Estimation of Behind-the-Meter Solar Generation by Integrating Physical with Statistical Models

Farzana Kabir, Nanpeng Yu
Electrical and Computer Engineering
University of California, Riverside
Riverside, CA 92521-0429, USA
fkabi001,nyu@ucr.edu

Weixin Yao
Department of Statistics
University of California, Riverside
Riverside, CA 92521-0429, USA
weixin.yao@ucr.edu

Rui Yang,Yingchen Zhang
National Renewable Energy Laboratory
Golden, CO 80401, USA
rui.yang,yingchen.zhang@nrel.gov

***Abstract*—Accurate estimation of solar photovoltaic (PV) generation is crucial for distribution grid control and optimization. Unfortunately, most of the residential solar PV installations are behind-the-meter. Thus, utilities only have access to the net load readings. This paper presents an unsupervised framework for estimating solar PV generation by disaggregating the net load readings. The proposed framework synergistically combines a physical PV system performance model with a statistical model for load estimation. Specifically, our algorithm iteratively estimates solar PV generation with a physical model and electric load with the Hidden Markov model regression. The proposed algorithm is also capable of estimating the key technical parameters of the solar PV systems. Our proposed method is validated against net load and solar PV generation data gathered from residential customers located in Austin, Texas. The validation results show that our method reduces mean squared error by 42% compared to the state-of-the-art disaggregation algorithm.**

I. Introduction

Solar generation is an economically attractive source of renewable energy. Residential and commercial PV adoptions are increasing rapidly around the world [1], [2]. Most residential solar PV systems are deployed behind the smart meters installed by the electric utilities. Hence, the utilities can only collect the net load data [3], which equals the difference between electric load and solar PV generation. An accurate estimation of the solar PV generation is crucial to an array of distribution system operation and planning activities, including hosting capacity analysis, feeder/substation net load forecasting, Volt VAR control [4] and distribution network reconfiguration. The lack of visibility into the behind-the-meter solar generation brings many operational and planning challenges to the distribution system operators.

A supervised net load disaggregation algorithm is a possible approach to estimate solar generation. However, it requires the historical solar generation and electric load data for individual customers which are generally not available to the electric utilities. Another possible approach is to directly calculate solar PV generation from PV system performance models. However, this approach is not realistic because the technical parameters of solar PV systems are often unknown and can change over time. Therefore, an unsupervised approach to net load disaggregation is the ideal solution for electric utilities.

A few research teams developed unsupervised net load disaggregation algorithms to estimate solar generation for individual customers. Smart meter data, substation monitoring data, and solar proxy information are leveraged to estimate solar generation of individual homes located on the same distribution feeder in the algorithm developed by Tabone et al. [5]. Although this approach has potential for real-time disaggregation, it requires the installation of smart meters at every house served by the distribution feeder and perfect knowledge of which customers are served by it. The net load disaggregation problem is formulated as a convex optimization problem in the consumer mixture model [6], where the load consumption behavior of a customer is modeled by a mixture of representative customers without solar PV systems. The 'SunDance' technique [7] not only disaggregates net load signal but also estimates the geometry of the solar PV systems. The 'SunDance' algorithm has two key modules which estimate a location's maximum clear sky solar generation potential and model the universal weather-solar effect.

Despite the fact that the existing models have shown great promise in estimating the behind-the-meter solar generation, there is still great room for improvement. Highly simplified solar PV generation models are used in the disaggregation algorithm of Tabone et al. [5] and the consumer mixture model [6], which are incapable of capturing the nonlinear relationships among technical parameters of the solar PV system, weather data, and solar generation. A large number of hyperparameters need to be jointly tuned in the disaggregation algorithm of Tabone et al. [5] making the algorithm impractical and brittle. Moreover, if metered solar generation data in nearby locations are used as solar proxies, the proposed correction of transposition errors by using multiple solar proxies from PV systems of different geometry may not work if such a variety of metered PV installations are not available. Furthermore, these two statistical approaches [5], [6] can not estimate the technical parameters of the solar PV systems, which are extremely useful for both real-time estimation and long-term forecasting. Finally, the clear sky generation model [7] relies heavily on the net load data of a house when it is unoccupied on a sunny day. However, such data may not always be available for all customers.

