Disaggregating Solar Generation Using Smart Meter Data and Proxy Measurements from Neighbouring Sites

Xinlei Chen
University of Alberta
Edmonton, Canada
xinlei1@ualberta.ca

Moosa Moghimi Haji
University of Alberta
Edmonton, Canada
moghimih@ualberta.ca

Omid Ardakanian
University of Alberta
Edmonton, Canada
ardakanian@ualberta.ca

ABSTRACT

This paper investigates the problem of disaggregating solar generation from smart meter data when historical disaggregated data from the target home is unavailable and deployment characteristics of the PV system are unknown. The proposed method takes advantage of solar generation data (aka proxy measurements) from a few sites located in the same area as the target home, and solar generation data synthesized using a physical PV model. We evaluate our method with 4 different proxy settings on around 140 homes in Australia, and show that the solar disaggregation accuracy is improved by 33.84% and 15.41% over two state-of-the-art methods using only one real proxy along with three synthetic proxies. Furthermore, we show that using the disaggregated home load instead of the net load measured by a smart meter could improve the accuracy of two popular non-intrusive load monitoring techniques by at least 22%.

CCS CONCEPTS

• **Computer systems organization** → *Sensors and actuators*; • **Computing methodologies** → **Modeling and simulation**.

KEYWORDS

Solar disaggregation, non-intrusive load monitoring, smart meter

ACM Reference Format:
Xinlei Chen, Moosa Moghimi Haji, and Omid Ardakanian. 2021. Disaggregating Solar Generation Using Smart Meter Data and Proxy Measurements from Neighbouring Sites. In *The Twelfth ACM International Conference on Future Energy Systems (e-Energy '21), June 28–July 2, 2021, Virtual Event, Italy.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3447555.3464856

1 INTRODUCTION

Solar photovoltaic (PV) generation is the fastest growing renewable energy source today [10]. Almost half of this growth is projected to be behind-the-meter (BTM) installations, which are typically PV systems on the roof of homes and buildings [8]. High penetration of PV systems introduces new challenges for planning and operation of power distribution networks, requiring the system operators and electric utilities to develop low-cost techniques for forecasting and monitoring the solar power injected into their systems.

Several methods have been proposed in the literature to estimate the solar power generated by BTM PV systems. They can be divided into three major categories: (a) methods that rely on satellite and aerial imagery [8] to identify PV systems and estimate their physical characteristics, (b) methods that rely on a few separately metered solar sites in a geographical area to estimate the total solar generation [16, 17], and (c) methods that apply source separation techniques to disaggregate solar generation from feeder-level measurement or smart meter data [5, 7]. Solar disaggregation methods can be categorized based on the type of models they use for estimating solar power [6]. Some methods use a physics-based model to estimate solar generation [2, 5, 11] while others use a data-driven, black-box model [4, 12, 21].

In this paper we propose a solar disaggregation method to accurately estimate the output of a BTM PV system in an offline fashion. Our method has two key advantages over prior work on solar disaggregation. First, it requires active power measurements with low temporal resolution, which is already collected by ordinary smart meters, and *proxy measurements* from only one or a few PV systems located in the same geographical area. Second, our method has a low computational overhead making it suitable for large scale implementation. Our contribution is threefold:

- We propose a solar disaggregation method that merely relies on net load data and just a small number of separately metered solar PV systems. We show that real proxy measurements can be replaced with synthetic ones to achieve comparable performance as long as there is one real proxy.
- We compare our method against other state-of-the-art methods on a large dataset. We find that the proposed method yields a higher accuracy than the baselines.
- We examine how the improved accuracy of solar disaggregation affects the accuracy of two popular non-intrusive load monitoring (NILM) methods, namely Factorial Hidden Markov Model (FHMM) [13] and Sequence-to-Point [22].

2 PROBLEM DEFINITION & METHODOLOGY

In this section we introduce the solar disaggregation problem, present models for home load and solar power, and propose an iterative algorithm for estimating solar generation and home load.

