A Two-Layer Approach for Estimating Behind-the-Meter PV Generation Using Smart Meter Data

Fankun Bu, *Graduate Student Member, IEEE*, Rui Cheng, *Graduate Student Member, IEEE*, and Zhaoyu Wang, *Senior Member, IEEE*

***Abstract*—As the cost of the residential solar system decreases, rooftop photovoltaic (PV) has been widely integrated into distribution systems. Most rooftop PV systems are installed behind-the-meter (BTM), i.e., only the net demand is metered, while the native demand and PV generation are not separately recorded. Under this condition, the PV generation and native demand are invisible to utilities, which brings challenges for optimal distribution system operation and expansion. In this paper, we have come up with a novel two-layer approach to disaggregate the unknown PV generation and native demand from the known hourly net demand data recorded by smart meters: 1) At the aggregate level, the proposed approach separates the aggregate PV generation time series from the aggregate net demand time series for customers with PVs. 2) At the customer level, the separated aggregate-level PV generation is allocated to individual PVs. These two layers leverage the spatial correlations of native demand and PV generation, respectively. One primary advantage of our proposed approach is that it is more independent and practical compared to previous works because it does not require PV array parameters, meteorological data and previously recorded solar power exemplars. This paper has verified our proposed approach using real native demand and PV generation data.**

***Index Terms*—Behind-the-meter, distribution system, PV generation estimation, rooftop photovoltaic, smart meter.**

I. Introduction

In the last decade, residential rooftop photovoltaic (PV) has been proliferating in distribution systems. In most cases, utilities only install a bi-directional smart meter to record the net demand of customers with PVs. This type of installation is referred to as behind-the-meter (BTM), in which case the net demand equals native demand minus PV generation. Therefore, the PV generation produced by solar array and the native demand consumed by appliances are unknown to utilities. Only metering the net demand can reduce the financial cost for utilities; however, as the penetration level of PV increases, the unobservability of notable PV generation and native demand brings significant challenges to distribution systems. We focus on three specific applications to elaborate the necessity of estimating the unknown BTM PV generation and native demand: *First*, the unavailability of native load and PV generation might cause unacceptable forecasting errors because some forecasters require reconstituting the generation and native demand time series [1], [2]. In contrast, knowing BTM PV generation and native load can help utilities forecast generation and load separately, thus provide utilities useful information regarding load/generation growth. *Second*, the invisibility of PV generation and native load can hinder designing optimal service restoration plans [3], [4]. During the restoration stage after an outage, the native demand might be several times higher than the pre-outage demand due to the simultaneous restarting of a large number of air-conditioning appliances. This anomalous demand should be estimated for optimal restoration plans because it can damage electric devices when simultaneously restoring a large number of customers. In practice, utilities usually multiply the normal native demand before outage by a ratio to estimate the anomalous demand during restoration. Also, utilities typically do not consider PVs as reliable restoration sources [3]. Therefore, separating normal native demand and generation is needed for designing optimal restoration plans. *Third*, the unobservability of native demand and solar generation might cause inaccurate reliability analysis. When evaluating a transmission system's reliability, each distribution system is generally simplified as a bus whose native load duration curve is constructed [5], [6]. For those utilities with a high-penetration PV integration, directly using the net demand to construct the load duration curve can significantly underestimate the actual native load [7]. This is because the net demand is typically smaller than the native demand due to the existence of PV generation. In contrast, using the native demand separated from the net demand can help construct more accurate load duration curves. In summary, disaggregating BTM PV generation and native demand from the recorded net demand can enhance distribution system observability and awareness and can also provide more accurate information for transmission system reliability analysis.

Previous works on BTM PV generation disaggregation can be categorized into two types: *Type I - Model-based approaches:* PV array performance model is employed to represent physical

Manuscript received October 30, 2021; revised February 27, 2022; accepted March 29, 2022. This work was supported in part by the National Science Foundation under Grant EPCN 2042314, and in part by the Grid Modernization Initiative of the U.S. Department of Energy (DOE) under Grant GMLC project 2.1.1 – FASTDERMS. Paper no. TPWRS-01692-2021. *(Corresponding author: Zhaoyu Wang.)*

The authors are with the Department of Electrical and Computer Engineering, Iowa State University, Ames, IA 50011 USA (e-mail: fbu@iastate.edu; ruicheng@iastate.edu; wzy@iastate.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TPWRS.2022.3164010.

Digital Object Identifier 10.1109/TPWRS.2022.3164010

PV arrays. In [8], a PV model is combined with a clear sky model to estimate customer-level solar generation. In [9], a virtual equivalent PV station model is utilized to represent the aggregate generation of BTM PVs within a region. In [10] and [11], a physical PV model and a statistical model are utilized to estimate BTM solar generation and native demand, respectively. One primary disadvantage of these model-based approaches is that detailed PV array parameters or accurate meteorological data are required. However, in practice, these parameters are typically unavailable to utilities. Also, acquiring meteorological data might cause additional costs to utilities. *Type II - Model-free approaches:* In [12] and [13], net demands under heterogeneous weather conditions are employed to estimate BTM PV capacity, which is then multiplied by a standard solar power time series to infer BTM PV generations. In [14], native demand and PV generation are estimated using 1-second net demand data by identifying appliances' states, which are then leveraged to estimate appliance demands and solar power. Based on the variation difference between load and solar power, in [15], an approach is proposed for estimating service transformer-level PV generation. In [16], regional-level generation is estimated by installing additional sensors to record typical PV generation profiles. In [17], feeder-level solar generation is estimated by utilizing net load measurements and a nearby PV farm's generation readings. Using known native loads for customers without PVs and the generations for a limited number of observable PVs, in [18], the authors formulate an optimization process to estimate the aggregated native load and PV generation. In [19], a federated learning-based framework is proposed to probabilistically estimate community-level BTM solar generation. In [20], an approach is developed to estimate the reactive power by taking advantage of the correlation between the weekly nighttime and daytime native reactive power demands. Furthermore, previously in [21] and [22], we have proposed two approaches for estimating the unknown BTM generation using measured solar power exemplars. One primary shortcoming of the model-free approaches is that they rely on contextual information, i.e., recorded solar power exemplars or meteorological data, which might bring additional costs to utilities.

Considering the shortcomings of previous approaches, this paper proposes a novel BTM PV generation and native demand estimation framework which does not require *previously* recorded solar power and meteorological measurements. Our approach is based on two findings from real data. The first finding is the spatial correlation of native load, i.e., the native demands of two sizeable residential customer groups are strongly correlated and have highly homogeneous shapes. The second finding is the spatial correlation of solar power generation, i.e., the generations for two PVs in a distribution system are significantly correlated and have highly similar profiles.

Our proposed approach contains **two** layers: (1) At the aggregate level, the total generation of all BTM PVs is estimated by leveraging our first finding. (2) At the customer level, utilizing our second finding, the estimated aggregate BTM PV generation is allocated to individual customers. Utilizing the two findings improves our approach's robustness against the customer-level load uncertainty [23]. The second layer contains three steps: first, our approach trains a model to produce multiple candidate generation time series, using solar power data generated by a publicly available tool. Second, our approach determines the peak generation for each PV. Finally, the allocating procedure is formulated as an optimization problem. The overall structure of our proposed approach is shown in Fig. 1. This paper has verified our proposed approach using real hourly native demand and PV generation data [24].

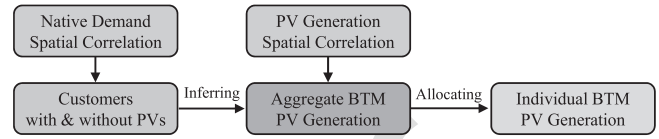


Fig. 1. Overall structure of the proposed BTM PV generation estimation approach.

Smart meters can record individual customers' demands at an interval of one hour or shorter. Such fine-grained temporal and spatial granularity can give us more details than traditional monthly bills. Many researchers have developed advanced approaches to mine useful information from smart meter data. For example, [25] utilizes smart meter measurement to perform state estimation for enhancing distribution system observability, [26] employs water consumption data recorded by smart water meters to train aggregate water demand forecasters, [27] utilizes high-resolution phasor measurement units' data to conduct false data detection, and data redundancy strengthening, [28] converts smart meter data into manageable load profiles via linearizing load patterns. Our proposed approach takes advantage of smart meter data's temporal and spatial granularity to perform BTM generation estimation.

The main contributions of our paper are summarized as follows: (1) This paper proposes an approach that does not rely on PV array parameters, historical meteorological data, and pre-recorded generation exemplars. This independence can significantly improve the viability of our approach because acquiring the above three types of information can bring challenges or additional costs for utilities. (2) Our approach only relies on the net demands of customers with PVs and the native demands of customers without PVs for estimating the aggregate-level PV generation. These two types of demands - net and native - are typically available to utilities, making our approach significantly practical. (3) Our approach innovatively estimates individual PV-installed customers' peak generations by mining net demand data. The peak generations are then utilized to estimate individual PV-installed customers' BTM generation time series.

