Permutation test and bootstrap methods for unsupervised detection and estimation of behind-the-meter photovoltaic generation

Chao Liu | Jing Shi | Hongkun Chen | Lei Chen | Guocheng Li

School of Electrical Engineering and Automation, Wuhan University, Wuhan, People’s Republic of China

Abstract
The penetration of unmonitored behind-the-meter photovoltaic systems has increased rapidly over the past decade. While unsupervised methods have been proposed to estimate behind-the-meter photovoltaic generation, their performance is significantly affected by hyperparameters. Existing methods for choosing hyperparameters require labelled data, which are unavailable to system operators. In this paper, a data-driven method is proposed to generate labelled data to optimize the hyperparameters for unsupervised estimation of behind-the-meter photovoltaic generation. First, a permutation-test based method is developed to detect photovoltaic installation in an unsupervised way. Second, consumers without PV are combined with limited monitored photovoltaic sites to simulate the original system through the bootstrap method. The simulated system provides labelled data for hyperparameter optimization. Finally, the near-optimal hyperparameters are searched on the simulated system and applied to the original system’s estimation. Through experiments on a smart meter dataset, the effectiveness of the proposed methodology is verified.

1 | INTRODUCTION

The cost of solar PV generation has decreased substantially in the past decade due to continuous technological innovation and policy support. The PV generation electricity prices can compete with grid-supplied electricity prices in some regions, showing immense economic and environmental benefits [1, 2]. As a result, PV installations are growing rapidly, with a considerable portion belonging to distributed PV systems [3]. Due to net metering policies, most of the distributed PV systems are located behind-the-meter (BTM) and seldom directly monitored by distribution system operators [4]. The generation of BTM PV appears as a decrease in the net load, making it difficult to distinguish the decrease of consumer consumption and the increase of BTM PV generation. Thus, high penetration of BTM PV generation can significantly complicate the system’s net load characteristics [5]. Moreover, the performance of network protection [6], net load forecasting [7], distribution network planning [8], and demand response [9] is also affected. The lack of visibility of BTM PV generation proposes a great challenge to the secure and economical operation of distribution networks. Theoretically, systems operators need to install additional meters to measure each PV site’s generation to implement smart grid technologies better. However, the enormous amount of BTM PV sites makes this approach not practically viable. And this may raise more concerns about the relevant privacy issues [10]. Therefore, it is essential to investigate effective methods for estimating BTM PV generation.

A traditional approach is to map the PV generation of visible sites to the region’s PV generation. Shaker et al. proposed a supervised method that constructs the mapping function of the measured output of representative sites to the aggregated generation of the whole region [11]. Another work of them modifies this model into an unsupervised method that uses fuzzy arithmetic to model uncertainties and the aggregation effect [12]. Saint-Drenan et al. construct the statistical distribution of PV configurations and estimate the region’s PV generation by sampling PV plants from the statistical distribution [13]. These studies are essentially amplifying the power measurements of visible sites by the ratio of the region’s total PV capacity to their...
capacity. Their prerequisite is that the total PV capacity of the region is known to system operators. While system operators may know the total capacity of a medium and large-sized system, it is difficult to obtain the detailed PV location distribution within a specific region, making these approaches not flexible for small-scale analysis. Besides, unauthorized deployment of PV systems has occurred in some places, making the total capacity imprecise [14]. Mapping the visible sites’ generation to the aggregated generation is therefore not widely applicable.

Recently, more work focuses on disaggregating BTM PV generation from net load measurements as advanced meter infrastructures have been rapidly deployed all over the world [10]. Some supervised methods have achieved excellent estimation performance on open datasets. Dinesh et al. formulate the estimation of BTM PV generation as a non-intrusive load monitoring problem [15]. Wang et al. apply K-means clustering to extract net load features under different generalized weather conditions and propose a two-stage estimation model based on support vector machines (SVM) [16]. Mason et al. utilize a robust deep neural network to estimate PV size, tilt, and azimuth [17]. However, these supervised methods require a lot of labelled data for training. As it is not feasible to install additional meters to obtain historical disaggregated measurements for training, the applicability of the supervised methods to BTM PV generation estimation is limited.

Many efforts have been devoted to unsupervised estimation methods to reduce the dependence on training datasets, which introduce exogenous features to model the load and PV generation. Zhang et al. develop a three-step strategy to detect and estimate BTM PV installation, where the change points in historical smart meter measurements are detected, verified, and estimated in an unsupervised fashion [18]. Kara et al. proposed an unsupervised estimation framework using reactive power measurements of micro-phasor measurement units and active power measurements from nearby monitored PV sites [19]. Cheung et al. extract consumption patterns of consumers without PV and use their combination to model consumers’ actual model [20]. Wang et al. proposed an estimation method that searches the equivalent PV capacity by minimizing the correlation between estimation residual and irradiance, and they build separate models for disaggregated components of the net load to improve net load forecasting [21]. Sossan et al. utilize the similarities of the net load and BTM PV generation in a certain frequency range to conduct estimation [22]. Bu et al. construct the exemplar library of typical load and PV generation profiles, and synthesize composite exemplars for semi-supervised estimation [23]. Kabir et al. combine a mixed hidden Markov model with a PV system performance model to jointly estimate the BTM PV generation of a group of consumers [24]. These unsupervised methods have achieved promising performance, and some of them can even compete with the state-of-the-art supervised methods. The problem of unsupervised methods is that they have several hyperparameters that are crucial to their estimation performance [18–24]. In previous studies, researchers apply cross-validation to find the optimal hyperparameters, which involves the use of labelled data. But it is difficult to implement cross-validation in practical applications because labelled data are unavailable. Otherwise, there would be no use for estimation methods. Therefore, system operators have to choose hyperparameters based on their experience, which may lead to significant errors.