2.1 Definition

Customer-level solar disaggregation concerns decomposing the customer's net load into home load and solar generation. Let $\mathbf{y} \in \mathbb{R}^T$ be a vector that collects the measured net load of one customer over T intervals. Similarly, let $\hat{\boldsymbol{\ell}}, \hat{\mathbf{s}} \in \mathbb{R}^T$ denote respectively the estimated home load and BTM solar generation in the same period. These quantities must satisfy this equality constraint: $\mathbf{y} = \hat{\boldsymbol{\ell}} - \hat{\mathbf{s}}$.

e-Energy '21, June 28–July 2, 2021, Virtual Event, Italy

ACM ISBN 978-1-4503-8333-2/21/06…$15.00
https://doi.org/10.1145/3447555.3464856

Let $\mathbf{X}^L \in \mathbb{R}^{T\times K_l}$ be the set of K_l features that determine the customer's home load, and $\mathbf{X}^S \in \mathbb{R}^{T\times K_s}$ be the set of K_s proxy measurements that can be used to approximate the customer's solar generation via a mixture model. We can train a non-linear model g to map the features to the home load, and a linear solar mixture model to estimate the customer's solar generation:

$$\hat{\boldsymbol{\ell}} = g(\mathbf{X}^L; \boldsymbol{\theta}), \tag{1}$$

$$\hat{\mathbf{s}} = \mathbf{X}^S \mathbf{w}, \tag{2}$$

where $\boldsymbol{\theta}$ is a vector that represents parameters of the load model, and $\mathbf{w} \in \mathbb{R}^{K_s}$ is the weight vector of the mixture model. Hence, $\mathbf{w}_k$ is a scalar that represents the weight assigned to the k^{th} proxy.

In our defined solar disaggregation problem, we assume no information is available about the customers except their approximate location (i.e., the city or district they are located in) and their net load data. Thus, the deployment characteristics (e.g., panel size, orientation, tilt) of BTM PV systems are also not known a priori. In addition to the smart meter data, solar irradiance, wind speed, and ambient temperature at the city scale can be downloaded via an API. We also assume there is at least one separately metered PV system (providing proxy measurements) in the same city or district as our target home. The deployment characteristics of this site could differ from the PV system installed at the target home. We argue this is not a strong assumption as utilities usually have access to direct solar measurements from several sites in a city.

2.2 Models

We now introduce our load estimation and solar mixture models described in (1) and (2) respectively.

Solar mixture model: We aim to approximate the solar power generated by the BTM PV system installed at the target home using a mixture of proxy measurements from PV systems located in the same city or district. The intuition behind this approximation is that PV systems in the same geographical area have more or less the same solar generation pattern regardless of their deployment characteristics.

There are two specific challenges that must be addressed to get a good approximation. First, we do not have control over the deployment characteristics of the PV systems that provide proxy measurements. If they had exactly the same orientation angle as the PV system installed at the target home, estimating the target home's solar generation would reduce to learning a single scaling factor. One way to address this challenge is to adopt a solar mixture model to approximate the target home's solar generation as a weighted sum of a number of proxy measurements, as shown in Equation (2). The second challenge is that proxy measurements from a large number of neighbouring PV systems might not be available in practice. To address this challenge, we combine proxy measurements from real PV systems with measurements synthesized by a physical PV model that takes into account solar irradiance data of an arbitrary location in the same city. The physical PV model that we use to obtain data for *synthetic proxies* is based on PVWatts [9]. We develop this model using the PV Performance Modeling Collaborative [20], which is described in Appendix A.1.

Home load model: To estimate the home load, we adopt a random forest regression model. We use the scikit-learn [15] library to train this model. Four explanatory variables are used as features, $\mathbf{X}^L$, for all customers. These variables include ambient temperature $\mathbf{c}$, exponentially weighted moving average of temperature over the last 24 hours $\mathbf{c}_{wmv}$, hour of the day $\mathbf{h}$, and a binary variable $\mathbf{d}$ that indicates if it is a weekday or weekend. Thus, we have $\mathbf{X}^L = [\mathbf{c}, \mathbf{c}_{wmv}, \mathbf{h}, \mathbf{d}]$.