Throughout the paper, vectors are denoted using bold italic letters, and matrices are represented as bold non-italic letters. In addition, we adopt the sign convention that the native demand consumed by customers and the power output from PVs are both positive.

The rest of the paper is organized as follows: Section II introduces our first and second findings regarding spatial correlation of native demand/generation. Section III presents how we estimate the aggregate generation for customers with PVs.

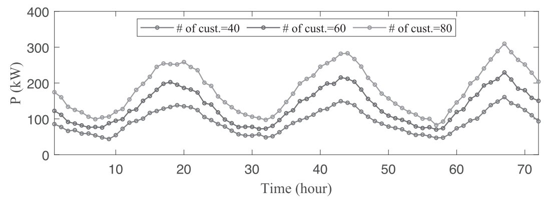


Fig. 2. Three-day actual native demand curves for three example groups with different customer numbers.

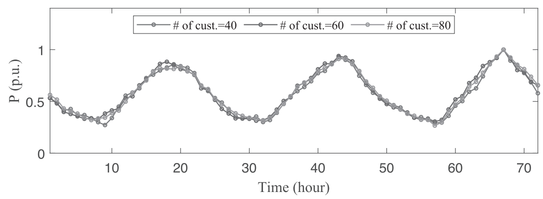


Fig. 3. Three-day normalized native demand curves for three example groups with different customer numbers.

Section IV presents the procedure of formulating and solving an optimization problem to allocate the estimated aggregate generation to individual PVs. In Section V, case studies are analyzed. Section VI concludes the paper.

II. Spatial Correlation of Native Demand/PV Generation

A. Finding 1: Native Demand Spatial Correlation Between Two Sizeable Groups

By examining real residential native demand data, we find that once the customer numbers for two groups reach a certain level, their native demands are highly correlated. This finding is leveraged for estimating the *aggregate* native demand time series for customers with PVs.

Specifically, we use native demand curves to illustrate the observed spatial correlation. Fig. 2 presents real native demand curves for three example groups with different customer numbers, i.e., 40, 60, and 80, respectively. We can observe that these three curves demonstrate almost identical shapes, although they have different magnitudes. The high shape similarity can also be corroborated by Fig. 3, which presents normalized native demand curves corresponding to the curves in Fig. 2. Note that the normalized curves are obtained by dividing the real curves in Fig. 2 by their peaks, respectively.

To stress the importance of Fig. 3, we first define **two** types of customer groups: the residential customers **with** and **without** PVs. These two customer groups are denoted as C_w and C_o, respectively. For C_o, its native demand is recorded by smart meters. For C_w, we only know its net demand, and we do not know its native demand. Our goal is to estimate C_w's unknown native demand and thus to estimate its PV generation. Therefore, Fig. 3 inspires us that given the known native demand curve of C_o, we can infer the unknown native demand curve of C_w by multiplying the native demand curve of C_o by a *ratio*, r.

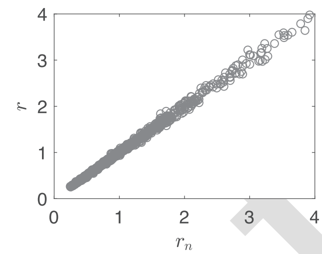


Fig. 4. The relationship between native demand ratio and the nocturnal native demand ratio between two example customer groups.

Since the native demands for the customers in C_o are directly recorded by smart meters, the native demand curve of C_o can be obtained by aggregating the native demand time series over the customers in C_o. The challenge for inferring the unknown native demand curve of C_w is that the ratio, r, is unknown and needs to be estimated. The unknown of r is caused by the unavailability of the native demand during the daytime for the customers in C_w. This is because PV generates power during the daytime, which masks the native demand in the case of net metering. Thus, we cannot use daytime native demand to compute r. Instead, we use the nocturnal native demand to estimate r because PV does not generate power during nighttime, and thus the nocturnal native demand for C_w is known. Based on the above inference, we propose first utilizing the nocturnal native demand to compute a nocturnal native demand ratio, r_n, and then approximating r as r_n.

One *pre-condition* for approximating r as r_n is that r should be close to r_n. To verify this condition, we randomly select two groups with different customer numbers ranging from 20 to 80. Then, for each group, the native demand time series are spatially aggregated over customers to obtain an aggregate native demand time series. After that, we compute r using the two groups' native demand time series throughout a certain period, and compute r_n using the two groups' native demand time series only during *nighttime* within that period. Finally, we plot r against r_n, as shown in Fig. 4. We can see that r is almost identical with r_n. Therefore, we can accurately estimate r by directly letting it equal r_n.

Once we obtain the estimate of r, we can compute the unknown native demand of C_w by multiplying the known native demand of C_o by the estimate of r. After that, estimating the unknown PV generation of C_w is straightforward, i.e., by subtracting the recorded net demand measurements from the estimated native demand.

B. Finding 2: Generation Spatial Correlation Between Two PVs

There are two primary factors that determine the generation spatial correlation: (1) In most cases, a distribution system is geographically bounded in a small district. (2) The most widely available sampling resolution for smart meters is 1-hour. Under these two conditions, different PV arrays are subject to nearly identical meteorological inputs. Thus, the identical inputs can result in highly similar shapes among PV generation curves.

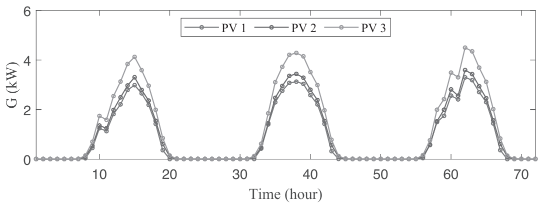


Fig. 5. Three-day real generation curves for three example PVs with different capacities.

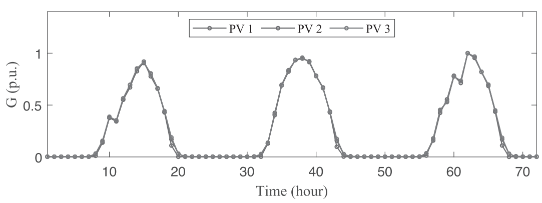


Fig. 6. Three-day normalized generation curves for three example PVs with different capacities.

Fig. 5 presents three example PV generation curves corresponding to different PV array capacities. Similar to the native demand curves for sizeable customer groups, these three generation curves also demonstrate significant spatial correlation, i.e., they possess highly similar shapes. This high similarity can also be corroborated by Fig. 6, where the normalized generation curves corresponding to the three curves in Fig. 5 overlap with each other. Most importantly, Figs. 5 and 6 inspire us that estimating a BTM PV generation curve comes down to two steps: first, determine the generation curve's shape, and then determine its magnitude. This two-step method can notably simplify the estimation of unknown BTM PV generation time series. This is because compared to model-based methods, our approach is developed on the foundation of high similarity among generation curves; therefore, it requires significantly less information.

III. Estimating Aggregate BTM PV Generation for Customers With PVs

As elaborated in Section II-A, the native demands of two sizeable customer groups are highly correlated. This high correlation inspires us that we can infer the unknown native demand of C_w by multiplying the known native demand of C_o by a ratio:

$$\hat{\boldsymbol{P}}_w = \boldsymbol{P}_o r, \tag{1}$$

where, $\hat{\boldsymbol{P}}_w = \{\hat{P}_w(t)\}$ and $\boldsymbol{P}_o = \{P_o(t)\}, t = 1, \ldots, T$, denote the estimated native demand time series for C_w and the actual native demand time series for C_o, respectively. T is the total number of native demands in a selected window (e.g., one month). $P_o(t)$ is computed by aggregating the measured native demands over customers without PVs:

$$P_o(t) = \sum_{i=1}^{N_o} P_{o,i}(t), \quad t = 1, \ldots, T, \tag{2}$$

where, N_o represents the total number of customers in C_o, i.e., customers without PVs. $P_{o,i}(t)$ denotes the measured *native* demand at time t for the i'th customer in C_o.

In (1), r denotes the native demand ratio between C_w and C_o, and is defined as follows:

$$r = \frac{\Sigma_{t=1}^{T} P_w(t)}{\Sigma_{t=1}^{T} P_o(t)}. \tag{3}$$

However, as presented in Section II-B, since the *diurnal* native demand for C_w is masked by PV generation and unavailable to utilities, we need to estimate r using *nocturnal* native demand measurements. This approximation method is based on the observation that PV does not generate power during nighttime and the verification that r and r_n are almost identical. Specifically, we use r_n to approximate r:

$$\hat{r} = r_n = \frac{\Sigma_{t \in I_n} P_w(t)}{\Sigma_{t \in I_n} P_o(t)}, \tag{4}$$

where, I_n denotes the set of nighttime hours. In our paper, I_n refers to the hours between 9:00 P.M. and 5:00 A.M. Note that for the hours in I_n, since PV does not generate power, $P_w(t)$ equals the known aggregate *net* demand, $P'_w(t)$. Therefore,

$$\hat{r} = \frac{\Sigma_{t \in I_n} P'_w(t)}{\Sigma_{t \in I_n} P_o(t)}, \tag{5}$$

where, $P'_w(t)$ is computed by aggregating the measured net demands over customers in C_w:

$$P'_w(t) = \sum_{i=1}^{N_w} P'_{w,i}(t), \quad t = 1, \ldots, T, \tag{6}$$

where, N_w represents the total number of customers in C_w. $P'_{w,i}(t)$ denotes the measured *net* demand at time t for the i'th customer in C_w.