To overcome the problem of hyperparameter selection and improve the practical performance of unsupervised estimation, we propose a novel hyperparameter optimization method requiring no labelled ground truth data in this paper. It should be noted that our method is not to accelerate the hyperparameter optimization, which has been widely explored in deep learning [25]. The proposed method aims to generate labelled data for hyperparameter optimization. It is motivated by the fact that if consumers without PV can be detected with high accuracy, it is possible to exploit the net load of consumers without PV and the active power of limited monitored PV sites to simulate the original system (i.e., consumers with PV). The load and PV generation of the simulated system are known and can be used for hyperparameter optimization. Accordingly, a permutation-test based method is proposed for unsupervised detection of BTM PV installation. Next, the detected consumers without PV are combined with monitored PV sites to simulate the original system by the bootstrap method. The optimal hyperparameters for the simulated system provide an approximation to the optimal hyperparameters for the original system. They are finally applied to the original system’s estimation.

The main contributions are summarized as follows:

1. A variation feature is proposed to extract the PV generation component of the net load.
2. A permutation-test based method is proposed to detect BTM PV installation by verifying the linear correlation between the variation features of the net load and PV generation, which requires no labelled data.
3. The bootstrap method is used to simulate the original system by sampling from consumers without PV and limited monitored PV sites, which generates labelled data for hyperparameter optimization.
4. The optimal hyperparameters searched on the simulated system are applied to the unsupervised estimation of the original system’s BTM PV generation.

The rest of this paper is organized as follows. Section 2 states the problem of hyperparameter optimization. Section 3 proposes the framework and details of the methodology. Section 4 illustrates the criteria to evaluate the performance of the proposed methods. Section 5 presents the case studies on a smart meter dataset. Finally, conclusions are provided in Section 6.

2 | PROBLEM STATEMENT

The original problem is to estimate BTM PV generation from net load measurements by using exogenous features. The most common exogenous feature is solar irradiance, which can be collected from ground-based measurements [15, 21, 22] or satellite data [20, 24]. However, this paper does not directly utilize solar irradiance because establishing solar monitoring stations
is costly, and the space-time resolution of satellite data is limited. As sources of publicly available PV monitoring data keep increasing, it is possible to take active power measurements from nearby publicly available PV sites as irradiance proxies [10, 26]. Moreover, these active power measurements can reflect the impact of tilt/azimuth configurations of PV panels, which is beneficial to reducing the estimation errors. Therefore, this paper adopts irradiance proxies as the exogenous feature. Considering that the penetration of publicly available sites is still relatively low, the number of irradiance proxies is controlled to be relatively small.

Suppose $N_S^i \in \mathbb{R}^T$ is the net load measurements for $T$ time intervals for a certain system $i$, $L_S^i \in \mathbb{R}^T$ is the systems’ load, and $P_S^i \in \mathbb{R}^T$ is the system’s PV generation. The net load $N_S^i$ is equal to the load $L_S^i$ minus the PV generation $P_S^i$:

$$N_S^i = L_S^i - P_S^i \quad (1)$$

A linear relationship is usually used to model the PV generation $P_S^i$:

$$P_S^i = \sum_{j=1}^{J} C_j I_{j} + \epsilon_{PV,i} \quad (2)$$

where $I_{j} \in \mathbb{R}^T$ denotes the irradiance proxy for $T$ time intervals for a certain PV site $j$, $j = 1, \ldots, J$ denotes the nearby publicly available PV sites, $C_j$ denotes the PV capacity coefficient that is to be estimated, and $\epsilon_{PV,i} \in \mathbb{R}^T$ denotes the error term of system $i$.

The net load $N_S^i$ and the irradiance proxy $I_{j}$ are the only known variables in the above equations. Existing studies have developed a variety of models to estimate the PV capacity coefficient $C_j$. Although these models are based on different assumptions, their process can be described as constructing a function:

$$C_j = f(N_S^i, I_{j}, X, \theta) \quad (3)$$

where $f(\cdot)$ is the function depending on a certain estimation model, $X$ denotes other optional exogenous features, and $\theta$ denotes the hyperparameter set that is determined before estimation.

Since the hyperparameter set determines the detailed estimation criterion, the estimation performance can be seriously affected by it. This is exemplified in Figure 1, where two different systems are estimated twice using both of their optimal hyperparameters. Significant errors occur when one system applies the other one's optimal hyperparameter. To improve the estimation performance, we transform the original problem into an optimization problem of hyperparameters:

$$\theta^*_j = \arg\min \psi(\epsilon_{PV,i}) \quad (4)$$

where $\theta^*_j$ denotes the optimal hyperparameters for system $i$, $\psi(\cdot)$ denotes the evaluation metric, and the constraints are Equations (1)–(3).

### Figure 1

Estimation results with different hyperparameters (night-time observations are omitted) (a) system A and (b) system B

### 3 | METHODOLOGY

The optimization of hyperparameters requires labelled data for evaluation, which are unavailable in many cases. Therefore, this paper proposes a method that simulates the original system to generate the necessary labelled data. The basic idea of the proposed method is that irradiance proxies and net load measurements from consumers without PV can provide labelled data for evaluation. While the penetration of PV has been increasing rapidly in the past decade, there are few regions where all consumers have PV installation. Thus, consumers without PV are available in many cases, which meet the prerequisite of our method. But the number of consumers without PV will decrease substantially in cases with very high PV penetration, which may affect the performance of the proposed method. We will discuss this issue in Section 5.2.