2.3 Solar Disaggregation Algorithm

Our solar disaggregation algorithm includes two main parts: a weight initialization technique and an iterative algorithm for updating the model parameters.

Weight initialization: The first step for implementing our method is to initialize the weight vector $\mathbf{w}$ of the solar mixture model using the net load data of the target home and the solar generation data collected from the proxies. A good initialization can enhance the performance of the disaggregation method and reduce its convergence time. Our weight initialization method has three main steps:

(1) Estimating the physical characteristics of PV systems installed at the target home and real solar proxies.
(2) Finding the maximum solar generation of each PV system.
(3) Solving an optimization problem to determine the initial weight vector $\mathbf{w}$.

We use an open source toolkit, SolarTK [2], to estimate the physical characteristics of PV systems including its tilt, orientation, and panel size. To estimate these parameters, the toolkit takes the real solar generation data as input and finds the maximum solar generation. We can run this toolkit on proxy measurements, but we lack the real solar generation data from the target home. To solve this problem, we approximate the solar generation of the target home given the net load data $\mathbf{y}$ from this expression $\hat{\mathbf{s}} \approx \max(\mathbf{0}, \ell_{base} - \mathbf{y})$, where ℓ_{base} is the target home's base power consumption calculated as the minimum consumption level at night time. SolarTK is then run on the estimated solar generation of the target home and the real solar generation of solar proxy/proxies to obtain the estimated parameters of all PV systems. Since we use the longitude and latitude of an arbitrary location in the city as the approximate location for all the PV systems in that city, the estimated parameters may not be highly accurate.

We then calculate the maximum solar generation for each proxy and target home using the estimated deployment characteristics obtained in Step 1. The maximum solar generation is the potential generation of a PV system under clear sky condition, that is determined by the system's physical characteristics, the ambient temperature and the location of PV. We denote the maximum solar generation of the k^{th} proxy by $\mathbf{m}_k^p \in \mathbb{R}^T$, and the maximum solar generation of the target home by $\mathbf{m}^c \in \mathbb{R}^T$.

In the last step, we determine the initial weight vector, $\mathbf{w}$, for the solar mixture model following the idea of [17, 19]; the solar generation of a target home with unknown deployment characteristics is estimated utilizing metered solar generation of sites with nonuniform deployment characteristics. Formally, we can write

$$\mathbf{s}_t^{target} = \alpha \cdot \mathbf{s}_t^{proxy} \tag{3}$$

where α depends on time, site location, and other site-specific factors. In our method, we simplify α to be a constant weight factor for

Algorithm 1: Solar disaggregation for one target customer

```
Input  : Net load of the target customer, $\mathbf{y} \in \mathbb{R}^T$;
         Proxy measurements from $K$ sites, $\mathbf{X}^S \in \mathbb{R}^{T\times K}$;
         Initial weights of the solar mixture model, $\mathbf{w} \in \mathbb{R}^K$;
         Load related features, $\mathbf{X}^L$.
Output : Estimated solar generation and home load of the
         target customer, $\hat{\mathbf{s}}$, $\hat{\boldsymbol{\ell}}$;
Init: $\mathbf{w}^0 \leftarrow \mathbf{w}/K$;
      $\mathbf{s}^0 \leftarrow \mathbf{X}^S\mathbf{w}^0$;
      Initialize parameters $\boldsymbol{\theta}^0$ for load model $g$;
1 while $iter < Max\ Iteration$ and $|\boldsymbol{w}^{iter} - \boldsymbol{w}^{iter-1}| > \epsilon$ do
2     $\boldsymbol{\ell}^{iter} \leftarrow \mathbf{s}^{iter} + \mathbf{y}$;
3     Incrementally train the model g with input feature $\mathbf{X}^L$
        and output $\boldsymbol{\ell}^{iter}$;
4     Update load $\boldsymbol{\ell}^{iter} \leftarrow g(\mathbf{X}^L, \boldsymbol{\theta}^{iter})$;
5     $\mathbf{s}^{iter} = \boldsymbol{\ell}^{iter} - \mathbf{y}$;
6     $\mathbf{w}^{iter} \leftarrow \mathrm{argmin}_{\mathbf{w}} \|\mathbf{X}^S\mathbf{w} - \mathbf{s}^{iter}\|_2$;
7     $\mathbf{s}^{iter} \leftarrow \mathbf{X}^S\mathbf{w}^{iter}$;
8 end
```

