Representative feeders for spatial scaling of stochastic PV hosting capacity

Arpan Koirala, Md Umar Hashmi, Dirk Van Hertem
ESAT - Electa, KU Leuven & EnergyVille
Genk, Belgium
arpan.koirala@kuleuven.be

Reinhilde D'hulst
VITO & EnergyVille,
Genk, Belgium
reinhilde.dhulst@vito.be

***Abstract*—The photovoltaics (PV) hosting capacity (HC) of the power system infrastructure is an important planning problem. The energy transition is happening now, resulting in the addition of new load types and generation in the low voltage distribution network. With these new loads and generation devices connected to the distribution network, a new planning approach is required that considers their stochastic nature. Computing the stochastic HC for all individual low voltage distribution feeders is challenging, as a small service area can have hundreds of feeders. This work aims to capture appropriate clustering schemes and the most relevant features of low voltage distribution feeders so that an accurate estimate of the hosting capacity of the full service area can be calculated by scaling up the hosting capacity of a small number of representative feeders. Case studies in actual feeders from suburban Spain showed that representative feeders obtained from feature reduction and appropriate cluster size could be used to scale the stochastic PV HC of an extensive service area. The case study showed that 3% of the total feeders could estimate the PV HC of the LVDS feeders in the large service area by selecting appropriate features.**

***Keywords*—distribution system, renewable hosting capacity, spatial scaling, representative feeders, low voltage, photovoltaics.**

I. Introduction and Motivation

The hosting capacity (HC) of a low voltage distribution system (LVDS) refers to "the amount of new generation or consumption that can be accommodated on a given feeder without impacting system operation under existing control and infrastructure configuration" [1]. In [2], it was seen that LVDS hosting capacity to new photovoltaics (PV) is a multidimensional stochastic problem due to many uncertainties, which can be classified into three types: a) power generation and load consumption uncertainties, b) PV scenario uncertainties, i.e., size, connection phase and type of the installations, and c) feeder uncertainties, i.e., feeder type and feeder size. The first type of uncertainty can be regarded as an operational uncertainty that depends on weather and consumer behaviour. The second is a planning level uncertainty, dealing with the size, type and location of PV installed. The third uncertainty emerges from the features of the feeder involved. The distribution network is the final capillary of the power grid. They are a significant part of the power network by length and spread. A small suburban town may have 160 such LV feeders [3], while the total number of LV feeders operated by a single operator can go up by tens of thousands. However, the variability in the feeder's features is usually neglected, and the HC is calculated only for a few selected feeders [2].

The easiest but computationally expensive solution for getting HC of a network, considering the uncertainty of the feeder features, is through individual analysis of each feeder. This feeder-wise calculation is an easy solution if the HC calculation method is deterministic or the service area is small. However, it might not be feasible if computationally expensive stochastic HC calculation methods [2] are used or the service area is quite large [4]. The limited availability of LVDS feeder data and load data also limits the feederwise analysis of the complete network. The other alternative is to estimate the HC of the whole network based on a smaller set of representative feeders [5].

PNNL has published 24 synthetic representative feeders [6] based on the statistical features of medium voltage (MV) grids. Similarly, EPRI has published six representative MV feeders sanitized for scientific use [7]. In [8], 9 MV and 8 LV prototype feeders are presented to represent Australian feeders after the analysis of 204 members of the MV and 8858 members of the LV database. In [5], a set of 383 feeders from England were analyzed and clustered to get 11 representative feeders. One of such representative feeders was also published as the IEEE European LV test feeder [9]. In [5], the usefulness of the representative feeders was shown to identify the group of feeders which can host 100% PV. It was shown that if a representative feeder has 100% PV HC capacity, then all the feeders in the subset represented by that feeder also have 100% PV HC. In [4], it was shown that the PV HC trend of an LVDS system with 50 000 feeders can be represented by a set of 2% of randomly chosen feeders. In [10], representative feeders for a set of 24,000 Austrian LVDS feeders were obtained to identify the limiting constraints of the group represented by those feeders, i.e., thermal or voltage, for a high PV penetration. These results were for simplified deterministic PV HC calculations with no other uncertainties considered [5], only the location of PV considered as uncertain [4], or only the limiting constraint is observed [10]. So far, the existing studies do not consider scaling stochastic PV HC in a spatial dimension using representative feeders. Furthermore, there are no studies regarding the selection of suitable features of low voltage feeders, such that the representative feeders represent the PV HC of their respective groups.