### 3.1 | Framework

Figure 2 shows the framework of the proposed methodology. First, the permutation-test based method is used to detect
consumers without PV by verifying the linear correlation between the variation features of the net load and irradiance proxies. Then, consumers without PV and irradiance proxies are randomly sampled through the bootstrap method to simulate a system with similar net load characteristics to the original system. The grid search algorithm is used to find the optimal hyperparameters of the simulated system. Finally, the near-optimal hyperparameters are searched on the simulated system and applied to the original system’s estimation.

3.2 Variation feature

While consumer behaviours are more consistent with a stable pattern, PV generation is primarily determined by meteorological factors. For this reason, the load and PV generation have different variability. The paper makes use of this difference and extracts a variation feature that leaves mainly the PV generation component. The consumption pattern of an individual customer can be obtained by averaging load measurements at each day of the week [27]. Therefore, the proposed method divides daytime into four segments, and for each month and each day of the week, the typical net load curve is calculated, which represents the typical net load variation feature. The treatment of segmentation allows a limited time shift in consumer behaviours. Suppose \( N^C_{typical,k} \in \mathbb{R}^{24} \) is the typical net load curve for a certain consumer \( k \), \( N^C_{sd,k} \in \mathbb{R}^{24W} \) is the consumer’s segmented daytime net load for \( W \) weeks, and \( F_N^k \in \mathbb{R}^{24W} \) is the consumer’s net load variation feature. The net load variation feature is calculated as follows:

\[
F_N^k(24(w - 1) + s) = N^C_{sd,k}(24(w - 1) + s) - N^C_{typical,k}(s) \tag{5}
\]

where \( w = 1, \ldots, W \) denotes the week of the month, and \( s = 1, \ldots, 24 \) denotes the segment of the week.

The same process is applied to the irradiance proxies to calculate the irradiance proxy variation feature:

\[
F^I_j(24(w - 1) + s) = I_{sd,j}(24(w - 1) + s) - I_{typical,j}(s) \tag{6}
\]

where \( F^I_j(24(w - 1) + s) \) denotes the irradiance proxy variation feature of PV site \( j \) at segment \( s \) in week \( w \), \( I_{sd,j}(24(w - 1) + s) \) represents the irradiance proxy variation feature.

The net load variation feature \( F_N^k \) and the irradiance proxy variation feature \( F^I_j \) both consist mostly of the PV generation component because of the variability difference between consumer behaviours and PV generation. Thus, the Pearson linear correlation coefficient \( \rho_{k,j} \) is used to detect PV installation:

\[
\rho_{k,j} = \frac{\text{Cov}(F_N^k, F^I_j)}{\sigma_{F_N^k}\sigma_{F^I_j}} \tag{7}
\]

where \( \text{Cov}(\cdot, \cdot) \) denotes the covariance, \( \sigma_{F_N^k} \) denotes the standard deviation of the net load variation feature \( F_N^k \), and \( \sigma_{F^I_j} \) denotes the standard deviation of the irradiance proxy variation feature \( F^I_j \).

The Pearson coefficient \( \rho_{k,j} \) ranges from \(-1\) to \(1\), with \(-1\) representing a negative linear correlation, \(1\) representing a positive linear correlation, and \(0\) representing no correlation. When the \( \rho_{k,j} \) is close enough to \(-1\), it indicates a high probability that consumer \( k \) has installed a PV system with similar configurations to publicly available PV site \( j \). Figure 3 shows an example of variation features. While the power of the net load and the negative irradiance proxy shows a similar periodic trend, significant differences occur when the net load is positive. In contrast, the variation features of the two signals display a remarkable linear correlation because the load component has been mostly removed.

3.3 Permutation test

To verify whether the linear correlation exists, we construct a statistical inference using the hypothesis test. The null hypothesis \( H_0 \) and the alternative hypothesis \( H_1 \) of the proposed hypothesis test are:

\[
\begin{cases}
H_0 : F_N^k \text{ and } F^I_j \text{ are not negatively linear correlated} \\
H_1 : F_N^k \text{ and } F^I_j \text{ are negatively linear correlated}
\end{cases} \tag{8}
\]

Since it is difficult to obtain the prior distribution of \( F_N^k \) and \( F^I_j \), parametric testing methods such as \( t \)-test cannot be implemented in this problem. The paper adopts the
permutation test, where the distribution of the test statistic is obtained by rearranging the observation points and calculating the test statistic.

As the testing process is conducted for a large number of consumers, the detection problem is a multiple hypothesis test, calling for the control of the family-wise error rate (FWER). However, strong control methods such as Bonferroni correction would be too conservative for this problem, since variation features of different consumers and irradiance proxies may be correlated to a certain degree. In an extreme situation, when all variation features are completely dependent, the multiple hypothesis test is equivalent to a univariate hypothesis test, which does not need FWER correction. Therefore, the paper adopts the max statistic method to adaptively correct FWER, which uses the most extreme statistic of each permutation to derive the $p$-values [28]. The permutation test in this paper contains the following five steps:

1. Choose a significance level $\alpha$, which is commonly set to 0.05 or 0.01 [29].
2. Construct a permutation by exchanging one or more pairs of observations, and the lowest $\rho_{k,j}$ of the permutation is calculated.
3. Repeat step (2) for a certain number of times and construct the empirical distribution under the null hypothesis using the lowest $\rho_{k,j}$ of each permutation.
4. Calculate $\rho_{k,j}$ of the original sequences and derive their $p$-values from the empirical distribution.
5. If the $p$-values are lower than $\alpha$, reject the null hypothesis $H_0$ in favour of the alternative hypothesis $H_1$.

For consumer $k$, once $H_0$ is rejected, the proposed method classifies it as a consumer with PV. Otherwise, the consumer is considered to have no PV installation.

