Probabilistic Estimation of PV Hosting Capacity in DER-Rich Feeders Using Smart Meter Data

Liming Liu, *Student Member*, *IEEE*, Naihao Shi, *Student Member*, *IEEE*, Zhaoyu Wang, *Senior Member*, *IEEE*, Matthew J. Reno, *Senior Member*, *IEEE*, Joseph A. Azzolini, *Member*, *IEEE*

***Abstract*—The growing penetration of photovoltaic (PV) systems in distribution networks (DNs) highlights the significance of PV hosting capacity (HC) estimation for system planning and operation. Compared to traditional model-based methods, data-driven HC estimation offers superior computational efficiency and scalability, gaining widespread attention. This paper aims to address the limitations of previous data-driven approaches for residential PV HC estimation in DER-rich feeders by proposing a probabilistic HC estimation framework. The framework first determines transformer pairings using available smart meter (SM) data and limited network information. Next, an optimization-based voltage sensitivity estimation model accurately estimates voltage sensitivity at customer nodes in the target low-voltage secondary network (SNet). Finally, a Gaussian mixture density network characterizes the head bus voltage of each SNet and derives the distribution of the HC at customer nodes. Numerical results and method comparisons on the EPRI Ck5 circuit validate the effectiveness of the proposed framework.**

***Index Terms*—Data-driven method, PV system, hosting capacity, voltage sensitivities, smart meter, distribution system.**

I. Introduction

With the increasing penetration of distributed energy resources (DER), particularly photovoltaic (PV) systems, in distribution networks (DN), hosting capacity (HC) estimation has become essential for distribution power system design, planning and operation. PV HC is typically defined as the maximum PV capacity that can be reliably integrated into the DN without violating operational constraints, such as voltage deviation limits. Traditionally, PV HC is estimated by iteratively running power flow on the system electric model until any constraint is violated. While this model-based approach provides accurate HC results, its high computational burden and reliance on a complete and accurate electrical model make it impractical for large DNs. This challenge is particularly significant in low-voltage secondary networks (SNets), where acquiring a detailed electrical model is often difficult. Given these limitations, data-driven HC estimation methods have gained significant attention, facilitated by the advancements in smart meters (SM). In previous studies, [1] proposed a spatial-temporal LSTM model to achieve online dynamic HC calculation. However, this approach still requires an accurate electric model to generate the node HC values and power flow data needed for offline training. Leveraging SM data, [2] developed a linear regression model (LRM) for fast HC estimation in low-voltage SNets. Although this model effectively estimates voltage sensitivities and HC, it may struggle with DER-rich feeders, where complex power flows and fluctuating voltages on the primary side of SNets pose more challenges. To mitigate the impacts of voltage variations in SNets, [3] designed a physics-inspired neural network for voltage estimation using only SM data (validated on partially synthetic SM data with voltage values generated from OpenDSS). However, this approach can be limited when incomplete SM coverage or data missing issues exist. Additionally, [4] focused on voltage-constrained HC by developing a neural network to capture power flow relationships among customer nodes within the same low-voltage network. [5] introduced a hybrid approach that integrates deep learning with nonlinear functions to compute node voltages for HC estimation. However, this method does not consider low-voltage SNets, where residential PV systems are typically connected.

As DER penetration increases, feeders become DER-rich, leading to complex power flows and voltage variations influenced by voltage regulators in the primary DN [6]. These factors, along with incomplete SM coverage and data quality issues, challenge accurate data-driven HC estimation in SNets. To fill the gap, a data-driven framework is proposed for residential HC estimation. In the framework, an optimization-based voltage sensitivity estimation model is developed to accurately estimate voltage sensitivity at customer nodes in the target SNet. Furthermore, a Gaussian mixture density network (GMDN) is employed to characterize the head bus voltage of each SNet and derive the distribution of HC at customer nodes. The final output includes the confidence interval (CI) of customers' HC, providing reliable estimation results for utilities.

II. Data-Driven Probabilistic HC Estimation Framework

The paper proposes a comprehensive probabilistic HC (PHC) estimation framework for DER-rich feeders, as illus-

This material is based upon work supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under the Solar Energy Technologies Office Award Number 38426. This article has been authored by an employee of National Technology & Engineering Solutions of Sandia, LLC under Contract No. DE-NA0003525 with the U.S. Department of Energy (DOE). The employee owns all right, title and interest in and to the article and is solely responsible for its contents. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this article or allow others to do so, for U.S. Government purposes. The DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan https://www.energy.gov/downloads/doe-public-access-plan. National Science Foundation ECCS 2042314. Corresponding Author: Zhaoyu Wang. Liming Liu, Naihao Shi, and Zhaoyu Wang are with the Department of Electrical and Computer Engineering, Iowa State University, Ames, IA 50011 USA; Matthew J. Reno and Joseph A. Azzolini are with Sandia National Laboratories, Albuquerque, NM 87123 USA.

