Solar Disaggregation: State of the Art and Open Challenges

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

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

ABSTRACT

Disaggregating solar power from net meter data has got traction in recent years as utility companies are seeking ways to identify behind-the-meter solar photovoltaics, improve their planning and operation practices, and apply variable pricing to distributed solar generation. In this notes paper we survey the literature on solar disaggregation and describe datasets that can be used for evaluating disaggregation methods. We identify limitations and threats to validity of this research, and discuss existing challenges and how they can possibly be addressed. These open challenges highlight the need for the development of advanced techniques and the use of other data sources to solve the solar disaggregation problem.

CCS CONCEPTS

• **Computer systems organization** → *Embedded and cyber-physical systems*; • **Computing methodologies** → *Machine learning approaches.*

KEYWORDS

Distributed solar generation, disaggregation, source separation.

ACM Reference Format:
Xinlei Chen and Omid Ardakanian. 2020. Solar Disaggregation: State of the Art and Open Challenges. In *The 5th International Workshop on Non-Intrusive Load Monitoring (NILM '20), November 18, 2020, Virtual Event, Japan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3427771.3429387

1 INTRODUCTION

Solar photovoltaic (PV) is currently enjoying the fastest growth of renewable sources worldwide thanks to declining costs of solar projects and government tax credits. Solar PV generation in end-use sectors in the U.S. is anticipated to increase more than fourfold, from 41 billion kilowatt-hours (1% of U.S. generation) to 182 billion kilowatt-hours (4% of U.S. generation) by 2050 [10]. The large-scale adoption of distributed PV systems and the rising number of prosumers present new challenges for planning and operation of power distribution grids, from the management of voltage to the configuration of protection systems and increased wear and tear on utility equipment. In light of this, utility companies seek innovative solutions to identify unregistered solar panels, and estimate their peak capacity and real-time production. Since most distributed PV systems are installed behind the meter (BTM), net meter data from advanced metering infrastructure (AMI) is the only type of data that is commonly available. In fewer cases, substation-level and feeder-level voltage and current phasor measurements can be obtained from the supervisory control and data acquisition system (SCADA) or distribution-level phasor measurement units (DPMUs) [16].

Several methods have been proposed to date to estimate the peak generation capacity and real-time production of BTM solar systems. For example, satellite and aerial imagery has been used to identify solar PV systems and estimate their physical deployment characteristics (size, tilt, orientation, etc.) [9]. This approach can provide only a rough estimate of the peak capacity of PV installations and cannot accurately estimate real-time solar generation. Another line of work relies on utilizing PV generation data from a small number of PV sites that are metered separately to estimate the total solar production in a given geographical area [22, 23]. These methods require the knowledge of the total installed capacity of PV systems in that area to estimate PV generation. In this paper, we survey and classify methods based on *source disaggregation*, a technique that has been widely used for non-intrusive load monitoring (NILM) [19]. These methods, if proven reliable, can eliminate the need for a separate meter (in addition to the standard meter) in jurisdictions where customers with installed solar systems are compensated using feed-in tariff and power purchase agreement.

Despite the growing literature on solar disaggregation there are still several key challenges, ranging from the lack of fine-grained data and an evaluation toolkit to latent flexibility, which hinder the application of these methods in practice. Furthermore, the increased BTM solar penetration brings in new challenges for NILM as the smart meter reading may not be equal to the sum of the power demand of individual appliances in the daytime. This indicates the need to disaggregate solar power from net meter data before trying to separate the household demand into the constituent appliances, and calls for incorporating solar disaggregation algorithms in NILM software, such as NILMTK [3].

This paper aims to elaborate on the open challenges in this area and put forward recommendations to address them in future work. These challenges underline the scope for developing new techniques for solar disaggregation and justify the effort to collect data from a larger set of solar installations where solar generation can be separately metered.