In this paper, we develop an unsupervised framework to disaggregate net load measurements into solar generation and electric load estimates for individual customers without

information about their exact location. Our proposed algorithm seamlessly integrates a physical solar PV system performance model with a statistical model for estimating electric load. The accurate physical solar PV system performance model not only improves the accuracy of solar generation estimation but also allows us to estimate the technical parameters of the solar PV systems. A hidden Markov model (HMM) regression [8] is adopted to accurately estimate electric loads for customers under different energy consumption states. The performance of our proposed method is compared with the state-of-the-art net load disaggregation algorithms on data gathered by Pecan street [9] for residential customers located in Austin, Texas. I The unique contributions of this paper are as follows:

1) This is the first algorithm to seamlessly integrate a physical PV system performance model and a statistical load estimation model to disaggregate net load data and to estimate the solar PV system technical parameters.
2) The synergistic combination of HMM regression model and solar PV performance model enables us to significantly reduce the electric load and solar generation estimation errors of existing algorithms.

The remainder of the paper is organized as follows: Section II describes the overall framework of the net load disaggregation algorithm. Section III presents the technical methods used in the solar estimation, electric load modeling, and post-aggregation adjustment steps of our proposed algorithm. The numerical study based on our proposed algorithm and benchmark algorithms is shown in Section IV. Finally, Section V concludes the study.

II. Overall Framework

The aim of the net load disaggregation algorithm is to decompose the net load readings of a residential customer with a solar PV system into the solar PV generation and electric load. In other words, given the net load measurements for a residential customer NL_t, for time intervals $t \in \{1, 2, ..., T\}$, we need to estimate the customer's electric load L_t, and solar generation S_t for each time interval t. We do not have information about their exact location, historical PV generation or consumption, solar panel configuration, or other solar PV system parameters. However, we have the city's approximate longitude and longitude, which can be used as a proxy for the locations of all customers. According to the net load definition, the net load, electric load, and solar generation of a customer satisfy the following equality constraints at any time t:

$$NL_t = L_t - S_t; \quad L_t \geq 0,\, S_t \geq 0 \quad \forall t \tag{1}$$

The overall framework of our proposed net load disaggregation algorithm for each customer is shown in Fig. 1. The net load includes two components: the electric load and solar generation. We estimate one of the two components one at a time while fixing the other component. The iterative estimation scheme ends when the stopping criteria are met. We discuss the algorithm in detail in Section III-D. The solar PV system technical parameters are estimated by a physical model based estimation method presented in Section III-A. Solar generation is estimated by a physical model that utilizes these estimated parameters. The physical model of solar generation, called the PV system performance model, is presented in III-B. The electric load of the customer is estimated based on a statistical model, which is described in Section III-C. Finally, we make a post-disaggregation adjustment described in Section III-D on the disaggregated signals to ensure that electric load minus solar generation equals net load measurement at all times.

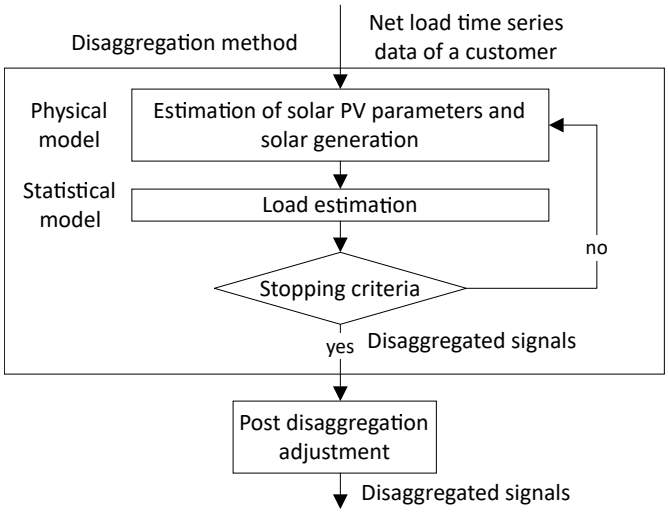

Fig. 1. The overall framework for disaggregating net load of residential customers with solar PV systems.

III. Technical Methods

In this section, we present our disaggregation methods and algorithms by integrating a physical solar PV generation model and a statistical electric load estimation model.

A. Estimation of Technical Parameters of Solar PV Systems

For the net load disaggregation algorithm, the solar PV generation, i.e., the AC output power of the PV array (P_{ac}), is estimated by a physical PV system performance model. If an estimate of the solar PV generation of a customer is available, we propose to estimate the technical parameters of the solar PV system by minimizing the sum of squared error between the input solar generation estimates and the calculated solar generation from the PV system performance model.

