Disaggregating solar generation behind individual meters in real time

Michaelangelo Tabone
Stanford University, Department of
Civil Engineering
Stanford, California
michaelangelo.tabone@gmail.com

Sila Kiliccote
SLAC National Accelerator
Laboratory
Menlo Park, California
silak@slac.stanford.edu

Emre Can Kara
SLAC National Accelerator
Laboratory
Menlo Park, California
eck@fastmail.com

ABSTRACT

Real-time photovoltaic (PV) generation information is crucial for distribution system operations such as switching, state-estimation, and voltage management. However, most behind-the-meter solar installations are not monitored. Typically, the only information available to the distribution system operator is the installed capacity of solar behind each meter; though in some cases even the presence of solar may be unknown. We present a method for disaggregating behind-the-meter solar generation using only information that is already available in most distribution systems: advanced metering infrastructure, substation monitoring, and generation monitoring at a few PV systems nearby the circuit. The proposed method accurately predicts which homes have solar in over 90% of cases, and recovers the 15-min resolution PV generation signals with root mean square errors between 20% and 50% of average daily PV generation both historically and real-time. A sensitivity analysis shows the method to be robust to the number of buildings and time span of data used to fit. However including more than 3 solar proxies can cause false positive of PV systems behind meters. We find that the proposed method performs better at homes that export electricity to the grid more often.

CCS CONCEPTS

• **Applied computing** → *Mathematics and statistics*;

KEYWORDS

Energy disaggregation, solar generation

ACM Reference format:

Michaelangelo Tabone, Sila Kiliccote, and Emre Can Kara. 2018. Disaggregating solar generation behind individual meters in real time. In *Proceedings of ACM International Conference on Systems for Energy-Efficient Built Environments, Shenzhen, China, November 2018 (BuildSys'18)*, 10 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

Incorporating large amounts of distributed energy resources (DER) within consumer buildings is challenging to the operation and planning of distribution systems. There are many proposals for

advanced control mechanisms of large numbers of DERs. However, these strategies often assume that a reliable model of the distribution system exists, and that the operational limits and states of DERs are observed. Achieving the level of observability that is required by these control schemes will require significant investment into sensor systems. In short, lack of visibility into the distribution grid is a limitation to widespread DER adoption.

Solar generators located behind-the-meter are of particular concern to distribution system operators. In Hawaii, a recent study found that 10% penetration of photovoltaic (PV) would cause voltage and protection issues [19], [9]. Distributed PV generation can affect system losses, transformer aging, mechanical wear on voltage control equipment, power quality, and reverse power flow [5]. Distribution system operators are preparing by assessing the ability of each circuit in their system to sustain a continued increase in distributed PV. Simulation models are typically used to study how solar generation will affect distribution systems in the future. Cohen and Callaway find that the effects of PV depend dramatically on its location in a distribution feeder, the layout of that feeder, and the prevailing climate [5]. To accurately make these assessments, distribution system operators need to understand the amount of PV currently in their systems, and the effects that it is already creating.

PV generation can also affect the real time operations of distribution systems. Downstream of substations, regions may be switched from one feeder to another in order to better balance load between phases, or in order to isolate lines for service. This switching often occurs with no real time measurements of power consumption within each region, thus consumption must be estimated based on historical or characteristic load profiles. Distributed PV generation will affect these decisions, as the amount of PV and the current conditions will alter the actual load in different regions from that expected from characteristic load models.

A common method for estimating the output of PV systems is a model-based approach, such as those used in [1, 6]. The critical input to these models is the plane-of-array (POA) irradiance which is calculated using solar-earth geometry models, knowledge of the geometry of the PV array, and site-specific measurements of direct normal irradiance (DNI), diffuse horizontal irradiance (DHI) and global horizontal irradiance (GHI). However, accurate measurements of both DNI and DHI are costly to obtain, typically only GHI is measured [7] [8]. In this case both the DNI and DHI need to be estimated from a given GHI measurement. A total of ten models were examined in [7] for estimating the POA irradiance, given site-specific GHI measurements, with typical root-mean-squared-error (RMSE) of $\approx 10\%$ for all sky conditions for a 40° tilted south-facing panel and an RMSE of $\approx 13\%$ for a two-axis tracking plane.

© 2018 Association for Computing Machinery. ACM acknowledges that this contribution was authored or co-authored by an employee, contractor or affiliate of the United States government. As such, the United States Government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for Government purposes only.
BuildSys '18, November 7–8, 2018, Shenzhen, China
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5951-1/18/11...\$15.00
<https://doi.org/10.1145/3276774.3276776>

The use of satellite-derived estimations appears to offer no significant improvement in performance [1] [10]. Jamaly et al. [10] estimate aggregate PV ramp rates for 86 DG systems using satellite irradiance data and demonstrate an RMSE of $\approx 20\%$. The authors conclude that the primary source of this error stemmed from the inaccuracy of the satellite-estimated irradiance, rather than the model employed, to estimate power production. This is in line with the conclusions of [7] and [8] regarding POA irradiance estimation inaccuracy. This inaccuracy, coupled with the low frequency of the data, 30-minute resolution for this study, leaves room for improvement for the purpose of real-time disaggregation. The inaccuracy in estimating the POA, compounded by the lack of site-specific GHI measurements, render the model-based approach suboptimal for real-time disaggregation of DG.

Recent work focuses on the estimating the generation from small-scale PV systems [18] and relies on the identification of a limited number of PV sites whose data could be utilized to estimate the power production of a much larger set of systems within an area. Such a methodology requires knowledge of total capacity installed within a region and fails to consider potential site-specific sources of error, irrespective of irradiance. One such source of error is the estimation of the soiling losses, arising from the accumulation of particulate matter on PV systems. These soiling losses averaged 0.051%/day during dry periods for a sample of 1286 residential and commercial PV sites in California [14]. There were significant differences, however, between rates with 26% of sites having losses above 0.1%/day while systems with a tilt angle $< 5^\circ$ had mean soiling losses of 0.18%/day. The authors identify no statistically significant differences by region, making such a phenomenon more difficult to model. Issues such as the degradation of PV arrays and shading of residential installations further impact the reliability of capacity based generation estimations.

