Disaggregation of Household Solar Energy Generation Using Censored Smart Meter Data

Joe Brown*, Alessandro Abate, Alex Rogers

firstname.lastname@cs.ox.ac.uk
Department of Computer Science, University of Oxford
**Corresponding Author*

Abstract

Quantifying small scale domestic solar (PV) generation from energy consumption is becoming increasingly important as the install base of small solar (PV) panels rapidly grows. Unfortunately, it is often the case that the only insight into the consumption and generation of energy within a house comes from smart-meter readings. The smart meter records the amount of energy the house takes from the grid, and does not independently measure and report the local generation that might be consumed by the home, or fed back to the grid. To address this issue, we propose a novel approach to disaggregate solar (PV) generation from energy consumption that also infers installed PV capacity. This is done by disaggregating PV generation from censored smart meter readings, and specifically by finding the most likely distribution for the energy consumption and using it to infer the solar generation. We extend this approach to propose the first technique to infer PV capacity without weather data or a solar proxy, using instead only smart meter readings given a group of houses in close proximity. We evaluate the algorithm on two datasets: (i) the US Pecan Street dataset is adapted so that net energy meter readings are censored; and (ii) a constructed dataset, combining smart meter readings from UK households and solar energy generation from locations across the UK. Our results show comparable accuracy at inferring PV capacity compared to existing approaches, which cannot deal with censored readings which represent over 50% of PV panel installations in the UK.

Keywords: Solar Energy, Smart Meters, Data Disaggregation, Residential Sector, Energy Generation

1. Introduction

Solar energy (PV) generation in the UK has increased by a factor of 130 between December 2010 and December 2019, with small installations (under 10 kW) increasing in number by 4000%[1], making up a growing proportion of the grids electricity supply. Thus, grid operator who must balance the supply and demand of electricity in real time in order to keep the grid at a stable frequency [1, 2, 3], have an increasing challenge to estimate the quantity of energy provided by these PV installations. Ideally, the location of these panels, and their capacity would be reported to the grid operator, and their generation would be measured and reported in real-time through smart meters. Unfortunately, in the UK, no reporting scheme currently operates, and the deployed base of smart meters (and those scheduled to be deployed to all UK homes in the future) do not separate local generation and consumption. Rather, they simply record the net energy taken from the grid. More significantly, they also fail to report net export to the grid, and simply report 0 kWh of consumption over each half hour, during these periods. In statistical terms, these observations are said to be censored. Thus, there is a need to develop approaches to identify homes with PV panels installed, and to estimate their real-time net contribution to overall grid generation, by disaggregating domestic energy generation using existing censored smart meter data.

In the setting where smart meters are uncensored, existing approaches often combine known physical models of PV panels with an analytical approach to identify the maximum generation of a PV panel to disaggregate the energy generated [4, 5]. Others take a supervised learning approach to the problem, using radial basis functions and wavelet kernel support vector machine that map weather metrics to a solar output [6]. Many approaches do not focus on disaggregation, but only on inferring PV generation using its relationship with weather [7, 8]. Alternative approaches have implemented solar disaggregation at a feeder or small-region level, which does not correspond directly to our work [9, 10, 11].

[1] https://www.gov.uk/government/statistics/solar-photovoltaics-deployment (Accessed 16/04/20)

The above approaches do not explicitly account for censored smart meter readings. This means that they would not work in over 50% of real world settings where censoring occurs due to incorrect installations[2]. As the number of PV panel installations grow, this creates a significant issue in balancing the grid. A new approach is required to deal with censored smart meter readings, otherwise PV panels will go undetected and their generation unaccounted for. Without a non-intrusive approach, costly interventions need to be taken, such as re-fitting smart meters or installing new hardware, to measure PV generation separately.

In this paper, an approach to disaggregate PV generation and energy consumption from censored smart meter readings is presented. The approach infers the maximum power input into the house that can be expected from the specific system of PV panels, this means the inverter efficiency is also accounted for. Then using a solar proxy, which is defined as the known solar generation from a local house, the PV generation can be inferred. Using a solar proxy to infer solar generation of a different house has been successfully demonstrated in other bodies of research [5]. To find the PV capacity, the most likely joint probability distribution of the PV capacity and the energy consumption for each time period across a year is found. Figure 1 provides a visualisation of this process.