each site, i.e., $\mathbf{w}_k$. Since we do not have the true solar generation from the target home in Equation (3), we use the maximum solar generations to determine the initial weight of each solar proxy. Specifically, the weight factor $\mathbf{w}_k$ for the k^{th} proxy can be determined by solving the following optimization problem:

$$\begin{aligned} \min_{\mathbf{w}_k} \quad & \|\mathbf{w}_k \cdot \mathbf{m}_k^p - \mathbf{m}^c\|_2 \\ \text{subject to} \quad & \mathbf{w}_k > 0 \end{aligned} \tag{4}$$

Disaggregation algorithm: In this step, we iteratively estimate the home load and solar generation until the parameters of our model converge. Algorithm 1 presents the pseudocode of the proposed solar disaggregation algorithm. After obtaining the initial weights $\mathbf{w}$ for the solar mixture model, we first estimate the solar PV generation $\mathbf{s}^{iter}$ using a linear combination of the solar proxies. Then, we use $\mathbf{y} = \hat{\boldsymbol{\ell}} - \hat{\mathbf{s}}$ to calculate the estimated home load $\boldsymbol{\ell}^{iter}$ (line 2) and incrementally train the load model using $\boldsymbol{\ell}^{iter}$ and load related features $\mathbf{X}^L$ (line 3). Based on the updated home load $\boldsymbol{\ell}^{iter}$ (line 4), we determine solar generation $\mathbf{s}^{iter}$ (line 5), update the weights for solar proxies (line 6), and recalculate the solar generation using the updated weights (line 7). We repeat the above steps until the solar proxy weights $\mathbf{w}$ converge or we reach the maximum number of iterations. In our experiments, it typically takes between 20 and 80 iterations for this algorithm to converge depending on the number of proxies and goodness of initial weights.

3 EVALUATION

3.1 Dataset

We use the Ausgrid [1] dataset to evaluate the estimation accuracy of different solar disaggregation methods. This dataset includes 30-minute resolution net load measurements in addition to direct measurements of home load and solar generation from homes with rooftop PV systems. It consists of 140 customers with rooftop PV systems in Sydney, Australia (in the southern hemisphere) with the latitude and longitude of -33.888575 and 151.187349 respectively. These locations are approximate since we only have the postal code of each customer. Since Sydney is a sprawling city, we cluster the customers into three clusters according to their latitude and longitude. Each cluster still spans a large area of the city. We consider two periods in two seasons, one from November 1, 2012 to November 30, 2012 in the summer season (T=1440) and the other one from May 1, 2013 to May 30, 2013 in the winter season (T=1440). A small number of customers are removed due to data quality issues in each season. For Sydney's weather data, we pull the solar radiation, wind speed, and outside air temperature with 30-minute temporal resolution using the Solcast API [18].

3.2 Variants of our Disaggregation Method

We implement our method with 4 different solar proxy settings.

- **3Proxies:** we directly use solar generation data for the same periods from 3 real rooftop PV systems in the same area.
- **1P+1SP:** we only use 1 real solar proxy combined with 1 synthetic proxy. In this case, the ideal orientation angle in the southern hemisphere is 0°.
- **1P+3SP:** we use 1 real solar proxy combined with 3 synthetic proxies with different orientation angles.
- **3SP:** we use 3 synthetic proxies with different orientation angles just like the previous setting.