Then, using the estimate of r and the known native demand time series for C_o, we can apply (1) to compute the estimated native demand time series for C_w. Finally, inferring the PV generation time series for C_w, $\hat{\boldsymbol{G}}_w = \{\hat{G}_w(t)\}$, $t = 1, \ldots, T$, is straightforward:

$$\hat{\boldsymbol{G}}_w = \hat{\boldsymbol{P}}_w - \boldsymbol{P}_w{}', \tag{7}$$

where, $\boldsymbol{P}'_w = \{P'_w(t)\}$, $t = 1, \ldots, T$, denotes the known net demand time series for C_w.

The above procedure for estimating the aggregate-level PV generation and native demand for C_w are illustrated in Fig. 7.

IV. Estimating BTM PV Generation for Each Individual PV

Knowing the aggregate BTM PV generation and native demand might not be sufficient for some applications [29], [30]. For example, some demand response schemes require known customer-level native demand [12]. Therefore, estimating individual customers' BTM native demand and PV generation is of significance.

To achieve this goal, we propose an approach to allocate the estimated aggregate PV generation/native demand time series to individual customers with PVs. As discussed in Section II-B,

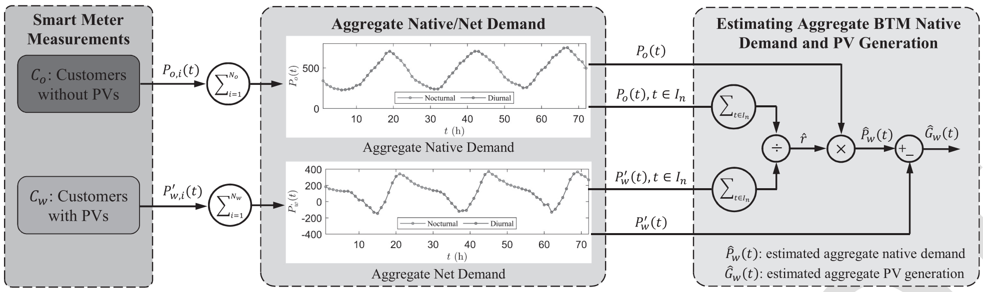


Fig. 7. Detailed structure of the proposed aggregate-level BTM PV generation/native demand estimation.

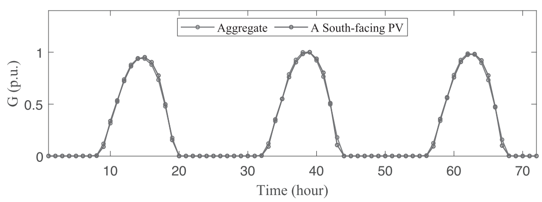


Fig. 8. Three-day normalized aggregate generation curve for all PVs and normalized generation curve for an individual PV facing south.

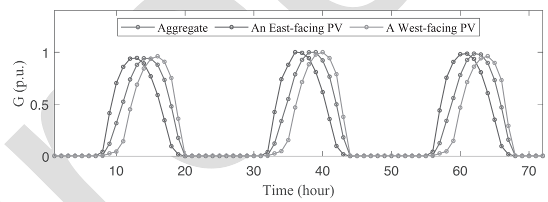


Fig. 9. Three-day normalized aggregate generation curve of all PVs and normalized generation curves for two example PVs facing east and west, respectively.

estimating an individual PV's generation curve boils down to determining the generation curve's shape and its magnitude. In this section, our approach has three steps to perform allocating: (Step-I): generate candidate generation curves for individual PVs; (Step-II): estimate the peak generation for each PV; and (Step-III): allocate the estimated aggregate PV generation time series to individual PVs by solving an optimization problem.

A. Generating Diverse Candidate Generation Curves for Individual PVs

As discussed earlier, in a geographically bounded distribution system, two primary factors determining a generation curve are the magnitude and shape. This subsection aims to generate candidate generation curves for those non-south-facing PVs. First, we train a regression model using the data generated by PVWatts Calculator. Then, we feed the estimated generation curve of a south-facing PV into the trained model to infer the targeted candidate generation curves for those non-south-facing PVs.

In Section III, we have obtained the estimated time series for the aggregate generation of all PVs. One question is whether we can use that shape to represent the unknown shapes of individual PVs. To answer this question, we have conducted a numerical experiment. First, we normalized the aggregate generation curve of all PVs by dividing the aggregate generation time series by its peak. Then, in the same way, we normalized the generation curve of an example PV facing *south*. The two normalized curves are plotted in Fig. 8. It can be seen that the normalized curve corresponding to the aggregate generation for all PVs is highly similar to the normalized curve for a south-facing PV. One primary reason for this similarity is that the majority of residential PVs face south because a south-facing PV can typically generate more power than PVs in other directions. Most importantly, Fig. 8 tells us that a south-facing PV's generation curve can be accurately represented by the normalized aggregate generation curve of all PVs.

Note that in distribution systems, in addition to the majority of south-facing PVs, there exist some residential PVs with other azimuths, such as east or west. These non-south-facing PVs' generation curves cannot be fully represented by the normalized aggregate PV generation curve in Fig. 8. Specifically, compared to the normalized aggregate PV generation curve, the normalized generation curves for an east-facing PV and a west-facing PV are somewhat "left-skewed" and "right-skewed," respectively, as shown in Fig. 9. Therefore, it is necessary to obtain candidate shapes for those non-south-facing PVs' generation curves. To achieve this goal, our basic idea is first to feed PV power data generated by PVWatts Calculator into a regression model to capture the relationship between the generations for a south-facing PV and a non-south-facing PV. Then, the aggregate generation curve estimated in Section III, which can accurately represent a south-facing PV's generation curve, is fed into the trained regression model to produce diverse generation curves corresponding to non-south azimuths. The overall structure is shown in Fig. 10:

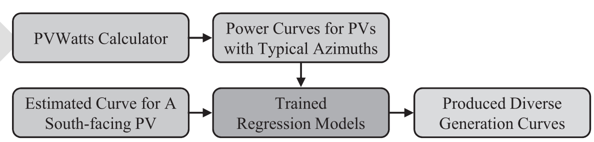


Fig. 10. Overall structure for producing diverse candidate PV generation curves using power output data generated by PVWatts Calculator.

1) Training a Gaussian Process Regression Model: Since the shape of a south-facing PV's generation curve can be approximated as the shape of the aggregate generation curve of all PVs, one intuitive way for inferring *non-south-facing* PVs' candidate shapes is to produce diverse shapes based on the *south-facing* PV's estimated generation curve. This idea is based on our observation that there exists a mapping between the generation curves for PVs with different azimuths. Therefore, one critical step for producing diverse candidate generation curves is to identify the relationship between a non-south-facing PV's generation curve and a south-facing PV's generation curve. To capture the relationship, first, we use PVWatts Calculator [31], an online application developed by the National Renewable Energy Laboratory (NREL), to generate power output data for PVs with typical azimuths, e.g., east, south, and west. Then, using the generated PV output power data, we train a Gaussian Process Regression (GPR) model to capture the relationship between the generation curve corresponding to a typical azimuth except for south (e.g., east) and the generation curve corresponding to the azimuth of the south. The primary reason for selecting GPR is that after running numerical tests, GPR demonstrated a relatively better performance when applied to our dataset than some other state-of-the-art nonlinear regression models, such as the Support Vector Machine model and the Polynomial regression model.