An alternative estimation approach exploits the requirement of metering for PV sites above a specified capacity. When present, however, these meters typically offer low-resolution measurements (often 15-min) and are not required on residential systems. Thus they offer only limited visibility. Although there is interest in real-time monitoring and control of solar inverters by utilities and distribution grid operators, it requires a higher bandwidth communications infrastructure, which requires greater implementation costs [17].

In this work, we present a means of estimating the generation of behind-the-meter photovoltaic (PV) systems at individual homes. The main contributions of this work are given as follows: (i) We only use weather data, and information that is already collected within distribution systems: advanced metering infrastructure (AMI), supervisory control and data infrastructure (SCADA), and some information on the solar resource, which we call the solar proxy. Hence, the proposed system prevents the need for installing additional sensors in the distribution grid. (ii) We expand on previous work that disaggregates solar generation at the distribution system feeder-level, and requires reactive power values [12]; we now estimate solar generation at individual homes, and we do not require reactive power measurements in our model.

This work is similar to previous work that uses AMI data to estimate behind the meter solar generation. Mohan et al. [15] adopt an unexplained proprietary method that uses weather data and AMI data to predict behind-the-meter solar generation. Chen and Irwin

Figure 1: Example distribution system where loads may or may not have behind-the-meter PV generators, and some PV generators are connected and metered separately from loads.

use AMI data to estimate the geometry of behind-the-meter PV systems. They use AMI and weather data to augment a geometric model of behind-the-meter solar production [4]. Our work differs from these methods by being part of an open source package, and not attempting to estimate solar geometry.

2 PROBLEM FORMULATION

Figure 1 presents an illustrative distribution system. The following paragraph describes this system and introduces the nomenclature for this paper. A set of buildings, indexed by i , are all connected in the same system. Each building consumes active power at all monitored times, $L_{t,i}$. A PV system may be located behind the meter at each building, and will produce active power, $S_{t,i}$. Meters at each building measure the net load, $Y_{t,i} = L_{t,i} - S_{t,i}$. Meters at substations measure aggregate downstream net load and line losses, $\bar{Y}_t = \sum_{i=1}^N Y_{t,i} + Z_{t,i}$; where $Z_{t,i}$ are the line losses resulting from power consumed at node i .

The distribution system operator receives AMI data, but only after a long lag—in actual systems today, this lag is often a day. It receives data measured at the substation every 4 seconds at close to real time, known as “streaming” data. The operator desires both historical and real time information about solar generation. In the historical context, the operator would like to know whether a PV system is located behind each meter, what is its effective capacity, and what is its historical generation profile. This historical data can identify reasons for various system events or to plan for more PV in the same system. In real time, the operator would like to know the actual load and PV generation within regions of nodes in the feeder; this will inform operational decisions like switching regions from one substation to another. We propose a disaggregation problem that is trained on historical AMI data, but only requires SCADA data for real time applications.

Problem (1) defines the training disaggregation, which is an adaptation of “contextually supervised source separation”, originally presented by Wytock and Kolter in [20]. In Problem (1), L_i , and S_i are the i^{th} columns of L and S , and signify the time series of load or solar generation at node i . $\bar{L} \in \mathbb{R}^T$ represents aggregate load among all nodes ($\bar{L} = \sum_{i=1}^N L_i$). $X_i \in \mathbb{R}^{T, K_i}$ is a set of observed contextual data that relate to solar generation at a node i , and $X_{\bar{L}} \in \mathbb{R}^{T, K_{\bar{L}}}$ is a set of observed contextual data that relate to aggregate load in the system. $\theta_{\bar{L}} \in \mathbb{R}^{K_{\bar{L}}}$, and $\theta_i \in \mathbb{R}^{K_i}$ are coefficients scale the data in the X matrices to predict aggregate load or behind-the-meter solar. As decision variables, we refer to all coefficients collectively as $\Theta = \{\theta_{\bar{L}}, \theta_1, \dots, \theta_N\}$. $A_{\bar{L}} \in \mathbb{R}^{T, T}$ and $A_i \in \mathbb{R}^{T, T}$, are diagonal

matrices of time varying weights for each signal at each time, where $\alpha_{i,t} = A_{i,t,t}$, and $\alpha_{L,t} = A_{L,t,t}$.

$$\begin{aligned} & \underset{L, S, \Theta}{\text{minimize}} && \|A_L(\bar{L} - X_L \theta_L)\|_2^2 + \sum_{i=1}^N \|A_i(S_i - X_i \theta_i)\|_2^2 \quad (1) \\ & \text{subject to} && L_i + S_i = Y_i, \quad \forall i \\ & && S_i \leq 0, \quad \forall i \\ & && L_i \geq 0, \quad \forall i \\ & && \theta_i \geq 0, \quad \forall i \end{aligned}$$

The problem itself is best explained in two parts, though they are performed simultaneously in the optimization. Eqs. (2) and (3) define models that predict solar generation at each home and aggregate load on the feeder; where the $\epsilon_i \in \mathbf{R}^T$ terms represents the time series of errors between the linear model and solar generation behind node i , and $\epsilon_L \in \mathbf{R}^T$ is analogous for aggregate load. Part 1 fits the parameters, Θ , by minimizing a weighted sum of square errors. Section 3.1 describes the form of these models that we use in this paper, and the exogenous data that we include in X_L and X_i . We fit one model to aggregate load instead of multiple models to individual loads because aggregate load is easier to predict.

$$S_i = \theta_i X_i + \epsilon_i \quad (2)$$

$$\forall i \bar{L} = \theta_L X_L + \epsilon_L \quad (3)$$

The operator does not observe S or L —if it did it could fit these models independently to find the parameters. Part 2 fits S and L as decision variables in the same optimization with same the objective of minimizing square errors. Solar generation is constrained to be negative, load is constrained to be positive, and the sum of load and solar behind each meter is constrained to equal the net load observed at that meter, Y_i . The weighting terms, α , in (1) weight the errors of each individual solar and load model at each time. These weights serve as a measure of confidence in the models: a large alpha indicates that one expects the model to have small prediction errors, thus any prediction error is heavily weighted in the optimization function, and vice-versa. In Section 2.1 we discuss these weights in more detail, and how one may tune them to optimize disaggregation.