The main contributions of this paper are as follows:

a) a methodology is proposed to obtain the appropriate number of the representative feeders, whose stochastic PV HC can be used to estimate the cumulative PV HC of the network,

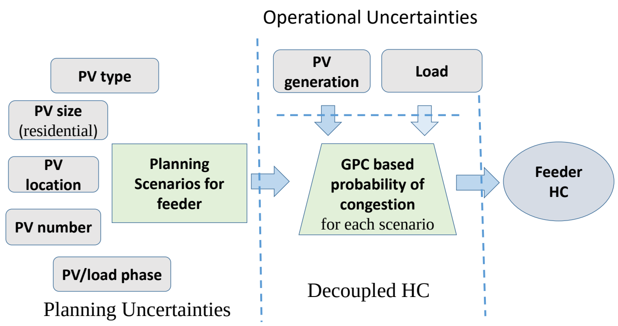

Fig. 1: Decoupling of planning and operational uncertainties to calculate PV HC [11]

b) a method to identify the most relevant features of LVDS feeder for stochastic PV HC scaling is proposed (Section II),
c) a case study is demonstrated for a) and b) using the set of European LV feeders from [3] (Section III).

II. Methodology

A. PV Hosting Capacity

The main objective of this paper is to capture representative feeders for a given LVDS so that an accurate estimate of the PV HC of the full service area can be calculated by scaling up the HC of a small set of those feeders. The representative feeders are obtained through clustering schemes with the relevant features of LVDS feeders. The PV HC calculation method used in this paper is the decoupled PV HC method proposed in [11]. This stochastic HC method splits the planning level uncertainties from the operational uncertainties (Fig. 1), and the best planning scenario out of the considered scenarios is considered to be the PV HC of the feeder. Defining PV HC itself is a well-debated topic [2], and not in the scope of this paper. Hence, a detailed discussion on the methodology to calculate PV HC of LVDS feeders is skipped in this paper and interested readers are referred to [2] and [11].

Large scale LVDS can have hundreds of small feeders whose PV HC needs to be calculated independently. Calculating HC is a computationally intensive process and can substantially be reduced by using representative feeders generated from the relevant features of the feeders in the network.

B. Representative Feeders

The first step for obtaining representative feeders is to create clusters of closely linked feeders, and crucial to this is how the *distance* between feeders is defined and which features of the feeders are taken into account to define this distance. In this paper, the following features are available for all feeders:

(1) number of consumers connected to a feeder,
(2) yearly consumption per customer in kWh,
(3) yearly reactive consumption per customer in kWh,
(4) total conductor length in km,
(5) main path length in km,
(6) average length to the customer in km,
(7) total line impedance in Ohm, and
(8) average path impedance in Ohm.

The steps involved in finding appropriate representative feeders suitable for scaling PV HC can be summarized as follows:

1) Identify clusters of closely related LVDS feeders.
2) Choose one representative feeder per cluster.
3) Calculate stochastic PV HC of representative feeders.
4) Scale up the HC of representative feeders to the full set LVDS feeders.

The first step of the process is clustering. There are several clustering methods, of which *k*-means is most often used for LVDS feeder classification [5], [10], [12]. For *k*-means, an unsupervised clustering algorithm, the number of clusters and hence the number of representative feeders is an input.

The next step is then to find the representative feeder of each cluster, i.e., the closest feeder to the centroid. A parameter $x_{i,\psi}$ is in standard form when

$$x_{i,\psi} = \frac{\gamma_{i,\psi} - \bar{\gamma}_k}{\text{std}(\gamma_\psi)}, \tag{1}$$

where $\gamma_{i,\psi}$ is the measurement value of the feature ψ for feeder i, $\bar{\gamma}_\psi$ denotes the mean value of that feature, and std(γ) denotes the standard deviation of the feature. One representative feeder from each of these clusters is selected using the parameter in the standard form. Consider, x_k^{min} and x_k^{max} denotes the minimum and maximum value of $x_{i,\psi}$ in a cluster. We define the normalized distance of a given feature and a feeder in that cluster as

$$P_{i,\psi} = \frac{x_{i,\psi} - x_\psi^{\text{min}}}{x_\psi^{\text{max}}}. \tag{2}$$

The representative feeders are selected by identifying ζ_{rep} given as (3) for all clusters.

$$\zeta_{\text{rep}} = \min_i \sum_\psi |P_{i,\psi}|. \tag{3}$$

The unknowns in these steps are: a) how many clusters (and representative feeders) do we need to get the scaled up PV HC as close to the fully calculated one? And b) which features are more relevant for PV HC scaling? This paper presents a case study in European LVDS feeders to answer these questions.

