Disaggregating Customer-level Behind-the-Meter PV Generation Using Smart Meter Data

Fankun Bu, Student Member, IEEE, Kaveh Dehghanpour, Member, IEEE, Yuxuan Yuan, Student Member, IEEE, Zhaoyu Wang, Member, IEEE, and Yifei Guo, Member, IEEE

Abstract—Customer-level rooftop photovoltaic (PV) has been widely integrated into distribution systems. In most cases, PVs are installed behind-the-meter (BTM) and only the net demand is recorded. Therefore, the native demand and PV generation are unknown to utilities. Separating native demand and solar generation from net demand is critical for improving grid-edge observability. In this paper, a novel approach is proposed for disaggregating customer-level BTM PV generation using low-resolution but widely available smart meter data. The proposed approach exploits the high correlation between monthly nocturnal and diurnal native demands. First, a joint probability density function (PDF) of monthly nocturnal and diurnal native demands is constructed for customers without PVs, using Gaussian mixture modeling (GMM). Deviation from the constructed PDF is leveraged to probabilistically assess the monthly solar generation of customers with PVs. Then, to identify hourly BTM solar generation for these customers, their estimated monthly solar generation is decomposed into an hourly timescale; to do this, we have proposed a maximum likelihood estimation (MLE)-based technique that takes advantage of hourly typical solar exemplars. Unlike previous disaggregation methods, our approach does not require native demand exemplars or knowledge of PV model parameters, which makes it robust against volatility of customers’ load and enables highly-accurate disaggregation. The proposed approach has been verified using real smart meter data.

Index Terms—Rooftop photovoltaic, distribution system, Gaussian mixture model, maximum likelihood estimation.

I. INTRODUCTION

In practice, customer-level rooftop PVs are integrated into distribution systems at behind-the-meter (BTM), where only the net demand is recorded. The measured net demand equals native demand minus the PV generation, which are unknown to utilities separately. The native demand refers to the original demand consumed by home appliances. The invisibility of native demand and BTM solar generation poses challenges in distribution network design [1], [2], operation [3], [4], [5] and expansion [6], [7]. Thus, disaggregating PV generation from net demand is of significance to utilities.

Previous works regarding PV generation disaggregation can be classified into two categories based on the scale of solar power: Class I - Customer-level approaches: In [8], customer PV generation is estimated by combining a PV performance model with a clear sky model, and using meteorological/geographical data. In [9], a non-intrusive load monitoring (NILM) approach is proposed to disaggregate customers’ PV generation from their net demand using measurements with 1-second resolution. In [7], [10], a data-driven method is proposed for estimating the capacity and power output of residential rooftop PVs using customers’ net load curve features. In [11], a physical PV performance model is combined with a statistical load estimation model, along with weather data to identify key PV array parameters. The disadvantages of previous customer-level approaches are as follows: dependency on the availability of accurate native demand exemplars, unavailability of detailed PV model parameters, requiring high-resolution sensors and weather data. These obstacles make the previous methods susceptible to the uncertainties of customer behavior and rooftop solar power generators, which result in a decline in disaggregation accuracy.

Class II - System-level approaches: Many previous works have proposed methods to disaggregate solar power from net demand at transformer-, feeder-, and regional-levels. In [12], a data-driven approach is presented for separating the aggregate solar power of groups of customers using their service transformer measurements. In [13], an exemplar-based disaggregator is proposed to separate the output power of an unobservable solar farm from the feeder-level μPMU measurements, using power measurements of nearby observable PV plants and irradiance data. In [6], a regional-scale equivalent PV station model is proposed to represent the total generation of small-scale PVs. The model parameters are optimized using known solar power data. In [14], a data-driven approach is proposed to estimate the total rooftop PV generation in a region by installing temporary sensors to measure representative solar arrays. Furthermore, previously in [15], we developed a game-theoretic data-driven approach for disaggregating the PV generation of sizeable groups of customers using solar and load exemplars. However, Class II approaches lack sufficient accuracy for performing customer-level PV disaggregation.