Appendix A shows that the optimization can be equivalently rewritten to remove the individual load signals as decision variables and including aggregate load in their place. Load at individual nodes can then be deduced from the observed net load and the disaggregated solar generation: $L_i = Y_i - S_i$.

Eq. (4) defines the streaming problem which relies only on the aggregate net load measured at the substation, $\bar{Y}$, and the exogenous variables X_L and $X_i \forall i$. In this optimization, we adopt the model parameters that are previously fit in the training problem, Θ , and the only decision variables are the aggregate load signal and the individual solar signals. The optimization assumes that $\bar{Y} = \sum_{i=1}^N Y_i$ which ignores losses within the distribution system because aggregate net load is measured at the substation and individual net loads are measured at consumer meters.

$$\begin{aligned} & \underset{\bar{L}, S}{\text{minimize}} && \|A_L(\bar{L} - X_L \theta_L)\|_2^2 + \sum_{i=1}^N \|A_i(S_i - \theta_i X_i)\|_2^2 \quad (4) \\ & \text{subject to} && \bar{L} + \sum_{i=1}^N S_i = \sum_{i=1}^N Y_i, \\ & && S_i \leq 0 \quad \forall i. \end{aligned}$$

2.1 Tuning α

The weighting terms in the A matrices, $\alpha_{i,t}$, normalize the square errors from each linear model at each time to make them directly comparable in the objective function. An intuitive choice for this normalization is the inverse of the expected variance of each model's errors: $\alpha_{i,t} = \text{Var}(\epsilon_{i,t})^{-1}$; and $\alpha_{L,t} = \text{Var}(\epsilon_{L,t})^{-1}$. Given this normalization, the expected value of each term is one. In [11], we show that normalizing to the true variance of model disturbances amounts to a maximum likelihood disaggregation under the conditions that model errors are independent and Gaussian. Unfortunately, estimating the expected variance of model errors requires that we observe the source signals, which we do not.

We present a tuning method of α in which we assume observations of solar generation are available at at least one location, which we refer to as the tuning systems, S_{tune_i} . Tuning is a multi-step process. This work requires a more complicated method than the one in [11] due to difficulties in estimating the variance of the solar models at unobserved systems. To overcome this, we first estimate the error variance of solar models at the tuning systems. Then, we scale this variance to all other systems using the coefficients from the untuned optimization. Finally, we subtract the variance of solar model errors from the variance of total net load model errors to estimate the variance of the aggregate load model. Since this can result in a negative variance in some cases, we also enforce lower bounds for all variances. The tuning steps are given as follows:

1. Fit (15a) under naive weights; collect estimates of S_i and C_i .
2. Fit a model predicting S_{tune_i} from X_i ; collect the residuals, $\epsilon_{\text{tune}_i} \in \mathbf{R}^T$.
3. Fit a model predicting the squared normalized residuals of the tuning system's model to the tuning data, $X_{\text{tune}} \in \mathbf{R}^{T, K_{\text{tune}}}$, shown in Eq. (5).
4. Estimate the variance of errors in each solar model by scaling predictions from Eq. (5), shown in Eq. (6).
5. Fit a linear model to predict total net load, $\bar{Y}$, from all data X_i and X_L ; collect the residuals, $\epsilon_{\bar{Y}} \in \mathbf{R}^T$.
6. Fit a linear model predicting the squared residuals of the model predicting aggregate net load, $\epsilon_{\bar{Y}}$, to X_{tune} shown in Eq. (7).
7. Estimate the variance of errors in the aggregate load model by subtracting the expected variances of each solar model from the expected variance of aggregate net load model, shown in Eq. (8).
8. Enforce a minimum expected variance on each model, and invert the variances to find tuned weights, α .
9. Scale all α values such that the largest is 1.

Eq. (5) describes the linear model relating normalized squared residuals to the tuning data, X_{tune} . In practice, X_{tune} only contains

indicators for the hour of day, and is used to prevent the model from assigning positive values to PV generation at night. Eq. (6) uses Eq. (5) to estimate the variance of errors in the solar model at a meter by re-scaling the output of the tuning model to the effective solar capacity at that meter. Eq. (6) also enforces a minimum standard error of 5% of average absolute solar generation (estimated in the naive disaggregation).

$$\left(\frac{\epsilon_{tune_i,t}}{C_{tune_i}}\right)^2 \approx X_{tune,t} \theta_{tune,S} \quad (5)$$

$$\sigma_{i,t}^2 = \max\left(0.05 \cdot T^{-1} \|S_i\|_1, C_i^2 X_{tune,t} \theta_{tune,S}\right) \quad (6)$$

Eq. (7) estimates the total variance of errors in predicting aggregate net load, $\sigma_{\bar{Y},t}^2$, by regressing the square residuals onto the tuning data. Eq. (8) estimates the variance in errors in the Aggregate load linear model by subtracting the solar variances from the net-load variance. This step also maintains a minimum variance of 5% of the average absolute net load.

$$\left(\epsilon_{\bar{Y},t}\right)^2 \approx \sigma_{\bar{Y},t}^2 = X_{tune,t} \theta_{tune,\bar{Y}} \quad (7)$$

$$\sigma_{\bar{L},t}^2 = \max\left(0.05 \cdot T^{-1} \|\bar{Y}\|_1, \sigma_{\bar{Y},t}^2 - \sum_i \sigma_{i,t}^2\right) \quad (8)$$

The time-varying variance estimates for each linear model provide inverse normalization constants for the objective function in (1), $\alpha_{i,t} = \sigma_{i,t}^{-2}$ and $\alpha_{\bar{L},t} = \sigma_{\bar{L},t}^{-2}$. In practice, large values of α cause convergence errors, so we form an equivalent problem by normalizing all α values to the maximum among them.