Our novel approach allows solar disaggregation to be performed in real-world situations where smart meter readings are censored. Furthermore, in line with current smart meter standards, only half-hourly readings are required for the algorithm to work. The algorithm can classify the presence of PV panels on small buildings and infer the solar generation at each time of the day. The algorithm is evaluated on the widely-used US Pecan Street dataset, and to demonstrate the viability of the approach in the UK setting a dataset is constructed combining a dataset of UK smart-meter readings and a dataset of energy generation recordings from small PV systems. Combining the two datasets is required as there are multiple publicly available datasets of small houses with labelled energy consumption, however there are currently no large scale datasets with local energy generation also recorded. Our experiments demonstrate the algorithms success at detecting houses with solar panels

[2] https://www.utilityweek.co.uk/sta-warns-smart-meters-solar-panels-decoupled (Accessed 16/04/20)

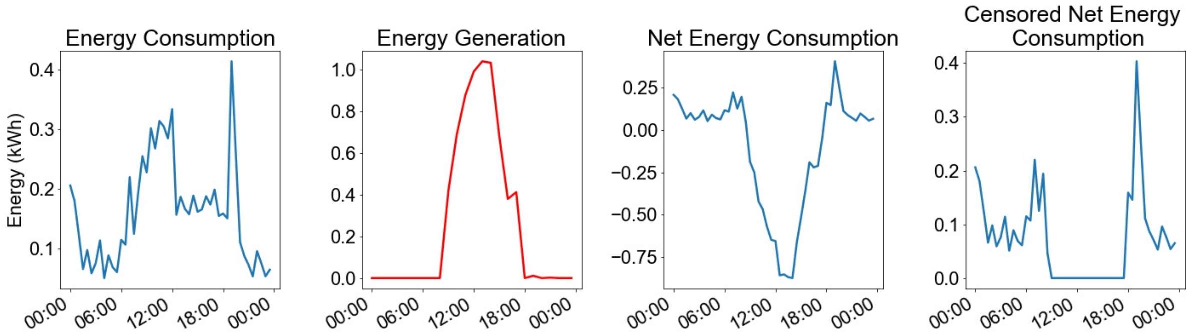

Figure 1: The leftmost plot shows the energy consumption for a random day from the Pecan Street dataset, whilst the second plot shows the corresponding energy generation for the day. The third plot displays the net meter reading, which most existing approaches to solar disaggregation are designed to use. The rightmost plot shows the corresponding readings of a censored smart meter - notice here the regions where net smart metering is censored, namely kept at zero when negative.

from censored smart meter readings and inferring their PV capacity and PV generation. This is the first approach to explicitly deal with censored smart meters and the results are comparable to other approaches, which do not address the issue of censored readings and assume net-metering. This approach will allow inference of PV capacity and generation from all energy generating customers in the UK, as the current methods rely on reporting to the feed-in tariff, inference from quarterly energy generation and national grid models which do not explicitly account for generation by individual homes.

The rest of the paper is structured as follows. Section 3 outlines the model to infer PV capacity from censored smart meter readings. Section 4 illustrates our implementation of an algorithm to infer PV capacity, which is extended in Section 5 to not require solar irradiance. Finally, Section 6 introduces the datasets, Section 7 outlines the experiments and the results and Section 8 summarises our contribution to the field and outlines future work.

2. Related Work

Recent work proposes a method for disaggregating solar PV generation behind-the-meter for individual buildings using historical advanced metering infrastructure, which records the net energy consumption for a house, feeder level net energy consumption, and a solar proxy (similar to what is used in this paper), however the approach still requires net meter data, which means that in its current form it would not work when the smart meter data is censored. [5]. Other work has looked at identifying PV generation on a larger geographic basis, estimating the electric generation from "invisible solar PV resources" by identifying a small number of solar PV sites and using it to estimate a larger set of sites via data dimension reduction and clustering techniques [12, 13]. The total capacity of the PV sites must be known a-priori and this approach does not work on a household-specific basis like our proposed approach, but only for regions.