We denote the latest solar PV generation estimates of a customer at time t as S_t. Let $g_t(\boldsymbol{\theta}_S)$ denote the estimate for solar PV generation at time t based on the PV system performance model g_t with the technical parameters $\boldsymbol{\theta}_S$. The technical parameters include the DC rating P_{dc0}, array tilt angle θ_t, array azimuth angle θ_{az}, nominal inverter efficiency η_{nom}, and loss of the PV array l. The parameters of the solar PV system $\boldsymbol{\theta}_S = [P_{dc0}, \theta_t, \theta_{az}, \eta_{nom}, l]$ can be estimated by the following constrained optimization:

$$\begin{aligned} &\arg\min_{\boldsymbol{\theta}_s} \sum_{t=1}^{T} \left(S_t - g_t(\boldsymbol{\theta}_S)\right)^2 \\ &\text{subject to } S_t \geq 0,\ \boldsymbol{\theta}_{S,min} \leq \boldsymbol{\theta}_S \leq \boldsymbol{\theta}_{S,max} \end{aligned} \tag{2}$$

where T is the net load time series length, $\boldsymbol{\theta}_{S,min}$ and $\boldsymbol{\theta}_{S,max}$ denote the lower and upper limits of the PV system technical parameters respectively, which will be discussed in Section IV-B. The highly nonlinear nature of the PV system performance model makes Equation (2) a nonlinear optimization problem, which we solve by an interior-point algorithm.

B. PV System Performance Model

The PV system performance model in this study is highly nonlinear and is based on the PV performance modeling collaborative [10] and PVWatts [11] for a fixed mount system. We calculate P_{ac} by using the DC output power of the PV array P_{dc}, DC-to-AC ratio, and the nominal inverter efficiency η_{nom}. If the DC output power of the PV array is greater than the effective inverter DC input power rating $P_{dc0,inv}$, then the AC output of the inverter is capped at the AC nameplate rating of the inverter (P_{ac0}), where $P_{ac0} = \frac{P_{dc0}}{\text{DC-to-AC ratio}}$. Otherwise, if $0 < P_{dc} < P_{dc0,inv}$,

$$P_{ac} = g\left(\boldsymbol{\theta}_S\right) = \eta\left(\eta_{nom}, P_{dc}\right) P_{dc} \tag{3}$$

where the inverter efficiency, η, can be calculated following PVWatts [11]. P_{dc} can be calculated by using the specified DC rating P_{dc0}, cell temperature T_c, transmitted irradiance E_{tr}, and the loss in the PV array system l as follows:

$$\begin{aligned} P_{dc} &= g'\left(P_{dc0}, \theta_t, \theta_{az}, l\right) \\ &= (1-l) \times \frac{E_{tr}\left(\theta_t, \theta_{az}\right)}{E_0} P_{dc0}\left[1 + \gamma\left(T_c\left(\theta_t, \theta_{az}\right) - T_0\right)\right] \end{aligned} \tag{4}$$

Here, γ and T_0 are known parameters. The operating cell temperature (T_c) can be calculated based on the Sandia module and cell temperature model [10]. The model estimates the operating cell temperature T_c from the plane of array irradiance (E_{POA}), wind speed, ambient air temperature (T_a), and the temperature difference between the module and cell. Calculation of the plane of array irradiance (E_{POA}) and transmitted irradiance (E_{tr}) requires solar irradiance data (direct normal irradiance, diffuse horizontal Irradiance, and global horizontal irradiance), solar PV installation geometry information (tilt and azimuth angle of the solar PV array), and solar position data (solar zenith and solar azimuth angle) [10]. Solar zenith and azimuth angle at time t can be calculated using the solar position algorithms [12] for known locations.

C. Hidden Markov Model Regression for Load Modeling

We propose to employ a statistical Hidden Markov model (HMM) regression [8] to improve the traditional linear regression model for the load modeling [13]. This is done by simultaneously modeling the dependence structure among the load data and incorporating the heterogeneity of the regression models over different time periods. The statistical regression model is widely used to incorporate the effects of explanatory variables, such as temperature, humidity, wind speed, hour, and day of the week to estimate the load. In this study, to capture the non-linear relationship between temperature and load, we use a 3rd-degree polynomial of temperature, denoted by $\boldsymbol{c}, \boldsymbol{c^2}$, and $\boldsymbol{c^3}$ following the proposal of Hagan et al. [14], along with the weighted moving average of the temperature of last 24 hours, $\boldsymbol{c}_{wmv}$. By empirical analysis, we use a 3rd-degree polynomial of the *hour* of the day denoted by $\boldsymbol{h}, \boldsymbol{h^2}$, and $\boldsymbol{h^3}$ to model the nonlinear relationship between *hour* and load. To capture the different effects of weekend and weekday, we introduce a dummy variable $\boldsymbol{d}$ to indicate weekend. We also include the interaction of temperature and hour of the day, $\boldsymbol{c} \times \boldsymbol{h}$, as an explanatory variable. Explanatory variables are denoted by $\boldsymbol{X} = \left[\boldsymbol{c}, \boldsymbol{c^2}, \boldsymbol{c^3}, \boldsymbol{c}_{wmv}, \boldsymbol{d}, \boldsymbol{h}, \boldsymbol{h^2}, \boldsymbol{h^3}, \boldsymbol{c} \times \boldsymbol{h}\right]$.