For the streaming problem, only steps 4, and 7-9 are needed as all tuning models will already be fit in the training problem. Streaming observations of the tuning data, X_{tune} , are required for the streaming problem.

3 DATA

From Pecan Street Inc., we obtained one complete year of data from 52 households with PV in Austin Texas. For each home, both PV generation and load are metered separately at one minute sampling rates, which Pecan Street averages into 15-minute time-aligned intervals. Pecan Street also provides hourly resolution outdoor temperature. In order to model homes with PV behind the meter, we subtracted the solar generation from electricity demand to create net load at the meter. For homes without PV behind the meter, we used only the electricity consumption as load. Figure 2 shows one example day from two households. Similarly to the systems in Figure 2, all systems metered by Pecan Street export electricity at some point throughout the year.

3.1 Preparing Contextual Information

Contextual information stored in the X variables drives solar disaggregation, where information in X_L should relate to aggregate load and provide any additional information that can predict solar generation beyond what is contained in X_i ; and vice versa.

To predict solar behind the meter at each home, we assume that there is some amount of metered solar generation installed nearby,

Figure 2: Load, solar generation, and net load measured at 2 households in the Pecan Street dataset.

which we include in each X_i and refer to as “solar proxies.” While we use metered solar generation as solar proxies in this paper, we could use any information that is related to PV generation—e.g, sky cover index or insolation measured with a pyrometer.

Solar proxies will likely be collected at systems with different tilts and azimuths than those of unmetered PV systems. Thus, we expect that there will be nonlinear differences between the proxies and the unmetered generation that cannot be captured by the linear models. These errors can be considered “transposition errors” because they result from the transposition of the irradiance beam from the proxy’s plane to the planes of the unobserved PV systems. Correcting for these transposition errors would require knowledge of the geometry of the unobserved systems, which we do not presume that the distribution system operator will have. Recent research suggests that system geometry can be recovered from net load data, in which case one could transpose generation from the proxy system to the plane of the unobserved system before disaggregation in order to prevent these errors [4].

We propose that having multiple proxies from PV systems of different geometries will help correct for these errors. Each proxy would have a separate coefficient, thus estimating unmetered PV generation to be a linear combination of the transposed proxy signals. Also, the disaggregation method used in this paper can correct for transposition errors on its own, we explain this process in the remainder of this paper.

Contextual information for aggregate load include time of day dependence and temperature dependence. Time of day is included as a load predictor because load patterns repeat themselves on a roughly daily schedule, and that the daily schedules of solar generation is fully predicted by the solar proxy. Solar generation is also dependent on temperature. We assume that this temperature dependence is effectively captured by the solar proxy, and any remaining temperature dependence is related to loads such as space heating/cooling and water heating.

Time of day dependence is included using hour of day indicator variables, $HOD \in \{0, 1\}^{T,24}$, where $HOD_t, i = 1$ if time t is in the i^{th} hour of the day.

Temperature dependence of load accounts for heating, ventilation, and air conditioning, which comprise to 40% of all residential energy use [16]. We adopt a method for identifying piece-wise linear relationship between outdoor temperature and load that was originally presented in Mathieu et al. [13]. This method expands

Table 1: Creation of regressors to predict aggregate load.

ϕ	Φ		
	$b_1 = 60$	$b_2 = 70$	$b_2 = 80$
55	-5	0	0
65	5	0	0
70	10	0	0
74	10	4	0
85	10	10	5
95	10	10	15

Figure 3: (top) Time series of temperature over one day. (bottom) Expansion of temperature into regressors representing variation within 10 degree bins, lines are colored to represent bins, labeled by lower bounds.

the temperature data vector, $\phi \in \mathbf{R}^T$, into a matrix, $\Phi \in \mathbf{R}^{T, N_b}$ where each regressor represents variation within a binned range. (9) presents this expansion mathematically; where b_j represents the j^{th} bin edge in ascending order; $b_{j+1} > b_j$. Similar to [13], we create bins with equally spaced edges at 10F such that all temperatures observed in the training set are within at least 10F of a bin edge. Temperature in the streaming data can be outside of this range. Table 1 and Figure 3 present examples encodings where bin boundaries are 60, 70, and 80.

$$\Phi_{t,j} = \begin{cases} \min(\max(\phi_t - b_j, 0), b_{j+1} - b_j) & i \in [2, \dots, N_b - 1] \\ \min(\phi_t - b_j, b_{j+1} - b_j) & i = 1 \\ \max(\phi_t - b_j, 0) & i = N_b \end{cases} \quad (9)$$

Table 2: Performance of one run the descriptive model fit to one region of 10 homes.

House ID	Ave.	$\alpha = 1$		tuned α		
		RMSE	MBE	ave(α)	RMSE	MBE
93	-0.84	0.61	0.37	1.04	0.56	0.25
8741	0.00	0.67	0.51	0.00	0.00	0.85
8084	-1.47	0.48	0.20	0.27	-0.03	0.19
7741	0.00	0.67	0.51	0.01	0.00	0.29
7627	-0.72	0.47	0.22	0.10	-0.05	0.30
5129	0.00	0.67	0.51	0.00	0.00	0.73
4336	0.00	0.67	0.51	0.01	0.00	0.35
26	-0.92	0.45	0.23	0.20	-0.08	0.22
Agg. Load	12.63	4.23	-3.05	0.89	-0.41	0.00

4 DESCRIPTIVE RESULTS

For descriptive purposes, we perform one run the training problem on a region of 8 households—randomly selected without replacement from the 52—using solar from 3 systems as proxies and using 10 randomly selected days of data for training and testing.