For the above settings that include synthetic proxies, we set the tilt angle to the absolute value of the city's latitude and use a uniform 3kW DC rating. We set orientation angle of the three synthetic proxies to 0°, 90°, and 270° respectively (i.e., N, E, and W facing panels). The tilt and DC rating have a similar effect on solar generation curve, i.e., they scale the curve up or down [5], whereas the orientation shifts the peak of the generation curve to earlier or later. Therefore, we can set the tilt angle and DC rating similarly for all the synthetic solar proxies because the elements of our weight vector $\mathbf{w}$ will be adjusted by Algorithm 1.

3.3 Baselines

We compare the performance of our solar disaggregation method with two methods that also use the data that is commonly available to the utility and outperform other solar disaggregation methods proposed in the literature. Specifically, we use the solar disaggregation methods proposed in [11] and [2] as our baselines; these methods are labelled "Baseline 1" and "Baseline 2", respectively. For a fair comparison, we implement the one-nearby-proxy-based solar estimation method in [2] that was adopted in the case study of computing the clear sky index. In each experiment, we use the same real solar proxy for our method and Baseline 2.

4 EXPERIMENTAL RESULTS

In this section we evaluate the performance of our methods in disaggregating BTM solar generation using root-mean-square error (RMSE) and normalized RMSE (nRMSE) metrics. nRMSE is the RMSE normalized by the mean value of the real solar generation. We then investigate the impact of running solar disaggregation on the performance of NILM methods.

4.1 Solar Disaggregation Performance

We compare the 4 variants of our method – 3Proxies, 1P+1SP, 1P+3SP, and 3SP – with the two baselines described in Section 3.2. For each variant, we evaluate the disaggregation performance for all customers with PV systems in the dataset. Since our method utilizes proxy measurements, we run the experiment 10 times for each target home with real solar proxies that are randomly selected from the same cluster as the target home.

Estimation accuracy: Table 1 shows the average nRMSE and RMSE of solar generation and home load estimation across all customers in two seasons. We observe that 3Proxies yields the lowest error compared to the other variants of our method and the two baselines. 1P+1SP and 1P+3SP also beat the two baselines in both seasons. On average, 1P+1SP reduces nRMSE by 31.09% and 11.61% and 1P+3P reduces nRMSE by 33.84% and 15.41% compared to Baseline 1 and Baseline 2, respectively. This observation suggests that by utilizing as few as only one directly measured solar generation site and synthetic proxies with different orientation settings, we can get a better estimate of solar generation and home load than the state-of-the-art solar disaggregation methods. Interestingly, 3SP has the worst performance among the four variants of our method, although it still beats both baseline methods in winter. This underscores the importance of having at least one real proxy for solar disaggregation to account for abrupt changes (e.g., due to a passing cloud) in the solar generation.

Table 1: Comparison of disaggregation methods. Each cell contains two slash-separated metrics: nRMSE and RMSE.

	Summer		Winter	
Method	Solar	Load	Solar	Load
3Proxies	0.469/0.0621	0.232/0.0593	0.771/0.0602	0.199/0.0582
1P+1SP	0.543/0.0723	0.273/0.0686	0.841/0.0677	0.226/0.0649
1P+3SP	0.525/0.0691	0.254/0.0650	0.797/0.0638	0.212/0.0617
3SP	0.643/0.0858	0.305/0.0780	0.854/0.0702	0.231/0.0669
Baseline 1	0.612/0.0778	0.287/0.0778	1.713/0.1210	0.375/0.1210
Baseline 2	0.631/0.0857	0.337/0.0857	0.927/0.0767	0.272/0.0767

Figure 1 in A.2 illustrates the disaggregated solar generation in two seasons by 1P+3SP and the two baseline methods. It can be seen that our estimated solar is generally closer to true solar generation in terms of estimating peak generation time and generation scale. Figure 2 in A.2 shows real solar generation profiles of a target home and the four proxy measurements used in 1P+3SP. Note that the peak generation of the target home and real solar proxy happen at different times as they have different orientations. In this case, the synthetic proxy with an orientation angle close to the target home's orientation angle gets a much higher weight compared to the other synthetic proxies and the real proxy. This highlights the advantage of incorporating the synthetic proxies.