Considering the shortcomings of previous approaches, in this paper we propose a novel customer-level solar power disaggregation technique that does not require native demand exemplars, which are difficult to obtain due to the high volatility of customers’ load [16]. Also, our approach is purely data-driven and only leverages widely-deployed smart meter data. Thus, compared with previous Class I methods, our method is more robust against customer-level data uncertainties and does not rely on the knowledge of PV parameters and weather
Our idea is to first estimate each customer’s monthly BTM PV generation and then decomposes it into hourly solar power using solar exemplars. Note that unlike native demand exemplars, solar exemplars can be easily constructed from observable PVs, due to the strong spatial correlation in irradiance in geographically-bounded distribution systems. The key in eliminating native demand exemplars from the disaggregation process is based on an observation from our real smart meter data: the monthly nocturnal and diurnal native demands are highly correlated; since customers with and without PV have very different diurnal smart meter readings (yet similar nocturnal records), the observed correlation can be used for identifying the monthly BTM solar generation.

The first step is to construct the joint probability density function (PDF) of monthly nocturnal and diurnal native demands for customers without PVs. This will be done using a Gaussian Mixture Model (GMM) technique [17], which has demonstrated significant flexibility in forming smooth approximations to arbitrarily-shaped PDFs. The constructed joint PDF captures the monthly load characteristics of customers without PVs; hence, this joint PDF serves as a benchmark for evaluating the deviations caused by monthly BTM solar generation for customers without observable PVs. The second step is to project the obtained customer-level monthly solar estimations onto hourly values; to do this, the monthly BTM solar generations are represented as a linear weighted summation of solar exemplars with hourly resolution. The weights are optimized using a constrained maximum likelihood estimation (MLE) process, and will be leveraged to disaggregate the hourly net demand of customers with BTM PV generators. To enhance the robustness of MLE against missing and bad data, a penalty term is integrated into the weight identification process. Throughout the paper, vectors are denoted using bold italic letters and matrices are denoted as bold uppercase letters.

The rest of the paper is organized as follows: Section II introduces the overall framework for customer-level BTM PV generation disaggregation and describes smart meter dataset. Section III presents the process for constructing joint PDF of monthly diurnal and nocturnal native demands. Section IV describes the procedure of formulating and solving MLE to perform disaggregation. In Section V, case studies are analyzed and Section VI concludes the paper.

II. OVERALL DISAGGREGATION FRAMEWORK AND DATASET DESCRIPTION

A. Overall Framework

In distribution systems, residential customers can be typically categorized into three types: (I) \( C_P \) is the set of customers without PVs whose native demand is recorded by smart meters. (II) \( C_G \) denotes the small group of customers with PVs whose PV generation and native demand are both observable separately. (III) \( C_N \) represents the set of customers with PVs whose net demand is recorded by smart meter, while their native demand and PV generation are not separately visible. Our goal is to disaggregate PV generation and native demand from the net demand of individual customers in \( C_N \).

The overall process is illustrated in Fig. 1. First, the known monthly nocturnal and diurnal native demands of customers in \( C_P \) are employed to construct a joint PDF using GMM modeling technique. This joint PDF is constructed based on a sizeable number of customers without PVs. Then, for each customer in \( C_N \), the unknown PV generation is optimally estimated by performing MLE, and using the constructed joint PDF, known monthly net demand and solar exemplars.

B. Dataset Description

The hourly native demand and PV generation data used in this paper are from Midwest U.S. utilities [18]. The time range of solar power is one year, and the time range of native demand of customers without PVs is three years. This system consists of 1120 customers, of which 480 are residential customers without PVs and 337 are residential customers with PVs. The nominal capacity of PVs ranges from 3 kW to 8 kW. Net demand data is obtained by aggregating customers’ PV generation and native demand data.