1) Appropriate number of representative feeder: The number of clusters required to represent the overall network is unclear. One of the ways is through the goodness of a cluster measured using the mean silhouette index. The silhouette coefficient of a node is a confidence indicator of its association in a cluster [13]. Rousseeuw in [14] proposed an interpretation based on the value of silhouette coefficient: where 0.71 to 1.0 implies a strong structure. However, there is no guarantee that the clusters based on the this coefficient will correctly represent the similarity in PV HC. Furthermore, the cluster size is increased to see if the silhouette index-based cluster size is appropriate choice for getting the best representative feeder for scaling LVDS PV HC.

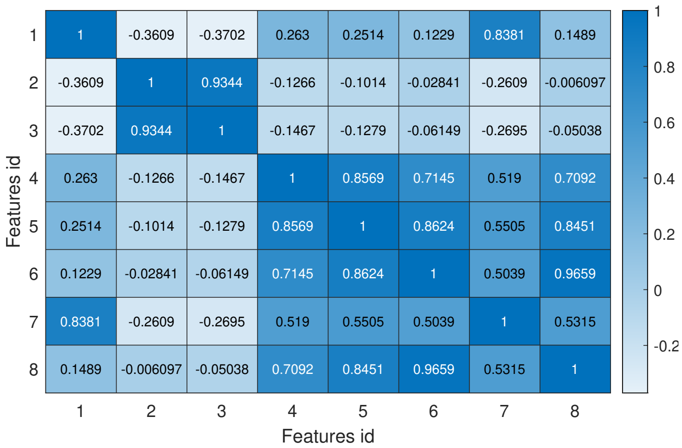

Fig. 2: Covariance matrix of normalized to standard form used for feature correlationship.

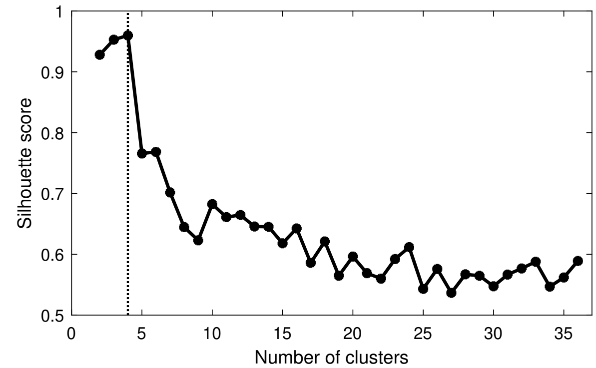

Fig. 3: Mean silhouette score for different sized feeder clusters for 160 Spanish distribution network feeders

2) Choosing appropriate features: The covariance matrix of the eight parameters in standard form is shown in the heatmap shown in Figure 2. Note the high correlationship between (i) active and reactive power, (ii) total number of consumers and total line impedance, and (iii) features (4), (5), (6), and (8). This covariance matrix can be used to reduce the features taken for clustering. However, it does not say anything about if these feature reduction improves the scaling of the PV HC. As seen in [5], the mapping of the PV HC may not be straightforward in all clusters. The clustering process is repeated with a reduced number of features and different combinations of features to fine-tune the result. The aim is to find the relevant set of features of LV feeders, which will assist the scaling of the stochastic PV HC spatially.

C. Scaling PV HC

Once a set of representative feeders is obtained using the features and defined number, the PV HC of those feeders is compared with the cumulative PV HC of the respective cluster. The term *scaling* here refers to the use of representative feeders' PV HC to estimate the PV HC of bigger service area in distribution network. The stochastic decoupled PV HC [11] is used. Herein, two terminologies used in the remaining paper are introduced: *total* HC, meaning the sum of PV size that can be installed in the feeder and *normalized* HC, meaning the average PV size per consumer. Both total and normalized HC is compared while scaling using the representative feeders.