To disaggregate solar energy generation from smart-meter readings a number of machine learning techniques have been proposed. One such approach is to use support vector machine models with different kernel functions to disaggregate the solar energy generation from smart meter readings [6, 14]. However, the approach requires procuring data at a granularity that is much finer than that available from a regular smart meter, as it uses dedicated meter hardware. Other approaches to disaggregate energy generation rely on physical models to infer solar irradiance and then use a physical model to map the solar irradiance to the predicted PV energy generation [4]. This approach works by predicting the base load energy used by a house and also finding the times in the year when there is no reduction in solar irradiance due to weather. By combining these two sources the solar generation of the house can be predicted. Deep learning approaches to real-time observability of PV generation behind the meter have also shown success [15]. However, none of the above approaches have been shown to deal with situations where smart meter readings are censored, which is the focus of this work.

Physical models can fail in real world settings, due to the models not correctly representing the true complexity of the actual system being described and due to the difficulty in forecasting solar irradiance that is incident on the ground [16]. Recent success has been achieved via data-driven or hybrid approaches, where the observed relationships between the actual solar irradiance reaching the ground and weather factors are used to train the models. This relinquishes the requirement on the physical models to accurately reflect the real world.

This paper provides a benchmark for common machine learning techniques applied to the problem of forecasting solar irradiance and guides towards which weather features are useful inputs for a predictive model [7]. Alternatively, neural networks have been used to predict solar irradiance incident on the ground [17, 18]. Intuitively we can attribute their success to the ability of deep learning to consider a large set of weak features (i.e. weather factors) as there is a non-trivial correlation between different weather factors affecting solar irradiance. Other approaches show the strong correlation that solar irradiance has in time, however these methods are limited as they are unable to forecast beyond the next 30 minutes [19].

Satellite and aerial imagery has been utilized in recent work to detect solar panels and to construct a dataset of the actual deployment of solar panels [20, 21, 22, 23]. Advances in deep learning have made it possible for image recognition techniques, using convolutional neural networks, to be deployed on large-scale tasks [24, 25]. One such deep learning framework, DeepSolar, finds the GPS location of PV panels and the size of the installation from 30cm-resolution satellite imagery data [20]. The framework has been used to create a dataset of PV panel locations around the US, locating 1.47m panels with a precision of 93.1% in residential areas. Extensions to this research have culminated in a predictive model that estimates the solar deployment density based on census data. The use of satellite imagery allows for a scalable and non-intrusive approach to detect the location of solar panels. However high-quality satellite data can be expensive, making this approach prohibitive in some applications. Whilst this approach can provide estimates of the solar PV generation based on panel size and location, there is no way to infer the real amount being generated, and it can only provide an upper bound on predicted generation. Other research has also shown the benefits of using satellite imagery to improve nowcasting from power data demonstrating the feasibility to use satellite imagery techniques in parallel with other approaches [26]. As such this is a complementary field of research and can be used in tandem with our proposed approach to validate results (cf. Future Work).

3. Censored Solar Disaggregation Model

We consider a dataset of smart-meter readings, r_{htd} [kWh], where $h \in [1...H]$, $t \in [1...T]$, $d \in [1...N]$, such that H denotes the number of households, T is the number of time steps the data readings are separated into, and N is the number of days of data. A smart meter records only the energy supplied to the house from the grid. This means if behind-the-meter energy generation, g_{htd} [kWh], is larger than the energy consumption, x_{htd} [kWh], there is a censored reading, $r_{htd} = \max(0, x_{htd} - g_{htd})$.

3.1. Energy Generation

The PV panel is assumed to be the only source of possible energy generation in the household. A PV panel generates energy according to $g_{htd} = \tau c_h f_{td}$, where τ [hours] is the constant time step size, c_h [kW] is the PV capacity, namely the maximum power input into the home from the PV panel. In particular, $c_h = 0$ kW implies that there is no PV panel present. In this paper we use a solar proxy as an input to the algorithm. The solar proxy factor, $0 \leq f_{td} \leq 1$, is the proportion of the solar proxy recorded divided by the maximum recorded solar proxy for the year. The accuracy with which the solar proxy correctly represents the solar energy generation of the house in question will improve as the scale of smart meter installations increases, as it will reduce the distance to the nearest PV panel.