Row 1 of Figure 4 shows model results for solar at two individual homes and the aggregate load of the region. The untuned model assigns a significant amount of solar to nighttime readings, and fits a small amount of solar to the home on which there isn't any; both of these problems are corrected by tuning the α parameters.

Within each plot, we show the Coefficient of Variation (CV) of the disaggregated signals—solar generation within each household and aggregate load in the region. Eq. (10) defines the coefficient of variation as the root mean square error divided by the average of the true signal. This normalized metric allows us to compare the disaggregation quality for signals of different magnitudes—I.e. generation from differently sized PV systems or load from variously sized regions and homes.¹ To provide an intuitive understanding of CV, we note that the average capacity factor of a residential PV system is about 20%. Thus, dividing the CV by 5 roughly yields the average error of our analysis as a percent of rated capacity. Therefore, a CV of 50% for PV generation equates to an average error of 10% of rated solar capacity. The CV metric fails when applied to homes without solar, for which the denominator is zero. For the remainder of this paper, we use RMSE to present error metrics for homes without solar, and CV to present results from homes with solar and aggregate load.

$$CV = \frac{\sqrt{\sum_{i=1}^T (\text{truth}_i - \text{estimate}_i)^2}}{\frac{1}{T} \sum_{i=1}^T \text{truth}_i} \quad (10)$$

$$= \frac{RMSE}{\text{Mean truth}}$$

¹Mean Average Percent Error, MAPE, is a more common normalized performance metric. In MAPE, the prediction error at each interval is normalized to the actual value at that interval. However both solar generation and load go through periods of being 0 or virtually 0; in which case the MAPE becomes arbitrarily large or infinite. These readings dominate the overall MAPE metric, making it inappropriate for this analysis.

Table 3: Description of sensitivity scenarios

Sensitivity	# of Homes	# of Days	# of Prox	# of PV	# of Runs/scen.
Rev. Power	8	10	3	4	100
# of Homes	2-9	10	3	50%	100
# of Days	8	5-25	3	4	100
# of Proxies	8	10	1-9	4	100

Figure 5 emphasizes the effects of the tuned hyper-parameters, α , on the model. Panel A displays the inverse weighting for one PV system for aggregate load over one day, this value is proportional the expected variance of model errors. The expected variance changes throughout the day because it is dependent on the tuning regressors, which in this case are the indicator matrix HOD . α_{AL}^{-1} is much larger than α_i^{-1} , signifying that more of total linear model disturbances are assigned to the the aggregate load signal than the solar signal. Panels B and C display solar generation and aggregate load respectively: both that predicted only by the linear model (e.g., $X_i\theta_i$), and that predicted by the disaggregation (e.g., S_i). The solar signal follows the linear model predictions closely, while the Aggregate load model varies from the linear predictions. Please see the original paper on solar disaggregation by Kara et al. for more explanation of how the α weightings affect deviations from the linear models [12].

Table 2 describes the model performance for identifying solar generation at each household in the region, and identifying aggregate load. Coefficients of variation are typically over 50%, as shown by comparing the RMSE to the average signal. Estimates are also very biased by between 20% to 65% for homes with solar. Homes without solar are even more biased, typically they are erroneously estimated to contain a small but significant amount of PV generation. In the following section, we discuss how tuning the weighting parameters can reduce these errors.

5 SENSITIVITY

We study the performance of this model under the sensitivity of four parameters: the fraction of time a PV system is feeding back into the grid, the number of days of data used to train, the number of homes included in a region, and the number of solar proxies. In each sensitivity case we also show the performance of both the initial and the tuned models. Table 3 describes the parameters used for each sensitivity analysis. For each permutation of parameters, we create a region of homes by randomly selecting them from the total pool, and then remove solar generation from 40% of them. Solar proxies are solar generation at homes randomly selected from those not included in the training set. Days used for fitting are also randomly selected from the 365 within 2015.

One of the most influential constraints on the model confines solar generation behind a meter to be greater than or equal to the amount of energy exported from that meter at any time. This constraint is outlined explicitly in the reduced training problem, (15a). The more often PV generation exceeds demand at a home, the more definite information we have about solar generation, and the better we expect the model to perform. Figure 6 shows sensitivity of solar disaggregation performance to the fraction of time power

is injected at a meter. Panel A presents results from the initial disaggregation, before tuning the α parameter, and Panel B presents results after tuning. Reverse power flow greatly improves model performance, particularly before tuning.

Load is difficult to predict at the individual home level, where variation is driven by large loads, e.g., air conditioning or refrigerators cycling on/off, coffee machines, irons, etc. Aggregating load from many homes together smooths out these variations and makes it easier to predict. Thus we expect that increasing the number of homes in a region increases the performance of the model overall.

Figure 7 shows error metrics for estimates of aggregate load, solar generation at individual homes. As expected, performance of load prediction increases with the number of homes in a region. Most of this benefit appears to come from better identification of homes without PV generation. The benefits of aggregation level off around 8 homes, which is the value we use for the remainder of the sensitivity analyses.

The number of days used for training also affect the results. Variations in solar generation and load are autocorrelated, these correlated errors are more likely to bias model parameters. Thus using data from more days adds variations that help uncover the true relationships between net-load and the regressors, instead of autocorrelated variations. Figure 8 shows error metrics as they vary with the amount of data used for training. Increasing the number of days greatly improves performance for the untuned model, however model tuning appears to mitigate some of the poor performance from a reduced number of days. Again, most of the performance increase comes from better identification of homes without solar.