### 3.4 Bootstrap

The bootstrap method is very effective in estimating statistics and has been applied to many areas in power systems such as data augmentation and probabilistic forecasting [15, 30, 31]. In this paper, it is used to simulate the original system. Consumers without PV and consumers with PV can be assumed to be independently drawn from the identical distribution. This assumption can be satisfied when most analysed consumers have similar consumption patterns (e.g. residential loads). The same assumption can be applied to PV generation from irradiance proxies and consumers with PV as well because PV generation is mainly determined by solar irradiance of that region. With the two assumptions satisfied, it possible to simulate a system that has similar net load characteristics to consumers with PV by bootstrap.

Considering that irradiance proxies remain zero output during night time, the paper splits the sampling process into two stages. In the first stage, consumers without PV are sampled, and their net load measurements are aggregated. This process repeats until the average night-time net load of the simulated system $\overline{SN}_n$ becomes higher than or equal to the average night-time net load of the original system $\overline{ON}_n$. In the second stage, irradiance proxies are sampled, and their power measurements are subtracted from the simulated system’s net load. This process repeats until the average daytime net load of the simulated system $\overline{SN}_d$ becomes lower than or equal to the average daytime net load of the original system $\overline{ON}_d$. Figure 4 shows an example of the average daily net load curves of the simulated system and the original system, which are nearly overlapped. The example shows that separately controlling the simulated system’s average net load in daytime and night-time enables us to sample a system with similar net load characteristics to the original system.

After the simulated system is obtained through sampling, unsupervised estimation methods with different hyperparameters are evaluated on it. The grid search method is used to find the optimal hyperparameters. The proposed sampling and evaluation process will repeat until the hyperparameter optimization result converges. This process provides an approximation to the optimal hyperparameters of the original system. Finally, the searched near-optimal hyperparameters are applied to the estimation of the original system.

![Figure 3](image-url)  
Variation features of the net load and the irradiance proxy (values of the irradiance proxy have been multiplied by $-1$) (a) normalized power and (b) normalized variation features.
3.5 Unsupervised estimation methods

Existing unsupervised estimation methods can be categorized into consumer-level methods and system-level methods according to the estimation object. As the proposed bootstrap-based method relies on sampling consumers to simulate the original system, its size should not be too small. Consequently, the proposed hyperparameter optimization method is more applicable to system-level estimation methods. Two representative system-level estimation methods are selected for validation. The first one is the consumer mixture model, which typifies a series of methods that assume the load to be a combination of exogenous features [20]. Another one is the frequency range dynamics model, which requires no exogenous features relevant to the load [22]. Both of the two methods are offline methods, implementing estimation with the time window of one day. And they require only net load measurements of smart meters and irradiance proxies as input.

3.5.1 Consumer mixture model

The hypothesis of the consumer mixture model is that the consumption behaviour of a consumer can be represented as a combination of latent representative behaviours. To identify these latent features, for each month, this model applies the $k$-medoids algorithm with dynamic time warping to the daily load curve of consumers without PV. The clustering centroids, denoted by $g_l \in \mathbb{R}^{m_1}$, $l = 1, \ldots, CN$, are extracted as features for estimation. For the best quality of the clusters, according to ref. [20], the cluster number $CN$ is recommended to be eight.

The model uses a linear relationship to model the system’s load:

$$L^S_i(t) \approx \sum_{l=1}^{CN} \omega^d_l g_l(t)$$  \hspace{1cm} \text{(9)}

where $\omega^d_l$ denotes the load coefficient for cluster $l$ on day $d$.

According to Equations (1), (2), and (9), the estimation process of the consumer mixture model includes three steps:

**Step 1:** Find $\omega^d_l$ using night-time data only

$$\min_{\omega^d_l} \left\{ \sum_{t \in \text{nighttime}} \left( \sum_{l=1}^{CN} \omega^d_l g_l(t) - N^S_l(t) \right)^2 + \lambda \sum_{l} \omega^d_l^2 \right\}$$

subject to $\omega^d_l \geq 0 \forall l$ \hspace{1cm} \text{(10)}

where $\lambda$ denotes the regularization parameter.

**Step 2:** Find $C_j$ using all data

$$\min_{C_j} \sum_{t} \left( \sum_{l=1}^{CN} \omega^d_l g_l(t) - \sum_{j=1}^{J} C_j I_j(t) - N^S_l(t) \right)^2$$

subject to $C_j \geq 0 \forall j$ \hspace{1cm} \text{(11)}

**Step 3:** Use post-optimization adjustment to obtain the disaggregated signals

$$\min_{\tilde{L}^S_i(t), \tilde{P}^S_i(t)} \sum_{t} \left( \tilde{L}^S_i(t) - \sum_{l=1}^{CN} \omega^d_l g_l(t) \right)^2 + \beta \left( \tilde{P}^S_i(t) - \sum_{j=1}^{J} C_j I_j(t) \right)^2$$

subject to $\tilde{L}^S_i(t) \geq 0 \tilde{P}^S_i(t) \geq 0$

$$\tilde{L}^S_i(t) - \tilde{P}^S_i(t) = N^S_l(t)$$ \hspace{1cm} \text{(12)}

where $\tilde{L}^S_i(t)$ denotes the estimated load of system $i$ at time interval $t$, $\tilde{P}^S_i(t)$ denotes the estimated PV generation of the system, and $\beta \in \mathbb{R}_+$ denotes the post-optimization coefficient.

For the consumer mixture model, the hyperparameters to be optimized are $\lambda$ and $\beta$.