2 PROBLEM DEFINITION

Solar disaggregation is the problem of estimating solar generation from net load measurements which can be obtained at different levels of aggregation. Customer-level solar disaggregation is to separate the power consumption measured by a smart meter into household (or business) demand and solar generation. Feeder-level

NILM '20, November 18, 2020, Virtual Event, Japan

ACM ISBN 978-1-4503-8191-8/20/11…$15.00
https://doi.org/10.1145/3427771.3429387

(or substation-level) solar disaggregation concerns separating the overall solar generation at the feeder from the total active power consumption of loads connected to that feeder.

Given a sequence of real power measurements $\mathbf{P} = \{P_1, P_2, ..., P_T\}$ from a smart meter or a DPMU, the problem concerns estimating aggregate solar generation S_τ and aggregate demand L_τ at time τ provided that $P_t = L_t - S_t + B_t + \sigma_t$ $(\forall t \in \{1, 2, ..., T\})$, where $L_t, S_t \geq 0$, σ_t represents the measurement noise and potential losses, and B_t represents battery charge ($B_t > 0$) or discharge ($B_t < 0$) power. Most related work only considers standalone PV systems without storage, hence B_t is assumed to be zero at all times. Note that the disaggregation algorithm is offline when $\tau < T$, and it is an online algorithm otherwise.

2.1 Differences with NILM

Solar disaggregation differs from NILM in that solar generation varies significantly and continuously over time. Most NILM approaches, on the contrary, assume that the operation of an appliance can be divided into a finite number of operating states (e.g., ON, OFF, standby), and the power draw in each state is known and constant. Thus, they cannot disaggregate the demand of Continuously Variable Devices from the net demand measured at the main electrical panel [19].

3 DATASETS

Solar disaggregation studies use net metering data collected at different levels of aggregation. The customer-level data is available through AMI while the feeder-level data is typically collected by DPMUs or the SCADA system if there is such instrumentation beyond the distribution substation. For validation purposes, additional gross meters need to be installed behind the utility meter to separately measure the solar inverter output.

The most popular dataset which has been used for evaluating solar disaggregation methods is released by *Pecan Street Inc.* [12]. It contains customer-level measurement of solar generation and household demand for a total of 73 homes located in three states in the U.S. (New York, California, and Texas). Measurements cover three granularity levels, namely 1 second, 1 minute and 15 minutes. This is useful to evaluate a disaggregation algorithm at different temporal resolutions as we do in this paper. Another relatively large dataset is released by *Ausgrid*, a utility company in Sydney, Australia [1]. The half-hour household consumption and solar output data are collected from 300 customers with rooftop solar PV systems. The dataset spans 3 years, from July 1, 2010 to June 30, 2013, and contains the actual solar panel capacity for each customer. Finally, the SunDance dataset [20] includes hourly net meter, solar generation, and weather data for 100 sites in North America.

To our knowledge, the above datasets are the only publicly available datasets that contain both solar generation and home load data. However, there are several residential load datasets which are published online [4, 11]. Real solar generation data from specific PV sites can be found in [20, 26]. Synthetic PV output can also be simulated using the System Advisor Model (SAM) developed by the National Renewable Energy Laboratory according to the real solar irradiation data and other weather data [18]. Therefore, a synthetic net metering dataset can be created easily by combining home load and solar generation from different sources.