Specifically, first, we use PVWatts Calculator to generate time-series data for a south-facing PV and a PV with other typical azimuth (e.g., east). Then, each time series is normalized so that the peak generation is 1 p.u. The two normalized time series corresponding to the south-facing PV and the non-south-facing PV are denoted as $\boldsymbol{G}_s^* = \{G_s^*(t)\}$ and $\boldsymbol{G}_{ns}^* = \{G_{ns}^*(t)\}$, $t = 1, \ldots, T$, respectively. $G_s^*(t)$ and $G_{ns}^*(t)$ denote the normalized generation at time t for a south-facing PV and a non-south-facing PV, respectively. Our goal is to use $G_s^*(t)$ to explain $G_{ns}^*(t)$ because PVs in a geographically bounded distribution system typically have highly correlated generations. By conducting numerical experiments, we find that in addition to $G_s^*(t)$, the hour-in-day, $H_d(t)$, and day-in-year, $D_y(t)$, are also related with $G_{ns}^*(t)$. Therefore, we use $G_s^*(t)$, $H_d(t)$, and $D_y(t)$ as the input variables and $G_{ns}^*(t)$ as the output variable, respectively, to train a GPR model. The function of GPR is to capture the relationship between $G_{ns}^*(t)$ and $G_s^*(t)$. The basic idea behind GPR is that if the distance between two explanatory variables is small, the difference between their corresponding dependent variables will also be relatively small. Specifically, the output, $G_{ns}^*(t)$, is denoted as a function of the input vector, $\boldsymbol{X}^*(t)$:

$$G_{ns}^*(t) = f(\boldsymbol{X}^*(t)), \tag{8}$$

where, $\boldsymbol{X}^*(t) = [G_s^*(t), H_d(t), D_y(t)]^\mathsf{T}$. For GPR, $f(\boldsymbol{X}^*(t))$ is assumed to be a random variable reflecting the uncertainty of functions evaluated at $\boldsymbol{X}^*(t)$. Specifically, the function $f(\boldsymbol{X}^*(t))$ is distributed as a Gaussian process:

$$f\left(\boldsymbol{X}^*(t)\right) \sim \mathcal{GP}\left(\mu(\boldsymbol{X}^*(t)), K(\boldsymbol{X}^*(t), \boldsymbol{X}^*(t'))\right), \tag{9}$$

where, $\mu(\boldsymbol{X}^*(t))$ represents the expected value of $f(\boldsymbol{X}^*(t))$, i.e., the value of $G_{ns}^*(t)$. The covariance function, $K(\boldsymbol{X}^*(t), \boldsymbol{X}^*(t'))$, represents the dependence between $G_{ns}^*(t)$'s at different times. In our problem, the covariance function, $K(\cdot,\cdot)$, is specified by the Squared Exponential Kernel function expressed as:

$$K\left(\boldsymbol{X}^*(t), \boldsymbol{X}^*(t')\right) = \sigma_f^2 \exp\left(-\frac{||\boldsymbol{X}^*(t) - \boldsymbol{X}^*(t')||_2^2}{2\sigma^2}\right), \tag{10}$$

where, $||\cdot||_2$ represents l_2-norm, σ_f and σ are hyper-parameters, which are determined using cross-validation. Intuitively, (10) measures the distance between $\boldsymbol{X}^*(t)$ and $\boldsymbol{X}^*(t')$, which can also reflect the similarity between $G_{ns}^*(t)$ and $G_{ns}^*(t')$.

Note that $G_s^*(t)$ and $G_{ns}^*(t)$ are generated solar powers using PVWatts Calculator; thus, they are known and a T-dimensional joint Gaussian distribution can be constructed as:

$$\begin{bmatrix} f\left(\boldsymbol{X}^*(1)\right) \\ \vdots \\ f\left(\boldsymbol{X}^*(T)\right) \end{bmatrix} \sim \mathcal{N}\left(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*\right), \tag{11}$$

where,

$$\boldsymbol{\mu}^* = \begin{bmatrix} \mu\left(\boldsymbol{X}^*(1)\right) \\ \vdots \\ \mu\left(\boldsymbol{X}^*(T)\right) \end{bmatrix}, \tag{12a}$$

$$\boldsymbol{\Sigma}^* = \begin{bmatrix} K\left(\boldsymbol{X}^*(1), \boldsymbol{X}^*(1)\right) & \cdots & K\left(\boldsymbol{X}^*(1), \boldsymbol{X}^*(T)\right) \\ \vdots & \ddots & \vdots \\ K\left(\boldsymbol{X}^*(T), \boldsymbol{X}^*(1)\right) & \cdots & K\left(\boldsymbol{X}^*(T), \boldsymbol{X}^*(T)\right) \end{bmatrix}. \tag{12b}$$

The joint Gaussian distribution formulated in (11) represents a trained non-parametric model, which captures the relationship between $G_{ns}^*(t)$ and $G_s^*(t)$.

2) Inferring a Non-South-Facing PV's Generation Curve: As shown in Fig. 8, the normalized generation curve for a south-facing PV, $\boldsymbol{G}_s = \{G_s(t)\}, t = 1, \ldots, T$, can be approximated as the normalized estimated aggregate generation curve for all PVs:

$$\boldsymbol{G}_s = \frac{\hat{\boldsymbol{G}}_w}{\hat{G}_m}, \tag{13}$$

where, $\hat{G}_m$ denotes the peak of $\hat{\boldsymbol{G}}_w$. To infer the unknown generation time series for a non-south-facing PV, $\boldsymbol{G}_{ns} = \{G_{ns}(t)\}$, $t = 1, \ldots, T$, we assume $G_{ns}(t)$ is a function of $G_s(t)$, i.e., $G_{ns}(t) = f(G_s(t))$. By appending $f(G_s(t))$ to the end of (11), an $(N+1)$-dimensional joint Gaussian distribution can be constructed as:

$$\begin{bmatrix} G_{ns}^*(1) \\ \vdots \\ G_{ns}^*(T) \\ G_{ns}(t) \end{bmatrix} = \begin{bmatrix} f(\boldsymbol{X}^*(1)) \\ \vdots \\ f(\boldsymbol{X}^*(T)) \\ f(\boldsymbol{X}(t)) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_* \\ \mu_1 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}^* & \boldsymbol{\Sigma}_{*1} \\ \boldsymbol{\Sigma}_{*1}^T & \Sigma_{11} \end{bmatrix}\right), \tag{14}$$

where, $\boldsymbol{X}(t) = [G_s(t), H_d(t), D_y(t)]^\mathsf{T}$ is a vector of explanatory variables. $\boldsymbol{\Sigma}_{*1}$ represents the training-test set covariances and Σ_{11} is the test set covariance. Since $G_{ns}^*(t)$, $\boldsymbol{X}^*(t)$, and $\boldsymbol{X}(t)$ are known, using the Bayes rule, the distribution of $G_{ns}(t)$

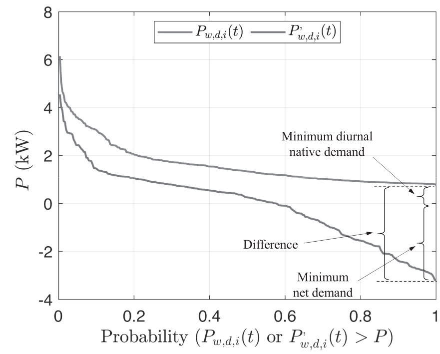


Fig. 11. Load duration curves for an example customer's diurnal native demand and diurnal net demand.

conditioned on $\boldsymbol{G}^*_{ns}$ can be computed as follows:

$$G_{ns}(t)|\boldsymbol{G}^*_{ns} \sim \mathcal{N}(\mu_1(t), \Sigma_1(t)), \tag{15}$$

where, $\mu_1(t) = \boldsymbol{\Sigma}^T_{*1}\boldsymbol{\Sigma}^{*-1}\boldsymbol{G}^*_{ns}$ and $\Sigma_1(t) = \Sigma_{11} - \boldsymbol{\Sigma}^T_{*1}\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{*1}$. Note that $\mu_1(t)$ denotes the most probable value of the estimated generation at time t for a non-south-facing PV. By conducting the above inferring procedure for all the t's, we can obtain a candidate generation time series corresponding to a *particular* typical PV azimuth. Since there are multiple typical azimuths, such as east, and west, we can infer multiple candidate PV generation time series:

$$\boldsymbol{G}^j_{ns} = \{G^j_{ns}(t)\}, \quad t = 1, \ldots, T, \quad j = 1, \ldots, N_{ns}, \tag{16}$$

where, $G^j_{ns}(t)$ denotes the inferred PV generation at time t, for the j'th typical non-south-facing azimuth. N_{ns} denotes the total number of typical non-south-facing PV azimuths and is determined by conducting numerical experiments.

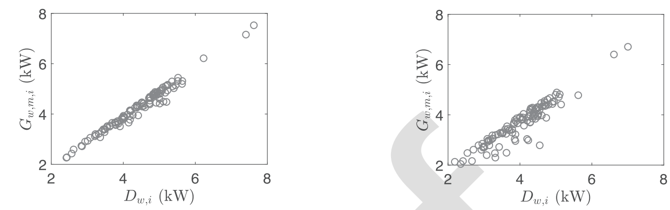
B. Estimating Peak Generation for Each Individual PV

Simply knowing the candidate shapes for unknown generation curves is insufficient for allocating the estimated aggregate generation to individual PVs. As discussed earlier, we should also know the magnitudes for the candidate generation curves. To estimate the peak generation, we employ our observation from real data that the peak generation is almost identical with the difference between the minimum diurnal native demand and the minimum net demand.