For this comparison, two indices are proposed:

Mean average absolute error (MAAE): In this index the absolute difference between HC of representative feeder (HC_κ^{rep}) and mean of HC of all feeders in the cluster κ ($\bar{HC}_\kappa$) is summed for all clusters and normalized by total number of cluster K:

$$\text{MAAE} = \sum_{\kappa} \frac{|HC_\kappa^{\text{rep}} - \bar{HC}_\kappa|}{K} \tag{4}$$

Difference from the representative feeder (DR): In this index, difference of HC of each feeder i in cluster c from their representative feeder is measured:

$$\text{DR}_{(i,\kappa)} = HC_\kappa^{\text{rep}} - HC_{(i,\kappa)} \tag{5}$$

The output will be in the form of distribution of error for a particular clustering scenario from which mean (DR_μ), and variance (DR_σ) of the difference can be obtained.

In the case-study following, increasing or decreasing the number of clusters and use of different combination of features is done to observe the error on scaling using indices MAAE and DR. The result will be a comment on using the representative feeders to scale the stochastic PV HC:

a) can the representative feeders using all features be used to obtain the cumulative PV HC of an extensive distribution service area?
b) if not, what are the most relevant features of LVDS when scaling the stochastic PV HC spatially for a large service area?
c) what is the influence of increasing or decreasing the number of clusters in scaling the PV HC?

III. Case Studies

In this section, feeders from an actual European LVDS [3] is used to show the case studies in scaling of the PV HC spatially. The mean silhouette score for different k values of clusters is shown in Figure 3. It can be observed that for $k = 4$, the mean silhouette score exceeds 0.95, and therefore, four clusters are to be formed according to this index. The identified representative feeders for clustering based on maximum silhouette score are shown in Table I. The increased cluster size has silhouette score above 0.51, which means they still have a reasonable structure. The different cluster size on both sides of the one with maximum silhouette score is evaluated to check the significance of this score in PV HC scaling. The aim is to obtain an appropriate number representative feeders for a given threshold of error in scaling.

The first part of this section will deal with impact of increasing the number of representative feeders scaling of HC. The second case study is about impact of scaling based on different features. In the third study, impact of both feature selection and increasing the number of representative feeders is shown. All the experiments are repeated for 100 times to get a mean value to reduce the error due to uncertainty in cluster centroids.

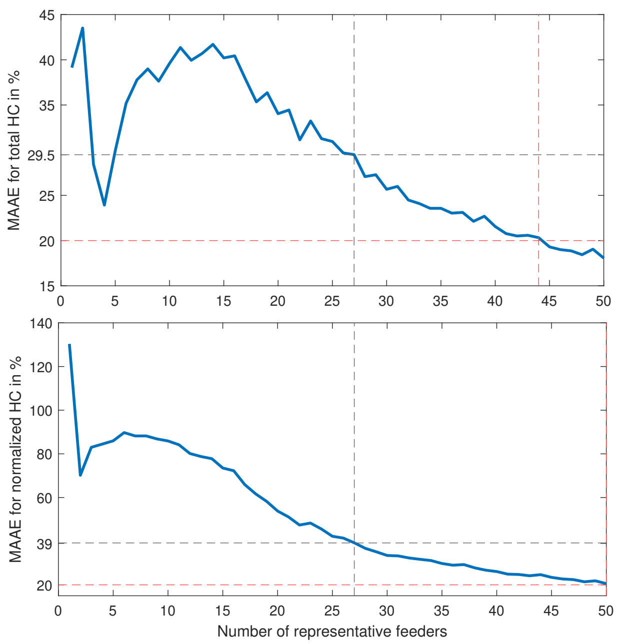

Fig. 4: MAAE of stochastic PV HC scaling from representative feeders using all features

A. *Role of number of representative feeders in scaling*

According to the silhouette score, the best suited number of clusters is chosen to be four while. This clustering considers all the eight features described in Section II(Fig. 3). On close observation, cluster 1 had the majority of residential feeders. In contrast, cluster 4 contained all the special feeders: such as feeders without any load on them or serving consumers like hospitals or sport-centres. For further calculations, the feeders of cluster 4 were removed to avoid unnecessary bias. The 136 feeders in clusters 1 to 3 were used for scaling PV HC.