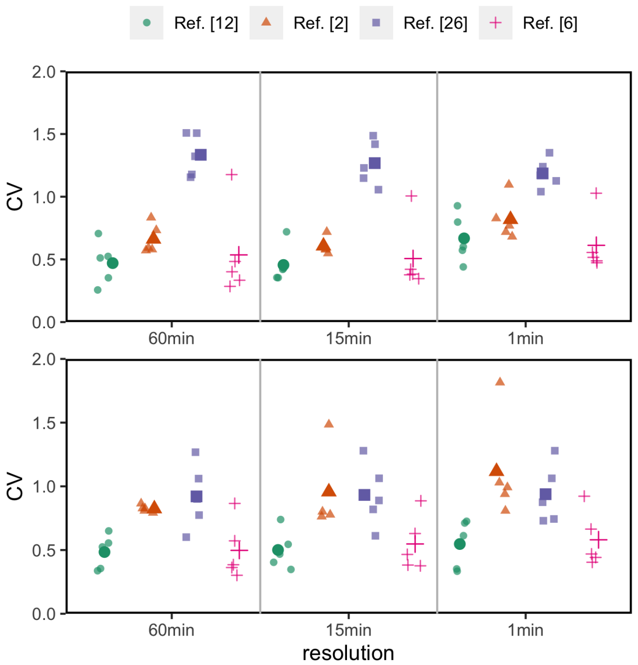

Figure 1: CV of 4 disaggregation methods at different temporal resolutions. Each small marker shows the result of a single home. Each large marker indicates the CV of one method averaged over 5 homes. The top figure shows performance results from 2018/06/01 to 2018/06/30 and the bottom one shows performance results from 2018/12/03 to 2018/12/30.

4 APPROACHES TO DATE

Several solar disaggregation techniques have been proposed to date drawing on algorithms from machine learning, signal processing, and state estimation. Table 1 summarizes these techniques. The vast literature on behind-the-meter solar disaggregation can be categorized into two classes based on how BTM PV systems are modelled. In particular, in model-based techniques PV systems are modelled using a physical PV model, while data-driven techniques develop a black-box PV model leveraging the training data.

4.1 Data-driven Approaches

With the growing adoption of different metering technologies in distribution grids, data-driven methods have become increasingly popular. Kara *et al.* [15] developed a linear proxy-based estimator to disaggregate feeder-level solar generation from the aggregate real power at the substation, using reactive power data measured by a DPMU and PV generation profiles from nearby metered PV systems. A similar method was proposed by Tabone *et al.* to tackle customer-level solar disaggregation [27]. But instead of using reactive power data to estimate the aggregate home load, they use temperature and time of the day. Both methods leverage a contextually supervised source separation model [29] with different features. But the lack

Table 1: Related work on disaggregation of behind-the-meter solar generation

Reference	Level	Rate	Category	Used Data	Approach
Ref. [6] (Ref. [2])	Customer	1 hour	Model-Based	Lon. & Lat. of target home, Weather	Physical model, Machine Learning
Ref. [15] (Refs. [16, 28])	Feeder	1 min	Data-Driven	NS, Reactive power at the feeder	Convex Optimization
Ref. [27] (Ref. [14])	Customer	15 min	Data-Driven	NS, Ambient temperature	Convex Optimization
Ref. [7] (Ref. [8])	Customer	15 min	Data-Driven	NL, Irradiance	Convex Optimization
Ref. [25]	Feeder	Multiple	Data-Driven	Local GHI observation	Convex Optimization
Ref. [13]	Customer	15 min	Model-Based	Weather, GHI, DNI, DHI	Physical model, Optimization
Ref. [17]	Customer	1 hour	Data-Driven	NS, SC, Weather	Machine Learning
Ref. [21]	Feeder	30 min	Data-Driven	Ambient temperature, GHI	State Estimation
Ref. [5]	Feeder	1 hour	Data-Driven	NS, NL	Optimization

NS: output of a nearby solar installation; **NL**: consumption of a nearby home; **SC**: output of a solar installation and its peak capacity (not necessarily in the same region);

of data from separately-metered nearby PV installations presents a barrier to the application of these methods at scale.

