Cost-Benefit Analysis of Phase Balancing Solution for Data-scarce LV Networks by Cluster-Wise Gaussian Process Regression

Wangwei Kong, Kang Ma, *Member, IEEE*, Lurui Fang, *Student Member*, Renjie Wei, *Student Member*, and Furong Li, *Senior Member, IEEE*

***Abstract*—Phase imbalance widely exists in the UK's low voltage (415V, LV) distribution networks. The imbalances not only lead to insufficient use of LV network assets but also cause energy losses. They lead to hundreds of millions of British pounds each year in the UK. The cost-benefit analyses of phase balancing solutions remained an unresolved question for the majority of the LV networks. The main challenge is data-scarcity – these networks only have peak current and total energy consumption that are collected once a year. To perform a cost-benefit analysis of phase balancing for data-scarce LV networks, this paper develops a customized cluster-wise Gaussian process regression (CGPR) approach. The approach estimates the total cost of phase imbalance for any data-scarce LV network by extracting knowledge from a set of representative data-rich LV networks and extrapolating the knowledge to any data-scarce network. The imbalance-induced cost is then translated into the benefit from phase balancing and this is compared against the costs of phase balancing solutions, e.g. deploying phase balancers. The developed CGPR approach assists distribution network operators (DNOs) to evaluate the cost-benefit of phase balancing solutions for data-scarce networks without the need to invest in additional monitoring devices.**

***Index Terms*—cost-benefit analysis, Gaussian process regression, low voltage, phase balancing, phase imbalance, power distribution, three-phase system**

I. Introduction

THREE-PHASE imbalance exists in the majority (>70%) of UK's low voltage (415V, LV) networks [1] because of the uneven load allocation and random load behavior [2], [3], [4]. Phase imbalance causes additional energy losses [5], [6] and extra network investment costs [7], [8]. The additional energy losses include losses caused by phase residual currents and imbalance-induced transformer copper losses. The additional network investment costs include the additional investments on both LV transformers and network feeders, because phase imbalance wastes network capacity.

Phase balancing solutions include phase swapping [9], [10], demand-side management [11] and deploying phase balancers based on power electronics [12]. To justify any phase balancing solution, it is important to perform a cost-benefit analysis of the solution before making any investment decision. However, up until now, no published work performs a cost-benefit analysis of phase balancing solutions for the majority of the UK's LV networks that only have a minimal amount of data, e.g. data collected only once a year. These networks are referred to as data-scarce LV networks.

A number of references investigate imbalance-induced energy loss, which is a key input for the cost-benefit analysis. Reference [6] improves the backward-forward sweeping method to calculate the power loss in an imbalanced distribution network. Reference [13] introduces an imbalance factor to evaluate line losses under the imbalanced situation. Reference [14] performs a loss analysis based on a power flow algorithm for imbalanced radial distribution networks. References [15] and [16] perform power loss analysis for PV penetrated systems with full data of the network topology, load and generation. Reference [17] developed a statistical approach as a combination of clustering, classification and range estimation to estimate imbalance-induced energy losses for data-scarce networks.

This paper addresses a different problem from [17]: Reference [17] estimates the imbalance-induced energy loss only, whereas this paper performs a cost-benefit analysis of any phase balancing solution on data-scarce networks. This paper significantly extends [17] by considering a comprehensive range of imbalance-induced costs, including the additional reinforcement cost (ARC), the imbalance-induced energy losses caused by phase residual currents, and the imbalance-induced transformer copper losses. Furthermore, this paper develops a completely different methodology from [17]: Reference [17] develops a combined approach of clustering, classification and range estimation, whereas this paper develops a regression methodology tailored for the cost-benefit analysis of phase balancing solutions.

This paper addresses a real need for the UK industries: to identify, among a mass population of LV networks, a subset of networks that are worth phase balancing, i.e. where the benefit from phase balancing outweighs its cost [18], [19]. However, existing solutions require full data from distribution networks. There is a gap in performing cost-benefit analyses of phase balancing on data-scarce LV networks. This paper directly

W. Kong, K. Ma and F. Li are with the Electronic & Electrical Engineering Department, University of Bath, Bath, BA2 7AY, UK. Correspondence author: K. Ma; email: K.Ma@bath.ac.uk.

addresses the industrial need by bridging the gap. This paper for the first time performs a cost-benefit analysis of phase balancing for any data-scarce LV network. To this end, this paper develops a new cost-benefit analysis framework for phase balancing on data-scarce LV networks. The core of the framework is a customized cluster-wise Gaussian process regression (CGPR) approach, which accounts for a full range of imbalance-induced costs. The approach estimates the total cost of phase imbalance for any data-scarce LV network by extracting knowledge from a set of representative data-rich LV networks and extrapolating the knowledge to any data-scarce LV network. The imbalance-induced cost is then translated into the benefit from phase balancing and is compared against the costs of candidate phase balancing solutions, e.g. deploying phase balancers.

The CGPR approach supports the distribution network operators (DNOs) to perform cost-benefit analyses of phase balancing solutions on data-scarce LV networks. In this way, DNOs can decide whether phase balancing is economically feasible and which phase balancing solution yields the greatest net benefit compared to alternatives.

The remainder of the paper is organized as follows: Section II presents an overview of the methodology; Section III introduces the formulas for calculating imbalance-induced costs; Section IV presents the cost-benefit analysis framework, including the CGPR approach; Section V performs a case study and Section VI concludes the paper.

II. Overview of Methodology

To perform an accurate cost-benefit analysis of a phase balancing solution, full time-series of phase voltage and current data are required as the input data. However, these data are not available from the majority of UK's LV networks. In this paper, we have the time-series of phase current and voltage data of 800 representative data-rich LV networks throughout a year. These networks are located within the business area of a UK DNO and the data are the deliverables of the "Low Voltage Network Templates" project [20]. When conducting the trial project and collecting network data, Western Power Distribution specifically chose networks of a diverse and heterogeneous nature so that the dataset is representative. These 800 networks cover various customer types (domestic, commercial and industrial customers) and geographical areas (urban, suburban, and rural areas). For example, Cardiff contains a large number of commercial customers and load; Monmouthshire is a representative for the rural area [20].

Fig. 1 presents an overview of the CGPR approach. The key to this approach is to evaluate the imbalance-induced cost (including the cost of additional energy losses and the cost of additional network investment) for data-scarce LV networks. The approach consists of three stages:

Stage I: The 800 data-rich networks are clustered into three groups, i.e., urban, suburban and rural, by applying the *k*-means clustering method.

Stage II: Input features are selected for regression and these features are available from data-scarce LV networks. Then, utilizing the data-rich LV networks, Gaussian process regression (GPR) models are trained for each cluster of the LV networks to model the relationship between the selected features and the two imbalance-induced costs, i.e. the ARC and additional energy loss cost (AELC). The trained models are applied to data-scarce networks that only have the aforementioned features to estimate the imbalance-induced costs. An advantage of the approach is that it only requires features that are available from the majority of UK's data-scarce LV networks. Cross-validations are performed to validate the estimated imbalance-induced costs.

Stage III: The total imbalance-induced cost is calculated based on the estimations of the ARC and AELC. The imbalance-induced cost is then translated into the potential benefit from phase balancing. This benefit is compared to the cost of the phase balancing solution. This leads to a conclusion of whether the phase balancing solution is economically feasible or not as well as which phase balancing solution yields the greatest net benefit compared to alternatives.