Specifically, to explain our observation regarding the correlation, we start with Fig. 11, showing the load duration curves for the i'th customer's diurnal *native* demand, $P_{w,d,i}(t)$, and diurnal *net* demand, $P'_{w,d,i}(t)$. Thus, we can compute the difference between the minimums of $P_{w,d,i}(t)$ and $P'_{w,d,i}(t)$:

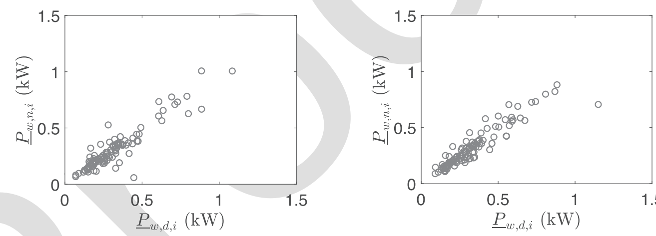
$$D_{w,i} = \underline{P}_{w,d,i} - \underline{P}'_{w,d,i}, \tag{17}$$

where, $\underline{P}_{w,d,i}$ and $\underline{P}'_{w,d,i}$ denote the minimums of $P_{w,d,i}(t)$ and $P'_{w,d,i}(t)$ during a selected window, respectively. Note that $\underline{P}_{w,d,i}$ is positive, and $\underline{P}'_{w,d,i}$ is negative. Then, our finding is that $D_{w,i}$ is highly similar to the peak generation, $G_{w,m,i}$, as shown in Fig. 12. This relationship inspires us to approximate



(a) Spring (b) Summer

Fig. 12. The relationship between peak generation and the difference between minimum diurnal *native* demand and minimum *net* demand. (a) Spring (b) Summer



(a) Spring (b) Summer

Fig. 13. The relationship between minimum *diurnal* native demand and minimum *nocturnal* native demand. (a) Spring (b) Summer

$G_{w,m,i}$ as $D_{w,i}$:

$$\hat{G}_{w,m,i} = D_{w,i}, \quad i = 1, .., N_w, \tag{18}$$

where, $\hat{G}_{w,m,i}$ is the estimate of $G_{w,m,i}$. However, one challenge is that $D_{w,i}$ depends on $\underline{P}_{w,d,i}$, which is unknown due to BTM PV generation. Therefore, we need to estimate $\underline{P}_{w,d,i}$, which is involved with another finding from real native demand data. Specifically, as shown in Fig. 13, the minimum *diurnal* native demand, $\underline{P}_{w,d,i}$, can be approximated as the minimum *nocturnal* native demand, $\underline{P}_{w,n,i}$:

$$\underline{P}_{w,d,i} \approx \underline{P}_{w,n,i}, \quad i = 1, .., N_w. \tag{19}$$

Note that since PV does not generate power during nighttime, $\underline{P}_{w,n,i}$ is known to utilities. Finally, using the estimate of $\underline{P}_{w,d,i}$ and the known $\underline{P}'_{w,d,i}$, we can compute $D_{w,i}$ using (17), and then compute $\hat{G}_{w,m,i}$ using (18).

C. Allocating the Estimated Aggregate PV Generation to Individual PVs

Sections III, IV-A, and IV-B provide the estimated aggregate generation time series of all PVs, inferred candidate generation curves for individual PVs, and estimated generation peaks for individual PVs, respectively. Therefore, estimating individual PVs' generation curves comes down to allocating the estimated aggregate generation time series to individual PVs. This allocating procedure is formulated as an optimization process:

$$\min_{\mathbf{K},\boldsymbol{\gamma}} \; ||\mathbf{G}_e * \mathbf{K} * \mathbf{1} - \hat{\boldsymbol{G}}_w||^2_2 + \lambda * ||\boldsymbol{\gamma}||^2_2 \tag{20a}$$

$$s.t. \;\; \mathbf{G}_e * \mathbf{K} \leq \mathbf{1} * (\hat{\boldsymbol{G}}_{w,m} + \boldsymbol{\gamma})^{\mathsf{T}}, \tag{20b}$$

$$\mathbf{0} \leq \boldsymbol{\gamma} \leq P_0 * \mathbf{1}, \tag{20c}$$

where, $\mathbf{G}_e = [\boldsymbol{G}_s, \boldsymbol{G}_{ns}^1, \ldots, \boldsymbol{G}_{ns}^{N_{ns}}]$ is a T-by-N_e matrix, which denotes a collection of candidate generation curves. $N_e = N_s + 1$ denotes the total number of candidate generation curves. $\mathbf{K} = [\boldsymbol{K}_1, \ldots, \boldsymbol{K}_{N_w}]$ is an N_e-by-N_w matrix of decision variables, which denote the weights assigned to candidate generation curves for individual PVs. $\boldsymbol{K}_i, i = 1, \ldots, N_w$, is an N_e-by-1 vector, which denotes the weights assigned to candidate generation curves for the i'th PV. The first $\mathbf{1}$ is an N_w-by-1 vector of ones. $\mathbf{G}_e * \mathbf{K}$ results in a T-by-N_w matrix, which is a collection of estimated generation time series for individual PVs. The first term in the objective function (20a) reflects the difference between the estimated aggregate PV generation, $\hat{\boldsymbol{G}}_w$, and the weighted summation of individual PV's estimated generations, $\mathbf{G}_e * \mathbf{K} * \mathbf{1}$. The second term in the objective function (20a) considers the estimation errors of peak generations. λ is a tuning parameter. $\boldsymbol{\gamma}$ is an N_w-by-1 vector with non-negative elements, which reflect the errors of approximating $G_{w,m,i}$ as $D_{w,i}$, as shown in (18). The second $\mathbf{1}$ is a T-by-1 vector of ones. $\hat{\boldsymbol{G}}_{w,m} = [\hat{G}_{w,m,1}, \ldots, \hat{G}_{w,m,N_w}]^\mathsf{T}$ denotes an N_w-by-1 vector of the estimated generation peaks for all PVs. $(\hat{\boldsymbol{G}}_{w,m} + \boldsymbol{\gamma})$ denotes the corrected generation peaks with consideration of estimation errors. $\mathbf{1} * (\hat{\boldsymbol{G}}_{w,m} + \boldsymbol{\gamma})^\mathsf{T}$ produces a T-by-N_w matrix, in which each column contains the same element. Constraint (20b) ensures that the estimated generation time series for each PV is smaller than its estimated peak generation. $\mathbf{0}$ is an N_w-by-1 vector of zeros. P_0 denotes the maximum error of approximating $G_{w,m,i}$ as $D_{w,i}$ for individual PVs. The third $\mathbf{1}$ is an N_w-by-1 vector of ones. Constraint (20c) ensures that the estimation errors for individual PVs are non-negative and smaller than an upper bound. The reason for constraining the elements of $\boldsymbol{\gamma}$ as non-negative is that $D_{w,i}$ typically under-estimates $G_{w,m,i}$, as shown in Fig. 13.

The optimization process represented in (20) is a convex quadratic programming problem, thus, we can obtain a unique solution for $\mathbf{K}$, i.e., $\mathbf{K}^* = [\boldsymbol{K}_1^*, \ldots, \boldsymbol{K}_{N_w}^*]$. Then, the estimated generation time series for the i'th PV, $\hat{\boldsymbol{G}}_{w,i} = \{\hat{G}_{w,i}(t)\}$, $t = 1, \ldots, T$, can be computed as:

$$\hat{\boldsymbol{G}}_{w,i} = \mathbf{G}_e * \boldsymbol{K}_i^*, \quad i = 1, \ldots, N_w. \tag{21}$$

Then, the estimated native demand time series for the i'th customer, $\hat{\boldsymbol{P}}_{w,i} = \{\hat{P}_{w,i}(t)\}, t = 1, \ldots, T$, can be computed as:

$$\hat{\boldsymbol{P}}_{w,i} = \boldsymbol{P}'_{w,i} + \hat{\boldsymbol{G}}_{w,i}, \quad i = 1, \ldots, N_w. \tag{22}$$

where, $\boldsymbol{P}'_{w,i} = \{P'_{w,i}(t)\}, t = 1, \ldots, T$, denotes the known net demand time series recorded by smart meter for the i'th customer with PVs.

Note that (20) can be solved for a selected window. The window size, T, can impact estimation accuracy and runtime, which will be examined in the Case Study Section. The detailed steps for estimating customer-level PV generation are illustrated in Fig. 14.

V. Case Study

In this section, the proposed two-layer BTM solar power and native demand estimation approach is verified using real PV generation and native demand data.

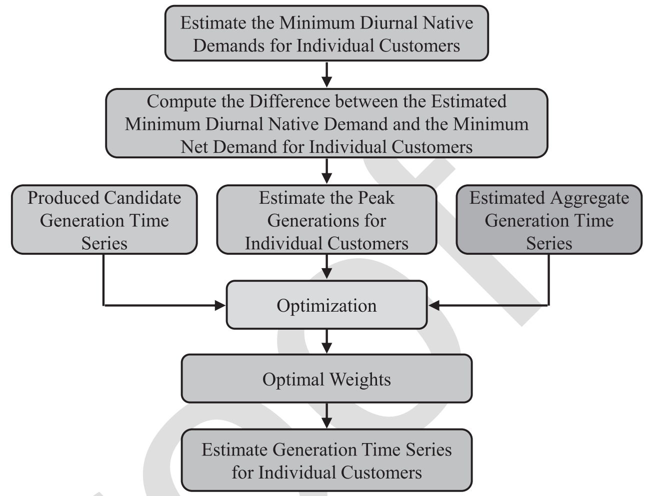
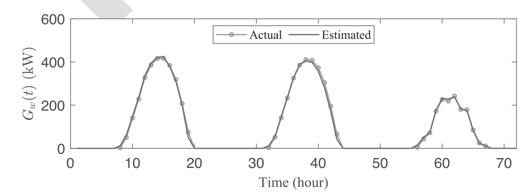
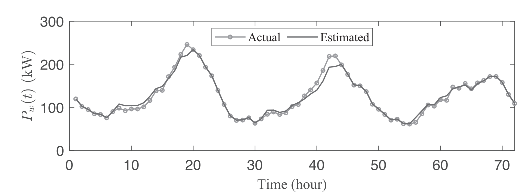


Fig. 14. Detailed steps of the individual customer-level BTM PV generation estimation.