Cheung *et al.* [7] model demands of customers with PV systems using a mixture model of representative customers without PV systems which are selected via clustering. Considering the presence of battery storage, the authors improve their method in [8] by adding a “hidden battery” model for the operation of BTM battery. Bu *et al.* [5] utilize separately metered load and PV generation data of some customers to disaggregate higher-level net load using a game theoretic approach. They formulate solar disaggregation as a nested bi-layer optimization problem. This method adaptively updates the estimation in each time step and shows robustness to unobserved events and abnormalities (e.g., PV system failures). Leveraging the fact that customer-level net load curves have different shapes under different weather conditions, Li *et al.* [17] extract multiple features and feed them to machine learning models to infer PV capacity and estimate solar generation.

In another line of work, Sosan *et al.* [25] disaggregate the solar generation in the frequency domain given that the spectral density of the aggregated power flow is similar to the measured PV generation. The result of this method is promising enough to encourage researchers to use frequency domain analysis to address the disaggregation problem especially at the feeder-level where higher resolution data might be available. Shaffery *et al.* [21] propose a Bayesian Structure Time Series (BSTS) model for solar disaggregation. Unlike other data-driven methods, it can provide probabilistic estimation of PV generation and load consumption, allowing the operator to determine necessary reserves. However, the slow training process limits its real-world application.

4.2 Model-based Approaches

Compared to data-driven approaches, model-based approaches require no or just a few data points for calibration. They learn the deployment characteristics of PV systems (e.g., capacity, tilt, orientation), which is useful for both real-time estimation and long-term forecasting [13]. Chen *et al.* [6] design a model-based solar disaggregation technique which requires the knowledge of the home’s location (latitude and longitude) and a small amount of historical net meter data. A clear sky generation model, which estimates the maximum generation under clear sky situation, is used together with a physical PV model to estimate PV generation. Since net load data is the only data available from the household, the clear sky generation model training relies on net load data collected from the target home when it is unoccupied on a sunny day (i.e., when the home load is at its minimum and solar generation is at its maximum). But this data is not always available for a customer.

To overcome the above shortcomings, Kabir *et al.* [13] develop an unsupervised solar disaggregation framework which does not rely on location information or net load data collected under certain conditions. They integrate a physical PV model with a Hidden Markov Regression model for estimating home load. Starting with an initial guess of the key parameters of the physical model, they iteratively estimate solar generation and home load given the measured net load. The main drawbacks of this method are that it takes many iterations to converge and is therefore quite slow, and that accurate disaggregation requires accurate weather and irradiance data to be collected from the vicinity of target homes.

5 CHALLENGES

Despite several efforts to date to disaggregate solar power from net meter data there are several challenges which need to be addressed. **Low temporal resolution of customer-level data:** Data collected by smart meters is coarse-grained (typically 1 sample every 15 minutes). This dramatically increases the difficulty of capturing inherent variability in solar generation and household demand, and prevents researchers from successfully separating frequency components of the two signals. High-frequency DPMU data can support the use of signal processing techniques. But since these sensors are typically installed at higher levels of aggregation in the distribution grid, fluctuations smooth out further due to aggregation.

To understand how the temporal resolution of input data could affect the performance of solar disaggregation algorithms when applied to data collected for individual customers, we run two model-based [2, 13] and two model-free [7, 27] approaches[1] proposed in the literature to disaggregate net meter data of five randomly selected homes in Austin, Texas. We run each algorithm three times for a given home, each time using the net meter data with a different resolution (1min, 15min, and 1hr). We obtain the net meter, home load, and PV generation data for a month in summer and a month in winter from the Pecan Street dataset and pull in 5-minute weather data (GHI, GNI, DNI, Temperature, etc.)[2] for the same periods for a

[1] We implemented [7] from scratch and [13] based on PV modelling code provided by the authors. We used the implementations of [2, 27] we found on authors’ websites.

[2] For the 1min scenario, we use weather data of the closest 5 minute interval.

location in Austin[3] (30.267°N,-97.743°E) from the Solcast API [24]. Note that a subset of these features are needed by each algorithm as specified in Table 1. We do not re-tune the (hyper)parameters of each algorithm across the three runs, and report coeffecient of variation (CV) of root-mean-square error (RMSE) which is RMSE normalized by the mean value of the measurements. For fair comparison, we only include the estimates at the top of the hour in the calculation of CV as they are available for all three runs. As shown in Figure 1, increasing the temporal resolution does not improve the reliability of estimates in general. We believe this is because the algorithms are not designed to take advantage of high-resolution net meter data as it cannot be obtained from AMI today.