3.5.2 Frequency range dynamics model

The Frequency Range Dynamics Model is based on the hypothesis that PV generation is the dominant component of net load measurements in a certain frequency range. It is motivated by the fact that BTM PV generation and the net load have similar frequency density in a certain frequency band. To exploit this similarity, the model initially uses a Butterworth bandpass filter to filter the net load $N^S_t$ and irradiance proxies $I_{j,t}$, where the lower cutoff frequency $CF_L$ and the upper cutoff frequency...
CF_U are predefined and reflect the frequency range where the frequency density similarity exists. Let FN_S_i and FI_j be the filtered signals of N_S_i and I_j, respectively. Based on the similarity hypothesis, a linear relationship exists between FN_S_i and FI_j:

\[
FN_S_i(t) \approx \sum_{j=1}^{J} C_j FI_j(t) \tag{13}
\]

Since the two signals are only similar but not the same in the frequency range, outliers exist in the filtered signals. To overcome this problem, the model uses a robust linear regression to estimate the capacity coefficient C_j:

\[
C_j = \arg\min \left\{ \sum_{t=1}^{T} \phi(FN_S(t) - \sum_{j=1}^{J} C_j FI_j(t)) \right\} \tag{14}
\]

where \(\phi(\cdot)\) is the bisquare loss function, a nonconvex relationship that adaptively tunes the weight of extreme values to be robust against outliers [32].

For the frequency range dynamics model, the hyperparameters to be optimized are the lower cutoff frequency CF_L and the upper cutoff frequency CF_U.

4 | EVALUATION CRITERIA

In this section, several criteria are proposed to evaluate the proposed method, including the dataset, the performance metrics, and the methods for comparisons.

4.1 | Dataset

To verify the proposed hyperparameter optimization methods, we adopt a smart meter dataset from Pecan Street Inc. [33]. One year of measurements of the net load, the load, and PV generation are analysed. Although this dataset provides finer resolution options, the paper adopts the 15 min resolution since it is the main way that smart meter data are collected [10]. Before the evaluation, data cleaning is conducted to eliminate invalid data from consumers. After removing consumers with a large quantity of bad data, 74 consumers without PV and 119 consumers with PV from Austin, Texas, are extracted. The distributions of peak loads and peak PV generation for the extracted consumers are shown in Figures 5 and 6, respectively. Most consumers’ peak loads are lower than 12 kW, indicating that these consumers are mainly residential loads. And most of their peak PV generation is lower than 6 kW, suggesting that the PV systems are mostly of small size. We have plotted the average daily power curves of 50 consumers in Appendix to display their diverse consumption patterns and PV capacities.

For each round of evaluation, a fixed number of consumers with PV are randomly sampled as irradiance proxies, leaving other consumers as the entire system for evaluation. The number of irradiance proxies is chosen as five because experiments show that increasing the number of irradiance proxies has little impact on the performance of detection and estimation.

4.2 | Evaluation of detection

Receiver operating characteristic (ROC) curves and the area under the ROC curve (AUC) values are used for the evaluation of the proposed detection method. The paper defines true positive rate (TPR) and false positive rate (FPR) as follows:

\[
TPR = \frac{n_{TP}}{n_{P}} \tag{15}
\]

\[
FPR = \frac{n_{FP}}{n_{N}} \tag{16}
\]

where \(n_{TP}\) denotes the number of consumers with PV that are correctly detected, \(n_{P}\) denotes the number of all consumers with PV, \(n_{FP}\) denotes the number of consumers without PV that are wrongly identified as consumers with PV, and \(n_{N}\) denotes the number of all consumers without PV.

The significance level \(\alpha\) is initially set as 0, where all consumers are identified as consumers without PV. Then by increasing \(\alpha\) gradually to 1, both TPR and FPR keep non-decreasing. For each \(\alpha\), TPR and FPR are calculated and drawn on the graph, which makes the ROC curve. The integral of the ROC curve provides the AUC value, which ranges from 0.5 to 1. The higher the AUC value, the better the classifier.

To verify the superiority of the proposed detection method, we compare it with distributed photovoltaic system detection (DPVSD), a supervised model based on SVM [16]. DPVSD
first generates four generalized weather classes by clustering and then extracts features based on four typical net load curves. The SVM trained with these features serves as the classifier for PV installation detection. The radial basis function is chosen as the kernel function. Four-fold cross-validation is used to validate the performance.

### 4.3 Evaluation of hyperparameter optimization

The performance metric of the original estimation problem is the normalized root mean squared error (nRMSE):

\[
\text{nRMSE}(\hat{P}_i^S(t)) = \sqrt{\frac{1}{T} \sum_{t=1}^{T} \left( \frac{P_i^S(t) - \hat{P}_i^S(t)}{\max(P_i^S(t))} \right)^2}
\]

where \(\hat{P}_i^S(t)\) denotes the estimated BTM PV generation of system \(i\).

To verify the effectiveness of the proposed hyperparameter optimization method, we compare the bootstrap-based method with three ideal situations:

1. Individually ideal situation: For each system, its hyperparameters are chosen to minimize its nRMSE:

\[
\theta^*_{\text{DIS},i} = \arg\min\{\text{nRMSE}(\hat{P}_i^S(t))\}
\]

where \(\theta^*_{\text{DIS},i}\) denotes the individually optimal hyperparameters for system \(i\).

   It is the best situation where all systems achieve optimal estimation.

2. Overall ideal situation: All systems are estimated with the same hyperparameters that minimize their average nRMSE:

\[
\theta^*_{\text{OIS}} = \arg\min \left\{ \frac{1}{Q} \sum_{i=1}^{Q} \text{nRMSE}(\hat{P}_i^S(t)) \right\}
\]

where \(\theta^*_{\text{OIS}}\) denotes the overall optimal hyperparameters, and \(Q\) denotes the number of systems for evaluation.