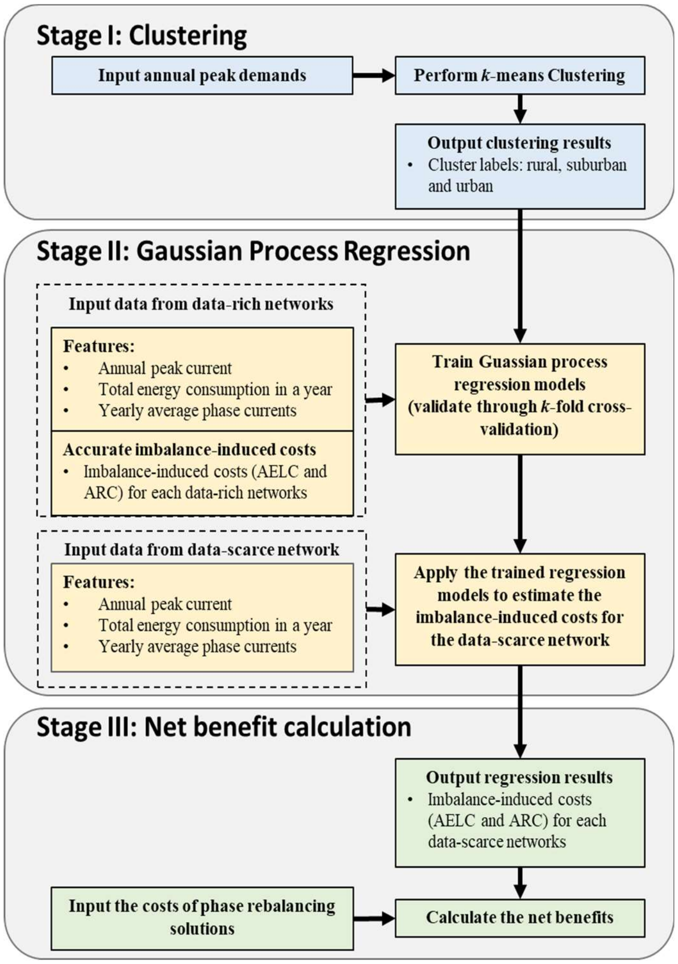

Stage I: Clustering
Input annual peak demands
Perform *k*-means Clustering
Output clustering results
• Cluster labels: rural, suburban and urban
Stage II: Gaussian Process Regression
Input data from data-rich networks
Features:
• Annual peak current
• Total energy consumption in a year
• Yearly average phase currents
Accurate imbalance-induced costs
• Imbalance-induced costs (AELC and ARC) for each data-rich networks
Train Guassian process regression models (validate through *k*-fold cross-validation)
Input data from data-scarce network
Features:
• Annual peak current
• Total energy consumption in a year
• Yearly average phase currents
Apply the trained regression models to estimate the imbalance-induced costs for the data-scarce network
Stage III: Net benefit calculation
Output regression results
• Imbalance-induced costs (AELC and ARC) for each data-scarce networks
Input the costs of phase rebalancing solutions
Calculate the net benefits

Fig. 1. Overview of the CGPR approach

III. Imbalance-Induced Cost for Individual Data-Rich Networks

This section presents the methods to calculate the components of the imbalance-induced cost for LV networks. The imbalance-induced cost consists of the ARC and the AELC. The AELC is broken down into the cost of energy losses caused by phase residual currents and the cost of transformer copper losses. The future cost is discounted back to form the present value. Then, the cost-benefit analysis is performed based on present values.

The present value of the ARC is detailed in [7]

$$ARC = f_{PV}(DIB) \approx 3k_f DIB_f + k_t DIB_t$$
$$\text{Subject to } k_\chi = Asset_\chi \cdot (1+d)^{\frac{logU_N}{\log(1+r)}} \cdot \frac{\log(1+d)}{\log(1+r)} \qquad (1)$$
$$\chi \in \{f, t\}$$

where DIB_f and DIB_t are the degrees of phase imbalance for main feeders and LV transformers, respectively. The mathematical definitions of DIB_f and DIB_t are given by (3) and (4), respectively. $Asset_\chi$ is the future asset reinforcement cost, where subscript χ can be either f (feeder) or t (transformer); d is the discount rate; U_N is the asset utilization rate and r is the load growth rate.

The factors U_N, DIB_f and DIB_t are given by (2), (3) and (4), respectively [7].

$$U_N = \frac{3 \cdot max\{P_\emptyset\}}{C_{asset}} \qquad \emptyset \in \{A, B, C\} \qquad (2)$$

where $P_\emptyset$ is the power on phase $\emptyset$ and C_{asset} is the asset capacity.

$$DIB_f = \frac{max\{P_\emptyset\} - \frac{P_t}{3}}{P_t} \qquad \emptyset \in \{A, B, C\}, \qquad (3)$$

where P_t is the total power of three phases when the maximum phase power occurs. $P_\emptyset$ is defined in (2).

$$DIB_t = \frac{P_N}{P_t} \qquad (4)$$

where P_N is neutral line power. P_t is defined in (3).

A. Imbalance-induced energy loss

The imbalanced-induced energy loss contains two components: the energy loss caused by a phase residual current [17] and the transformer copper loss.

1) Energy loss caused by phase residual current

The energy loss caused by phase residual current is calculated considering different earthing systems [21], e.g., Terre-Neutral-Combined (TN-C) and Terre-Neutral (TN-S) systems [22]. The majority of the UK's LV distribution networks follow the TN-S earthing system [22]. Therefore, this paper considers the TN-S earthing system.

The estimation of energy loss caused by the phase residual current is given in [17]

$$E_{loss} = \sum_{t=1}^{N_t} I_{prc}{}^2(t) \cdot R_n \cdot \Delta t \qquad (5)$$
$$\text{where } I_{prc}(t) = [I_A^2(t) + I_B^2(t) + I_C^2(t) - I_A(t)I_B(t) - I_B(t)I_C(t) - I_A(t)I_C(t)]^{1/2}$$

where $I_A(t), I_B(t)$ and $I_C(t)$ are current values for the phases A, B and C at time t, respectively; $I_{prc}(t)$ denotes the phase residual current at time t; R_n denotes the neutral wire resistance. N_t is the number of hours within the year.

The neutral line energy loss for the Nth year is

$$E_{loss_N} = E_{loss} \cdot (1+r)^{2(N-1)} \qquad (6)$$

where N represent the Nth year; r is defined in (1); and E_{loss} is defined in (5).

2) Transformer copper loss cost

Phase imbalance increases the transformer copper loss beyond that under the phase balanced scenario. The transformer copper loss under the balanced case is given in [23]:

$$E_{trans} = 3\sum_{t=1}^{N_t} I^2(t) \cdot R_w \cdot \Delta t \qquad (7)$$

where $I(t)$ is the balanced phase current at time t and R_w is the resistance of the transformer winding; N_t is the number of hours within a year.

The transformer copper loss under the imbalanced case is also given in [23]

$$E_i = \sum_{t=1}^{N_t} \left(I_A{}^2(t) + I_B{}^2(t) + I_C{}^2(t)\right) \cdot R_w \cdot \Delta t \qquad (8)$$

where $I_A(t), I_B(t)$ and $I_C(t)$ are current values for the phases A, B and C at time t, respectively; R_w and N_t are defined in (7).

As a result, the imbalance-induced transformer copper loss is:

$$E_{t\_i} = E_i - E_{trans} \qquad (9)$$

where all variables are defined in (7) and (8).

The transformer copper loss for the Nth year is

$$E_{t\_iN} = E_{t\_i} \cdot (1+r)^{2(N-1)} \qquad (10)$$

where r is the load growth rate; all other variables are defined in (7), (8) and (9).

B. The present value of the total imbalance-induced cost

As stated above, the total additional energy loss is the sum of losses caused by the phase residual current and transformer copper. Therefore, the total imbalance-induced energy loss in year N is given by

$$E_{tot_N} = E_{loss_N} + E_{t\_iN} \qquad (11)$$

where E_{loss_N} and E_{t_iN} are defined in (6) and (10), respectively.

The total AELC of the Nth year is transferred to the present value.

$$AELC = f_{PV}\left(E_{tot_N}\right) = \frac{E_{tot_N} \cdot \pi}{(1+d)^N} \qquad (12)$$

where π is the energy price; d is the discount rate; and E_{tot_N} is defined in (11).

The imbalance-induced energy losses incur costs every year until the three phases are rebalanced. In contrast, the ARC is a one-off investment when the asset capacity is reached. Therefore, the present value of the total imbalance-induced cost is given by

$$f_{PV\_N} = f_{PV}(DIB) + \sum_{n=1}^{N} f_{PV}\left(E_{tot_n}\right) \qquad (13)$$

where the function $f_{PV}(DIB)$ is defined in (1); the function $f_{PV}\left(E_{tot_n}\right)$ is defined in (12).