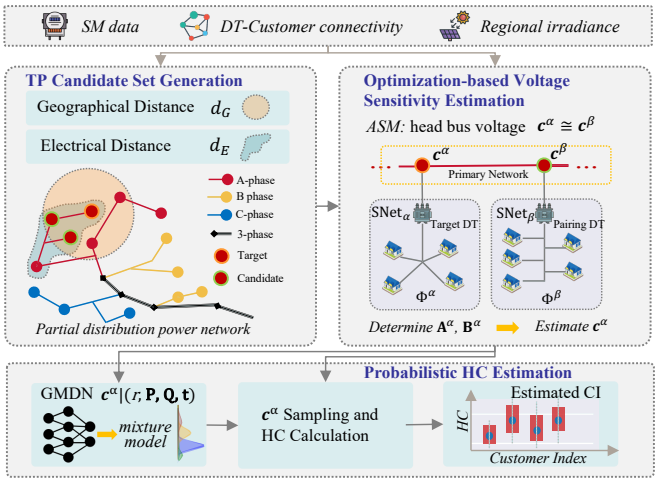

Fig. 1. The structure of the proposed probabilistic HC estimation framework

trated in Fig. 1, comprising three main steps.

A. Transformer-Paring Candidate Set Generation

The basic idea that we utilize to mitigate the impacts from varying voltage in the head bus of SNet is to assume the two adjacent distribution transformers (DT) approximately share the primary winding voltage. By jointly estimating the voltage sensitivities of two SNets, the influence of head bus voltage can be eliminated. Since the electrical model could be unavailable in data-driven hosting capacity estimation, a rule-based transformer-pairing (TP) method is designed to identify the adjacent DTs. The TP is determined based on two criteria: selected geographical distance (d_G) and specified electrical distance (d_E). Given the lack of a network model, each DT's location is approximated using the centroid of its connected end users, while the d_E between two DTs is defined as the average Pearson correlation coefficient between the voltage measurements of any two users connected to them. For a given DT i, the TP candidate set is defined as:

$$S_i = \left\{(i,j) \mid d_G(j) \leq d_G^{(k)}, d_E(j) \leq d_E^{(u)}\right\}$$

where $d_G^{(k)}$ and $d_E^{(u)}$ represent the k-th smallest geographical distance and the u-th smallest electrical distance, respectively. This TP candidate set serves as the input for the next steps.

B. Optimization-based Voltage Sensitivity Estimation

The joint estimation of SNet voltage sensitivities is achieved by the designed optimization model for each TP from S_α, which represents the set of TP candidates for the target DT α. The optimization model aims to determine the active power voltage sensitivity $\mathbf{A}^\alpha$ and reactive power voltage sensitivity $\mathbf{B}^\alpha$ of the target SNet and the SNet connecting to the pairing DT β by minimizing the difference between their primary-side voltages. In this work, we focus on single-phase SNets, a common configuration in North America, where voltage sensitivities exhibit properties such as non-positivity and symmetry. While the current study is limited to single-phase SNets, the underlying framework is general and can be extended to more complex low-voltage SNets, including two-phase and three-phase networks, which is part of our ongoing research. The constraints incorporate the LinDistFlow model for SNets and the inherent characteristics of voltage sensitivities [3]. The SM data collect bus injection loads and voltages at the customer end. The optimization formulation is as follows:

$$\min_{\mathbf{A}^x, \mathbf{B}^x \ \forall x \in \{\alpha,\beta\}} \sum_{t=1}^{T} \left(\sum_{l=1,k=1}^{|\Phi^\alpha|,|\Phi^\beta|} \|c_{t,l}^\alpha - c_{t,k}^\beta\|_2 \right) / T \quad (1)$$

subject to

$$-c_{t,l}^\alpha = \boldsymbol{p}_{t,.}^\alpha \mathbf{A}_{.,l}^\alpha + \boldsymbol{q}_{t,.}^\alpha \mathbf{B}_{.,l}^\alpha - \boldsymbol{v}_{t,l}^\alpha, \quad \forall l \in \Phi^\alpha \quad (2a)$$