The linear regression model is widely used to model the dependence of the load L_t on the explanatory variables $\boldsymbol{X}_t$ assuming a homogeneous relationship between $\boldsymbol{X}_t$ and L_t for all time periods. However, the load data can exhibit quite different patterns depending on whether a customer is present at home or not. For example, when a customer is at home, the load consists of heating, ventilation, air-conditioning, or appliance usage. On the other hand, when the customer is not at home, the load can be very low and consists of the power usages from the refrigerator and some appliances like water heaters, routers, modems, cable TV boxes, and TVs in standby mode. This change of the load pattern over time periods can be modeled by an HMM regression analysis.

At time t, let s_t be a *latent* state variable, $s_t = 1$ if the customer is home and $s_t = 2$, if not. Let $\boldsymbol{X}_t$ be the explanatory variables for the load L_t. In HMM regression, given s_t,

$$L_t = \boldsymbol{X}_t^T \boldsymbol{\beta}_{s_t} + \varepsilon_{s_t}, \quad \varepsilon_{s_t} \sim N\left(0, \sigma_{s_t}^2\right) \tag{5}$$

Note that the regression parameters $\boldsymbol{\beta}_{s_t}$ are allowed to be different for different states s_t. The dependence structure of the latent time series $\{s_t\}$ is modeled by an underlying Markov chain, $p_{ij} = P\left(s_t = j | s_{t-1} = i, s_{t-2} = k, \ldots\right) = P\left(s_t = j | s_{t-1} = i\right)$ where $i, j = 1, 2$ and p_{ij} is the transition probability from state i to state j satisfying $\sum_{j=1}^{2} p_{ij} = 1$ for each i. The HMM regression has been applied in many fields [15] including econometrics, where it is known as the Markov switching regression model [16] or regime switching model [17] with exogenous explanatory variables. A general HMM regression model can be estimated by maximum likelihood or Bayesian inference [8], [18]. We use the MS_Regress package of Matlab [19] to perform the maximum likelihood estimation.

D. Disaggregation Algorithm

We propose to disaggregate net load measurements into electric load $\hat{\mathbf{L}}$ and solar PV generation $\hat{\mathbf{S}}$ at individual homes by integrating the physical PV system performance model introduced in Section III-B and the statistical HMM regression introduced in Section III-C. The algorithm for disaggregating net load for an individual customer is shown in Algorithm 1. For an initial value of the solar PV system parameters $\boldsymbol{\theta}_S$, we first estimate the solar PV generation $\hat{\mathbf{S}}$ using the PV system performance model g. For any fixed solar generation estimate $\hat{\mathbf{S}}$, we can calculate the customer's electric load $\hat{\mathbf{L}} = \mathbf{NL} + \hat{\mathbf{S}}$ and then fit an HMM regression model to $\hat{\mathbf{L}}$. Based on the updated estimation of $\hat{\mathbf{L}}$, we can calculate the solar PV generation $\hat{\mathbf{S}} = \hat{\mathbf{L}} - \mathbf{NL}$ and estimate $\boldsymbol{\theta}_S$ by running

Algorithm 1 Algorithm for the disaggregation of net load of each customer and estimation of solar PV parameters

Input: Net load of a customer from AMI measurement, $\mathbf{NL}$
Output: User consumption $\hat{\mathbf{L}}$, solar generation $\hat{\mathbf{S}}$, and solar PV parameters $\boldsymbol{\theta}_S$

Initialization: Determine M initial solar PV system technical parameters $(\boldsymbol{\theta}_S)_1^{(0)}, \ldots, (\boldsymbol{\theta}_S)_M^{(0)}$