In Fig 4, it is seen that the MAAE of total HC of the network while using the representative feeder based on maximum silhouette score is 24%, when all features were used. The MAAE increases on increasing the number of representative feeders till some point and then starts decreasing. The MAAE of normalized HC is 60% at the maximum silhouette score. Similar to total HC, it increases by increasing the number of representative feeders first and then decreases. This inconsistent behaviour means we need to specify a threshold for MAAE of total HC and normalized HC to get the required number of representative feeders. For a threshold of 20%, the number of needed representative feeders is 44 and 50, respectively. If the maximum of these two is taken, 50 representative feeders (≈35% of total feeders) are required with an error threshold of 20% MAAE.

Alternatively, MAAE is observed for a fixed number of representative feeders. In this case, 20% of the total feeders were selected, i.e., 27 representative feeders. The MAAE is 29.5% and 39% for the total and normalized HC, respectively. Hence, the conclusion obtained from this study is mixed. Yes, the representative feeder can estimate the PV HC of a larger network. However, better estimation is obtained only by increasing the number of representative feeders. However, the result might differ for more extensive datasets with thousands of feeders, which we cannot say as the extensive network data are scarce for research. Furthermore, using all features to obtain representative feeders is not efficient enough, as we need a detailed analysis of 35% of feeders for MAAE less than 20%.

B. *Feature Selection*

The role of different features in scaling the stochastic PV HC is studied in this section. The correlation analysis in Fig. 2 shows the relation among the features, namely features (1) and (4), features (2) and (3), and features (4), (5), (6), and (8) are correlated but does not show their role in PV HC scaling. Three clustering scenarios are considered to get a better understanding of these features in PV HC scaling:

a) C1: Clustering using all features,
b) C2: Clustering using the individual features only, and
c) C3: Clustering using a combination of selected features.

A rank-based method is proposed to find better features for PV HC scaling. Eight different indices based on MAAE and DR are calculated:

a) **I1**: the number of representative feeders when an error threshold of 20% is fixed for the MAAE of total HC,
b) **I2**: the number of representative feeders when an error threshold of 20% is fixed for the MAAE of normalized HC
c) **I3**:MAAE of total HC when 20% of feeders are taken as representative feeders,
d) **I4**: MAAE of normalized HC when 20% of feeders are taken as representative feeders,

TABLE I: Representative feeders selected

	Feeder ID	Cluster 1 144	Cluster 2 52	Cluster 3 27	Cluster 4 23
(1)	No. of consumer	44	20	2	0
(2)	Avg active/year [kWh]	2068.79	4256.73	5424.98	0
(3)	Avg reactive/year	202.84	415.14	484.25	0
(4)	Total line length [km]	0.38	0.15	0.31	0.25
(5)	Main path [km]	0.21	0.13	0.29	0.22
(6)	Avg length to costumer [km]	0.15	0.13	0.16	0.0
(7)	Total line impedence [ohm]	4.60	2.52	0.30	0.67
(8)	Average path impedence [ohm]	0.10	0.13	0.15	0
	Indices	0.17	1.19	1.65	1.62

TABLE II: Rank of different indices when taking representative feeders, considering all features (C1) and individual features (C2)

Features	I1	I2	I3	I4	I5	I6	I7	I8	Rank sum
All	9	9	9	9	9	9	6	3	63
(1)	2	4	2	3	4	5	7	2	**29**
(2)	7	2	8	2	8	6	8	8	49
(3)	2	2	6	7	6	8	9	9	49
(4)	2	5	3	5	1	2	4	5	**27**
(5)	2	7	5	4	5	1	3	4	31
(6)	2	1	1	8	2	3	1	7	**25**
(7)	7	6	4	1	3	7	5	1	34
(8)	1	8	7	6	7	4	2	6	41

TABLE III: Rank of different indices when taking representative feeders considering all features (C1), individual features (C2), and combinations of features (C3)

Features	I1	I2	I3	I4	I5	I6	I7	I8	Rank sum
All	13	13	13	12	13	11	10	5	90
(1)	3	6	2	4	6	6	11	3	41
(2)	8	2	10	3	10	7	12	12	64
(3)	3	2	8	9	8	9	13	13	65
(4)	3	7	3	7	3	3	8	8	42
(5)	3	9	7	5	7	2	5	6	44
(6)	3	1	1	10	4	4	3	11	**37**
(7)	8	8	5	1	5	8	9	1	45
(8)	1	10	9	8	9	5	4	10	56
(4), (6)	1	2	4	6	1	10	2	9	**35**
(1), (4)	11	11	12	11	11	12	6	4	78
(1), (6)	12	12	11	13	12	13	1	7	81
(1), (4), (6)	10	5	6	2	2	1	7	2	**35**

e) **I5**: absolute value of DR_μ of total HC when 20% of feeders are taken as representative feeders,
f) **I6**: absolute value of DR_μ of normalized HC when 20% of feeders are taken as representative feeders,
g) **I7**: DR_σ of total HC when 20% of feeders are taken as representative feeders,
h) **I8**: DR_σ of normalized HC when 20% of feeders are taken as representative feeders,