In this paper, the present value of the total imbalance-induced cost is referred to as the imbalance-induced cost for simplicity.

IV. Methodology

A. Clustering

In this section, a CGPR approach is presented as a combination of clustering and a Gaussian process regression (GPR). As mentioned in the previous section, the imbalance-induced cost includes two parts: ARC and AELC. Fig. 2 shows the relation between the annual peak currents and the ARCs for the 800 LV networks. It can be seen that three distinctive relationships exist.

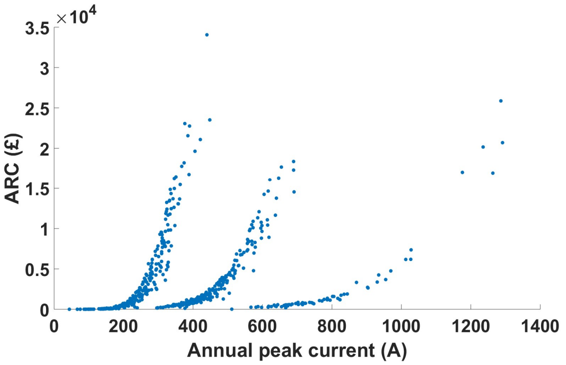


Fig. 2. The relationship between annual peak current and ARC

The underlying reason is that the ARCs are strongly correlated to the type of the LV networks, i.e. urban, suburban, and rural types. The three different relationships justify the development of a cluster-wise regression as opposed to a simple regression. Cluster-wise regression is an effective way of addressing problems with multiple regression models [24] [25].

As shown in Fig. 1, k-means clustering is used to cluster the networks into 3 groups (rural, suburban and urban) by the annual peak demands. This corresponds to Stage I in Fig. 1. The direct output of the clustering is which cluster each LV network belongs to (i.e. the cluster label for each LV network). From the outputs, it is straightforward to derive the range of annual peak currents for each cluster of the LV networks. In this way, given any LV network, determine which range its annual peak current falls into. This reveals the cluster to which the network in question belongs, i.e. whether the network is an urban, suburban, or rural one.

B. Gaussian process regression

The output of Stage I is used to train Gaussian process regression (GPR) models to model the relationship between the selected features and the imbalance-induced costs (i.e., AELC and ARC). The imbalance-induced costs are calculated using data from data-rich networks.

Then, the networks are treated as data-scarce networks and the selected features are used as the input to the trained GPR models. The GPR models output estimated imbalance-induced costs.

The regression process consists of the following steps:

1) Feature selection

For the majority of the UK's LV networks, the annual peak current ($\hat{I}$) and annual total energy consumption (E_{total}) are readily available. According to [17], the average phase current values can be obtained with minimal efforts from either the per-phase energy meters or the protection system for data-scarce networks. The average phase current values are transformed into a virtual phase residual current:

$$\bar{I}_{prc} = \sqrt{\bar{I}_a^{\,2} + \bar{I}_b^{\,2} + \bar{I}_c^{\,2} - \bar{I}_a\bar{I}_b - \bar{I}_b\,\bar{I}_c - \bar{I}_a\bar{I}_c} \qquad (14)$$

where $\bar{I}_a$, $\bar{I}_b$ and $\bar{I}_c$ are the yearly average phase current values for phases A, B, and C, respectively.

Two input feature vectors are defined to suit different levels of data availability in data-scarce networks. The first feature vector (ν_{f1}) contains two features ($\hat{I}$ and E_{total}):

$$\nu_{f1} = [\hat{I}\,, E_{total}] \qquad (15)$$

This feature vector is applicable in the absence of the average phase current values. The second feature vector (ν_{f2}) contains three features ($\hat{I}$, E_{total} and $\bar{I}_{prc}$):

$$\nu_{f2} = [\hat{I}, E_{total}, \bar{I}_{prc}] \qquad (16)$$

This feature vector requires that the data-scarce network have the average phase current data.

2) Gaussian Process Regression model training

In this step, regression models are trained for each cluster of LV networks. The regression models map the feature vectors defined in step 1) to the ARC and AELC (the ARC and AELC are calculated in Section III) separately. In this paper, the Gaussian process regression (GPR) is adopted. The reasons why the GPR is adopted are: 1) Gaussian process models allow the quantification of uncertainty, considering both intrinsic noises in the problem and parameter errors in estimation [26]; 2) the case studies confirm that the GPR achieves the best performance among classical regression models.

Take the GPR that maps the feature vectors to the ARC as an example. The GPR model is given by

$$p\left(ARC_* \mid ARC, \nu_f, \nu_{f_*}\right) \sim \mathcal{N}(\mu^*, \Sigma^*) \qquad (17)$$

where $\mu^* = K(\nu_f, \nu_f)\left(K(\nu_f, \nu_f) + \sigma^2 I\right)^{-1} ARC$

$$\Sigma^* = K\left(\nu_{f_*}, \nu_{f_*}\right) + \sigma^2 I - K\left(\nu_{f_*}, \nu_f\right)\left(K(\nu_f, \nu_f) + \sigma^2 I\right)^{-1} K\left(\nu_f, \nu_{f_*}\right)$$

where $p\left(ARC_* \mid ARC, \nu_f, \nu_{f_*}\right)$ is the probability distribution for ARC estimation; ν_f and ARC are the feature vector and the ARC for the data-rich networks, respectively; ν_{f_*} and ARC_* are the feature vector and the predicted ARC for the data-scarce network, respectively; the ARC is given by (1); $\mathcal{N}(\mu^*, \Sigma^*)$ denotes a Gaussian distribution with the mean μ^* and covariance Σ^*; K is a kernel matrix given by the squared exponential kernel function [26]; σ^2 is the noise variance; and I is the identity matrix. The feature vector ν_f could be ν_{f1} and ν_{f2} as given in (15) and (16), depending on the choice of features.

The GPR is detailed in [26]. The above GPR model is developed for each cluster of the LV networks. The GPR is detailed in [26]. The GPR model for the AELC estimation is the same as that for the ARC estimation as shown in (17), except that the ARC is replaced by the AELC. The results are compared with linear regression, which is detailed in [27] and which is not repeated in this paper.

C. *Cross-validation*

The CGPR approach is validated through k-fold cross-validation. This is a popular validation method as explained in [28]. The cross-validation is detailed as follows: the full dataset of 800 data-rich LV networks, including the features and the accurate ARC and AELC results, are randomly separated into k (k=10 in this paper) equal-sized groups. In each iteration of the k-fold cross-validation, one group of the LV networks are reserved as the validation set, whereas the remaining nine groups serve as the training set. The CGPR model is trained using the training set only. Then, the trained CGPR model predicts the imbalance-induced costs on the validation set, which are treated as if they were data-scarce. The outputs are estimated imbalance-induced costs for the LV networks in the validation set. These results are compared against the accurate imbalance-induced cost (the calculated costs from data-rich networks) results so that the CGPR model is validated. Each group is selected as the validation set once and there are ten iterations. It should be emphasized that throughout the process, the validation set and the training set are strictly separated from each other and the validation set is not used for training. The k-fold cross-validation is detailed in Fig. 3.