for each starting point $m \in M$ **do**
  Initialize solar generation, $\hat{\mathbf{S}}_m^{(0)} = g\left((\boldsymbol{\theta}_S)_m^{(0)}\right)$
  for j=1 to maxiter **do**
    Estimate user consumption, $\hat{\mathbf{L}}_m^{(j)} = \mathbf{NL} + \hat{\mathbf{S}}_m^{(j-1)}$
    Fit HMM regression model, denoted by $f(\boldsymbol{X}, \boldsymbol{\theta}_L)$, to $\hat{\mathbf{L}}_m^{(j)}$ and calculate parameters $(\boldsymbol{\theta}_L)_m^{(j)}$
    Update user consumption, $\hat{\mathbf{L}}_m^{(j)} = f\left(\boldsymbol{X}, (\boldsymbol{\theta}_L)_m^{(j)}\right)$
    Update solar generation, $\hat{\mathbf{S}}_m^{(j)} = \hat{\mathbf{L}}_m^{(j)} - \mathbf{NL}$
    Determine $(\boldsymbol{\theta}_S)_m^{(j)}$ from Equation (2) using $(\boldsymbol{\theta}_S)_m^{(j-1)}$ as initial value
    Update solar generation, $\hat{\mathbf{S}}_m^{(j)} = g\left((\boldsymbol{\theta}_S)_m^{(j)}\right)$
    Estimate net load, $\hat{\mathbf{NL}}_m^{(j)} = \hat{\mathbf{L}}_m^{(j)} - \hat{\mathbf{S}}_m^{(j)}$
    Calculate MSE of the net load, $E_m^{(j)}$
    if $\left|(\boldsymbol{\theta}_S)_m^{(j)} - (\boldsymbol{\theta}_S)_m^{(j-1)}\right| \leq \varepsilon$ **then**
      Break
    end if
  end for
end for
Determine $m^*, j^* = \underset{m,j}{argmin}\, E_m^{(j)}$
return $\hat{\mathbf{L}} = \hat{\mathbf{L}}_{m^*}^{(j^*)}$, $\hat{\mathbf{S}} = \hat{\mathbf{S}}_{m^*}^{(j^*)}$, and $\boldsymbol{\theta}_S = (\boldsymbol{\theta}_S)_{m^*}^{(j^*)}$

a constrained numerical optimization following Equation (2) using $\hat{\mathbf{S}}$ and g. We continue the above two processes until the estimated PV system parameters converge or the maximum number of iterations is reached. We repeat the same process for many initial values and select the solution that provides the lowest mean squared error for the estimated net load. Selection of initial values is discussed in Section IV-B.

Post-Disaggregation Adjustment: We propose to further improve the disaggregation performance by enforcing the constraint that the electricity consumption minus solar generation must be equal to the net load reading at any time. At any time t, we perform the following optimization inspired from [20] for each customer using the disaggregated signals $\hat{L}_t$ and $\hat{S}_t$ from Algorithm 1 to obtain the improved estimates:

$$\begin{aligned} &\underset{L_t, S_t}{\arg\min} \sum_{t=0}^{T} \alpha\Big(L_t - \hat{L}_t\Big)^2 + \beta\Big(S_t - \hat{S}_t\Big)^2 \\ &\text{subject to } L_t \geq 0,\ S_t \geq 0, \quad L_t - S_t = NL_t \end{aligned} \tag{6}$$

Here, α and β are user-specified parameters that put the weights for the error in the load and solar generation model, respectively. We propose two methods for determining the values of α and β. In the first variation, if the ground truth solar generation and load data are available for some customers, we determine α and β by the inverse of the variance of the error of the estimated load and solar generation for these customers:

$$\alpha = 1/Var\left(\varepsilon_{Load}\right), \quad \beta = 1/Var\left(\varepsilon_{PV}\right) \tag{7}$$

For the second variation, if the ground truth solar PV generation or load data are not available, we estimate the ground truth by the load and solar generation from steps 4 and 7 of Algorithm 1.

E. Error Metric

We measure the performance of net load disaggregation algorithms with three metrics: the mean squared error (MSE), mean absolute scaled error (MASE), and the coefficient of variation (CV). MASE is used instead of the mean absolute percentage error (MAPE) because many electric load and solar generation measurements are zero or close to zero, which makes MAPE extremely high even with small estimation errors. MASE scales mean absolute errors with the errors of a naive forecasting model which simply uses the last observation as prediction. Thus, it is scale invariant and treats all customers equally. CV is another normalized error metric defined as the root mean squared error divided by the mean actual signal. Let $y_{i,t}$ and $\hat{y}_{i,t}$ be the actual and estimated values of customer i at time t, respectively. Then, the mean MSE, MASE, and CV of N customers over a period T can be expressed as follows:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}\sum_{t=1}^{T}\left(y_{i,t} - \hat{y}_{i,t}\right)^2 \tag{8}$$

$$MASE = \frac{1}{N}\sum_{i=1}^{N}\frac{T-1}{T}\frac{\sum_{t=1}^{T}\left|y_{i,t} - \hat{y}_{i,t}\right|}{\sum_{t=2}^{T}\left|y_{i,t} - y_{i,t-1}\right|} \tag{9}$$

$$CV = \frac{1}{N}\sum_{i=1}^{N}\left(\sqrt{\sum_{t=1}^{T}\left(y_{i,t} - \hat{y}_{i,t}\right)^2}\Big/\frac{1}{T}\sum_{i=1}^{T} y_{i,t}\right) \tag{10}$$