At first, cases C1 and C2 are taken, and the features with the first three ranks are taken to form C3. Finally, a new ranking is done, taking C1, C2 and C3 together. The rank matrix for when taking only C1 and C2 is shown in Table II, from which features (1), (4) and (6) had the top three sums of the rank. The C3 scenarios are built from the combination of (1), (4) and (6), and a new rank table is formed using all C1, C2, and C3.

From Table III, it is seen that feature (6) and the combinations (4),(6) and (1), (4), (6) had better performance based on the ranking of all eight indices.

We can reasonably conclude that taking all features with equal weight would not necessarily lead to better representative feeders. However, the features such as the number of consumers connected to a feeder (feature 1), total conductor length in km (feature 4), and average length to the consumer (feature 6) have a lower error on scaling PV HC when taken individually. Similarly, the combination of these three features had a better ranking among all other tested combinations.

C. Impact of feature selection

In this section, the three top ranking feature combination from the previous study is evaluated to compare the improvement from the first case study with all features in terms of MAAE. The two factors considered were the reduction in MAAE on taking 20% of total feeders as representative feeders and the number of representative feeders required for a threshold of 20% of MAAE of both the total and normalized HC.

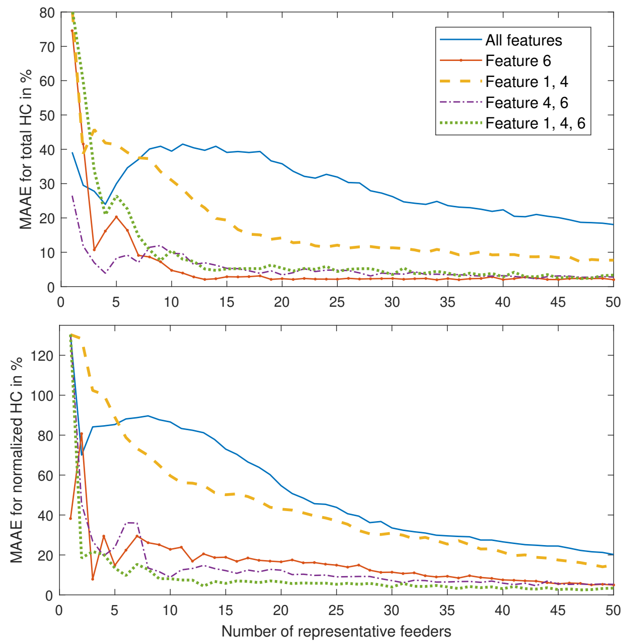

Fig. 5: MAAE of stochastic PV HC from representative feeders after feature reduction

From Table IV and Fig. 5, two main conclusions can be drawn. First, the feature reduction reduces the number of representative feeders required by 94% if the threshold of MAAE is fixed to 20%. On the other hand, if 20% of feeders were taken as representative feeders, the MAAE threshold for both normalized and total HC reduces to 5.29% compared to 39.2% when using all features (86.5% reduction).

Representative feeders based on average length to the consumer (Feature 6) have better PV HC scaling results alone and combined with total conductor length and number of consumers connected (Features 1, 4 and 6). However, as seen in Fig. 5, the combination of Features 1 and 4 only does not lead to better scaling of PV HC. It can be fairly concluded that using the right features for clustering the LVDS feeders means 3% of the total feeders can be used as representative feeders to scale the HC of the whole network while having 20% MAAE.