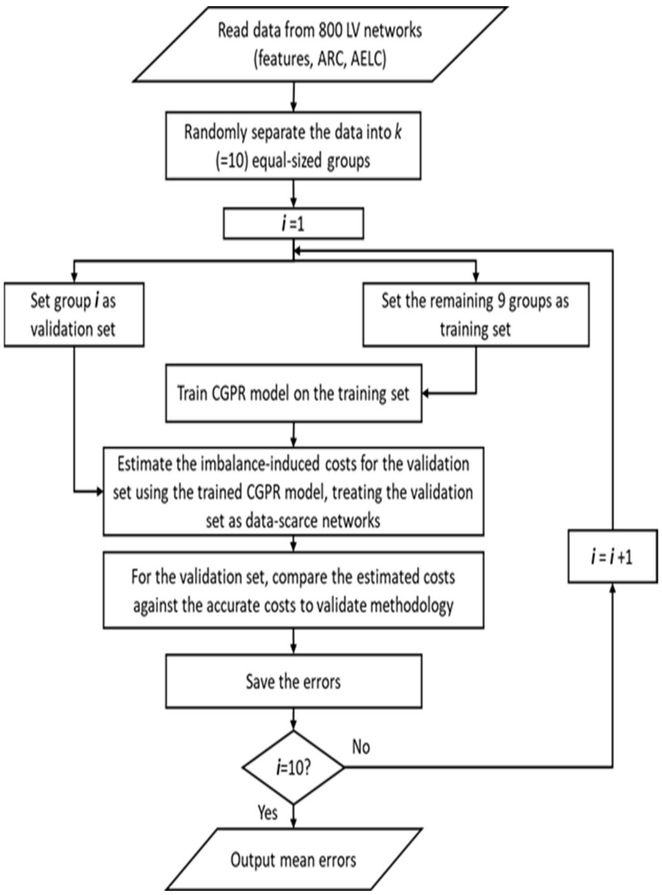

Read data from 800 LV networks (features, ARC, AELC)
Randomly separate the data into k (=10) equal-sized groups
i =1
Set group i as validation set
Set the remaining 9 groups as training set
Train CGPR model on the training set
Estimate the imbalance-induced costs for the validation set using the trained CGPR model, treating the validation set as data-scarce networks
For the validation set, compare the estimated costs against the accurate costs to validate methodology
i = i +1
Save the errors
i=10?
No
Yes
Output mean errors

Fig. 3. The flow chart of k-fold cross-validation

D. *Removal of outliers*

Following the cross-validation, 11% of the networks are identified as the outliers and are removed. This percentage is derived by using Chebyshev's inequality. Chebyshev's inequality is a widely adopted method for removing outliers [29]. When the distribution of the data is unknown, the Chebyshev's inequality is given by:

$$P(|X-\mu| \le k\sigma) \ge 1-\frac{1}{k^2} \tag{18}$$

where X is the set of sample data, μ is the mean of the sample data, σ is the standard deviation and k is a factor.

It is common practice to regard data samples that occur beyond 3σ (i.e., k=3) from the mean as outliers [30], [31]. Therefore, the outliers account for approximately 11% of the whole population of networks. Note that outliers are an objective existence and they can be identified and removed from consideration for better performance.

E. *Net benefit calculation*

The trained CGPR model takes the features of any given data-scarce network as the input and outputs the estimated imbalance-induced cost.

Note that the phase balancing solutions may not be able to fully rebalance the three phases. Therefore, the benefit from phase balancing is given by the difference of the total imbalance-induced costs before and after phase balancing

$$B^{ds}_{PV\_N} \approx f^{before}_{PV_N} - f^{after}_{PV\_N} \tag{19}$$

where $f^{before}_{PV_N}$ and $f^{after}_{PV\_N}$ are the estimated total imbalance-induced cost before and after phase balancing, respectively; the superscript ds means data-scarce; the subscript PV_N represents present value for N years.

Then, the benefit is compared with the cost of the phase balancing solution to determine whether it is beneficial to apply the phase balancing solution in question. Hence, the net benefit of applying the phase balancing solution is given by

$$B^{ds} \approx B^{ds}_{PV\_N} - f_{pb} \tag{20}$$

where $B^{ds}_{PV\_N}$ is the total benefit of phase balancing for the data-scarce networks; f_{pb} is the cost of applying a phase balancing solution.

Note that the net benefit B^{ds} can be negative, which means that it is not economically feasible to deploy the phase balancing solution.

V. Case Studies

This section presents case studies. The input data are shown in Section V-A. The results from the cluster-wise regression model are presented in Section V-B. Section V-C gives the discussions. Section V-D gives the cost-benefits analysis for two phase balancers (ZM-SPC [32] and EQU18 [33]) and active network management (ANM) scheme, respectively. The case study is based on the time-series phase current and phase voltage data from the 800 data-rich LV networks throughout a year.

A. *Imbalance-induced cost for data-rich networks*

This sub-section presents the calculation of imbalance-induced cost for data-rich networks. To derive the additional energy losses (defined (5) - (10)), the neutral wire resistance (R_n) is set as 0.244 Ω/km [17]. The winding resistances (R_w) are calculated from [34] and presented in Table I.

To derive the additional reinforcement costs (defined in (1) - (4)), the investment costs of the feeder and transformer are given in Table I. The discount value (d) is set as 5.0% [1] and [35]. The load growth rate (r) is set as 0.82% [36].

TABLE I. PARAMETERS FOR DIFFERENT AREAS [34], [37]

Assets Area	Urban	Suburban	Rural
Transformer investment cost (k£)	26.4	16.1	5.8
Main feeder investment cost (k£/km)	67.2	16.4	15.0
Main feeder length (km)	0.2	0.3	0.4
No. of feeders connected from transformers	5	3.5	1.5
Winding resistance (Ω)	0.0163	0.0265	0.0413

This paper assumes that the phase currents are 120° apart from each other. This is because there is hardly any LV network that has phasor measurements, as distribution network operators cannot justify the investment in phasor measurements in terms of the return on investment. Therefore, it is valid to assume that the phase currents are 120° apart from each other while phasor measurements are absent.

The phase residual current is the minimum, under the assumption that the phase currents are 120° apart from each other. Therefore, this assumption corresponds to a conservative cost-benefit analysis. If the actual phase currents are not 120° apart, the phase residual current will increase, so will the imbalance-induced energy losses and the associated cost. This means that the potential benefit from phase balancing will also increase, hence the net benefit will increase.

In this paper, a power factor of 0.9 is assumed and the harmonic distortion is not considered. The harmonic distortion results in the decrease of power factor and eventually increases the ARC. Besides, the harmonic currents cause additional energy losses which lead to higher AELC. Therefore, it shows that the estimation of the imbalance-induced costs is conservative, resulting in conservative net benefits, i.e. the lower bounds of the net benefits. The actual net benefits can be higher than the estimated value.

Fig. 4 shows the present values of AELC and ARC for urban, suburban and rural networks. The average AELC is approximately twice as much as the average ARC. The rural networks correspond to the least AELC and the greatest ARC among all three types of networks. In contrast, the urban networks correspond to the greatest AELC and the least ARC.

The reason for this is that the rural networks have the largest DIB (degree of imbalance) values, which causes the greatest ARC, in both LV transformers and main feeders among the three types of networks. However, the rural networks have the lowest loading levels, which lead to the lowest energy losses on the neutral lines and LV transformers. As a result, the rural networks have the largest average ARC but least average AELC. On the contrary, urban networks have the lowest DIB, which leads to the lowest ARC. They have the highest energy loss because of their high loading levels. Therefore, the urban network has the least AELC but largest ARC.

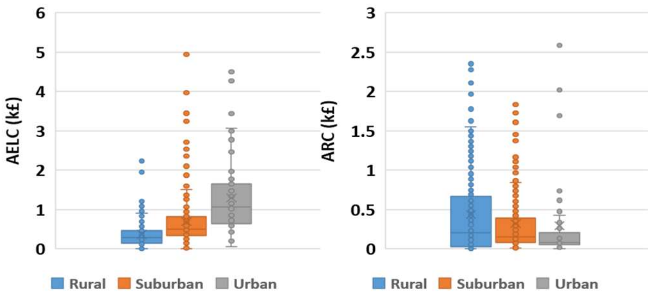


Fig. 4. The AELC and ARC for the 800 LV networks

B. Cluster-wise Gaussian Process Regression

In this section, the CGPR results are shown, where the cost-benefit analyses are performed over a time horizon of 10 years.

The ARC and AELC estimation are calculated using four regression methods: linear regression (LR), cluster-wise LR (CLR), GPR and CGPR. Results from all methods are validated through 10-fold cross-validations. The results obtained by applying these four regression methods are compared with each other in terms of the root mean squared error (RMSE). As mentioned in Section IV-A, two feature vectors are used as input, the first vector ν_{f1} contains two features ($\hat{I}$ and E_{total}), while the second vector ν_{f2} contains three features ($\hat{I}$, E_{total} and $\bar{I}_{prc}$). Therefore, the performances of different regression methods are compared with each other.