IV. Numerical Study

A. Dataset for Numerical Study

The 15-minute interval net load, customer load, and solar PV generation data gathered by Pecan Street [9] are used in the numerical study. The customers are located in Austin, Texas with an approximate longitude and latitude of $(30.29^oN, -97.69^oE)$. The study period is from October 3, 2015, to October 30, 2015. We select this specific period so that we can directly compare our results with that of the consumer mixture models [6]. Within the study period, we have 197 customers with valid solar generation and electric load data. The ground truth tilt and azimuth angle of the solar PV installations are not available. The DC rating of solar PV panels is available for 90% of the customers. The DC ratings are later used to validate the accuracy of our algorithm.

The solar irradiance and other weather data are gathered from the National Solar Radiation Database [21]. The solar irradiance data covers the entire US with a $4km \times 4km$ grid resolution and 30-minute granularity. We select the data from the closest grid box to the approximate location of

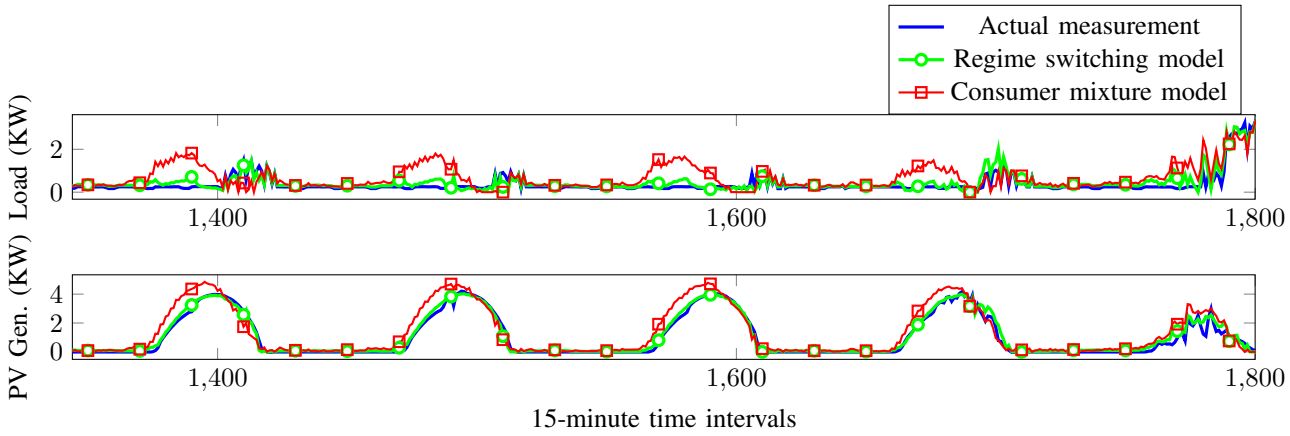

Fig. 2. Comparison of disaggregated load and solar PV generation with actual values for a customer from 2015/10/14 to 2015/10/19

all customers $(30.29^oN, -97.69^oE)$. The 30-minute interval data are converted into 15-minute interval ones with linear interpolation. In the first version of our proposed net load disaggregation algorithm, the hyperparameters α and β in Equation (7) are calculated using 10% of the customers' actual electric load and solar PV generation data.

B. Experimental Setup

The technical parameters of the solar PV system are estimated by solving a constrained optimization problem. The upper and lower bounds constraining the technical parameters are set as follows. The feasible range of solar PV array tilt angle θ_T is set as $[5^o, 50^o]$. The feasible range of PV array azimuth angle θ_{AZ} is assumed to be $[0^o, 360^o]$. The feasible range of solar panel's DC rating is selected to be $[1\text{KW}, 15\text{KW}]$. The feasible range of inverter nominal efficiency η_{nom} is between 0.92 and 0.99. Finally, we select the range of PV array loss to be $[9\%, 38\%]$. The range of solar PV panel loss is obtained from the derate factor ranges provided in [22] . The DC-to-AC ratio is roughly the same for all customers. It is fixed at 1.1, the default value in PVWatts.

Note that the upper and lower bounds for solar PV panel tilt and azimuth angles are determined based on the maximum and minimum angles of 160,000 solar PV panels in California from the California Solar Initiative working data set [23]. Also, we assume that the rooftop solar PV installations being studied are fixed array systems with no tracking system.

We select 8 initial solar PV system technical parameter sets for step 1 in Algorithm 1 by gradually increasing P_{dc0} in 7 steps from 1 KW to 8 KW. The other initial solar PV system parameters $[\theta_T, \theta_{AZ}, \eta_{nom}, l]$ are set at their most common values $25^o, 180^o, 0.96$, and 14% respectively.