III. STATISTICAL MODELING OF MONTHLY NATIVE DEMAND

A. Findings from Real Smart Meter Data

One key finding which sets the foundation for the proposed disaggregation approach is that the monthly nocturnal native demand and the monthly diurnal native demand are highly correlated, as shown in Fig. 2a. The importance of this observation is that it can be leveraged to reveal the monthly BTM generation of customers with PVs. For instance, consider two customers, one with PV and one without PV. These two customers can have statistically-similar monthly nocturnal net demand, however, their monthly diurnal net demand will be significantly different from a statistical perspective due to
BTM PV generation at daytime. Specifically, Fig. [2b] shows the nocturnal-diurnal net demand distribution for customers with PV which is significantly different from Fig. [2a]. Thus, the distribution shown in Fig. [2a], which represents the behavior of customers without PV, can be used as a benchmark to determine whether a customer has BTM PV generation and estimate the monthly solar power. These findings have inspired us to construct a joint distribution of monthly nocturnal and diurnal native demands of customers without PVs to evaluate the deviation caused by the BTM PV generation of customers with PVs. These deviations correspond to monthly BTM solar generation.

B. Constructing the Nocturnal-Diurnal Native Demand PDF

We use a parametric PDF estimation technique known as GMM to construct the joint distribution of known monthly nocturnal and diurnal native demands of customers without PVs. A GMM is a linear combination of Gaussian components, and has demonstrated high flexibility and robustness in modeling arbitrary distributions [19]. Since utilities have access to a large amount of native demand data, the constructed GMM-based joint PDF is able to probabilistically describes the quantitative relationship between the monthly nocturnal native demand and monthly diurnal native demand for customers without PVs. The native demand of customers with PVs also follow this joint PDF, while their observed monthly net demands can deviate from the joint distribution. Compared with empirical histograms, the GMM-based PDF only has a limited number of parameters, therefore, it can be conveniently leveraged for estimating the BTM PV generation of the customers with PVs. In our problem, the GMM approximation model can be described as follows:

\[
f(P_{m,n}, P_{m,d} | \Lambda) = \sum_{j=1}^{S} \theta_j g_j(P_{m,n}, P_{m,d} | \mu_j, \Sigma_j),
\]

where, \( f(\cdot, \cdot) \) denotes the approximated joint PDF, \( P_{m,n} \) and \( P_{m,d} \) denote the monthly nocturnal and diurnal native demands of customers without PVs (i.e., customers belonging to \( C_P \)), respectively. \( \Lambda \) denotes the parameter collection, \{\( S, \theta_j, \mu_j, \Sigma_j \)\}, which needs to be learned based on known native demand data. \( S \) denotes the total number of Gaussian components. \( \theta_j \)'s are the weights corresponding to the bi-variate Gaussian components \( g_j(Z | \mu_j, \Sigma_j) \) with \( Z = [P_{m,n}, P_{m,d}] \), which satisfy \( \sum_{j=1}^{S} \theta_j = 1 \) and \( 0 \leq \theta_j \leq 1 \). The bi-variate Gaussian component is defined as

\[
g_j(Z | \mu_j, \Sigma_j) = \frac{1}{(2\pi)^{S_j/2} |\Sigma_j|^{1/2}} \exp \left\{ -\frac{1}{2} (Z - \mu_j)^\top \Sigma_j^{-1} (Z - \mu_j) \right\},
\]

where, \( \mu_j \) and \( \Sigma_j \) are the Gaussian component mean vector and covariance matrix, respectively.

To learn \( \Lambda \), first, a dataset is constructed based on smart meter measurements of customers in \( C_P \). In practice, \( P_{m,n} \) and \( P_{m,d} \) of customers in \( C_P \) are known to utilities and can be obtained from hourly smart meter readings in each month:

\[
P_{m,n} = \sum_{t \in I_n} P_h(t),
\]

\[
P_{m,d} = \sum_{t \in I_d} P_h(t),
\]

where, \( P_h(t) \) denotes the native demand reading at the \( t \)'th hour, \( I_n \) and \( I_d \) denote the sets of nighttime and daytime hours, respectively. Then, we can obtain the matrix of monthly demands by concatenating all customers’ monthly native demand pairs:

\[
Z = [Z(1), \cdots, Z(N_c)]^\top
\]

where, \( N_c \) denotes the total number of customers, and \( Z(j) \) denotes a matrix of monthly nocturnal and diurnal native demand pairs of the \( j \)'th customer which is organized as follows:

\[
Z(j) = \begin{bmatrix}
P_{m,n}(j, 1) & P_{m,d}(j, 1) \\
P_{m,n}(j, 2) & P_{m,d}(j, 2) \\
\vdots & \vdots \\
P_{m,n}(j, N_m) & P_{m,d}(j, N_m)
\end{bmatrix}
\]

where, \( N_m \) is the total number of months. Then, we can obtain a dataset of observed monthly demand samples, \{\( Z(1), \cdots, Z(N') \)\}, through partitioning \( Z \) by rows, where, \( N' = N_c \times N_m \).

Thus, the problem of GMM approximation boils down to finding optimal parameter collection \( \Lambda^* \) that best fits the obtained dataset of monthly native demands, \( Z \), by assuming that the data samples are drawn independently from the underlying distribution. The most well-established idea for learning GMM parameters is to solve an optimization problem [17], [20], whereby the objective function can be formulated to maximize data likelihood, as follows:

\[
\max_{\Lambda} \prod_{i=1}^{N'} f(Z(i) | \Lambda),
\]

By taking the logarithm of objective function, (6) is rewritten as follows:

\[
\max_{\Lambda} \sum_{i=1}^{N'} \ln \left\{ f(Z(i) | \Lambda) \right\}.
\]

The optimization problem in (7) is solved using the expectation-maximization algorithm [17].

Based on the identified optimal GMM parameter collection from (7), \( \Lambda^* \), the joint PDF of monthly nocturnal and diurnal native demands can be specifically written as

\[
f(P_{m,n}, P_{m,d}) = \sum_{j=1}^{S^*} \theta^*_j g_j^*(P_{m,n}, P_{m,d}),
\]
The power generation profile of an individual PV is primarily determined by PV array capacity and orientation. The orientation determines the trajectory distortion of generation profile [12]. Therefore, the unknown BTM PV generation can be reliably represented using known generation profiles of BTM PVs (belonging to \( C_G \)) with typical orientations that serve as exemplars:

\[
G_{m,d} = \sum_{i=1}^{N} \omega_i G_{m,i}^{E} = \omega^T G_{m}^{E},
\]

where, \( N \) is the total number of solar exemplars, \( \omega = [\omega_1, \cdots, \omega_N]^T \) denotes an unknown weight vector to be optimized, and \( G_{m,i}^{E} = [G_{m,i,1}^{E}, \cdots, G_{m,i,N}^{E}]^T \) denotes the PV generation vector of solar exemplars, where, \( G_{m,i} \) is obtained by converting the known hourly diurnal PV generation into monthly diurnal solar power exemplars:

\[
G_{m,i}^{E} = \sum_{t \in M} G_{m,i}^{E}(t),
\]

where, \( G_{m,i}^{E}(t) \) is the PV generation of the \( i \)th exemplar at the \( t \)th hour. Therefore, disaggregating BTM PV generation of each customer in \( C_N \) comes down to finding optimal coefficients assigned to known solar exemplars. To do this, first, we represent the unknown monthly diurnal native demand using the known monthly demand and monthly PV generation of solar exemplars:

\[
P_{m,d} = P_{m,d}^\prime - \omega^T G_{m}^{E},
\]

where, \( P_{m,d}^\prime \) is the known monthly net demand which can be obtained as follows:

\[
P_{m,d}^\prime = \sum_{t \in M} P_h(t),
\]

where, \( P_h(t) \) denotes the recorded net demand at the \( t \)th hour.

Since the monthly nocturnal and diurnal native demands of customers with PVs probabilistically follow the constructed GMM-based joint PDF, by substituting [12] into [8], we can represent the distribution function for customers with BTM PVs as follows:

\[
f(P_{m,n}, P_{m,d}^\prime - \omega^T G_{m}^{E}),
\]

Then, the exemplar weight optimization is formulated as an MLE problem described as follows:

\[
\omega^* = \max_{\omega} \left\{ \prod_{i=1}^{M} f(P_{m,n}(i), P_{m,d}(i), G_{m}(i)|\omega) \right\},
\]

where, \( M \) is the total number of months.