(a) Aggregate PV generation



(b) Aggregate native demand

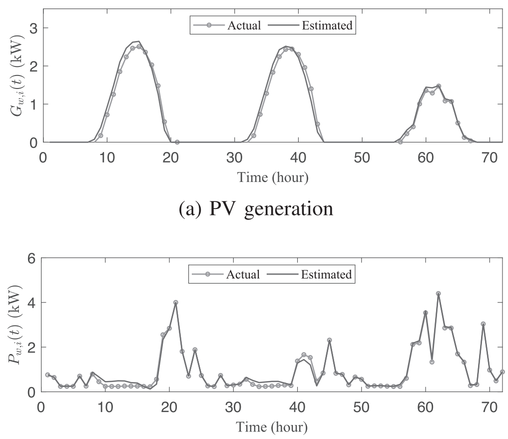
Fig. 15. Three-day actual and estimated aggregate PV generation and native demand curves. (a) Aggregate PV generation (b) Aggregate native demand

A. Dataset Description

The hourly native demand and PV generation data used in this paper are from a public dataset [24]. The time range of native demand and solar power is one year. This dataset contains a total number of 100 customers with PVs and 115 customers without PVs. For the customers with PVs, the net demand is obtained by subtracting PV generation from native demand.

B. Aggregate-Level BTM PV Generation Estimation Validation

Fig. 15 shows three-day actual and estimated aggregate PV generation/native demand curves. It can be seen that the estimated curves can accurately follow the actual curves. To quantitatively evaluate the estimation accuracy, we compute the mean



(a) PV generation

(b) Native demand

Fig. 16. Three-day actual and estimated PV generation and native demand curves for an example customer with PV. (a) PV generation (b) Native demand

absolute percentage error (MAPE) as follows:

$$MAPE = \frac{100\%}{N_d} \sum_{t \in I_d} \left| \frac{\hat{Y}_w(t) - Y_w(t)}{Y_{w,m}} \right|, \quad (23)$$

where, $\hat{Y}_w(t)$ represents $\hat{G}_w(t)$ or $\hat{P}_w(t)$. $Y_w(t)$ represents $G_w(t)$ or $P_w(t)$. $Y_{w,m}$ represents $G_{w,m}$ or $P_{w,m}$, where $G_{w,m}$ and $P_{w,m}$ denote the actual peaks of PV generation and native demand, respectively. I_d denotes the set of daytime hours. N_d denotes the total number of hours in I_d.

To comprehensively evaluate the performance of our approach, we also compute the mean squared error (MSE) and coefficient of variation (CV):

$$MSE = \frac{1}{N_d} \sum_{t \in I_d} \left(\hat{Y}_w(t) - Y_w(t) \right)^2, \quad (24)$$

$$CV = \frac{\sigma}{\mu}, \quad (25)$$

where,

$$\mu = \frac{1}{N_d} \sum_{t \in I_d} (\hat{Y}_w(t) - Y_w(t)), \quad (26a)$$

$$\sigma = \sqrt{\frac{1}{N_d - 1} \sum\nolimits_{t \in I_d} \left((\hat{Y}_w(t) - Y_w(t)) - \mu \right)^2}. \quad (26b)$$

The computed $MAPE$'s for PV generation and native demand are 1.21% and 1.28%, respectively. The computed MSE's for PV generation and native demand are about 58.09. Note that the actual peaks for the PV generation and native demand are 462.5 and 437.1 kW, respectively. The computed CV's for PV generation and native demand are about -3.48. The above error metrics reflect the high accuracy of our proposed approach.

C. Customer-Level BTM PV Generation Estimation Validation

1) Estimation Performance: Fig. 16 shows three-day actual and estimated PV generation and native demand curves for an example customer with PV. We can see that the estimated curves can accurately fit the actual curves. To comprehensively examine the performance of our approach, we compute the $MAPE$ for all customers with PVs. Specifically, the $MAPE$'s for the i'th customer are computed as follows:

$$MAPE_i = \frac{100\%}{N_d} \sum_{t \in I_d} \left| \frac{\hat{Y}_{w,i}(t) - Y_{w,i}(t)}{Y_{w,m,i}} \right| \quad (27)$$

where $Y_{w,i}(t)$ represent $G_{w,i}(t)$ or $P_{w,i}(t)$, $\hat{Y}_{w,i}(t)$ represent $\hat{G}_{w,i}(t)$ or $\hat{P}_{w,i}(t)$, and $Y_{w,m,i}$ represent $G_{w,m,i}$ or $P_{w,m,i}$. $G_{w,m,i}$ and $P_{w,m,i}$ denote the actual generation and native demand peaks for the i'th customer, respectively. We also compute the MSE and CV for each PV-installed customer:

$$MSE_i = \frac{1}{N_d} \sum_{t \in I_d} \left(\hat{Y}_{w,i}(t) - Y_{w,i}(t) \right)^2, \quad (28)$$

$$CV_i = \frac{\sigma_i}{\mu_i}, \quad (29)$$

where,

$$\mu_i = \frac{1}{N_d} \sum_{t \in I_d} (\hat{Y}_{w,i}(t) - Y_{w,i}(t)), \quad (30a)$$

$$\sigma_i = \sqrt{\frac{1}{N_d} \sum\nolimits_{t \in I_d} \left((\hat{Y}_{w,i}(t) - Y_{w,i}(t)) - \mu_i \right)^2}. \quad (30b)$$

TABLE I
EMPIRICAL CDF OF ESTIMATION ERROR METRICS

Empirical CDF	0.1	0.2	0.5	0.7	0.9
$MAPE$ of $\hat{G}$ (%)	2.84	4.05	4.96	6.38	8.80
$MAPE$ of $\hat{P}$ (%)	1.63	2.15	2.80	3.67	4.92
MSE of $\hat{G}$	0.04	0.06	0.10	0.19	0.33
MSE of $\hat{P}$	0.03	0.05	0.09	0.18	0.29
CV of $\hat{G}$	-11.80	-5.13	-2.60	2.37	16.12
CV of $\hat{P}$	-11.30	-4.65	-2.59	1.77	10.90

Table I summarises the empirical cumulative distribution functions (CDFs) for the estimation $MAPE$, MSE, and CV, which are constructed using all the computed $MAPE$'s, MSE's, and CV's, respectively. As can be seen, for the estimated hourly PV generation, 70% of the $MAPE$'s are less than 6.38%. Regarding the estimated hourly native demand, 70% of the $MAPE$'s are less than 3.67%. This effectively verifies the estimation accuracy of our proposed approach. We also provide the percentiles of MSE and CV based on all the PV-installed customers' generation and native demand estimates, which can more comprehensively evaluate the performance of our approach.

Note that the above results are obtained under the conditions that (1) five produced candidate generation curves are employed ($N_e = 5$), (2) the tuning parameter in (20a) is 100 ($\lambda = 100$), and (3) the optimization process specified in (20) is executed for individual windows with a time length of one month ($T = 720$ hours, the entire year is divided into 12 windows).

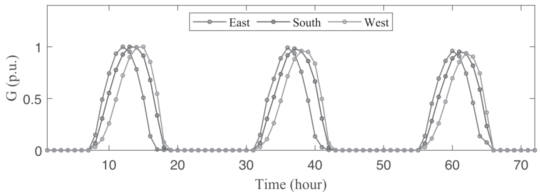


Fig. 17. Three-day produced candidate generation curves corresponding to three typical azimuths, i.e., east, south, and west.

TABLE II
IMPACT OF CANDIDATE GENERATION CURVES

Case	I	II	III
Average $MAPE$ of $\hat{G}$ (%)	5.677	5.474	5.473
Average $MAPE$ of $\hat{P}$ (%)	3.924	3.086	3.086
Runtime (s)	40	125	194

2) Testing the Candidate Generation Curves: As elaborated in Section IV-A, diverse candidate generation curves are produced for representing the unknown BTM generation. Thus, it is of interest to examine the effectiveness of producing candidate curves. Fig. 17 shows three produced candidate generation curves corresponding to three typical azimuths, i.e., east, south, and west, respectively. We can observe that compared to the generation curve corresponding to the south, the produced curve corresponding to the east is "left-skewed," and the produced curve corresponding to the west is "right-skewed". Therefore, the produced curves demonstrate diversity, which is consistent with our observation on real PV generation curves shown in Fig. 9.