3.2. Energy Consumption

It is assumed that the energy consumption, x_{htd}, for a given time step (e.g., 10:00-10:30, with $\tau = 0.5$ h) comes from a probability distribution, P, parameterised by the set Θ_{ht}, valid for all days, d, as:

$$X_{htd} \sim \mathrm{P}(\Theta_{ht}).$$

To select the probability distribution that best represents the energy consumption of residential households, multiple distributions that can model asymmetric data have been empirically evaluated. From the evaluation the gamma distribution has been selected as the probability distribution that best model energy consumption for a given time step in a residential household. It has been chosen as it produced the maximum likelihood, and equivalently, minimized the Kullback–Leibler divergence with the energy consumption data across all houses tested. Regions with different energy consumption patterns may benefit from re-evaluating which distribution provides the best fit, however this was the best fit for the houses we tested from the UK and the US. Mixture distributions may show improved results, however in order to keep the run time as low as possible, distributions with less parameters to infer have been preferred.

An indicator function is defined as $\mathbb{I}_{htd} = 1$ if $r_{htd} > 0$, and equal to 0 if censored for the respective time step, day, and household. A point is defined as censored if $r_{htd} = 0$ kWh. For the uncensored observations we have that:

$$x_{htd} = r_{htd} + \tau c_h f_{td}, \tag{1}$$

whereas in the censored case we know that the unknown energy consumption, x_{htd}, is bounded above by the energy generated, as:

$$x_{htd} \leq \tau c_h f_{td}. \tag{2}$$

In order to succinctly describe quantities over whole populations, we define sets of data corresponding to each house and time step. We introduce sets comprising values of indicator functions, energy consumption values, smart meter readings and solar proxy factors for a time step, as: $\mathbb{I}_{ht} = \{\mathbb{I}_{htd}, d \in [1...N]\}$, $\boldsymbol{X}_{ht} = \{x_{htd}, d \in [1...N]\}$, $\boldsymbol{R}_{ht} = \{r_{htd}, d \in [1...N]\}$, and $\boldsymbol{F}_t = \{f_{td}, d \in [1...N]\}$, respectively. We similarly define the set of alpha and beta parameters for the gamma distribution, PV capacity, indicator function values, energy consumption values and smart-meter readings for a specific time step and day across houses, as follows: $\boldsymbol{\alpha}_t = \{\alpha_{ht}, h \in [1...H]\}$, $\boldsymbol{\beta}_t = \{\beta_{ht}, h \in [1...H]\}$, $\boldsymbol{C} = \{c_h, h \in [1...H]\}$, $\mathbb{I}_{td} = \{\mathbb{I}_{htd}, h \in [1...H]\}$, $\boldsymbol{X}_{td} = \{x_{htd}, h \in [1...H]\}$ and $\boldsymbol{R}_{td} = \{r_{htd}, h \in [1...H]\}$.

4. PV Capacity Inference with Solar Proxy

We present a novel algorithm to infer the PV capacity, c_h, using a smart meter with censored readings, r_{htd}. For a given house, our algorithm aims to find the maximum power input into the home that can be expected from the installed PV panels. Notice that there is a counter-intuitive nature about the algorithm, as the PV capacity, which is our main point of interest, is found in the process of estimating the most likely distributions for energy consumption, X_{htd}.

Our algorithm finds the most likely gamma distribution representing the energy consumption for each time slot. If we first consider dealing with uncensored smart-meter readings, namely when all the smart-meter readings are known, then by finding the most likely energy consumption, we also find the most likely PV generation, g_{htd}, via Equation (1). This approach is extended to when the smart-meter readings are censored and we do not have a direct relation between the energy consumption and PV generation, the only relation we know is the inequality seen in Equation (2).