$$-c_{t,k}^\beta = \boldsymbol{p}_{t,.}^\beta \mathbf{A}_{.,k}^\beta + \boldsymbol{q}_{t,.}^\beta \mathbf{B}_{.,k}^\beta - \boldsymbol{v}_{t,k}^\beta, \quad \forall k \in \Phi^\beta \quad (2b)$$

$$\mathbf{A}^\alpha - \mathbf{A}^{\alpha\top}, \mathbf{A}^\beta - \mathbf{A}^{\beta\top} \leqslant \xi \quad (2c)$$

$$\mathbf{B}^\alpha - \mathbf{B}^{\alpha\top}, \mathbf{B}^\beta - \mathbf{B}^{\beta\top} \leqslant \xi \quad (2d)$$

$$\mathbf{A}^\alpha, \mathbf{A}^\beta, \mathbf{B}^\alpha, \mathbf{B}^\beta \leqslant 0 \quad (2e)$$

where $(\cdot)^\alpha$ and $(\cdot)^\beta$ denote the variables and parameters related to the SNets connected to the target DT and pairing DT. For instance, $\boldsymbol{p}_{t,.}^\alpha$, $\boldsymbol{q}_{t,.}^\alpha$, and $\boldsymbol{v}_{t,.}^\alpha$ represent the active power, reactive power, and squared voltage measurements of customers under the target DT at time t; Φ^α is the customer set connected to the target DT, and $|\Phi^\alpha|$ denotes the cardinality of the set; $c_{t,.}^\alpha$ and $c_{t,.}^\alpha$ are the squared head bus voltages of the SNets at time t. In this optimization model, the LinDistFlow model is applied to both SNets to estimate the head bus voltages based on the SM data, as represented by constraints (2a) and (2b). The objective of the optimization is to minimize the difference between the primary voltages of the two SNets. To ensure the estimated voltage sensitivities satisfy non-positivity and symmetry properties, additional constraints (2c) to (2e) are introduced. Here, $\leqslant$ denotes element-wise inequality, and ξ is a small constant. For each pair (α, β) from S_α, the estimated $\mathbf{A}^\alpha$ and $\mathbf{B}^\alpha$ are obtained. The estimation results of these pairs with the smallest optimization loss will be selected as the final values for SNet α. Then, the c^α will be calculated, serving as the inputs of the next HC estimation step. This process is repeated for all DTs that require estimation.

C. Probabilistic HC Estimation

Previous studies on HC estimation have considered various uncertainties, such as PV generation and loads [7]. In contrast, this work specifically addresses uncertainties arising from voltage variations at the head bus of each SNet, while retaining the flexibility to incorporate other uncertainties. Voltage variations in the primary network can stem from multiple factors, including rapid DER generation changes, downstream load fluctuations, voltage regulator operations, and capacitor bank switching. Additionally, voltage variations can propagate from the upstream transmission network, further complicating the analysis. The combined influence of these factors makes it challenging to implicitly model the voltage. The proposed approach accounts for the limitations of SM data and partial datasets in accurately modeling complex power flows on the primary side via data-driven methods. In such cases, PHC estimation offers more informative insights for utilities than

Fig. 2. HC estimation results of the proposed method and the comparative LRM for customers 1108 to 1160 are shown in the figure. A deterministic model-based approach is used as the benchmark. Customers between the two vertical dashed lines are connected to the same SNet. The selected SNets include between one and six customers, covering a wide range of scenarios.

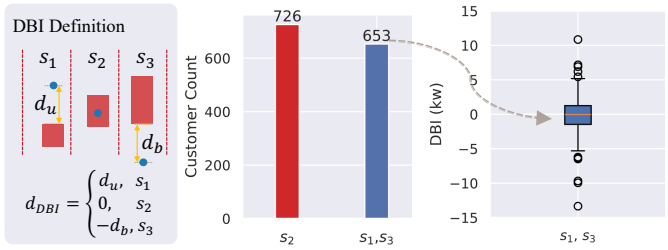

Fig. 3. Left diagram shows the DBI definition for three cases (s_1, s_2, s_3), where the distance is zero if the SHC lies within the interval, positive (d_u) if above, and negative ($-d_b$) if below. The middle figure displays the number of customers from EPRI Ckt5 system per case, and the right boxplot presents the distribution of distances for s_1 and s_3 (second case in middle figure)