In addition, we have also quantitatively examined the effectiveness of producing diverse candidate generation curves. Specifically, we test the impact of the number of candidate generation curves, i.e., we solve (20) separately for three cases with different numbers of candidate curves: (I) one candidate generation curve corresponding to the azimuth of south; (II) three candidate generation curves corresponding to the east, south, and west, respectively; and (III) five candidate generation curves corresponding to the east, southeast, south, southwest, and west, respectively. The other conditions for the three cases are the same: $\lambda = 100$ and $T = 720$ hours. To evaluate the impact of candidate number, we compute the average $MAPE$ over all PVs' $MAPE$'s obtained from (27). The results are summarized in Table II. We can see that as the candidate number increases, the estimation error decreases, and the execution time increases. In addition, the $MAPE$ for Case I is relatively greater than Case II and III, and Case II and Case III provide nearly identical $MAPE$'s. This is because three candidate curves - corresponding to the east, south, and west - can comprehensively represent the unknown BTM generation curve; adding extra candidate curves simply result in a slight accuracy improvement.

3) Testing the Tuning Parameter λ: As discussed in Section IV-C, λ in (20) reflects the confidence of estimating peak generations for individual PVs. One general principle for determining λ is that the largest element in $\boldsymbol{\gamma}$ is a couple of kilo-watts. In addition, the solutions for (20) should not be sensitive to λ, i.e., (20) should be robust to λ. To verify the robustness of our proposed approach, we solve (20) based on different values of λ, and then compute the corresponding average $MAPE$'s for the estimated PV generation and native demand. Other conditions are that $T = 720$ hours and five candidate generation curves - corresponding to the south, southeast, south, southwest, and west - are employed. The results show that for the λ's ranging from 100 to 500 with an interval of 100, the average $MAPE$'s for PV generation and native demand do not change (5.47% and 3.09%). The invariant average $MAPE$'s demonstrate the robustness of our proposed approach.

TABLE III
IMPACT OF WINDOW SIZE T

T (month)	1	2	3	4
Average $MAPE$ of $\hat{G}$ (%)	5.47	5.30	5.18	5.08
Average $MAPE$ of $\hat{P}$ (%)	3.09	2.99	2.92	2.87

4) Testing the Window Size T: Since our proposed approach can be conducted for each divided window, it is of importance to examine the impact of window size on estimation accuracy. To do this, we perform our approach for windows with different lengths and then compute the estimation $MAPE$. In Table III, it can be seen that the average $MAPE$ decreases as T increases. This is because for a wider window, the probability for the minimum diurnal native demand, $\underline{P}_{w,d,i}$, equaling the minimum nocturnal native demand, $\underline{P}_{w,n,i}$, is larger. Thus, we have a smaller estimation error for $\underline{P}_{w,d,i}$, as seen in (19). Then, based on (17) and (18), it can be seen that the smaller estimation error for $\underline{P}_{w,d,i}$ results in a more accurate $D_{w,i}$, which then brings a more accurate estimate for $G_{w,m,i}$. Finally, more accurate peak generation estimates result in smaller estimation errors for the PV generation and native demand time series.

D. Performance Comparison

This paper compares our proposed approach with previous works from two perspectives, qualitatively and quantitatively.

1) Qualitative Analysis: From a qualitative point of view, one primary advantage of our approach is that it does not require meteorological data and solar generation exemplars. For the aggregate level, our approach can perform PV generation estimation by only using recorded net demand data. For the customer level, our approach can also work by only relying on recorded smart meter data, although leveraging PVWatts Calculator's generated data can improve the estimation accuracy.

2) Quantitative Comparison: For the customer level, we have also compared our approach with previous works. Specifically, we focus on comparing our approach with the method presented in [22] and [11], which demonstrate better performance compared to previous works. Table IV summarizes the computed $MAPE$'s for our approach and the compared approach. Note that the average $MAPE$'s for our approach have lower and upper bounds because the considered window size, T, ranges from one month to four months. As can be seen, the approach in [22]

TABLE IV
AVERAGE $MAPE$ (%) COMPARISON

Approaches	Our Approach	Approach in [11]	Approach in [22]
$\hat{G}$	[5.08, 5.47]	7.38	5.24
$\hat{P}$	[2.87, 3.09]	9.94	2.95

TABLE V
AGGREGATE-LEVEL ESTIMATION $MAPE$ (%)

	W/O noise	Case 1	Case 2	Case 3	Case 4	Case 5
$\hat{G}$	1.21	1.17	1.22	1.38	1.53	1.73
$\hat{P}$	1.28	1.28	1.33	1.43	1.58	1.76

TABLE VI
AVERAGE CUSTOMER-LEVEL ESTIMATION $MAPE$ (%)

	W/O noise	Case 1	Case 2	Case 3	Case 4	Case 5
$\hat{G}$	5.47	5.84	5.86	5.64	5.54	5.62
$\hat{P}$	3.09	3.53	3.68	3.62	3.63	3.80

demonstrates a similar estimation accuracy as our approach does. However, our approach does not require solar exemplars, which makes it more independent and practical. The approach in [11] employs a statistical model and a physical model to represent the native load and the PV generation, respectively. Table IV shows that our approach has a better performance than the approach in [11] in terms of the average $MAPE$.

E. Robustness Against Measurement and Communication Noises

To test the robustness of our proposed approach, we add measurement and communication noises to the net demand measurements of customers with PVs and the native demand measurements of customers without PVs. For the measurement noise, we consider the Class 0.5 (having $\pm 0.5\%$ error) specified by ANSI C12.20. For the communication noise, we test five different packet loss rates considering that the packet loss rate depends on the communication bandwidth and data volume. For example, we purposely change 1% of the measurements to zero to achieve a 1% packet loss rate. To comprehensively evaluate our approach's performance, we set up five cases: Case 1 − 1% measurement lost + 0.5% random noise, Case 2 − 2% measurement lost + 0.5% random noise, Case 3 − 3% measurement lost + 0.5% random noise, Case 4 − 4% measurement lost + 0.5% random noise, and Case 5 − 5% measurement lost + 0.5% random noise. Then, we apply our approach to the above five cases and compute the $MAPE$ for evaluating the robustness. The results are summarized in Tables V and VI. We can observe that the $MAPE$'s slowly increase while the noise level increases, demonstrating the robustness of our approach.

F. Limitations of the Proposed Approach

Every method has its limitations, and there is no omnipotent method that can apply to all cases. The limitation of our proposed approach is that it requires time-series smart meter data with a temporal granularity that can distinguish daytime and nighttime. This is because our approach innovatively utilizes the temporal correlation between the aggregate *nocturnal* native demand and the aggregate *diurnal* native demand. Under this condition, only having access to the monthly demands of those PV-installed customers brings challenges to our approach because it cannot split the monthly demand into two parts, the diurnal and nocturnal demands, for computing the nocturnal native demand ratio. We intend to address this challenge in our future work.

VI. CONCLUSION

This paper is dedicated to proposing an independent and practical BTM solar power/native demand estimation approach. Our proposed approach contains two interconnected layers. The aggregate level leverages the spatial correlation of native demand to perform the aggregate PV generation/native demand estimation. The customer level utilizes the spatial correlation of PV generation to allocate the estimated aggregate PV generation/native demand to individual customers. The Case Study verifies that our approach can accurately estimate BTM PV generation/native demand, significantly enhancing distribution system observability and situation awareness. The numerical experiments also demonstrate that our approach does not require meteorological data and measured solar power exemplars. Therefore, our approach is more independent and thus is practical for utilities to implement.

REFERENCES

[1] J. Black and V. Rojo, "Long-term load forecast methodology overview," Sep. 2019. [Online]. Available: https://www.iso-ne.com/static-assets/documents/2019/09/p1_load_forecast_methodology.pdf Q3

[2] F. Wang, Z. Xuan, Z. Zhen, K. Li, T. Wang, and M. Shi, "A day-ahead PV power forecasting method based on LSTM-RNN model and time correlation modification under partial daily pattern prediction framework," *Energy Convers. Manage.*, vol. 212, no. 112766, pp. 1–14, May 2020.

[3] R. Seguin, J. Woyak, D. Costyk, J. Hambrick, and B. Mather, "High penetration PV integration handbook for distribution engineers," Nat. Renew. Energy Lab., Golden, CO, USA, Tech. Rep. NREL/TP-5D00-63114, 2016.

[4] T. A. Short, *Electric Power Distribution Handbook*. Boca Raton FL, USA: CRC Press, 2014.

[5] V. Krishnan and J. D. McCalley, "Building foresight in long-term infrastructure planning using end-effect mitigation models," *IEEE Syst. J.*, vol. 11, no. 4, pp. 2040–2051, Dec. 2017.

[6] W. Buehring, C. Huber, and J. Marques, "Expansion planning for electrical generating systems," Int. Atomic Energy Agency, Vienna, Austria, Tech. Rep. STI/DOC/10/241, 1984.