   Note that only one group of hyperparameters are used in the overall ideal situation. Due to the difference between systems’ optimal hyperparameters, the performance of the overall ideal situation will be worse than that of the individually ideal situation.

3. Learning-based ideal situation: All unique optimal parameters from the individually ideal situations are collected to construct an optimal hyperparameter set \(\theta_{\text{RIS}}\). Then, for each system, all the elements in \(\theta_{\text{RIS}}\) are used for estimation. The performance metric is obtained by averaging the nRMSE of all optimal hyperparameters. The frequency distribution of optimal hyperparameters is not considered here. It is motivated by the fact that system operators can only learn from limited monitored systems, and the frequency distribution of optimal hyperparameters is often unavailable.

Similar to the overall ideal situation, the hyperparameters are fixed for all systems in the learning-based ideal situation. Therefore, the performance of the learning-based ideal situation is also affected by the difference between systems’ optimal hyperparameters. And it will be worse than the overall ideal situation's performance. The reason is that the overall ideal situation utilizes the sensitivity of each system’s performance to the hyperparameters for optimization, whereas the learning-based ideal situation does not.

   The individually ideal situation and the overall ideal situation represent the optimal situations with varying and fixed hyperparameters for each system, respectively. They are not feasible in practical applications due to the unavailability of ground truth data. The learning-based ideal situation is more likely to occur than the other ideal situations as system operators may learn the optimal hyperparameters from open datasets. It reflects the traditional approach that hyperparameters are selected based on the learning experience. But the practical performance of the traditional approach will be worse than that of the learning-based ideal situation in this paper. The reason is that, for the learning-based ideal situation, the systems to learn from (i.e., the training set) and the systems to test (i.e., the test set) are the same in this paper. In contrast, system operators have to learn from datasets from other regions. Based on the above analysis, if the proposed method can compete with the learning-based ideal situation, it will demonstrate that the proposed method can improve the estimation performance in practical applications. Our ultimate goal is to make the performance as close as possible to that of the individually ideal situation and the overall ideal situation.

### 5 CASE STUDIES

#### 5.1 BTM PV detection

For the reliability of the results, the sampling of irradiance proxies is repeated 100 rounds. And for each round, the performance metrics are calculated and recorded. Figure 7 illustrates the average ROC curves of the permutation-test based method and DPVSD. Table 1 describes the mean and standard deviation of the AUC values, where the best method is shown in boldface. The experiment results show that, for the one-year dataset, the permutation-test based method outperforms DPVSD. As the AUC value of the permutation-test based method is 1, it can be considered as an ideal classifier in the proposed case.

   Considering that the installation of BTM PV may happen in the middle of the year, we investigate the performance’s sensitivity on different choices of the start time and the time window length in terms of AUC values. In Figure 8, the AUC values of
the two methods with time windows of three, six, nine, and 12 months under a specific choice of the start time are provided. The results show that, except when the start time is winter and the time window length is three months, the permutation-test based method consistently outperforms DPVSD. With the time window length increasing, the performance of the two methods displays an overall improving trend. However, the performance of DPVSD is significantly affected by the start time. Its AUC value in summer is greatly lower than in other seasons, implying that DPVSD cannot adapt to the seasonal effect of consumer behaviours. In contrast, the permutation-test-based method displays high adaptability to the seasonal effect. The AUC values of the permutation-test-based method keep higher than 0.98 for all the choices of the start time and the time window length, suggesting that its performance is not very sensitive to these factors. It should be noted that the proposed method does not require labelled data for training, whereas DPVSD does. The results demonstrate the superiority of the proposed permutation-test-based detection method.

5.2 | Hyperparameter optimization

Since the experiment results of BTM PV detection show that the proposed detection method can achieve the detection task in an almost ideal way, we assume that all consumers are correctly classified in the experiments of the hyperparameter optimization. We randomly sample 10, 60, and 114 consumers from all consumers with PV to form the original systems for estimation. For each system size, this sampling procedure is run for 50 rounds. Then the proposed bootstrap-based method and the three ideal situations are evaluated on these systems.

Figures 9 and 10 display the sensitivity analysis of the estimation performance to hyperparameters for a system of 60 consumers and its simulated system. Although the optimal nRMSE values differ, it can be observed that the optimal hyperparameters for the original system and the simulated system are almost in the same region. The results show that the assumption of the bootstrap method is satisfied, at least for the adopted smart meter dataset.

In Figures 11 and 12, the estimation performance of the bootstrap-based method and the three ideal situations over 50 rounds are compared, where a point above the diagonal means that the bootstrap-based method achieves better performance. Tables 2 and 3 provide the average nRMSE. As can be seen, the relationship between the three ideal situations’ performance is consistent with the qualitative analysis in Section 4.3: the individually ideal situation achieves the best, followed by the overall ideal situation and the learning-based ideal situation. The bootstrap-based method performs significantly better than the learning-based ideal situation, providing a 17.8% and 20.0% decrease in the average nRMSE for the consumer mixture model and the frequency range dynamic model, respectively. This demonstrates the effectiveness of the proposed method in improving the estimation performance in practical applications. And the bootstrap-based method’s average nRMSE is very close to that of the overall ideal situation, implying that the proposed method achieves a kind of near-optimal performance. Besides, the difference between the bootstrap-based method, the individually ideal situation and the overall ideal situation decreases as the system size increases. The reason is that the accompanying smoothing effect reduces the variability in electricity consumption and PV generation. Consequently, the system simulated by the bootstrap method becomes closer to the original system with the system size growing. It indicates that the larger the system size is, the better is the proposed method’s performance.