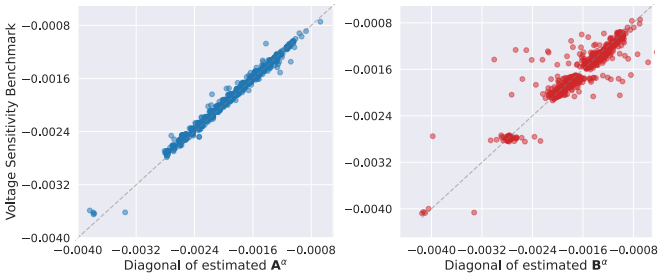

Fig. 4. Comparison of voltage sensitivities from the perturb-and-observe method and the proposed optimization method. Points closer to the dashed line indicate higher accuracy.

deterministic data-driven approaches. To achieve this, the first step is modeling the voltage $\boldsymbol{c}$. Given its potential multimodal distribution, GMDN model is introduced due to its capability to model complex distributions by parameterizing a mixture of Gaussian distributions conditioned on input features. Unlike traditional parametric models, GMDN leverages neural networks to learn the mixture parameters, e.g., the weights, means, and covariances, thereby providing a more flexible tool to capture the underlying uncertainty and multimodal nature of $\boldsymbol{c}$. In our case, the conditional probability distribution of $\boldsymbol{c}$ given external information $\mathbf{x}$ is approximated as a Gaussian mixture presented below:

$$p(\boldsymbol{c} \mid \mathbf{x}) \approx \hat{p}(\boldsymbol{c} \mid \mathbf{x}) = \sum_{g=1}^{G} \pi_g(\mathbf{x}) \mathcal{N}\left(\boldsymbol{c} \mid \boldsymbol{\mu}_g(\mathbf{x}), \boldsymbol{\Sigma}_g(\mathbf{x})\right), \quad (3)$$

where, $\pi_g(\mathbf{x})$ represents the mixing coefficient of the g-th component, satisfying $\sum_{g=1}^{G} \pi_g(\mathbf{x}) = 1$; $\boldsymbol{\mu}_g(\mathbf{x})$ denotes the mean of the g-th Gaussian component, dependent on $\mathbf{x}$. $\boldsymbol{\Sigma}_g(\mathbf{x})$ is the covariance matrix of the g-th Gaussian component, also parameterized by $\mathbf{x}$. Given inputs $\mathbf{x}$, the GMDN generates the mixture parameters $\{\pi_g(\mathbf{x}), \boldsymbol{\mu}_g(\mathbf{x}), \boldsymbol{\Sigma}_g(\mathbf{x})\}$ as outputs, and the whole network is trained by minimizing the negative log-likelihood loss [8]:

$$\mathcal{L} = -\log\left(\sum_{g=1}^{G} \pi_g(\mathbf{x}) \mathcal{N}\left(\boldsymbol{c} \mid \boldsymbol{\mu}_g(\mathbf{x}), \boldsymbol{\Sigma}_g(\mathbf{x})\right)\right). \quad (4)$$

By leveraging the flexibility of GMDN, the conditional probability distribution of $\boldsymbol{c}$ given $\mathbf{x}$ can be effectively captured, allowing for better handling of multimodal behaviors and improved uncertainty quantification.

Utilizing the estimated $\mathbf{A}^\alpha$ and $\mathbf{B}^\alpha$, the HC of customer z ($z \in \Phi^\alpha$) given SNet head bus voltage c^α, sampled from $\widehat{p}_{\boldsymbol{c}|\mathbf{x}}$, and the voltage constraint v_{ct} can be calculated as follows:

$$\kappa = \min_t([v_{\text{ct}}^2 - c_{t,z}^\alpha - \boldsymbol{p}_{t,\cdot}^\alpha \mathbf{A}_{\cdot,z}^\alpha - \boldsymbol{q}_{t,\cdot}^\alpha \mathbf{B}_{\cdot,z}^\alpha] \cdot {\mathbf{A}_{z,z}^\alpha}^{-1}). \quad (5)$$

Then, for the computed values of κ across all sampled voltages, m-th quantile $F_\kappa^{-1}(m)$ of it can be defined as $\inf\{z \in \mathbb{R} : F_\kappa(z) \geq m\}$. The HC CI for the target customer can then be determined. For example, the 95% CI is given by $\left[F_\kappa^{-1}(2.5\%), F_\kappa^{-1}(97.5\%)\right]$.

III. Numerical Results

To validate the proposed method, the EPRI Ckt5 circuit, consisting of 591 DTs and 1379 customers, is utilized [2]. A DER-rich scenario is simulated by integrating three utility-scale PV plants (1.1 MW) and two large EV charging stations (280 kW). Voltage regulators, e.g., load tap changers and capacitor banks, are also deployed. One year of hourly partially synthetic SM data for all customers, with voltage values generated from OpenDSS, is used as input.