4.2 NILM Performance

We use 1-minute resolution data from Pecan Street [14] to study the effect of BTM solar generation on the performance of NILM techniques. We choose this dataset because it has both solar generation and individual appliance consumption data. We select 6 homes from this dataset and apply two benchmark NILM techniques (i.e., FHMM [13] and Seq2Point [22]) implemented in NILMTK [3].

We evaluate the performance of these two NILM methods in disaggregating the loads of 7 appliances, including the washing machine, microwave, air conditioner, furnace, fridge, dryer and dish washer, in each of the 6 homes. Since a dryer is not present in 4 of these homes, we only report the results for the remaining 2 homes for this appliance. We train appliance models using real home load and individual appliance load data collected between June 1 and June 22, 2018. We then calculate the error of disaggregating each appliance's load in the test data (from June 23 to June 30, 2018) with 5 different sets of input data, including the true home load, the net load (i.e., home load - BTM solar generation), and 3 versions of the estimated home load obtained by applying our disaggregation method (1P+3SP), and the two baseline methods described earlier.

Figure 3 in A.2 shows the average RMSE for each appliance and the overall RMSE for all appliances in these homes. Two important observations can be made. First, solar disaggregation will improve the overall NILM accuracy but different appliances are affected to a different extent. The overall RMSE for all appliances will be 0.446 if we directly apply Seq2Point (the best performing NILM method) on the net load data. However, the RMSE will be 0.150 if we apply it to the home load estimated by 1P+3SP, an impressive 66.2% improvement in disaggregation accuracy. We also observed a 22.0% improvement for FHMM. Our second observation is that a higher accuracy in solar disaggregation leads to a better NILM performance, especially for appliances with more variable power usage patterns. The overall RMSE of Seq2Point when it runs on the home load estimated by the 1P+3SP method is 22.3% and 9.7% lower than when it runs on the home load estimated by Baseline 1 and Baseline 2, respectively.

5 CONCLUSION

Solar disaggregation will be essential for the reliable operation of a nimble and transparent power distribution system. In this work, we proposed a method for disaggregating solar generation of BTM PV systems in an offline fashion. This method relies on net load data and proxy measurements from a few PV systems located in the same geographic area as the target home. We evaluated our proposed method on a publicly available dataset from Australia. We found that the solar estimation accuracy improves by 15.41% on average over the best baseline (i.e., Baseline 2) when one real proxy and three synthetic proxies are used. Finally, we investigated whether a more accurate disaggregation technique could lead to higher accuracy in NILM. Our results suggest that using the disaggregated home load rather than the net load improves the overall accuracy of NILM by 22.0% to 66.2%.

We plan to test our method in a climate where snow accumulation can greatly affect the output of PV systems. Another direction we will pursue in future work is designing a disaggregation algorithm that can be used in the presence of BTM battery storage.

ACKNOWLEDGMENTS

As a part of the University of Alberta Future Energy Systems (FES), this research was made possible in part thanks to funding from the Canada First Research Excellence Fund (CFREF).

REFERENCES

[1] Ausgrid. [n.d.]. Solar home electricity data. https://www.ausgrid.com.au/Industry/Our-Research/Data-to-share/Solar-home-electricity-data.

[2] Noman Bashir et al. 2019. Solar-TK: A Data-Driven Toolkit for Solar PV Performance Modeling and Forecasting. In *Proc. 16th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*. IEEE, 456–466.

[3] Nipun Batra et al. 2019. Towards Reproducible State-of-the-Art Energy Disaggregation. In *Proc. 6th International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. ACM, 193–202.

[4] Fankun Bu et al. 2020. A Data-Driven Game-Theoretic Approach for Behind-the-Meter PV Generation Disaggregation. *IEEE Transactions on Power Systems* 35, 4 (2020), 3133–3144.

[5] Dong Chen et al. 2017. SunDance: Black-box Behind-the-Meter Solar Disaggregation. *Proc. 8th International Conference on Future Energy Systems* (2017), 45–55.