IV. Conclusion

The conclusion of this stochastic PV HC scaling exercise in this paper is two folds. The first one is that, yes, representative feeders with specific cluster size and features can be used to

TABLE IV: Comparison after feature selection

	no of rep. feeders MAAE ≤20%					MAAE for 20% total feeder			
Features	total HC	norm. HC	max	% of total feeder	% reduction	total HC	norm. HC	max	% reduction
All	44	50	50	37%	-	29.50	39.20	39.20	-
(6)	3	3	3	2.2%	94%	2.20	14.90	14.90	62%
(4),(6)	2	4	4	3%	92%	3.98	9.21	9.21	76.5%
(1), (4), (6)	7	5	7	5.22%	86%	5.20	5.29	5.29	86.5%

scale the HC of the representative feeder to get a cumulative HC of a larger service area. The final number would be nearly accurate as obtained by adding the HC of individual feeders, thereby reducing the need to calculate PV HC of all feeders. Proper feature selection can reduce the number of representative feeders required to scale the PV HC of the LVDS. Concretely, only 3% of feeders could estimate the PV HC of the whole service area with 20% of error. On the other hand, this type of estimation neglects all the variation in different feeders, meaning the HC of any random feeder in a cluster cannot be extracted precisely. The only solution in such a case is to have HC of all feeders independently.

For the test network, it was seen that maximum silhouette score based clustering was not good enough to scale the PV HC. However, it could be used to separate the special feeders. On increasing the cluster size, the MAAE starts saturating at one level, which can be used to select the appropriate number of clusters. Similarly, for the features selection, it was seen that using all the features can cause overfitting. In contrast, the representative feeders based on individual features had better performance on scaling PV HC. The salient features identified are: the number of consumers connected to the feeder, total conductor length, and average length to the consumer had less error in the scaling, both individually and in combination. The selection of salient features reduce the error in scaling by 86.5% in the test case and number of representative feeders required by 94% for 20% threshold on MAAE.

ACKNOWLEDGEMENT

The work is supported by the energy transition funds project BREGILAB organized by the FPS economy, S.M.E.s, Self-employed and Energy.

REFERENCES

[1] M. H. J. Bollen and S. K. Rönnberg, "Hosting capacity of the power grid for renewable electricity production and new large consumption equipment," *Energies*, vol. 10, no. 9, 2017.

[2] A. Koirala, T. Van Acker, R. D'hulst, and D. Van Hertem, "Hosting capacity of photovoltaic systems in low voltage distribution systems: A benchmark of deterministic and stochastic approaches," *Renewable and Sustainable Energy Reviews*, vol. 155, p. 111899, 2022.

[3] A. Koirala, L. Suárez-Ramón, B. Mohamed, and P. Arboleya, "Non-synthetic european low voltage test system," *International Journal of Electrical Power & Energy Systems*, vol. 118, p. 105712, 2020.

[4] R. Torquato, D. Salles, C. Oriente Pereira, P. C. M. Meira, and W. Freitas, "A comprehensive assessment of pv hosting capacity on low-voltage distribution systems," *IEEE Transactions on Power Delivery*, vol. 33, no. 2, pp. 1002–1012, 2018.

[5] 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.

[6] K. P. Schneider, Y. Chen, D. Engle, and D. Chassin, "A taxonomy of north american radial distribution feeders," in *2009 IEEE Power Energy Society General Meeting*, 2009, pp. 1–6.

[7] EPRI. (2017) EPRI test circuits.

[8] Y. Li and P. J. Wolfs, "Taxonomic description for western australian distribution medium-voltage and low-voltage feeders," *IET Generation, Transmission Distribution*, vol. 8, no. 1, pp. 104–113, 2014.

[9] TFWG, "IEEE PES distribution systems analysis subcommittee radial test feeders." 2016. [Online]. Available: http://sites.ieee.org/pes-testfeeders/

[10] B. Bletterie, S. Kadam, and H. Renner, "On the classification of low voltage feeders for network planning and hosting capacity studies," *Energies*, vol. 11, no. 3, p. 651, 2018.

[11] A. Koirala, M. U. Hashmi, R. D'hulst, and D. Van Hertem, "Decoupled probabilistic feeder hosting capacity calculations using general polynomial chaos," *PSCC*, 2022.

[12] J. Dickert, M. Domagk, and P. Schegner, "Benchmark low voltage distribution networks based on cluster analysis of actual grid properties," in *2013 IEEE Grenoble Conference*. IEEE, 2013, pp. 1–6.

[13] J. Ding, Q. Zhang, S. Hu, Q. Wang, and Q. Ye, "Clusters partition and zonal voltage regulation for distribution networks with high penetration of pvs," *IET Generation, Transmission & Distribution*, vol. 12, no. 22, pp. 6041–6051, 2018.

[14] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.