We work with half-hourly smart meter data ($\tau = 0.5$ h) and only use the time steps when the sun is shining, since a solar proxy factor equal to 0 does not help with the task of inferring PV generation. Our algorithm is trained using 365 days of half-hourly readings from historic smart meters and recorded solar irradiance. It is worth reiterating that whilst the smart meter readings are known, the energy consumption values are unknown and are being inferred. Once the PV capacity, c_h, is known, it can be used to calculate the PV energy generation for any time step with the solar proxy, f_{htd}. Also, note that the inferred value of the PV generation is constrained as we are inferring a scalar value (the PV capacity), which is shared across all time steps for the house, assuming the PV capacity remains constant, and this value is multiplied by the solar proxy, f_{htd}, for that specific time step to give the energy generation.

In this work the energy consumption is assumed to follow a gamma distribution across days for each half-hour period in a specific house. Other distributions, such as the log-normal distribution, lend themselves to analytical solutions and may prove to be better assumptions when working with different datasets where energy consumption patterns are different. The unknown parameters of the gamma distribution, denoted as α_{ht} and β_{ht}, are found by maximum likelihood estimation. When the true value in some cases is not observed due to the smart meter censoring readings below 0 kWh, a specific likelihood function can be used, which deals with data that is censored below [27]. To find both the PV capacity and the parameters of the gamma distribution representing the energy consumption, these parameters are iteratively updated in turn to maximise the likelihood.

4.1. Log-Likelihood of Censored Smart Meter Readings

A standard likelihood function is the product of probability density functions (PDF) for all the given data points. The PDF at a specific value corresponds to the probability that a data point with that value is observed. When the smart meter reading is 0 kWh, the energy consumption is unknown and we do not have a data point to calculate the corresponding PDF. However, we know that for the censored data points the energy generated by the PV panel is an upper bound on the energy consumption. Hence, the cumulative distribution function (CDF) of the upper bound is used when the smart meter reading is censored, as it describes the probability that the energy consumption value is less than the upper bound [27]. In summary, the proposed model uses the CDF when the smart meter reading is censored and the PDF when it is uncensored. The PDF, ϕ, and CDF, Φ, for the gamma distribution are defined respectively as follows where the gamma function is represented as Γ, whereas γ denotes the lower gamma function:

$$\phi(x|\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x),$$

$$\Phi(x|\alpha,\beta) = \frac{\gamma(\alpha,\beta x)}{\Gamma(\alpha)}.$$

For a specific house and time, we infer the gamma distribution parameters, α_{ht} and β_{ht}, parameterising the set of energy consumption values, $\boldsymbol{X}_{ht}$, given the calculated set of indicator functions $\mathbb{I}_{ht}$, the smart meter readings $\boldsymbol{R}_{ht}$, and the solar proxy readings $\boldsymbol{F}_t$. We can then find the likelihood of the energy consumption data by factorising across the days as follows:

$$\mathcal{L}(\alpha_{ht},\beta_{ht},c_h|\boldsymbol{R}_{ht},\mathbb{I}_{ht},\boldsymbol{F}_t,\tau) = \\ \prod_{d=1}^{N} \phi(x_{htd}|\alpha_{ht},\beta_{ht})^{\mathbb{I}_{htd}} \Phi(\tau c_h f_{td}|\alpha_{ht},\beta_{ht})^{(1-\mathbb{I}_{htd})}. \tag{3}$$

Note that we have the PDF of $x_{htd} = r_{htd} + \tau c_h f_{td}$ (as per Equation (1)) when the smart-meter reading is uncensored and the CDF of predicted PV generation when censored. By maximising this likelihood function the most likely gamma parameters representing half-hourly energy consumption, α_{ht} and β_{ht}, and PV capacity, c_h, are found. The likelihood function is factorised across all days (N) in the year.

As the log function is monotonically increasing, finding the maximum of the likelihood is equivalent to finding the maximum of the following log-likelihood expression:

$$\begin{aligned}\log \mathcal{L}(\alpha_{ht}, \beta_{ht}, c_h | \boldsymbol{R}_{ht}, \mathbb{I}_{ht}, \boldsymbol{F}_t, \tau) = \\ \sum_{d=1}^{N} \mathbb{I}_{htd} \log(\phi(x_{htd}|\alpha_{ht}, \beta_{ht})) \\ + \sum_{d=1}^{N} (1 - \mathbb{I}_{htd}) \log(\Phi(\tau c_h f_{td}|\alpha_{ht}, \beta_{ht})).\end{aligned} \tag{4}$$