Fig. 5 presents the RMSE values of using LR, CLR, GPR and CGPR with two and three features. In Fig. 5 - a) (i.e., the ARC estimation using two features), the GPR model performs better than the LR model and the CGPR model performs better than the CLR model in terms of RMSE. The RMSE of CLR is 2,537.94, while the RMSE of CGPR is 1,443.24.

In Fig. 5 - b) (i.e., the AELC estimation using two features), the GPR model has a similar performance to the LR model and the CGPR model also has a similar performance to the CLR model. The RMSE of CLR is 4,885.80 while the RMSE of CGPR is 4,752.92.

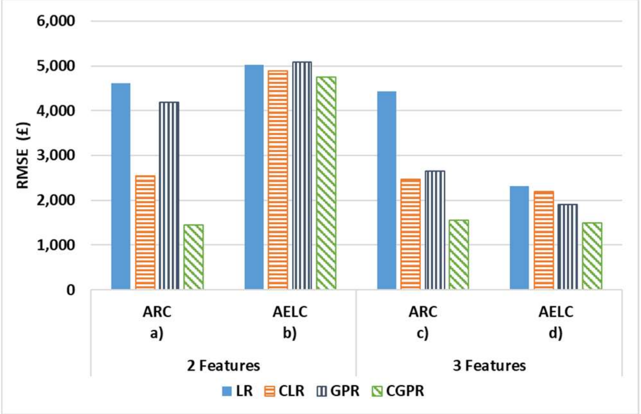


Fig. 5. Comparison of RMSEs of ARC and AELC estimation with different regression methods

In Fig. 5 - c) (i.e., the ARC estimation using three features), the GPR method performs better than LR; the CGPR method performs better than the CLR method. The RMSE of the CLR is 2,466.06, while the RMSE of CGPR is 1,554.89.

In Fig. 5 - d) (i.e., the AELC estimation using three features), the GPR method performs better than the LR; the CGPR method performs better than the CLR method. The RMSE of CLR is 2,199.55, while the RMSE of CGPR is 1,487.71. As a result, CGPR has the best performance among all methods.

For the CGPR model with three features as input, with 95% confidence, the range of RMSEs are [910.84, 1,309.20], [913.59, 1,184.83] and [1,916.03, 3,291.87] for rural, suburban and urban networks, respectively. The suburban networks have the smallest range of the RMSE while the urban networks have the largest range of the RMSE. Therefore, the GPR model performs the best on the imbalance-induced cost estimation for suburban networks among the three types of networks.

C. *Discussions*

Using Chebyshev's inequality, 11% of the networks are identified as outliers. Fig. 6 shows the comparison of the mean average percentage error (MAPE) before and after the removal of outliers. When using two features, the MAPE of the ARC drops from 29.95% to 23.76% and the MAPE of AELC decreases from 53.86% to 40.75%. When using three features, the MAPE of the ARC drops from 30.06% to 23.32% and the MAPE of AELC decreases from 53.87% to 21.33%.

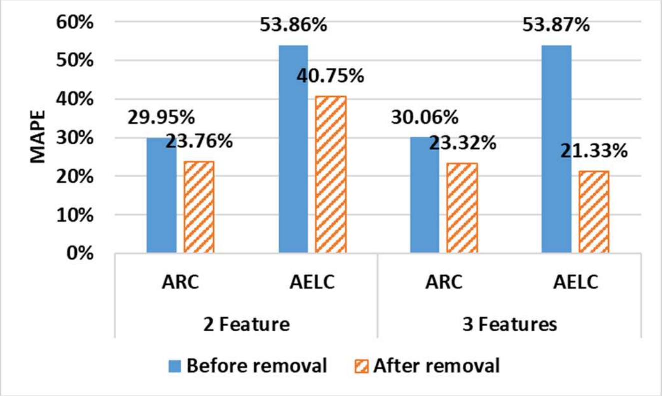


Fig. 6. Comparison of results before and after removing outliers

One of the main reasons why the MAPE is approximately 23% is that the CGPR approach only requires two or three features from data-scarce LV networks. Another reason is that only one year's data is used to estimate the imbalance-induced costs over the future 10 years (or 30 years), resulting in an accumulation of errors over the years. Among the three types of networks, the MAPE values for suburban networks are the lowest. In other words, the cost estimations for suburban networks demonstrate the best performance among the three types of networks. On the other hand, the cost estimations for rural networks demonstrate the worst performance among the three types of networks.

In general, there is a lack of monitoring in the UK's millions of LV networks. The two sets of features are chosen in this paper because they are either routinely collected by distribution network operators or are readily available to be collected. Using these features leads to a feasible cost for data collections and the feasibility of the cost-benefit analyses, if scaled up from individual networks to a mass population of networks. Therefore, the features are chosen to best suit the existing level of monitoring in the UK's LV networks and making the methodology scalable to the whole LV networks.

Utilities use load factors to estimate loss factors, which are then used to determine the energy losses of the system. Reference [38] discussed the ways of determining the energy losses using load factor and loss factor. Two values for the coefficient 'a' are suggested by [38], i.e., a = 0.16 and a = 0.3. Both the values are adopted and the lower error of using this method to the estimate energy loss cost is 67.09%. The reason for the large error is that it is difficult to determine the values of 'a' for a data-scarce distribution system. Besides, the distributions system has multiple branches connected to the main feeder which results in a higher estimation error. However, the developed CGPR approach only incurs an error of 21.33% when estimating the AELC. The developed CGPR approach performs better than the method adopted by utilities.

The estimated AELCs using CGPR are compared with the actual values for validation. A random selection of the comparison results (10 networks out of 800 ones) are presented because of the page limitation. As shown in Fig. 7, the estimation results follow a similar trend to the actual results.

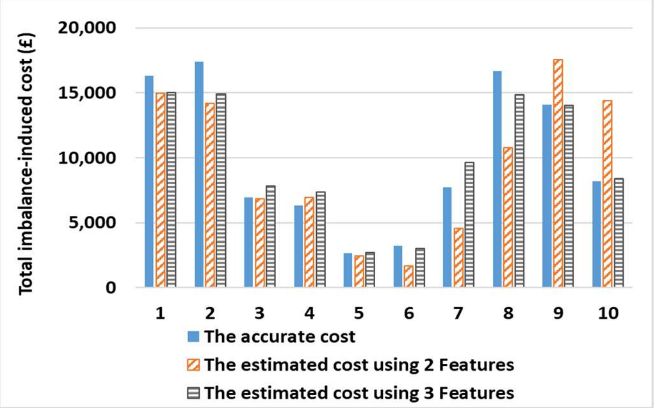


Fig. 7. Comparison of actual and estimated total imbalance-induced cost

The use of two features and three features are compared with each other. The latter is highly recommended as it incurs a much lower error. However, the use of two features still has its value just in case some LV networks do not have three features (i.e. they only have yearly peak current and total energy consumption). In the absence of the third feature (i.e. the yearly average phase currents), one way to perform cost-benefit analyses is to use the two-feature-version of the methodology; an alternative way is to collect the third feature from the networks, but this incurs a data collection cost. This cost can be prohibitively high when the cost-benefit analyses are to be scaled up to a mass population of networks. Therefore, a trade-off should be made between the data collection cost and the accuracy of the methodology for cost-benefit analyses.

Within the dataset of 800 LV networks, 11.2%, 44.4%, and 44.4% are urban, suburban, and rural networks, respectively. The same dataset was used to: 1) develop 11 representative LV substation load profiles [20], [25]; 2) classify four types of phase imbalance in terms of the imbalance direction [39]; 3) estimate the imbalance-induced energy losses in the neutral and ground for data-scarce LV networks [17]. These publications prove the diversity and heterogeneity within the dataset.

Furthermore, the dataset corresponds to a geographical area of a similar size and is of a similar nature (a mixture of urban, suburban, and rural networks) to that used in [40].