For benchmarking the performance of the voltage sensitivity estimation model, the perturb-and-observe method is used as a benchmark, denoted as the "voltage sensitivity benchmark". This method estimates voltage sensitivities by perturbing a system input and observing the corresponding change in bus voltages [9]. Since self-voltage sensitivities play a key role in the following HC calculations, the self-sensitivity results from our method and the benchmarks are shown in Fig. 4. As shown in the left subfigure, most points lie along the diagonal dashed line, indicating small differences between the estimated values and the benchmark. Although some errors

in the estimation of reactive power voltage sensitivity (right subfigure) are more significant than those in active power sensitivity, the overall performance remains strong, demonstrating the effectiveness of the proposed method. For HC performance assessment, the LRM from [2] is used for comparison, with simulation-based HC (SHC) as the benchmark. Parts of the HC estimation results are shown in Fig. 2. As a deterministic scenario, the SHC falls within the CI of the proposed probabilistic method for most customers, demonstrating its ability to capture the potential scenarios. While some estimates fall outside the 95% interval due to voltage sensitivities estimation and distribution modeling error, the deviations are minor compared to the significant discrepancies observed with the LRM for certain customers. To quantify the errors, the distance beyond interval (DBI) is introduced, as shown in Fig. 3. The PHC achieves a mean absolute DBI of 0.87 kW across all customers, whereas the LRM method exceeds 7.49 kW mean absolute error, highlighting the superior accuracy of the proposed model in handling the scenarios. The performance of PHC estimation can be further improved by optimizing transformer-pairing for voltage sensitivity, enlarging the historical dataset, and refining the GMDM structure in future work.

IV. Conclusion

This paper addressed the limitations of previous data-driven approaches for residential PV HC estimation in DER-rich scenarios by proposing a probabilistic HC estimation framework. Numerical results and method comparisons validated its effectiveness. Future work will aim to extend the proposed method to more complex low-voltage SNets, evaluate its performance using real feeder models with actual SM data, and analyze how large-scale data loss affects estimation accuracy.

References

[1] J. Wu, J. Yuan, Y. Weng, and R. Ayyanar, "Spatial-temporal deep learning for hosting capacity analysis in distribution grids," *IEEE Transactions on Smart Grid*, vol. 14, no. 1, pp. 354–364, 2022.

[2] J. A. Azzolini, M. J. Reno, J. Yusuf, S. Talkington, and S. Grijalva, "Calculating pv hosting capacity in low-voltage secondary networks using only smart meter data," in *2023 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, pp. 1–5, IEEE, 2023.

[3] L. Liu, N. Shi, D. Wang, Z. Ma, Z. Wang, M. J. Reno, and J. A. Azzolini, "Voltage calculations in secondary distribution networks via physics-inspired neural network using smart meter data," *IEEE Transactions on Smart Grid*, 2024.

[4] V. Bassi, L. F. Ochoa, T. Alpcan, and C. Leckie, "Electrical model-free voltage calculations using neural networks and smart meter data," *IEEE Transactions on Smart Grid*, vol. 14, no. 4, pp. 3271–3282, 2022.

[5] L. Su, X. Pan, X. Sun, J. Guo, and A. Anvari-Moghaddam, "Research on pv hosting capacity of distribution networks based on data-driven and nonlinear sensitivity functions," *IEEE Transactions on Sustainable Energy*, 2024.

[6] Z. Wang and J. Wang, "Time-varying stochastic assessment of conservation voltage reduction based on load modeling," *IEEE Transactions on Power Systems*, vol. 29, no. 5, pp. 2321–2328, 2014.

[7] S. Wang, Y. Dong, L. Wu, and B. Yan, "Interval overvoltage risk based pv hosting capacity evaluation considering pv and load uncertainties," *IEEE transactions on smart grid*, vol. 11, no. 3, pp. 2709–2721, 2019.

[8] H. Zhang, Y. Liu, J. Yan, S. Han, L. Li, and Q. Long, "Improved deep mixture density network for regional wind power probabilistic forecasting," *IEEE Transactions on Power Systems*, vol. 35, no. 4, pp. 2549–2560, 2020.

[9] S. Maharjan, R. Cheng, and Z. Wang, "Generalized analytical estimation of sensitivity matrices in unbalanced distribution networks," *IEEE Transactions on Power Systems*, 2024.