C. Result and Analysis

We implemented our proposed net load disaggregation method following Algorithm 1 using two variations of the post-disaggregation adjustment described in III-D with known and unknown error variance. The performance of our proposed model is compared against two state-of-the-art benchmark algorithms, the unsupervised consumer mixture model [6] and the SunDance algorithm [7]. Following the implementation of the consumer mixture model, the electric load of customers without solar PVs are clustered using the K-Medoids algorithm. As the cluster medoids can change based on the initial choice of medoids, we perform 100 simulations with different initial medoids and take their mean. When implementing the SunDance model, we used neural network to approximate the universal weather-solar effect model. The solar generation data from August 2015 to October 2015 of a house in the pecan street dataset is used for network training. The geometry of the solar PV installation is not known at the house.

TABLE I
COMPARISON OF VARIOUS DISAGGREGATION METHODS

Error Metric	Variable	HMM reg. model (known error var)	HMM reg. model (unknown error var)	Consumer Mixture Model	SunDance Model
MSE	Load	**0.23**	0.24	0.38	0.49
	Solar	**0.25**	0.25	0.43	0.54
MASE	Load	0.58	**0.55**	0.74	0.81
	Solar	3.08	**2.84**	3.91	3.74
CV	Load	**0.36**	0.37	0.46	0.57
	Solar	**0.62**	0.62	0.79	0.85

Table I shows the MSE, MASE, and CV for load and solar generation estimations based on our proposed methods and the benchmark algorithms. The HMM regression model with either known or unknown error variance yields smaller errors than both the consumer mixture model [6] and the SunDance model [7]. Our proposed model with known error variance performs slightly better than the model with unknown error variance in terms of MSE and CV. However, if we consider MASE, the later model performs better.

Even without using a validation data set with actual solar and load, our proposed model without known error variance reduces the MSE by 42% compared to the consumer mixture model [6]. The improvement of our proposed model is pronounced for customers who are absent from home for a period of time as the HMM regression model is well suited

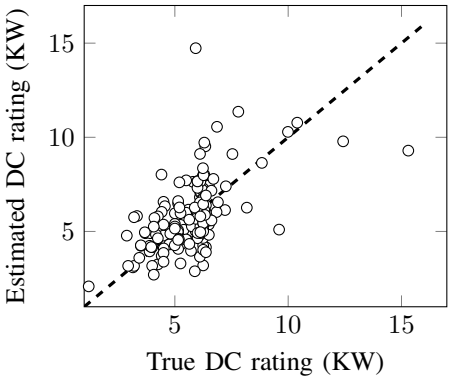

Fig. 3. Comparison of true and estimated DC rating of PV array

to capture load behavior in different regimes. In our study, 25 out of 197 customers are absent from their residence for an extended period of time. Our proposed model reduces MSE for these customers by 55% compared to the consumer mixture model. Fig. 2 illustrates the disaggregated load and solar PV generation signals of our proposed model and the benchmark consumer mixture model for a customer who is periodically absent from home. As shown in the figure, the load estimate from our proposed model follows the actual load data significantly more closely than the consumer mixture model during the periods of absence. As a result, the solar generation estimate from our proposed model is also considerably more accurate during these periods.

Our proposed model outperforms the SunDance model [7] in terms of estimation accuracy mostly due to the adoption of the more accurate physical solar PV system performance model. The SunDance model, on the other hand, relies heavily on accurate estimation of maximum solar generation and cloud cover. If there is a lack of lower consumption periods on sunny days, then the maximum solar generation estimation will be rather unreliable. Furthermore, the cloud cover measurement data typically do not have sufficient spatial resolution at the household level. Finally, the performance of our proposed model in estimating the DC size of the solar PV systems is illustrated in Fig. 3. As shown in the figure, the estimated solar DC ratings and the actual ones are quite similar. The four outliers in Fig. 3 arise from an error in the dataset where the net load is not equal to load minus solar generation.

V. Conclusion

We developed an unsupervised algorithm to disaggregate net load signals into solar PV generation and electric load consumption for residential customers with solar PV systems. The iterative net load disaggregation algorithm synergistically combines a physical solar PV system performance model for solar PV generation estimation with a statistical HMM regression model for load estimation. This unique approach results in a significant reduction in solar generation and electric load estimation errors. Further improvement in our net load disaggregation algorithm can be achieved by leveraging more granular solar irradiance data. The proposed algorithm can be used for online applications with the help of real-time prediction of solar irradiance data.