Abrupt and gradual changes in the output of PV systems: A number of different events can change the power output of a PV system over time. For example, inverter failure and panel cleaning will cause the PV output to change rapidly in a short period of time; while soiling on solar panels (i.e., the dry deposition of dust and light absorbing particles on the surface of the panel) will gradually decrease its power output. Traditional PV models cannot track these changes to adjust the estimated PV generation.

Latent flexibility: The continued rise in the adoption of BTM energy storage, demand-side management technologies, and smart thermostats in homes and buildings creates problems for most data-driven solar disaggregation models. This is because these components can change the load profile, but control policies used or price signals sent to them are not typically known when the disaggregation problem is solved. This latent flexibility is ignored in most previous work on solar disaggregation. An exception is [8] which addresses the disaggregation problem in the presence of a BTM battery. It also shows that state-of-the-art approaches cannot accurately perform disaggregation when there is a BTM battery. Thus, future work should focus on developing disaggregation algorithms that can remove effects of latent components from net meter data through identification of control signals and actions.

Lack of datasets containing different types of households: Most datasets that can be used for evaluating solar disaggregation methods do not include information about households that participate in demand-response (DR) programs, installed smart thermostats or BTM solar-plus-storage systems in the metadata. Specifically, from the three datasets described in Section 3, only the Ausgrid dataset records whether the customer participated in a DR program. Even when this information is available, it is unclear what control or price signals were sent to those households.

6 RESEARCH OPPORTUNITIES

Several interesting research questions are still to be addressed, some of which will require overcoming the challenges described in the previous section. We outline these opportunities below.

Fusion of customer-level and feeder-level data: In recent years high frequency (up to hundreds of Hz) measurements of real and reactive power have become available thanks to the deployment of DPMUs in the secondary side of the substation. Fusing this data with smart meter data and incorporating pseudo-measurements in the disaggregation method is an interesting direction that can improve the overall accuracy and allow for running the disaggregation method in an online fashion.

Dynamic disaggregation algorithms: As noted in the previous section, the power output of a PV system can change over time due to various reasons. Thus, the PV models trained/calibrated using historical data can no longer provide accurate estimates of solar generation. Dynamic disaggregation enables researchers to simultaneously solve system identification and solar disaggregation problems, presumably at two different timescales. This way the changes in the power output of a PV system can be detected and the PV models can be updated accordingly.

Single-channel blind source separation (BSS) which separates a set of source signals from a single mixed signal is a perfect fit for the solar disaggregation problem. Hence, different techniques used in BSS can be modified and applied to this problem.

Utilizing various data sources: As the Internet of Things (IoT) devices are becoming ubiquitous across the building sector, there is a huge potential to use the data collected by their embedded sensors to better predict the home load. For example, the occupant presence and actions have an impact on the household demand. Hence, utilizing the occupancy data which is recorded by smart thermostats or plug load energy use collected by submetering devices can improve the accuracy of several data-driven disaggregation techniques.

Building a framework for evaluating solar disaggregation algorithms: The lack of publicly available datasets that contain mixed and unmixed signals, open-source implementation of benchmark algorithms, and consensus on the evaluation metrics[4] has made it difficult to compare the methods proposed for solar disaggregation. We believe a comprehensive evaluation toolkit, similar to NILMTK [3], can greatly facilitate research in this area. Alternatively, benchmark solar disaggregation algorithms can be added to existing NILM evaluation frameworks; this would make sense since solar generation must be separated from net meter data before running NILM algorithms.

