Factors." pp. 1-6.