Further, the optimization solution should be subject to multiple constraints to enforce the identified PV generation to be non-positive and the estimated native demand to be non-negative. Finally, by taking logarithm of [15] and introducing the constraints, the complete optimization problem is elaborated as follows:

\[
\max_{\omega} \left\{ \sum_{i=1}^{M} \ln \left[f(P_{m,n}(i), P_{m,d}(i), G_{m}(i)|\omega)\right] \right\} - \frac{1}{2} \lambda \| \beta \|^2_2,
\]

(16a)
Algorithm 1 Disaggregating BTM PV generation and native demand from net demand for each customer

1: Classify residential customers into three types: $C_P$, $C_G$, and $C_N$
2: procedure DATA CONVERSION
3: For customers in $C_P$:
4: \[ P_{m,n} \leftarrow \sum_{t \in I_{k}} P_h(t), \ P_{m,d} \leftarrow \sum_{t \in I_d} P_h(t) \]
5: For customers in $C_G$:
6: \[ G^E_{m,i} \leftarrow \sum_{t \in I_{k}} G^E_{h,i}(t) \quad i = 1, \cdots, N \]
7: For customers in $C_N$:
8: \[ P_{m,n} \leftarrow \sum_{t \in I_{d}} P'_{h}(t), \ P'_{m,d} \leftarrow \sum_{t \in I_d} P'_h(t) \]
9: end procedure
10: procedure CONSTRUCT NOCTURNAL-DIURNAL NATIVE DEMAND PDF
11: For customers in $C_P$:
12: \[ \Lambda \left\{ \theta_j, \mu_j, \Sigma_j \right\} \quad j = 1, \cdots, S \]
13: \[ \Lambda^* \leftarrow \max_{\Lambda} \sum_{i=1}^{N} \ln\{ f(P_{m,n}, P_{m,d} | \Lambda) \} \]
14: end procedure
15: procedure PERFORM MLE FOR OPTIMIZING WEIGHTS
16: For customers in $C_N$:
17: \[ P_{m,d} \leftarrow I_{m,d} - \omega^T(G^E_{m}) \]
18: Solve optimization in (16) to obtain $\omega^*$
19: end procedure
20: procedure ESTIMATE HOURLY BTM PV GENERATION AND NATIVE DEMAND
21: For customers in $C_N$:
22: \[ \hat{G}_h \leftarrow (\omega^*)^T G^E_{h}, \hat{P}_h \leftarrow P'_h - \hat{G}_h \]
23: end procedure

s.t. \[ (\omega^T G^E_{h})^T \leq 0, \] (16b)
\[ P'_h - (\omega^T G^E_{h})^T \geq \beta, \] (16c)
\[ \beta \leq 0. \] (16d)

where, $G^E_{h} = \{G^E_{h}(1), \cdots, G^E_{h}(N_h)\}$ denotes a matrix of hourly PV generation solar exemplars’ time series, $G^E_{h}(k) = \{G^E_{h,1}(k), \cdots, G^E_{h,N}(k)\}^T$, $k = 1, \cdots, N_h$ denotes the vector of solar exemplars’ generation readings at the $k$'th hour, $N_h$ denotes the total number of hourly demand readings, $P'_h$ denotes the time-series hourly net demand readings and 0 represents a zero vector. In addition to maximizing the likelihood function shown in (15), a $l_2$-norm penalty term, \[ -\frac{1}{2} \lambda \| \beta \|_{2}^{2}, \] is added into the objective function, where, $\lambda \geq 0$ is a tuning parameter and $\beta$ is a vector with non-positive elements. Constraint (16b) ensures that the estimated hourly PV generation is non-positive. Constraints (16c) and (16d) ensure that the estimated time-series native demand is larger than a non-positive vector whose $l_2$-norm is penalized in the objective function. This penalty term is based on the following consideration: In practice, it is common for the solar generation to have data quality problems. For example, PV arrays can stop running due to solar panel failures. For the customers whose PV generation is supposed to be disaggregated from the known net demand, the unwanted PV failure does not cause significant disaggregation error. This is because the missing or zero PV generation samples cause an erroneous rise in the net demand readings only for a limited number of samples. These larger net demand readings can still give us positive estimated native demand values, since the native demand is estimated by subtracting the disaggregated BTM PV generation from net demand. In comparison, the zero readings of solar exemplars can cause a negative estimated native demand, which brings significant estimation errors. This is because removing a zero PV generation from a negative net demand measurement gives us a negative estimated native demand value. Thus strictly constraining the estimated native demand to be non-negative can cause unwanted errors. Therefore, we have leveraged a soft margin to penalize the effect of bad or missing data. The MLE problem in (16) is solved via numerical optimization using interior-point methods.