Finally, we test the bootstrap-based method with different numbers of consumers without PV to investigate how fewer consumers without PV affect its performance. Only the frequency range dynamics model is analysed here because the consumer mixture model’s performance will degrade as the number of consumers without PV decreases. The original system is chosen as 60 consumers with PV. Three scenarios are considered, with 20, 40, and 60 consumers without PV (33.3%, 66.7%, and 100% of the original system’s size, respectively), which correspond to very high, high, and normal penetration levels, respectively. Note that the three ideal situations are not affected by the number of consumers without PV as they directly use ground truth data of the original systems to optimize hyperparameters. The average result of 50 runs is plotted in Figure 13. The result shows that the bootstrap-based method consistently performs better than the learning-based ideal situation. In other words, although the performance of the bootstrap-based method will deteriorate at the very high PV penetration level, in the proposed case, the bootstrap-based method is effective as long as the number of consumers without PV is larger than 20. Moreover, the bootstrap-based method with 40 and 60 consumers without PV perform rather similarly, suggesting that 40 consumers without PV are enough to achieve its near-optimal performance. The results demonstrate the wide applicability of the proposed method.
FIGURE 8  AUC plots for different start times and time window lengths (a) start in spring, (b) start in summer, (c) start in autumn, and (d) start in winter

FIGURE 9  Sensitivity analysis for the consumer mixture model (a) simulated system and (b) original system

FIGURE 10  Sensitivity analysis for the (a) frequency range dynamics model simulated system and (b) original system

TABLE 2  Estimation performance for the consumer mixture model

|                  | Bootstrap | Individually ideal situation | Overall ideal situation | Learning-based ideal situation |
|------------------|-----------|------------------------------|-------------------------|-------------------------------|
| Average nRMSE    | 0.0259    | 0.0241                       | 0.0250                  | 0.0304                        |
TABLE 3  Estimation performance for the frequency range dynamics model

|                  | Bootstrap | Individually ideal situation | Overall ideal situation | Learning-based ideal situation |
|------------------|-----------|------------------------------|-------------------------|-------------------------------|
| Average nRMSE    | 0.0296    | 0.0255                       | 0.0279                  | 0.0349                        |

FIGURE 11  nRMSE comparisons for the (a) consumer mixture model individually ideal situation, (b) overall ideal situation, and (c) learning-based ideal situation

FIGURE 12  nRMSE comparisons for the frequency range dynamics model, (a) individually ideal situation, (b) overall ideal situation, and (c) learning-based ideal situation

6 1 CONCLUSION

This paper presents a hyperparameter optimization method for unsupervised estimation of BTM PV generation. The proposed method is based on a detection and simulation strategy. In the detection stage, the variation features of the net load and PV generation are extracted and used to detect PV installation through the permutation-test based method. In the simulation stage, the original system is simulated by sampling from consumers without PV and limited monitored PV sites through the bootstrap method, and the simulated system is tested with unsupervised estimation methods to search the near-optimal hyperparameters. Case studies on a smart meter dataset demonstrate the effectiveness of our proposed method. For the application of BTM PV detection, the proposed permutation-test based method consistently outperforms the baseline supervised method with different seasons and time window lengths. For the application of hyperparameter optimization, the proposed bootstrap-based method achieves near-optimal performance without knowing the labelled ground truth data, and it is still effective at high PV penetration levels.

The analysed consumers mostly consist of residential loads in this paper. However, when our method is applied to industrial loads, its performance may decline due to the distinctive consumption patterns of industrial consumers. In the future, we intend to address this challenge.
FIGURE 13 nRMSE plots for different numbers of consumers without PV

**Nomenclature**

**Abbreviations**

| Abbreviation | Definition |
|--------------|------------|
| AUC          | area under the ROC curve |
| BTM          | behind-the-meter |
| DPVSD        | Distributed PV System Detection |
| FPR          | false positive rate |
| FWER         | family-wise error rate |
| nRMSE        | normalized root mean square error |
| PV           | photovoltaic |
| ROC          | receiver operating characteristic |
| SVM          | support vector machine |
| TPR          | true positive rate |

**Functions and hypotheses**

| Function | Description |
|----------|-------------|
| Cov(·,·) | covariance |
| f(·)     | estimating function |
| H₀       | null hypothesis |
| H₁       | alternative hypothesis |
| 𝜙(·)     | bisquare loss function |
| 𝜙(·)     | evaluation metric |

**Indices**

| Index | Description |
|-------|-------------|
| i     | index of systems |
| j     | index of publicly available PV sites |
| J     | number of publicly available PV sites |
| k     | index of consumers |
| l     | index of clusters |
| Q     | number of systems |
| s     | index of segments of the week |
| t     | index of time intervals |
| T     | number of time intervals |
| w     | index of weeks of the month |
| W     | number of weeks of the month |

**Parameters**

| Parameter | Description |
|-----------|-------------|
| CF₁       | lower cutoff frequency |
| CF₂       | upper cutoff frequency |

**Variables**

| Variable | Description |
|----------|-------------|
| C_j      | PV capacity coefficient for PV site j |
| F_j      | irradiance proxy variation feature of PV site j |
| F_k      | net load variation feature of consumer k |
| F_I      | filtered irradiance proxy of PV site j |
| FN_i     | filtered net load of system i |
| g_i      | clustering centroid of cluster l |
| I_j      | irradiance proxy of PV site j |
| I_{sd,j} | segmented daytime irradiance proxy of PV site j |
| I_{typical,j} | typical irradiance proxy curve of PV site j |
| L_i      | load of system i |
| L_i      | estimated load of system i |
| n_{FP}   | number of consumers without PV that are wrongly identified as consumers with PV |
| n_{N}    | number of consumers without PV |
| n_p      | number of consumers with PV |
| n_{TP}   | number of consumers with PV that are correctly detected |
| N_i      | net load of system i |
| N_{sd,k} | segmented daytime net load of consumer k |
| N_{typical,k} | typical net load curve of consumer k |
| ON_{sd}  | average daytime net load of the original system |
| ON_{sd}  | average night-time net load of the original system |
| P_i      | PV generation of system i |
| P_i      | estimated PV generation of system i |
| SN_{sd}  | average daytime net load of the simulated system |
| SN_{sd}  | average night-time net load of the simulated system |
| X        | optimal exogenous features |
| σ_{N,k}  | standard deviation of load variation features |
| σ_{F,j}  | standard deviation of irradiance proxy variation features |
| ε_{PV,i} | estimation error of system i |
| ρ_{k,j}  | Pearson coefficient between the variation features of consumer k and PV site j |
| ω_l      | load coefficient for cluster l on day d |