Additional solar proxies add to the number of possible cloud patterns and solar panel geometries (discussed in Section 3) that are available to identify solar behind the meter. If homes span a large region, one expects that solar generation relates more to the cloud patterns at nearby systems than at those further away, thus having a diversity of locations for solar proxies should also be important. Pecan Street does not provide locations for the households that we monitor, so we cannot evaluate how the distance between homes affects the relationship between PV power generated at each home. We expect that, to some extent, increasing the number of proxies increase the ability of the model to disaggregate solar generation. However, Figure 9 shows that additional proxies have a slight positive effect on performance on homes with solar and a slight negative effect on performance of homes without solar. Including additional solar proxies appears to allow for more spurious fitting of home without solar to at least one proxy signal. This effect can outweigh the benefit of better disaggregating solar at true solar homes depending on the total penetration on the feeder.

With respect to tuned versus untuned models, we find that tuned models perform better in most cases, particularly for meters without any solar generation, and when applied to homes with a low fraction of reverse power flow. In cases where there is solar behind meters, where there is a large fraction of reverse power flow, and/or there is a large set of training data, the initial modes may perform as well or even better than the tuned models. This could be a result of over-fitting the model to the tuning systems in absence of information from the unobserved systems.

In order to compare our results to previous work, we also compute MAPE for all daylight hours and MAPE for all hours between

Figure 4: Descriptive results from model fitting. (Top row) Performance of the model before and after tuning alpha for (left) a home with solar, where points along the $y=x$ are preferred because they indicate that the results of disaggregation match actual solar production; (center) a home without solar, where a lower RMSE is preferred; (right) aggregate load where, again, points along the $y=x$ line are preferred. (Bottom row) Trace plots of disaggregated and actual signals for one day.

Figure 5: Panel A shows the progression of the α^{-1} hyperparameter for Aggregate load and for one solar system over the course of one day. Panel B shows solar generation from one system predicted by the linear model, and disaggregated by the optimization. Panel C shows aggregate load predicted by the linear model, and disaggregated by the optimization.

Figure 6: CV (lower preferred) of solar disaggregation binned by the fraction time a home reverses power flow and injects power into the grid. Each home's amount of reverse power flow is defined by the load profiles and the power generation during the days used for training.

11am and 3pm for the standard conditions (Row 1 of Table 3). We find a 10th percentile, median, and 90th percentile of 14%, 36%, and 118% for MAPE calculated during all daylight hours; and 7%, 20%, and 91%, for MAPE calculated between 11am and 3pm, respectively.

6 STREAMING DISAGGREGATION

We approximate the streaming optimization with household level data instead of substation level data. (4) shows our approximation which uses the sum of individuals loads as an approximate for load measured at the substation. This effectively assumes that there are no line losses between consumer meters and the substation: $\tilde{Y} = \sum_{i=1}^N Y_i$. In the following section, we discuss how this assumption affects our analysis and means of overcoming it.

Comparing (4) to (15a), one can see that the individual net loads—measured by AMI—only influence one constraint: solar generation

Figure 7: Results from analysis of sensitivity to the number of homes in a region. Lower CV and RMSE are always preferred.

Figure 8: Results from analysis of sensitivity to number of readings in the training set. Reading are collected at 15 minutes resolution. Lower CV and RMSE are always preferred.

behind a meter cannot be greater than the net load observed at that meter. Thus, by running the optimization without AMI, we are only differing from the same optimization by this AMI constraint.

We evaluate the effects of this constraint on disaggregation by running the streaming optimization on the same data to which the parameters are fit, and comparing it to the results achieved in fitting. We then test the ability for the model to perform disaggregation on the same houses during a future time period.

Figure 9: Results from analysis of sensitivity to the number of proxies. Lower CV and RMSE are always preferred.

Figure 10: Performance of streaming problem compared to the training problem. Lower CV and RMSE are always preferred.

Figure 10 shows results for each scenario. We observe that, on average, the real time problem has little difference in performance than the training problem, regardless of whether it is run on historical or future data. We note that this result is likely influenced by the results of α tuning for this particular data set. As described in Section 2.1, the linear models fit to solar perform much better than the linear models fit to aggregate load. As a result, the α parameters for solar generation are relatively large, and the disaggregated signal is almost identical to the output of the linear model, which is not affected by the AMI constraint. Should the model be run in a case where the linear models for solar are not as accurate, one would expect greater deviation between the source signals generated by the streaming and training problems.

6.1 Accounting for System Losses

One means of approximating losses is using a linear or quadratic scaling of individual loads as perceived by the substation. A quadratic scaling is equivalent to assuming that loads only affect the real power consumption along the lines; and a linear scaling is a first order Taylor approximation of this. We do not implement disaggregation with losses in this paper, however this section discusses how losses could be included as another source to be separated in the streaming problem.

Eq. (11) presents an equation that uses linear scaling coefficients to relate the results of the training problem to the net load measured at the substation, $\bar{Y}$. Appendix B explicitly shows the approximations needed to reduce a quadratic loss model to this formulation. The parameters in (11), μ , can be found by fitting the aggregate load and solar signals—which result from the training optimization—to net load measurements at the substation, $\bar{Y}$. Notably this approximation will also have some error, $\epsilon_{z,t}$, reflecting losses due to reactive power, second order effects, line resistances that vary with climate, the aggregation of load, and unmonitored power draws.

$$Z_t = \mu_0 + \mu_L \bar{L}_t + \sum_{i=1}^N \mu_i S_{i,t} + \epsilon_{z,t} \quad (11)$$

When integrating this equation into the disaggregation problem, the losses' errors must be included. Leaving them out would force them to be included in the underlying solar of aggregate load signal, which is what we aim to recover.

(12) presents the streaming optimization with system losses, Z , included. Z is predicted by the aggregate load signal and the individual load signals, but not perfectly. It can absorb errors between load measured at household meters and load measured at the substation that otherwise must be absorbed by the aggregate load and solar signals. The weight, α_Z , determines how much of this error will be assigned to system losses, and it should reflect the inverse variance of model errors identified in (11).

$$\begin{aligned} \underset{\bar{L}, S, Z}{\text{minimize}} \quad & \|A_L(\bar{L} - X_L \theta_L)\|_2^2 + \sum_{i=1}^N \|A_i(S_i - \theta_i X_i)\|_2^2 + \quad (12) \\ & \|A_Z(Z - \mu_L \bar{L} - \sum_{i=1}^N \mu_i S_i - \mu_0)\|_2^2 \\ \text{subject to} \quad & \bar{L} + \sum_{i=1}^N S_i = \sum_{i=1}^N Y_i, \\ & S_i \leq 0 \quad \forall i. \end{aligned}$$

7 DISCUSSION

This paper presents a method for disaggregating the generation of behind-the-meter solar photovoltaic (PV) generation at individual buildings from both historical data and streaming data.