[6] Xinlei Chen et al. 2020. Solar Disaggregation: State of the Art and Open Challenges. In *Proc. 5th International Workshop on Non-Intrusive Load Monitoring*. ACM, 6–10.

[7] Chung M. Cheung et al. 2018. Behind-the-Meter Solar Generation Disaggregation using Consumer Mixture Models. In *International Conference on Communications, Control, and Computing Technologies for Smart Grids*. IEEE, 1–6.

[8] Julian de Hoog et al. 2020. Using Satellite and Aerial Imagery for Identification of Solar PV: State of the Art and Research Opportunities. In *Proc. 11th International Conference on Future Energy Systems*. ACM, 308–313.

[9] Aron Dobos. 2014. *PVWatts Version 5 Manual*. Technical Report. Research Organization: National Renewable Energy Lab. (NREL), Golden, CO (United States).

[10] International Energy Agency. 2020. Global Energy Review 2020. https://www.iea.org/reports/global-energy-review-2020. Accessed 01.02.2021.

[11] Farzana Kabir et al. 2019. Estimation of Behind-the-Meter Solar Generation by Integrating Physical with Statistical Models. In *International Conference on Communications, Control, and Computing Technologies for Smart Grids*. IEEE, 1–6.

[12] Emre Kara et al. 2018. Disaggregating solar generation from feeder-level measurements. *Sustainable Energy, Grids and Networks* 13 (2018), 112–121.

[13] Hyungsul Kim et al. 2011. Unsupervised Disaggregation of Low Frequency Power Measurements. *Proc. SIAM Conference on Data Mining* 11, 747–758.

[14] Pecan Street Inc. [n.d.]. Dataport. https://www.pecanstreet.org/dataport/.

[15] Fabian Pedregosa et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, Oct (2011), 2825–2830.

[16] Hamid Shaker et al. 2016. A Data-Driven Approach for Estimating the Power Generation of Invisible Solar Sites. *IEEE Transactions on Smart Grid* 7, 5 (2016), 2466–2476.

[17] Hamid Shaker et al. 2016. Estimating Power Generation of Invisible Solar Sites Using Publicly Available Data. *IEEE Transactions on Smart Grid* 7, 5 (2016), 2456–2465.

[18] Solcast. [n.d.]. Weather Dataset. https://toolkit.solcast.com.au/.

[19] Fabrizio Sossan et al. 2018. Unsupervised Disaggregation of Photovoltaic Production From Composite Power Flow Measurements of Heterogeneous Prosumers. *IEEE Transactions on Industrial Informatics* 14, 9 (2018), 3904–3913.

[20] Joshua Stein. 2012. The Photovoltaic Performance Modeling Collaborative (PVPMC). In *38th Photovoltaic Specialists Conference*. IEEE, 003048–003052.

[21] Michaelangelo Tabone et al. 2018. Disaggregating Solar Generation behind Individual Meters in Real Time. In *Proc. 5th Conference on Systems for Built Environments*. ACM, 43–52.

[22] Chaoyun Zhang et al. 2018. Sequence-to-Point Learning With Neural Networks for Non-Intrusive Load Monitoring. In *AAAI*.

A APPENDIX

A.1 Physical Solar Model

The output power of the PV system with the specified rating P_{dc0} can be computed given the transmitted plane of array (POA) irradiance I_{tr} and cell temperature T_{cell}:

$$P_{dc} = \frac{I_{tr}}{E_{ref}} P_{dc0}(1 + \gamma(T_{cell} - T_{ref})) \tag{5}$$

Here γ represents the temperature coefficient, E_{ref} represents the reference irradiance, and T_{ref} represents the reference cell temperature. We set them respectively to -0.47%/$^\circ C$, 1000W/m^2, and 25$^\circ C$ to create synthetic proxies. I_{tr} is determined by solar irradiance data (direct normal irradiance, diffuse horizontal irradiance, global horizontal irradiance), PV system characteristics (e.g., tilt, orientation), and its location. T_{cell} is a function of wind speed, ambient air temperature, and solar irradiance data. We use different preset technical parameters for different synthetic proxies.