B. Estimating Hourly PV Generation and Native Demand

By solving the optimization (16), we can obtain the optimized weight vector, $\omega^*$, which is utilized to estimate the unknown hourly BTM PV generation of customers with PVs:

\[ \hat{G}_h = (\omega^*)^T G^E_{h}. \] (17)

Further, the hourly native demand can be estimated by subtracting the disaggregated BTM PV generation from known net demand readings:

\[ \hat{P}_h = P'_h - \hat{G}_h. \] (18)

An algorithmic overview of the aforementioned steps of BTM PV generation disaggregation is summarized in Algorithm 1.
V. CASE STUDY

In this section, the proposed customer-level rooftop BTM solar power separation approach is verified using real smart meter data described in Section II.

A. Assessing Statistical Behavior of Customers

The empirical histogram and the GMM-based estimation of $f(P_{m,n}, P_{m,d})$ are shown in Fig. 4a and Fig. 4b, respectively. Comparing these two figures, it can be seen that GMM is able to accurately model the joint distribution of monthly nocturnal and diurnal native demands using smooth parametric Gaussian density functions. Also note that the joint PDF surface is quite narrow, i.e., the data is highly concentrated around the linear representative of nocturnal and diurnal demands. This corroborates the high correlation between monthly nocturnal and diurnal native demands observed in Fig. 2a.

B. BTM PV Generation Disaggregation Validation

Using the constructed GMM-based joint PDF, along with the known monthly net demand of customers with PVs and PV generation of solar exemplars, we can solve the MLE problem described in (16). When selecting solar exemplars, it is demonstrated that on average, three exemplars can sufficiently represent the PV generation profiles, and introducing additional solar exemplars does not bring further disaggregation accuracy improvement [15]. Thus, we have selected three typical solar power curves from $C_G$ corresponding to PVs facing east, south and west, respectively. Fig. 5 shows disaggregated PV generation and native demand curves of one customer over two weeks, along with corresponding actual profiles. In Fig. 5a, it can be seen that the disaggregated curve closely fits the actual profile, regardless of the solar volatility on some days. This shows the accurate diaggereation capability of our proposed method and also corroborates our observation that PV generation profiles with similar PV array orientations are highly correlated. Fig. 5b shows the disaggregated and actual native demand profiles. It can be observed that despite the uncertain and volatile pattern of native demand, the disaggregated curve can still fit the real profile.

B. BTM PV Generation Disaggregation Validation

Using the constructed GMM-based joint PDF, along with the known monthly net demand of customers with PVs and PV generation of solar exemplars, we can solve the MLE problem described in (16). When selecting solar exemplars, it is demonstrated that on average, three exemplars can sufficiently represent the PV generation profiles, and introducing additional solar exemplars does not bring further disaggregation accuracy improvement [15]. Thus, we have selected three typical solar power curves from $C_G$ corresponding to PVs facing east, south and west, respectively. Fig. 5 shows disaggregated PV generation and native demand curves of one customer over two weeks, along with corresponding actual profiles. In Fig. 5a, it can be seen that the disaggregated curve closely fits the actual profile, regardless of the solar volatility on some days. This shows the accurate diaggereation capability of our proposed method and also corroborates our observation that PV generation profiles with similar PV array orientations are highly correlated. Fig. 5b shows the disaggregated and actual native demand profiles. It can be observed that despite the uncertain and volatile pattern of native demand, the disaggregated curve can still fit the real profile.