Given that the model is trained on the dataset from South Wales, UK, the model is applicable to networks within the region of a similar nature to South Wales (a mixture of urban areas like Cardiff, suburban and rural areas like Momonthshire). Caution has to be exercised when applying the trained model on substantially different areas, e.g. central London which is extremely urban and which is unlike anywhere else in the UK. The CGPR methodology is generic. If it is to be applied to other countries or the central London area, it should be trained on the dataset representative of the area in question.

D. Net benefit calculation

Given any data-scarce network, its imbalance-induced costs calculated through CGPR are used for a net benefit calculation. These costs are translated into the benefits of phase balancing for the data-scarce network using (13).

Table II shows the two selected types of phase balancers, along with their costs and lifetimes. The net benefits by applying two phase balancers are calculated using (20).

TABLE II. COSTS OF PHASE BALANCERS

Type	ZM-SPC [32]	EQU18 [33]	ANM [41] [42]
Lifetime (Years)	>10	>30	>20
Total costs (£)	4,890	2,381	73,600

The net benefits from phase balancing for data-scarce networks are estimated over the respective lifetime of the two phase balancers and the ANM scheme, i.e. 10 years for ZM-SPC, 30 years for EQU18 and 20 years for the ANM scheme. This paper assumes that power-electronics-based phase balancers and the ANM scheme can achieve full phase balancing because they can perform high-resolution real-time balancing.

As stated in the previous section, it is highly recommended using three features as the input for the proposed CGPR approach. In this section, the net benefits are estimated using three features. Fig. 8, Fig. 9 and Fig. 10 show the distribution of the estimated net benefits using three features from phase balancing by ZM-SPC for the rural, suburban and urban networks, respectively. Results show that approximately 70% of rural networks, 80% of suburban networks and 90% of urban networks benefit from ZM-SPC.

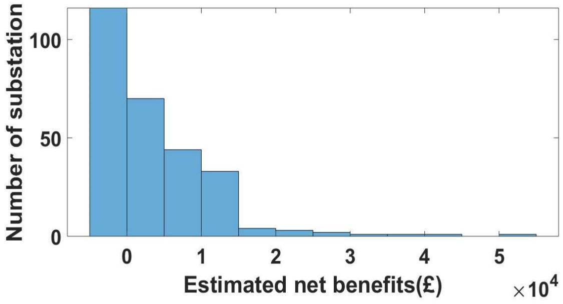


Fig. 8. The distribution of mean net benefits for rural networks from phase balancing by ZM-SPC

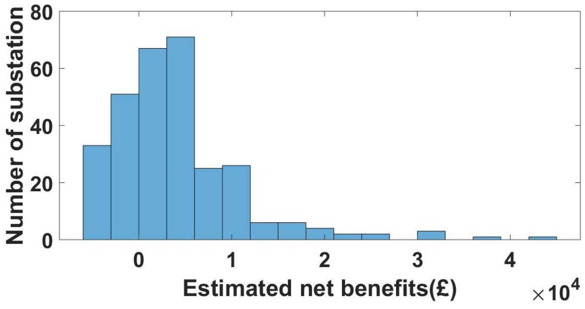


Fig. 9. The distribution of mean net benefits for suburban networks from phase balancing by ZM-SPC

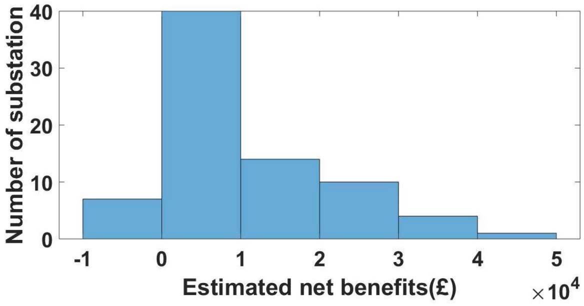


Fig. 10. The distribution of mean net benefits for urban networks from phase balancing by ZM-SPC

With 95% confidence, the range of net benefits from ZM-SPC for rural, suburban and urban networks are [£2,814.66, £5,106.51], [£3,461.50, £5,346.27] and [£7,591.93, £12,977.50], respectively.

Fig. 11, Fig. 12 and Fig. 13 show the distribution of the estimated net benefits using three features from phase balancing by EQU18 for the rural, suburban and urban networks, respectively. Results show that approximately 94% of rural networks, 97% of suburban networks and 99% of urban networks benefit from EQU18.

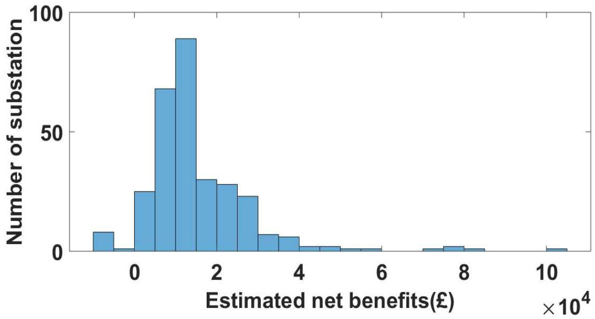


Fig. 11. The distribution of mean net benefits for rural networks from phase balancing by EQU18

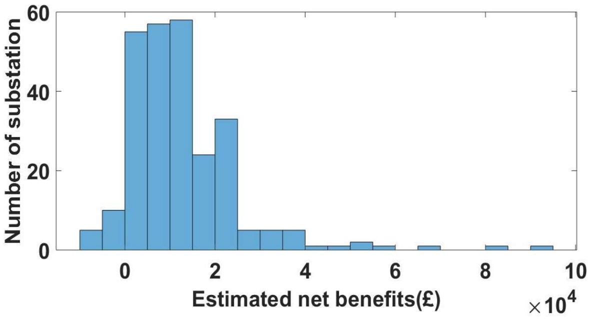


Fig. 12. The distribution of mean net benefits for suburban networks from phase balancing by EQU18

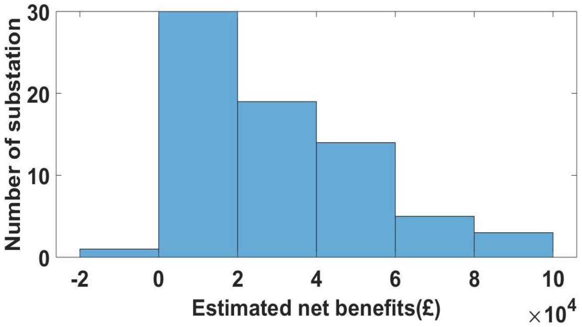


Fig. 13. The distribution of mean net benefits for urban networks from phase balancing by EQU18

With 95% confidence, the range of net benefits from EQU18 for rural, suburban and urban networks are [£11,153.87, £14,975.80], [£15,218.09, £18,974.98] and [£26,926.63, £39,441.18], respectively.

Fig 14, Fig 15 and Fig 16 show the distribution of the estimated net benefits using three features from phase balancing using ANM for the rural, suburban and urban networks, respectively. Results show that approximately 1% of rural networks, 1% of suburban networks and no urban network benefit from the ANM scheme.