A.2 Supplementary Figures

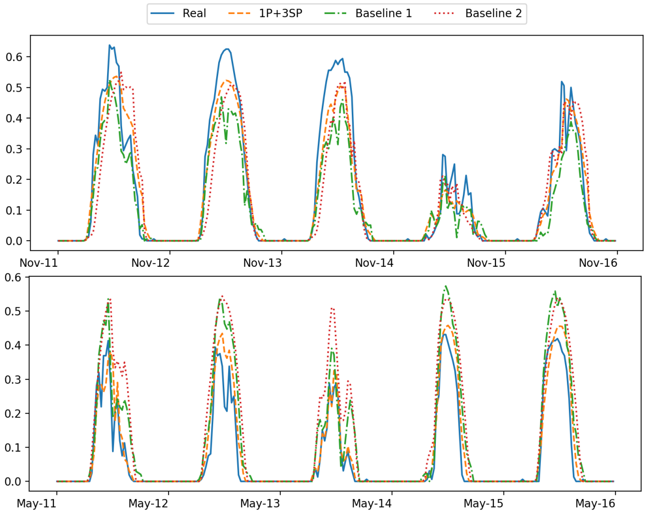

Figure 1: Comparison of disaggregated solar generation in summer (top) and winter (bottom) for a customer.

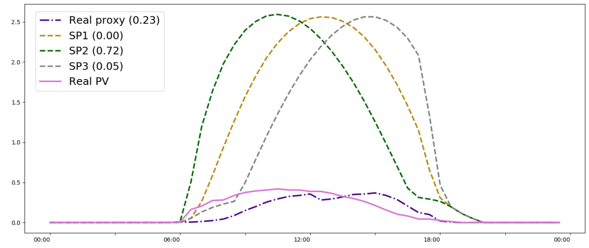

Figure 2: PV generation (kW) of a sample home. Dashed curves show proxy measurements. The relative weights (normalized to sum to 1) of proxies are given in the legend.

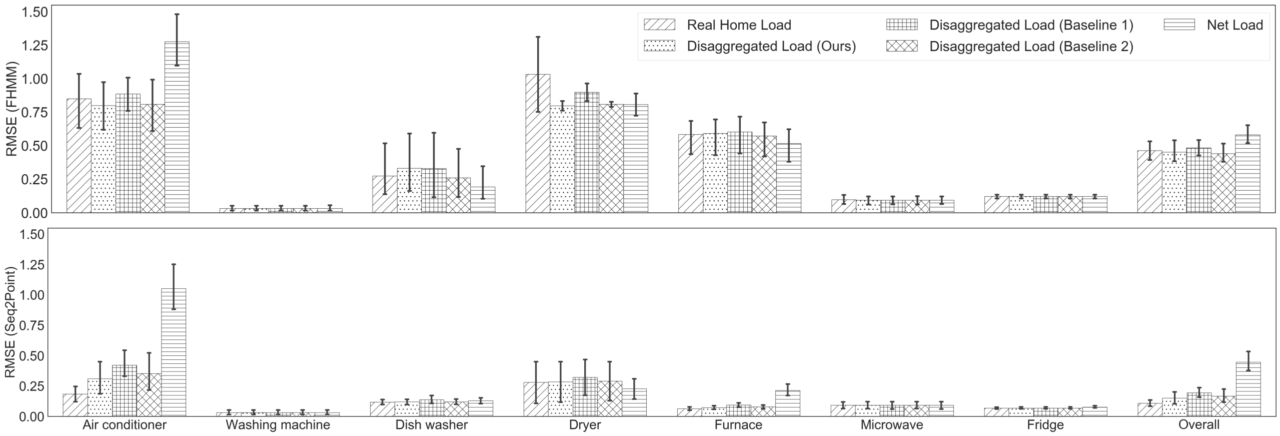

Figure 3: Average RMSE of each appliance among all selected homes. The top plot shows the disaggregation performance of FHMM and the bottom one shows the performance of Seq2Point. Error bars show the 95% confidence interval.