[7] W. Liu, D. Guo, Y. Xu, R. Cheng, Z. Wang, and Y. Li, "Reliability assessment of power systems with photovoltaic power stations based on intelligent state space reduction and pseudo-sequential Monte Carlo simulation," *Energies*, vol. 11, no. 6, 2018, Art. no. 1431.

[8] D. Chen and D. Irwin, "SunDance: Black-box behind-the-meter solar disaggregation," in *Proc. 8th Int. Conf. Future Energy Syst.*, 2017, pp. 16–19.

[9] Y. Wang, N. Zhang, Q. Chen, D. S. Kirschen, P. Li, and Q. Xia, "Data-driven probabilistic net load forecasting with high penetration of behind-the-meter PV," *IEEE Trans. Power Syst.*, vol. 33, no. 3, pp. 3255–3264, May 2018.

[10] F. Kabir, N. Yu, W. Yao, R. Yang, and Y. Zhang, "Estimation of behind-the-meter solar generation by integrating physical with statistical models," in *Proc. IEEE Int. Conf. Commun. Control Comput. Technol. Smart Grids*, 2019, pp. 1–6.

[11] F. Kabir, N. Yu, W. Yao, R. Yang, and Y. Zhang, "Joint estimation of behind-the-meter solar generation in a community," *IEEE Trans. Sustain. Energy*, vol. 12, no. 1, pp. 682–694, Jan. 2021.

[12] K. Li, F. Wang, Z. Mi, M. Fotuhi-Firuzabad, N. Duić, and T. Wang, "Capacity and output power estimation approach of individual behind-the-meter distributed photovoltaic system for demand response baseline estimation," *Appl. Energy*, vol. 253, 2019, Art. no. 113595.

[13] F. Wang *et al.*, "A distributed PV system capacity estimation approach based on support vector machine with customer net load curve features," *Energies*, vol. 11, no. 7, Jul. 2018, Art. no. 1750.

[14] C. Dinesh, S. Welikala, Y. Liyanage, M. P. B. Ekanayake, R. I. Godaliyadda, and J. Ekanayake, "Non-intrusive load monitoring under residential solar power influx," *Appl. Energy*, vol. 205, pp. 1068–1080, Aug. 2017.

[15] F. Sossan, L. Nespoli, V. Medici, and M. Paolone, "Unsupervised disaggregation of photovoltaic production from composite power flow measurements of heterogeneous prosumers," *IEEE Trans. Ind. Informat.*, vol. 14, no. 9, pp. 3904–3913, Sep. 2018.

[16] H. Shaker, H. Zareipour, E. Muljadi, and D. Wood, "A data-driven approach for estimating the power generation of invisible solar sites," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2466–2476, Sep. 2016.

[17] E. C. Kara, C. M. Roberts, M. D. Tabone, L. Alvarez, D. S. Callaway, and E. M. Stewart, "Disaggregating solar generation from feeder-level measurements," *Sustain. Energy, Grids Netw.*, vol. 13, pp. 112–121, 2018.

[18] K. Li, J. Yan, L. Hu, F. Wang, and N. Zhang, "Two-stage decoupled estimation approach of aggregated baseline load under high penetration of behind-the-meter PV system," *IEEE Trans. Smart Grid*, vol. 12, no. 6, pp. 4876–4885, Nov. 2021.

[19] J. Lin, J. Ma, and J. Zhu, "A privacy-preserving federated learning method for probabilistic community-level behind-the-meter solar generation disaggregation," *IEEE Trans. Smart Grid*, vol. 13, no. 1, pp. 268–279, Jan. 2022.

[20] S. Talkington, S. Grijalva, M. J. Reno, and J. A. Azzolini, "Solar PV inverter reactive power disaggregation and control setting estimation," *IEEE Trans. Power Syst.*, to be published, doi: 10.1109/TPWRS.2022.3144676.

[21] F. Bu, K. Dehghanpour, Y. Yuan, Z. Wang, and Y. Zhang, "A data-driven game-theoretic approach for behind-the-meter PV generation disaggregation," *IEEE Trans. Power Syst.*, vol. 35, no. 4, pp. 3133–3144, Jul. 2020.

[22] F. Bu, K. Dehghanpour, Y. Yuan, Z. Wang, and Y. Guo, "Disaggregating customer-level behind-the-meter PV generation using smart meter data and solar exemplars," *IEEE Trans. Power Syst.*, vol. 36, no. 6, pp. 5417–5427, Nov. 2021.

[23] F. Bu, K. Dehghanpour, Y. Yuan, and Z. Wang, "Quantifying load uncertainty using real smart meter data," in *Proc. IEEE Smart Grid Commun*, 2020, pp. 1–6.

[24] K. Nagasawa, C. R. Upshaw, J. D. Rhodes, C. L. Holcomb, D. A. Walling, and M. E. Webber, "Data management for a large-scale smart grid demonstration project in austin, texas," in *Proc. 6th Int. Conf. Energy Sustain.*, 2012, pp. 1027–1031.

[25] M. A. Khan and B. Hayes, "Smart meter based two-layer distribution system state estimation in unbalanced MV/LV networks," *IEEE Trans. Ind. Informat*, vol. 18, no. 1, pp. 688–697, Jan. 2022.

[26] A. A. Nasser, M. Z. Rashad, and S. E. Hussein, "A two-layer water demand prediction system in urban areas based on micro-services and LSTM neural networks," *IEEE Access*, vol. 8, pp. 147647–147661, 2020.

[27] Q. Wang, W. Tai, Y. Tang, M. Ni, and S. You, "A two-layer game theoretical attack-defense model for a false data injection attack against power systems," *Int. J. Electr. Power Energy Syst.*, vol. 104, pp. 169–177, 2019.

[28] Z. A. Khan, D. Jayaweera, and M. S. Alvarez-Alvarado, "A novel approach for load profiling in smart power grids using smart meter data," *Electr. Power Syst. Res.*, vol. 165, pp. 191–198, 2018.

[29] Q. Zhang, Y. Guo, Z. Wang, and F. Bu, "Distributed optimal conservation voltage reduction in integrated primary-secondary distribution systems," *IEEE Trans. Smart Grid*,vol. 12, no. 5, pp. 3889–3900, Sep. 2021.

[30] R. Cheng, Z. Wang, Y. Guo, and F. Bu, "Analyzing photovoltaic's impact on conservation voltage reduction in distribution networks," Oct. 2021, *arXiv:2110.14777*.

[31] A. P. Dobos, "PVWatts Version 5 Manual," Nat. Renew. Energy Lab., Golden, CO, USA, Tech. Rep. NREL/TP-6A20-62641, Sep. 2014.

Fankun Bu (Graduate Student Member, IEEE) received the B.S. and M.S. degrees from North China Electric Power University, Baoding, China, in 2008 and 2013, respectively. He is currently working toward the Ph.D. degree with the Department of Electrical and Computer Engineering, Iowa State University, Ames, IA, USA. From 2008 to 2010, he was a Electrical Commissioning Engineer for NARI Technology Company, Ltd., Nanjing, China. From 2013 to 2017, he was an Electrical Engineer for State Grid Corporation of China, Nanjing, China. His research interests include power system data analytics, behind-the-meter PV generation estimation, load and PV generation forecasting, distribution system modeling, renewable energy integration, and power system relaying.

Rui Cheng (Graduate Student Member, IEEE) received the B.S. degree in electrical engineering from Hangzhou Dianzi University, Hangzhou, China, in 2015 and the M.S. degree in electrical engineering from North China Electric Power University, Beijing, China, in 2018. He is currently working toward the Ph.D. degree with the Department of Electrical & Computer Engineering, Iowa State University, Ames, IA, USA. From 2018 to 2019, he was an Electrical Engineer with State Grid Corporation of China, Hangzhou, China. His research interests include power distribution systems, voltage/var control, transactive energy markets, and applications of optimization and machine learning methods to power systems.

Zhaoyu Wang (Senior Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from Shanghai Jiaotong University, Shanghai, China, and the M.S. and Ph.D. degrees in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA. He is currently the Northrop Grumman Endowed Associate Professor with Iowa State University, Ames, IA, USA. His research interests include optimization and data analytics in power distribution systems and microgrids. He was the recipient of the National Science Foundation CAREER Award, the Society-Level Outstanding Young Engineer Award from IEEE Power and Energy Society (PES), the Northrop Grumman Endowment, College of Engineering's Early Achievement in Research Award, and the Harpole-Pentair Young Faculty Award Endowment. He is the Principal Investigator for a multitude of projects funded by the National Science Foundation, the Department of Energy, National Laboratories, PSERC, and Iowa Economic Development Authority. He is the Chair of IEEE PES PSOPE Award Subcommittee, the Co-Vice Chair of PES Distribution System Operation and Planning Subcommittee, and the Vice Chair of PES Task Force on Advances in Natural Disaster Mitigation Methods. He is an Associate Editor for IEEE TRANSACTIONS ON POWER SYSTEMS, IEEE TRANSACTIONS ON SMART GRID, IEEE OPEN ACCESS JOURNAL OF POWER AND ENERGY, IEEE POWER ENGINEERING LETTERS, and *IET Smart Grid.*