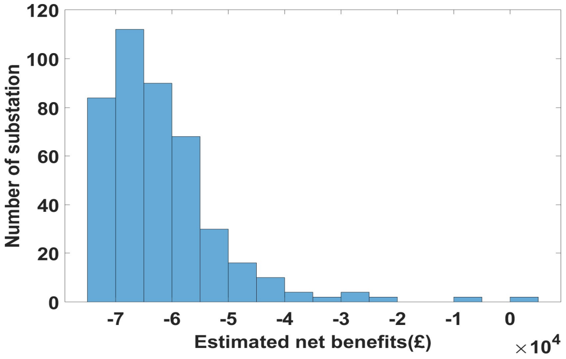


Fig. 14. The distribution of mean net benefits for rural networks from phase balancing by the ANM scheme

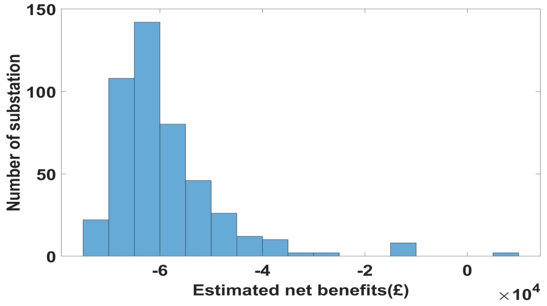


Fig. 15. The distribution of mean net benefits for suburban networks from phase balancing by the ANM scheme

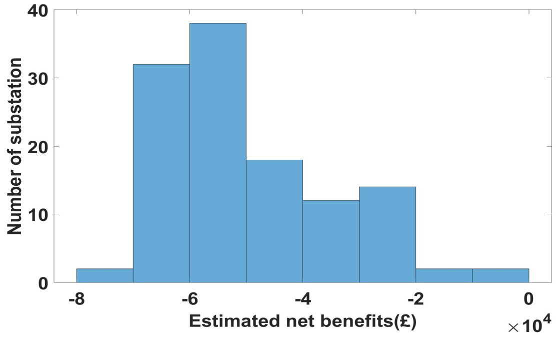


Fig. 16. The distribution of mean net benefits for urban networks from phase balancing by the ANM scheme

With 95% confidence, the range of net benefits from applying the ANM scheme for rural, suburban and urban networks are [£-63,127.49, £-60,249.45], [£-60,396.22, £-57,564.91] and [£-53,102.38, £-45,313.21], respectively. The net benefits are negative, meaning that adopting the ANM scheme for phase balancing is not cost-effective. However, it is worth mentioning that the ANM scheme typically brings other benefits such as relieving thermal overloads and voltage violations, apart from phase balancing.

Comparing the RMSEs (given in Section V-B) with the net benefits from phase balancing, it can be found that the RMSEs are insignificant.

Fig. 17, Fig. 18 and show the probability that the phase balancing solutions by ZM-SPC and EQU18 would produce a positive net benefit for any data-scarce LV network with 95% confidence, respectively. The probability of having positive net benefit assist DNOs to make the decision on whether to invest in phase balancing.

For example, the CGPR is used to estimate the net benefit for a data-scarce network 10036 from ZM-SPC. The network 10036 is a rural network and its estimated net benefit is £5001. Thus, with 95% confidence, the corresponded probability of network 10036 having a positive net benefit is 96.6%. If the DNO set the acceptable probability as 90%, the network 10036 is therefore identified as worth for phase balancing.

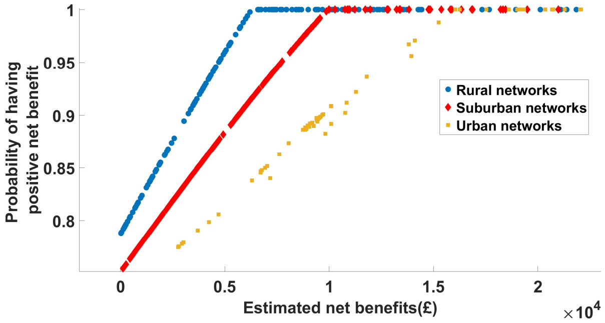


Fig. 17. The probability of having positive net benefits from phase balancing by ZM-SPC

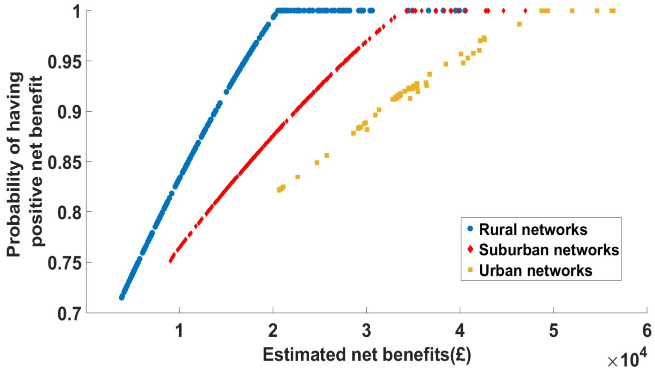


Fig. 18. The probability of having positive net benefits from phase balancing by EQU18

There is a way to strengthen the robustness of the CGPR model. The CGPR model already outputs the data-scarce LV networks where it is highly likely that a given phase balancing solution will deliver more benefit than cost. In this way, the

CGPR model serves as a filter. For these networks (which are a subset of the whole population of networks) that the CGPR model identifies as being worthy of phase balancing, the DNO can further check the cost-benefit of phase balancing on these networks by collecting time-series data from these networks and performing accurate cost-benefit analysis.

VI. Conclusions

This paper addresses a previously unresolved problem faced by the distribution network operators (DNOs), i.e., the cost-benefit analysis of phase balancing solutions for the vast majority of the low voltage (LV) networks that are data-scarce. To this end, this paper develops a new cluster-wise Gaussian process regression (CGPR) approach.

The approach is validated by the case studies considering two types of phase balancers and the active network management (ANM) scheme. The phase balancers are ZM-SPC and EQU18 with different costs and lifetime. The maximum potential net benefits for all types of LV networks are calculated for each phase balancer. Given any data-scarce network and phase balancing solution, the probability that the solution will produce a positive net benefit is quantified.

A major advantage of the approach is that it only requires the annual peak current and the total energy consumption throughout a year – these data are collected only once a year. The developed approach offers a cost-effective and efficient way to help DNOs understand: 1) whether a phase balancing solution is economically feasible for any data-scarce network; 2) if yes, the maximum potential net benefit from the solution.

References

[1] K. Ma, R. Li, and F. Li, “Utility-Scale Estimation of Additional Reinforcement Cost From Three-Phase Imbalance Considering Thermal Constraints,” *IEEE Transactions on Power Systems,* vol. 32, no. 5, pp. 3912-3923, 2017.

[2] G. Mokryani, A. Majumdar, and B. C. Pal, “Probabilistic method for the operation of three-phase unbalanced active distribution networks,” *IET Renewable Power Generation,* vol. 10, no. 7, pp. 944-954, 2016.

[3] M. W. Siti, D. V. Nicolae, A. A. Jimoh, and A. Ukil, “Reconfiguration and Load Balancing in the LV and MV Distribution Networks for Optimal Performance,” *Power Delivery, IEEE Transactions on,* vol. 22, no. 4, pp. 2534-2540, 2007.

[4] S. Yan, S. C. Tan, C. K. Lee, B. Chaudhuri, and S. Y. R. Hui, “Electric Springs for Reducing Power Imbalance in Three-Phase Power Systems,” *IEEE Transactions on Power Electronics,* vol. 30, no. 7, pp. 3601-3609, 2015.

[5] OFGEM. "Electricity distribution units and loss percentages summary," https://www.ofgem.gov.uk/sites/default/files/docs/2010/08/distribution-units-and-loss-percentages-summary.pdf.

[6] T. Alinjak, I. Pavic, and K. Trupinic, “Improved three-phase power flow method for calculation of power losses in unbalanced radial distribution network,” *CIRED - Open Access Proceedings Journal,* vol. 2017, no. 1, pp. 2361-2365, 2017.

[7] 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, 2016.

[8] J. Zhu, M. Y. Chow, and F. Zhang, “Phase balancing using mixed-integer programming [distribution feeders],” *Power Systems, IEEE Transactions on,* vol. 13, no. 4, pp. 1487-1492, 1998.

[9] A. Kavousi-Fard, T. Niknam, and M. Fotuhi-Firuzabad, “A Novel Stochastic Framework Based on Cloud Theory and Modified Bat Algorithm to Solve the Distribution Feeder Reconfiguration,” *IEEE Transactions on Smart Grid,* vol. 7, no. 2, pp. 740-750, 2016.

[10] C. H. Lin, C. S. Chen, H. J. Chuang, M. Y. Huang, and C. W. Huang, “An Expert System for Three-Phase Balancing of Distribution Feeders,” *IEEE Transactions on Power Systems,* vol. 23, no. 3, pp. 1488-1496, 2008.