7 CONCLUSION

Distributed PV generation is growing rapidly in most developing and developed countries. This will create new challenges for utility companies and complicate planning and operation of residential distribution networks. In the absence of elaborate instrumentation beyond the distribution substation, solar disaggregation techniques can help utilities get a better understanding of the overall solar generation, identify hotspots, and upgrade their equipment. In this paper, we surveyed prior work in this area, identified key challenges that must be overcome, and presented opportunities for improving the accuracy of solar disaggregation techniques. We hope that this work contributes to a broader research effort in this area.

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).

[3] The Pecan Street dataset does not include the home address, hence we queried weather data for a randomly selected location in downtown Austin.

[4] The following metrics have been used in related work to report the accuracy of a solar disaggregation algorithm: RMSE, MAPE, and coefficient of variation of RMSE.

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] N. Bashir, D. Chen, D. Irwin, and P. Shenoy. 2019. Solar-TK: A Data-Driven Toolkit for Solar PV Performance Modeling and Forecasting. In *2019 IEEE 16th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*. 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. 2019. A Time-Series Distribution Test System Based on Real Utility Data. *2019 North American Power Symposium (NAPS)* (2019), 1–6.

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

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

[7] C. 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] Chung M. Chueng et al. 2020. Disaggregation of Behind-the-Meter Solar Generation in Presence of Energy Storage Resources. In *IEEE Conference on Technologies for Sustainability (SusTech)*. IEEE.

[9] 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 ACM International Conference on Future Energy Systems*. ACM, 308–313.

[10] Energy Information Administration. 2020. Annual Energy Outlook 2020. *US Department of Energy* (2020).

[11] ISO New England Inc. [n.d.]. Zonal Information. https://www.iso-ne.com/isoexpress/web/reports/pricing/-/tree/zone-info.

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

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

[14] Emre Kara et al. 2016. Estimating Behind-the-Meter Solar Generation with Existing Measurement Infrastructure: Poster Abstract. In *Proc. 3rd ACM International Conference on Systems for Energy-Efficient Built Environments*. ACM, 259–260.

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

[16] Emre Can Kara et al. 2016. Towards Real-Time Estimation of Solar Generation From Micro-Synchrophasor Measurements. *ArXiv* abs/1607.02919 (2016).

[17] Kangping Li et al. 2019. Capacity and output power estimation approach of individual behind-the-meter distributed photovoltaic system for demand response baseline estimation. *Applied Energy* 253 (2019), 113595.

[18] Blair Nate et al. 2018. System Advisor Model (SAM) General Description (Version 2017.9.5).

[19] Oliver Parson et al. 2012. Non-Intrusive Load Monitoring Using Prior Models of General Appliance Types. In *Proc. 26th AAAI Conference on Artificial Intelligence*. AAAI Press, 356–362.

[20] UMass Trace Repository. [n.d.]. Smart* dataset. http://traces.cs.umass.edu/index.php/Smart/Smart.

[21] P. Shaffery et al. 2020. Bayesian Structural Time Series for Behind-the-Meter Photovoltaic Disaggregation. In *Innovative Smart Grid Technologies Conference*. IEEE, 1–5.

[22] H. 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.

[23] H. 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.

[24] Solcast. [n.d.]. weather dataset. https://toolkit.solcast.com.au/.

[25] F. 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.

[26] California Distributed Generation Statistics. [n.d.]. IOU solar PV net energy metering (NEM) interconnection data. https://www.californiadgstats.ca.gov/downloads/.

[27] M. 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.

[28] E. Vrettos et al. 2019. Estimating PV power from aggregate power measurements within the distribution grid. *Journal of Renewable and Sustainable Energy* 11, 2 (2019), 023707.

[29] Matt Wytock and J. Zico Kolter. 2014. Contextually Supervised Source Separation with Application to Energy Disaggregation. In *Proc. 28th AAAI Conference on Artificial Intelligence*. AAAI Press, 486–492.