References

[1] "Annual energy outlook 2019 with projections to 2050," US Energy Information Administration, Office of Energy Analysis, U.S. Department of Energy, Washington, DC 20585, Tech. Rep., January 2019.

[2] W. Wang, N. Yu, and R. Johnson, "A model for commercial adoption of photovoltaic systems in California," *Journal of Renewable and Sustainable Energy*, vol. 9, no. 2, p. 025904, 2017.

[3] N. Yu, S. Shah, R. Johnson, R. Sherick, M. Hong, and K. Loparo, "Big data analytics in power distribution systems," in *2015 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*. IEEE, 2015, pp. 1–5.

[4] K. Baker, A. Bernstein, E. DallAnese, and C. Zhao, "Network-cognizant voltage droop control for distribution grids," *IEEE Transactions on Power Systems*, vol. 33, no. 2, pp. 2098–2108, 2018.

[5] M. Tabone, S. Kiliccote, and E. C. Kara, "Disaggregating solar generation behind individual meters in real time," in *Proceedings of the 5th Conference on Systems for Built Environments*. ACM, 2018, pp. 43–52.

[6] C. M. Cheung, W. Zhong, C. Xiong, A. Srivastava, R. Kannan, and V. K. Prasanna, "Behind-the-meter solar generation disaggregation using consumer mixture models," in *2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, Oct. 2018, pp. 1–6.

[7] D. Chen and D. Irwin, "SunDance: Black-box behind-the-meter solar disaggregation," in *Proceedings of the Eighth International Conference on Future Energy Systems*. ACM, 2017, pp. 45–55.

[8] M. Fridman, "Hidden Markov model regression," Institute of Mathematics, University of Minnesota, Tech. Rep., 1993.

[9] C. Holcomb, "Pecan street inc.: A test-bed for NILM," in *International Workshop on Non-Intrusive Load Monitoring*, 2012.

[10] J. S. Stein, "The photovoltaic performance modeling collaborative (PVPMC)," in *38th IEEE Photovoltaic Specialists Conference*. IEEE, 2012, pp. 003 048–003 052.

[11] A. P. Dobos, "PVWatts version 5 manual," National Renewable Energy Laboratory, Golden, CO, USA, Tech. Rep., 2014.

[12] I. Reda and A. Andreas, "Solar position algorithm for solar radiation applications," *Solar Energy*, vol. 76, no. 5, pp. 577–589, 2004.

[13] J. Hinman and E. Hickey, "Modeling and forecasting short-term electricity load using regression analysis," *Journal of Institute for Regulatory Policy Studies*, pp. 1–51, 2009.

[14] M. T. Hagan and S. M. Behr, "The time series approach to short term load forecasting," *IEEE Transactions on Power Systems*, vol. 2, no. 3, pp. 785–791, Aug 1987.

[15] I. L. MacDonald and W. Zucchini, *Hidden Markov and other models for discrete-valued time series*. Chapman & Hall, 1997.

[16] S. M. Goldfeld and R. E. Quandt, "A Markov model for switching regressions," *Journal of Econometrics*, vol. 1, no. 1, pp. 3–15, 1973.

[17] J. D. Hamilton, "Regime switching models," in *Macroeconometrics and Time Series Analysis*. Palgrave Macmillan UK, 2010, pp. 202–209.

[18] O. Cappé, E. Moulines, and T. Rydén, *Inference in hidden Markov models*, ser. Springer series in statistics. Springer, 2005.

[19] M. Perlin, "MS_regress - the MATLAB package for Markov regime switching models," *Available at SSRN 1714016*, 2015.

[20] E. C. Kara, M. Tabone, C. Roberts, S. Kiliccote, and E. M. Stewart, "Estimating behind-the-meter solar generation with existing measurement infrastructure: Poster abstract," in *Proceedings of the 3rd ACM International Conference on Systems for Energy-Efficient Built Environments*. ACM, 2016, pp. 259–260.

[21] M. Sengupta, Y. Xie, A. Lopez, A. Habte, G. Maclaurin, and J. Shelby, "The National Solar Radiation Data Base (NSRDB)," *Renewable and Sustainable Energy Reviews*, vol. 89, pp. 51–60, 2018.

[22] B. Marion, J. Adelstein, K. e. Boyle, H. Hayden, B. Hammond, T. Fletcher, B. Canada, D. Narang, A. Kimber, L. Mitchell *et al.*, "Performance parameters for grid-connected PV systems," in *Conference Record of the Thirty-first IEEE Photovoltaic Specialists Conference, 2005.* IEEE, 2005, pp. 1601–1606.

[23] C. D. G. Statistics, "The California Solar Initiative - CSI working data set," 2019, https://www.californiadgstats.ca.gov/downloads/.