It is of importance to examine the representative feature of typical solar exemplars. In (10), the unknown BTM PV generation is represented using known generation profiles of solar exemplars. Therefore, these PV generation profiles which serve as exemplars should be distinguishable, otherwise, multiple solutions of weights with the same losses can be derived in the MLE optimization process. We have employed a dimensionality reduction technique known as t-SNE to visualize the dissimilarities among PV generation profiles of solar exemplars. Note that each time point is treated as one dimension in our problem. The dimensions of hourly and monthly PV generation time series are reduced for convenient visualization, as shown in Fig. 6. Fig. 6a shows the reduced two-dimensional solar power exemplars based on the hourly PV generation of PVs facing east, south and west. As can be seen, the solar exemplars are demonstrated to be distinct. Similarly, the monthly PV generation of solar exemplars also demonstrate distinguishable features, as shown in Fig. 6b. This is consistent with our observation that solar generation
Time (hour)
East South West

(a) Solar exemplars

(b) A PV facing east

(c) A PV facing south

(d) A PV facing west

Fig. 7. The proposed approach can correctly track proper solar exemplars to perform disaggregation.

profiles are primarily determined by PV panel orientations in geographically-bounded distribution systems.

It is of significance to test whether the proposed approach can track the appropriate exemplars (east, south or west) in the disaggregation process. Fig. 7a shows PV generation curves of the three exemplars facing east, south and west. We can see that PVs with different orientations show distinct profile distortions. Fig. 7b shows the disaggregated and real PV generation curves of a PV facing east, along with the optimized weights assigned to the three solar exemplars. It can be seen that the weight corresponding to the first exemplar (i.e., PV facing east) is much larger compared to the other two weights, which validates the tracking ability of our proposed approach. This verification can also be observed in Fig. 7c and 7d which show the weights, disaggregated and actual PV generation curves of PVs facing south and west, respectively. In all cases, our method has accurately detected the correct underlying BTM PV panel orientations.

The proposed customer-level BTM solar separation approach is applied to all 337 customers with PVs, and the disaggregation accuracy is evaluated in terms of mean absolute percentage error (MAPE), which is calculated as follows:

\[
MAPE = \frac{100}{NH} \sum_{t=1}^{NH} \left| \frac{\hat{O}_h(t) - O_h(t)}{O_h(t)} \right|
\]

where, \(O_h\) can be \(P_h\) or \(G_h\). Fig. 8 shows the distribution of disaggregation error for all customers in terms of MAPE. As can be seen, majority of the MAPEs are less than 20%. This effectively demonstrates the generalization ability of our proposed method. Table I summarises the empirical cumulative distribution function (CDF) of disaggregation MAPE. As can be seen, for the disaggregated hourly PV generation, 80% of the MAPEs are less than 13.5%. Regarding the disaggregated hourly native demand, 80% of the MAPEs are less than 14.9%. This effectively verifies the disaggregation accuracy of our proposed approach.

| Empirical CDF | 0.2 | 0.4 | 0.6 | 0.8 | 1.0 |
|---------------|-----|-----|-----|-----|-----|
| MAPE of \(G_h\) (%) | 2.5 | 4.8 | 9.7 | 13.5 | 33.4 |
| MAPE of \(P_h\) (%) | 3.1 | 8.3 | 12.3 | 14.9 | 29.1 |

C. Testing the Robustness of the Proposed Approach

To tackle the missing and erroneous PV generation samples, a penalty term has been introduced into (16) for enhancing the proposed algorithm’s robustness. Fig. 9 compares one-week disaggregated PV generation and native demand curves using the proposed approach with and without penalty term. The actual solar power and native demand curves are also plotted.
as benchmarks. In Fig. 9a it can be seen that the disaggregated PV generation curve using the proposed approach with penalty can closely fit the actual curve, while the disaggregated PV generation curve using the approach without penalty significantly deviates from actual benchmark. The overestimation of PV generation is due to the constraint that the estimated native demand should be strictly non-negative, which causes overshoot in presence of erroneous data. The same conclusion can be derived in Fig. 9b. To sum up, the introduction of penalty into the MLE optimization significantly enhances the robustness of our proposed approach against missing or bad data.