The historical problem relies on data from advanced metering infrastructure (AMI), and data from a proxy signal that is contemporaneously related to solar generation behind the meters. In this paper, we use generation from one or more nearby PV systems as solar proxies. On a set load and generation data from 52 homes within the Pecan Street dataset, we find that the historical method

is able to accurately predict which homes have solar in over 90% of cases. The historical problem is able to recover the 15-min resolution PV generation signals with root mean square errors between 20% and 50% of average daily PV generation. A sensitivity analysis shows the method to be robust to the number of buildings and time span of data used to fit. However including more than 3 solar proxies can cause false positive solar generation due to overfitting of solar proxies to unrelated noise in consumption data. Also, the method work better at homes that export electricity to the grid more often. We acknowledge that our work only focuses on a narrow geographical region, and it does not evaluate the robustness to different climate zones and solar generation conditions. We will expand our work to include different datasets [2] from different regions to address this drawback in the future.

Although the disaggregated signal is useful in distribution system operations, the real-world implications of the accuracy results obtained here requires further study of the proposed methods in a power systems operations simulation. We leave such simulations for future work. We believe that the proposed disaggregation methodology would prove useful in understanding sources of voltage problems due to over-generation in low-voltage distribution grids, and detecting systems that are offline in real-time.

A variety of previous research focusing on solar disaggregation makes different assumptions on data availability and there are no agreed upon performance evaluation metrics, or reference datasets. For example, some [4] assume that there is no solar proxy data available and cloud cover can be used for disaggregation, whereas others use data obtained from a pyrometer. We believe that our work can be extended to include a clear-sky model coupled with weather information for such cases. Furthermore, we believe that more work is needed to generate unified datasets and evaluation scenarios similar to [3].

We show that once the historical problem is fit to a set of homes, it can be applied in real time relying only on aggregate net load data from a substation, instead of net load measurements from individual buildings' AMI. We find that the streaming problem performs with the same accuracy as the historical data problem. Finally, we describe a methodology to incorporate distribution system line losses to our construction of the streaming disaggregation problem. We leave the performance assessment of line loss estimation to future work, since it requires additional power system simulations using Pecan Street data.

8 ACKNOWLEDGEMENTS

The authors acknowledge funding from the U.S. Department of Energy contract DE-AC02-76-SF00515 and from the the TomKat Center for Sustainable Energy at Stanford University.

REFERENCES

- [1] J Alonso-Montesinos, FJ Batlles, and JL Bosch. 2015. Beam, diffuse and global solar irradiance estimation with satellite imagery. *Energy Conversion and Management* 105 (2015), 1205–1212.
- [2] Sean Barker, Aditya Mishra, David Irwin, Emmanuel Cecchet, Prashant Shenoy, and Jeannie Albrecht. 2012. Smart*: An open data set and tools for enabling research in sustainable homes. *SustKDD, August* 111, 112 (2012), 108.
- [3] Nipun Batra, Jack Kelly, Oliver Parson, Haimonti Dutta, William Knottenbelt, Alex Rogers, Amarjeet Singh, and Mani Srivastava. 2014. NILMTK: an open source toolkit for non-intrusive load monitoring. In *Proceedings of the 5th international conference on Future energy systems*. ACM, 265–276.

- [4] Dong Chen and David Irwin. 2017. Sundance: Black-box behind-the-meter solar disaggregation. In *Proceedings of the Eighth International Conference on Future Energy Systems*. ACM, 45–55.
- [5] Michael A. Cohen and Duncan S. Callaway. 2015. Physical Effects of Distributed PV Generation on California's Distribution System. *arXiv:1506.06643 [physics]* (June 2015). <http://arxiv.org/abs/1506.06643> arXiv: 1506.06643.
- [6] Peter Gotseff, James Cale, Murali Baggu, David Narang, and Kevin Carroll. 2014. Accurate power prediction of spatially distributed PV systems using localized irradiance measurements. In *PES General Meeting| Conference & Exposition, 2014 IEEE*. IEEE, 1–5.
- [7] Christian A Gueymard. 2009. Direct and indirect uncertainties in the prediction of tilted irradiance for solar engineering applications. *Solar Energy* 83, 3 (2009), 432–444.
- [8] Christian A Gueymard. 2010. Progress in direct irradiance modeling and validation. In *Solar 2010 Conf., Phoenix, AZ, American Solar Energy Soc.*
- [9] IER. 2015. Hawaiians Learning About the Pitfalls of Solar Power. (September 2015). <http://instituteeforenergyresearch.org/analysis/hawaiians-learning-about-the-pitfalls-of-solar-power/> [Online; 11-September-2015].
- [10] Mohammad Jamaly, Juan L Bosch, and Jan Kleissl. 2013. Aggregate ramp rates of distributed photovoltaic systems in San Diego County. *Sustainable Energy, IEEE Transactions on* 4, 2 (2013), 519–526.
- [11] Emre C. Kara, Ciaran M. Roberts, Michaelangelo Tabone, Lilliana Alvarez, Duncan S. Callaway, and Emma M. Stewart. 2018. Disaggregating solar generation from feeder-level measurements. *Sustainable Energy, Grids and Networks* 13 (2018), 112 – 121. <https://doi.org/10.1016/j.segan.2017.11.001>
- [12] Emre Can Kara, Michaelangelo Tabone, Ciaran Roberts, Sila Kiliccote, and Emma M Stewart. 2016. 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, 259–260.
- [13] Johanna L Mathieu, Phillip N Price, Sila Kiliccote, and Mary Ann Piette. 2011. Quantifying changes in building electricity use, with application to demand response. *IEEE Transactions on Smart Grid* 2, 3 (2011), 507–518.
- [14] Felipe A Mejia and Jan Kleissl. 2013. Soiling losses for solar photovoltaic systems in California. *Solar Energy* 95 (2013), 357–363.
- [15] Rahul Mohan, Tim Cheng, Abhay Gupta, Vivek Garud, and Ye He. 2014. Solar energy disaggregation using whole-house consumption signals. In *NILM Workshop*.
- [16] Luis PÁrez-Lombard, JosÁl Ortiz, and Christine Pout. 2008. A review on buildings energy consumption information. *Energy and Buildings* 40, 3 (2008), 394 – 398. <https://doi.org/10.1016/j.enbuild.2007.03.007>
- [17] Emerson Reiter, Kristen Ardani, Robert Margolis, and Ryan Edge. 2015. *Industry Perspectives on Advanced Inverters for US Solar Photovoltaic Systems: Grid Benefits, Deployment Challenges, and Emerging Solutions*. Technical Report. NREL (National Renewable Energy Laboratory (NREL)).
- [18] H. Shaker, H. Zareipour, and D. Wood. 2016. Estimating Power Generation of Invisible Solar Sites Using Publicly Available Data. *IEEE Transactions on Smart Grid* 7, 5 (Sept 2016), 2456–2465. <https://doi.org/10.1109/TSG.2016.2533164>
- [19] E. Wesoff. 2014. How Much Solar Can HECO and Oahu Grid Really Handle? (February 2014). <https://www.greentechmedia.com/articles/read/How-Much-Solar-Can-HECO-and-Oahu-Grid-Really-Handle> [Online; 10-February-2014].
- [20] Matt Wytock and J Zico Kolter. 2013. Contextually supervised source separation with application to energy disaggregation. *arXiv preprint arXiv:1312.5023* (2013).