**ORCID**

Chao Liu https://orcid.org/0000-0003-0060-0965

**REFERENCES**

1. Yan, J., et al.: City-level analysis of subsidy-free solar photovoltaic electricity price, profits and grid parity in China. Nat. Energy 4(8), 709–717 (2019)
2. Senatla, M., et al.: Estimating the economic potential of PV rooftop systems in South Africa’s residential sector: A tale of eight metropolitan cities. IET Renew. Power Gener. 14(4), 506–514 (2020)
3. Feldman, D., et al.: Solar industry update. National Renewable Energy Laboratory, Golden, CO (2019)
4. Bayram, LS, Ustun, TS: A survey on behind the meter energy management systems in smart grid. Renew. Sustain. Energy Rev. 72, 1208–1232 (2017)
5. Haupte, S.E., et al.: Blending distributed photovoltaic and demand load forecasts. Sol. Energy 157, 542–551 (2017)
6. Mohammadi, P., Mehta, A.: Challenges of PV integration in low-voltage secondary networks. IEEE Trans. Power Deliv. 32(1), 525–535 (2017)
7. Van Der Meer, D.W., et al.: Probabilistic forecasting of solar power, electricity consumption and net load: Investigating the effect of seasons, aggregation and penetration on prediction intervals. Sol. Energy 171, 397–413 (2018)
8. Abad, M.S., et al.: Probabilistic assessment of hosting capacity in radial distribution systems. IEEE Trans. Sustain. Energy 9(4), 1935–1947 (2018)
9. Li, K., et al.: Capacity and output power estimation approach of individual behind-the-meter distributed photovoltaic system for demand response baseline estimation. Appl. Energy (2019), 253, p.1–12
10. Wang, Y., et al.: Review of smart meter data analytics: Applications, methodologies, and challenges. IEEE Trans. Smart Grid 10(3), 3125–3148 (2019)
11. Shaker, H., et al.: A data-driven approach for estimating the power generation of invisible solar sites. IEEE Trans. Smart Grid 7(5), 2466–2476 (2016)
12. Shaker, H., et al.: Estimating power generation of invisible solar sites using publicly available data. IEEE Trans. Smart Grid 7(5), 2465–2466 (2016)
13. Saint-Drenan, Y., et al.: An approach for the estimation of the aggregated photovoltaic power generated in several European countries from meteorological data. Adv. Sci. Res. 15, 51–62 (2018)
14. HECO warns customers to disconnect rogue solar systems. https://www.civilbeat.org/2014/09/heco-warns-customers-to-disconnect-rogue-solar-systems (2020). Accessed September 2020
15. Dinesh, C., et al.: Non-intrusive load monitoring under residential solar power influx. Appl. Energy 205, 1068–1080 (2017)
16. Wang, F., et al.: A distributed PV system capacity estimation approach based on support vector machine with customer net load curve features. Energies 11, 1–9 (2018)
17. Mason, K., et al.: A deep neural network approach for behind-the-meter residential PV size, tilt and azimuth estimation. Sol. Energy 196, 260–269 (2020)
18. Zhang, X., Grijalva, S.: A data-driven approach for detection and estimation of residential PV installations. IEEE Trans. Smart Grid 7(5), 2477–2485 (2016).
19. Kara, E.C., et al.: Disaggregating solar generation from feeder-level measurements. Sustain. Energy, Grids Networks 13, 112–121 (2018)
20. Cheung, C.M., et al.: Behind-the-meter solar generation disaggregation using consumer mixture models. IEEE Smart. Grid. Commun., Aalborg (2018), 1–7
21. Wang, Y., et al.: Data-driven probabilistic net load forecasting with high penetration of behind-the-meter PV. IEEE Trans. Power Syst. 33(3), 3255–3264 (2018)
22. Sossan, F., et al.: Unsupervised disaggregation of photovoltaic production from composite power flow measurements of heterogeneous prosumers. IEEE Trans. Ind. Informatics 14(9), 3904–3913 (2018)
23. Bu, F., et al.: A data-driven game-theoretic approach for behind-the-meter PV generation disaggregation. IEEE Trans. Power Syst. 35(4), 3133–3144 (2020)
24. Kabir, F., et al.: Joint estimation of behind-the-meter solar generation in a community. IEEE Trans. Sustain. Energy (12(1), 682–694 2020), https://doi.org/10.1109/TSTE.2020.3016896
25. Bergstraa, J., et al.: Algorithms for hyper-parameter optimization. In Proceedings of the 24th International Conference on Neural Information Processing Systems NY (2011), 2546–2554
26. Zhang, X., et al.: A solar time based analog ensemble method for regional solar power forecasting. IEEE Trans. Sustain. Energy 10(1), 268–279 (2019)

APPENDIX A

Figure A.1 depicts the average daily load curves and average daily PV generation curves of 50 consumers from the dataset.

FIGURE A.1 Average daily power curves of 50 consumers from the dataset (a) load curves and (b) PV generation curves