VI. CONCLUSION

This paper presents a novel robust approach to disaggregate invisible customer-level BTM PV generation and native demand from net demand using smart meter data. The proposed method employs a limited number of typical solar power exemplars, and does not rely on native demand exemplars. Also, the proposed approach innovatively leverages the significant correlation between nocturnal and diurnal native demands in the timescale of month to alleviate native demand volatility. In addition, a penalty term is innovatively integrated into the estimation problem to tackle missing or bad data. The numerical experiments verify that the approach is able to perform disaggregation with satisfactory accuracy and robustness, which further improves utilities’ situational awareness of grid-edge resources. The key findings of the paper are summarized as follows:

- Despite the uncertainty of hourly native demand, the monthly nocturnal and diurnal native demands are highly correlated. This has inspired us to first estimate the monthly PV generation, then decompose it into hourly solar power.
- Missing or bad data of PV generation is common in practice, and can cause significant disaggregation error. This has motivated us to introduce a penalty term into MLE to reduce the impact of missing and bad data.

REFERENCES

[1] F. Ding and B. Mather, “On distributed PV hosting capacity estimation, sensitivity study and improvement,” IEEE Trans. Sustain. Energy, vol. 8, no. 3, p. 10101020, Jul. 2017.
[2] Y. Zhang, J. Wang, and Z. Li, “Uncertainty modeling of distributed energy resources: Techniques and challenges,” Current Sustain. Energy Rep., vol. 6, no. 2, pp. 42–51, Jun. 2019.
[3] R. Seguin, J. Woyak, D. Costyk, J. Hambrick, and B. Mather, High Penetration PV integration handbook for distribution engineers. Golden, CO, USA: Nat. Renewable Energy Lab., 2016.
[4] B. Chen, C. Chen, J. Wang, and K. L. Butler-Purry, “Sequential service restoration for unbalanced distribution systems and microgrids,” IEEE Trans. Power Syst., vol. 33, no. 2, pp. 1507–1520, Mar. 2018.
[5] K. Sun, Y. Hou, W. Sun, and J. Qi, Renewable and Energy Storage in System Restoration. Hoboken, NJ, USA: Wiley-IEEE Press, 2019.
[6] 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.
[7] K. Li, F. Wang, Z. Mi, M. Fotuhi-Firuzabad, N. Dui, 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, p. 115595, 2019.
[8] D. Chen and D. Irwin, “Sundance: Black-box behind-the-meter solar disaggregation,” in e-Energy, pp. 16–19, May 2017.
[9] 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.
[10] F. Wang, K. Li, X. Wang, L. Jiang, J. Ren, Z. Mi, M. Shafie-khah, and J. Catalo, “A distributed PV system capacity estimation approach based on support vector machine with customer net load curve features,” Energies, vol. 11, no. 7, p. 1750, Jul. 2018.
[11] 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 IEEE SmartGridComm, pp. 1–6, Oct. 2019.
[12] 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.
[13] 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.
[14] 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.
[15] 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., pp. 1–1, 2020.
[16] S. Wang, Y. Dong, L. Wu, and B. Yan, “Interval overvoltage risk based PV hosting capacity evaluation considering PV and load uncertainties,” IEEE Trans. Smart Grid, pp. 1–1, 2019.
[17] C. M. Bishop, Pattern Recognition and Machine Learning. New York: USA: Springer, 2009.
[18] F. Bu, Y. Yuan, Z. Wang, K. Dehghanpour, and A. Kimber, “A time-series distribution test system based on real utility data,” 2019 North American Power Symposium, pp. 1–6, Oct. 2019.
[19] D. A. Reynolds, “Gaussian mixture models,” in Encyclopedia of biometrics, 2nd Edition, 2015, pp. 827–832.
[20] I. Friedman, T. Hastie, and R. Tibshirani, The Elements of Statistical Learning. Springer: Springer Series in Statistics, 2001.