A REDUCTION OF TRAINING PROBLEM

Fitting individual load signals as decision variables in (1) is a redundant process because they are not included in the objective function and they are fully defined by the first constraint which states $L_{i,t} = NL_{i,t} - S_{i,t}$. In this appendix we derive an equivalent optimization which replaces the individual loads, $L_{i,t}$, with aggregate load AL_t .

(13) combines the first and third constraints to define a new one which states that solar generation behind each meter must be less than the net load measured at that meter.

$$L_i = Y_i - S_i \geq 0 \implies S_i \leq Y_i \quad (13)$$

(14) redefines the fourth constraint in terms of solar and net load. The result constrains the sum of all solar generation and aggregate load to equal the sum of net loads measured at AMI.

$$\sum_{i=1}^N L_i = \bar{L} \implies \bar{L} \sum_{i=1}^N S_i = \sum_{i=1}^N Y_i \quad (14)$$

After rewriting the third and fourth constraints, $L_{i,t}$ only remains in the first. Because $L_{i,t}$ isn't included in the objective function, its value is arbitrary and thus it can take any value that satisfies the first constraint without penalty. Thus, we can remove this constraint and the decision variables from the optimization. (15a) shows the result. $L_{i,t}$ can be determined following this optimization by evaluating $L_{i,t} = NL_{i,t} - S_{i,t}$.

$$\underset{L, S, \Theta}{\text{minimize}} \quad \|A_{\bar{L}}(\bar{L} - X_{\bar{L}}\theta_{\bar{L}})\|_2^2 + \sum_{i=1}^N \|A_i(S_i - \theta_i X_i)\|_2^2 \quad (15a)$$

$$\text{subject to} \quad \bar{L} + \sum_{i=1}^N S_i = \sum_{i=1}^N Y_i, \quad \forall i \quad (15b)$$

$$S_i \leq 0, \quad \forall i \quad (15c)$$

$$S_i \leq Y_i, \quad \forall i \quad (15d)$$

$$\theta_i \geq 0, \quad \forall i \quad (15e)$$

B LOSSES

In a direct current power system, losses along lines are defined by the square of current, which in turn is defined by power consumption at all all downstream nodes. (16)-(17) shows this derivation for those who are not otherwise familiar; V_i is the voltage at node i which we assume not to vary in time, and R_{ℓ} is the total resistance of line segment ℓ , and $D(\ell)$ is the set of nodes downstream of line segment ℓ .

$$NL_{i,t} = I_{i,t} V_i \quad (16)$$

$$Z_t(NL_t) = \sum_{\ell \in L} R_{\ell} \sum_{i \in D(\ell)} \frac{NL_{i,t}^2}{V_i} \quad (17)$$

(18) gives a first order taylor approximation of (17), where all errors of this approximation are captured in the error term $\epsilon_{Z,t}$. (20) replaces the terms that have physical significance with ambiguous variables that will be fit by a regression without the physical model. (21) breaks the net load reading into components of solar and load. (22) applies an assumption that $\sum_{i=1}^N \mu_i L_{i,t} = \mu_{AL} \sum_{i=1}^N L_{i,t}$, the errors of which may also be included in the disturbance term.

$$Z_t(NL_t) = Z_{i,t}(\bar{NL}) + \sum_{i=1}^N \frac{\partial Z_t}{\partial NL_{i,t}} (NL_{i,t} - \bar{NL}_i) + \epsilon_{z,t} \quad (18)$$

$$= \mu_0 + \sum_{i=1}^N \mu_i NL_{i,t} + \epsilon_{z,t} \quad (20)$$

$$= \mu_0 + \sum_{i=1}^N \mu_i S_{i,t} + \sum_{i=1}^N \mu_i L_{i,t} + \epsilon_{z,t} \quad (21)$$

$$\approx \mu_0 + \sum_{i=1}^N \mu_i S_{i,t} + \mu_{AL} AL_t + \epsilon_{z,t} \quad (22)$$