[11] P. Lico, M. Marinelli, K. Knezović, and S. Grillo, "Phase balancing by means of electric vehicles single-phase connection shifting in a low voltage Danish grid." pp. 1-5.

[12] T. S. Win, Y. Hisada, T. Tanaka, E. Hiraki, M. Okamoto, and S. R. Lee, “Novel Simple Reactive Power Control Strategy With DC Capacitor Voltage Control for Active Load Balancer in Three-Phase Four-Wire Distribution Systems,” *IEEE Transactions on Industry Applications,* vol. 51, no. 5, pp. 4091-4099, 2015.

[13] T. Chen, “Evaluation of line loss under load unbalance using the complex unbalance factor,” *IEE Proceedings - Generation, Transmission and Distribution,* vol. 142, no. 2, pp. 173-178, 1995.

[14] L. Ochoa, R. Ciric, A. Padilha-Feltrin, and G. Harrison, *Evaluation of distribution system losses due to load unbalance*, 2005.

[15] M. Baggu, J. Giraldez, T. Harris, N. Brunhart-Lupo, L. Lisell, and D. Narang, "Interconnection assessment methodology and cost benefit analysis for high-penetration PV deployment in the Arizona Public Service system." pp. 1-6.

[16] F. M. Camilo, R. Castro, M. E. Almeida, and V. F. Pires, “Assessment of overvoltage mitigation techniques in low-voltage distribution networks with high penetration of photovoltaic microgeneration,” *IET Renewable Power Generation,* vol. 12, no. 6, pp. 649-656, 2018.

[17] L. Fang, K. Ma, R. Li, Z. Wang, and H. Shi, “A Statistical Approach to Estimate Imbalance-Induced Energy Losses for Data-Scarce Low Voltage Networks,” *IEEE Transactions on Power Systems*, pp. 1-1, 2019.

[18] U. P. Networks. "Phase Switch System," https://innovation.ukpowernetworks.co.uk/projects/phase-switch-system.

[19] S. E. Networks, *HV and LV Phase Imbalance Assessment* 7640–07–D4, September 2015.

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

[21] A. A. Sallam, and O. P. Malik, *Electric Distribution Systems*, Hoboken, NJ, USA: Hoboken, NJ, USA: John Wiley & Sons, Inc., 2012.

[22] G. Cronshaw, *EARTHING: YOUR QUESTIONS ANSWERED*, IEE Wiring Matters, 2005.

[23] E. J. O, O. S.O, and I. S. A;, “Evaluation Of Distribution System Losses Due To Unbalanced Load In Transformers A Case Study Of Guinness 15MVA, 33/11KV, Injection Substation And Its Associated 11/0.415kv Transformers In Benin City, Nigeria,” *International Journal of Engineering Research & Technology,* vol. 2, no. 3, March, 2013.

[24] H. Ying-Yi, and C. Zuei-Tien, “Development of energy loss formula for distribution systems using FCN algorithm and cluster-wise fuzzy regression,” *IEEE Transactions on Power Delivery,* vol. 17, no. 3, pp. 794-799, 2002.

[25] R. Li, C. Gu, F. Li, G. Shaddick, and M. Dale, “Development of Low Voltage Network Templates—Part II: Peak Load Estimation by Clusterwise Regression,” *IEEE Transactions on Power Systems,* vol. 30, no. 6, pp. 3045-3052, 2015.

[26] C. B. Do, “Gaussian processes,” 2008.

[27] H. J. Seltman, *Experimental Design and Analysis*, 2018.

[28] T. Wong, and N. Yang, “Dependency Analysis of Accuracy Estimates in k-Fold Cross Validation,” *IEEE Transactions on Knowledge and Data Engineering,* vol. 29, no. 11, pp. 2417-2427, 2017.

[29] B. G. A. T. A. F. S. K. Cooley, “Data outlier detection using the Chebyshev theorem,” in 2005 IEEE Aerospace Conference, Big Sky, MT, USA, 2005, pp. 3814-3819.

[30] 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.

[31] Y. Zhang, J. Liu, and H. Li, “An Outlier Detection Algorithm Based on Clustering Analysis,” in 2010 First International Conference on Pervasive Computing, Signal Processing and Applications, 2010, pp. 1126-1128.

[32] "ZM-SPC Series Three phase unbalanced device," March, 2019; http://zmgs.com/en/index.php?r=article/Content/index&content_id=596.

[33] "EQUI8 : THREE-PHASE LOW VOLTAGE NETWORK BALANCER," March, 2019; http://cmetransformateur.com/equi8-en/.

[34] S. Electric. "HV/LV distribution transformers," http://mt.schneider-electric.be/main/tfo/catalogue/an_iec.pdf.

[35] A. S. Sidhu, M. G. Pollitt, and K. L. Anaya, “A social cost benefit analysis of grid-scale electrical energy storage projects: A case study,” *Applied Energy,* vol. 212, pp. 881-894, 2018/02/15/, 2018.

[36] "Pathways for the GB Electricity Sector to 2030," https://www.energy-uk.org.uk/publication.html?task=file.download&id=5722.

[37] Y. Zhang, F. Li, Z. Hu, and G. Shaddick, “Quantification of low voltage network reinforcement costs: A statistical approach,” *IEEE Transactions on Power Systems,* vol. 28, no. 2, pp. 810-818, 2013.

[38] K. M. P. K. Sen, “A Better Understanding of Load and Loss Factors ” in 2008 IEEE Industry Applications Society Annual Meeting, Edmonton, AB, Canada, 2008.

[39] W. Kong, K. Ma, and Q. Wu, “Three-Phase Power Imbalance Decomposition Into Systematic Imbalance and Random Imbalance,” *IEEE Transactions on Power Systems,* vol. 33, no. 3, pp. 3001-3012, 2018.

[40] V. Rigoni, L. F. Ochoa, G. Chicco, A. Navarro-Espinosa, and T. Gozel, “Representative Residential LV Feeders: A Case Study for the North West of England,” *IEEE Transactions on Power Systems,* vol. 31, no. 1, pp. 348-360, 2016.

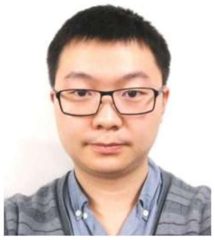
[41] *Low Carbon London Project Closedown Report*, UK Power Networks, 2015.

[42] Z. Hu, and F. Li, “Cost-Benefit Analyses of Active Distribution Network Management, Part II: Investment Reduction Analysis,” *IEEE Transactions on Smart Grid,* vol. 3, no. 3, pp. 1075-1081, 2012.

Wangwei Kong received the B.Eng degrees in Electrical Engineering at University of Bath (U.K.) and North China Electrical Power University (China), in 2015, and MSc degree in Electrical Power System from University of Bath in 2016. She is currently a PhD student at University of Bath. She is investigating into phase imbalance in distribution networks.

Kang Ma is working as a lecturer at University of Bath. His research focuses on the operation and planning of electrical distribution networks. In particular, he has expertise in the research of distribution network phase imbalance. He worked as an R&D engineer at China Electric Power Research Institute (Beijing) from 2011 to 2014. He received his PhD degree in Electrical Engineering from the University of Manchester (U.K.) in 2011 and his B.Eng. degree from Tsinghua University (China) in 2007.

Lurui Fang received his B.Eng. degree in Electrical Power Engineering at Chongqing University of Science and Technology in 2013. He received the MSc degree in Electrical Power Engineering at University of Southampton in 2014. He is currently a PhD student at University of Bath. His research focuses on phase imbalance in low voltage (0.4kV) distribution networks.



Renjie Wei is a PhD student at University of Bath, UK. He received his B.Eng degree in Electrical Engineering from North China Electric Power University (China) and University of Bath in 2018, following a “2+2” joint undergraduate program.

Furong Li (SM’09) was born in Shannxi province, China. She received the B.Eng. degree in electrical engineering from Hohai University, Nanjing, China, in 1990 and the Ph.D. degree from Liverpool John Moores University, Liverpool, U.K., in 1997. She is a Professor and the director of Center for Sustainable Power Distribution, University of Bath, U.K. Her major research interest is in the area of power system economics and markets.