The golden section search is a standard technique used to find the extrema of a function. Since the log-likelihood of the gamma distribution with respect to each variable is uni-modal (see Figure 2), we use this technique to find the parameters that maximise the log-likelihood, with the other parameters being fixed. The benefit of using the golden section search is that we do not need to revise the approach if we change the distribution, as we only need to be able to calculate the PDF and CDF for that distribution. The golden section search finds the parameters of the gamma distribution, α_{ht} and β_{ht}, maximising the likelihood as follows:

$$(\alpha^*_{ht}, \beta^*_{ht}) = \underset{\alpha_{ht}, \beta_{ht}}{\operatorname{argmax}} \log \mathcal{L}(\alpha_{ht}, \beta_{ht}, c_h | \boldsymbol{R}_{ht}, \mathbb{I}_{ht}, \boldsymbol{F}_t, \tau). \tag{5}$$

Furthermore, also using the golden section search the PV capacity can be found since it is independent of the time of the day, we find the value of c_h that maximises the likelihood across all gamma distributions, corresponding to single time intervals of the day, namely:

$$c^*_h = \underset{c_h}{\operatorname{argmax}} \sum_{t=1}^{T} \log \mathcal{L}(\alpha_{ht}, \beta_{ht}, c_h | \boldsymbol{R}_{ht}, \mathbb{I}_{ht}, \boldsymbol{F}_t, \tau). \tag{6}$$

Algorithm 1: Iterative golden section search to find PV panel capacity via likelihood maximisation

```
Result: Returns $c_h$
input: $\tau,\ r_{htd}\ \forall\ h \in [1, ..., H],\ t \in [1, ..., T],\ d \in [1, ..., N],\ f_{td}$ (if known) $\forall\ t \in [1, ..., T],\ d \in [1, ..., N]$
$c_h = 1\ \forall\ h$
$f_{td} = 1\ \forall\ t, d$
for $h$ in 1:H; $t$ in 1:T; $d$ in 1:N do
    if $r_{htd} > \tau c_h f_{td}$ then
        $\mathbb{I}_{htd} = 1$
    else
        $\mathbb{I}_{htd} = 0$
    end
end
$\mathcal{LL} = \sum_{t=1}^{T} \log \mathcal{L}(\alpha_{ht}, \beta_{ht}, c_h | \boldsymbol{R}_{ht}, \mathbb{I}_{ht}, \boldsymbol{F}_t, \tau)$
while $\mathcal{LL}$ is not converged do
    for $h$ in 1:H; $t$ in 1:T do
        $(\alpha^*_{ht}, \beta^*_{ht}) = \underset{\alpha_{ht}, \beta_{ht}}{\operatorname{argmax}} \log \mathcal{L}(\alpha_{ht}, \beta_{ht}, c_h | \boldsymbol{R}_{ht}, \mathbb{I}_{ht}, \boldsymbol{F}_t, \tau)$
        $(\alpha_{ht}, \beta_{ht}) \leftarrow (\alpha^*_{ht}, \beta^*_{ht})$
    end
    for $h$ in 1:H do
        $c^*_h = \underset{c_h}{\operatorname{argmax}} \sum_{t=1}^{T} \log \mathcal{L}(\alpha_{ht}, \beta_{ht}, c_h | \boldsymbol{R}_{ht}, \mathbb{I}_{ht}, \boldsymbol{F}_t, \tau)$
        $c_h \leftarrow c^*_h$
    end
    for $t$ in 1:T; $d$ in 1:D do                    // Ignore if solar irradiance is known
        $f^*_{td} = \underset{f_{td}}{\operatorname{argmax}} \log \mathcal{L}_{\text{irr}}(f_{td} | \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t, \boldsymbol{C}, \boldsymbol{R}_{td}, \mathbb{I}_{td}, \tau)$
        $f_{td} \leftarrow f^*_{td}$
    end
    $\mathcal{LL} = \sum_{t=1}^{T} \log \mathcal{L}(\alpha_{ht}, \beta_{ht}, c_h | \boldsymbol{R}_{ht}, \mathbb{I}_{ht}, \boldsymbol{F}_t, \tau)$
end
```

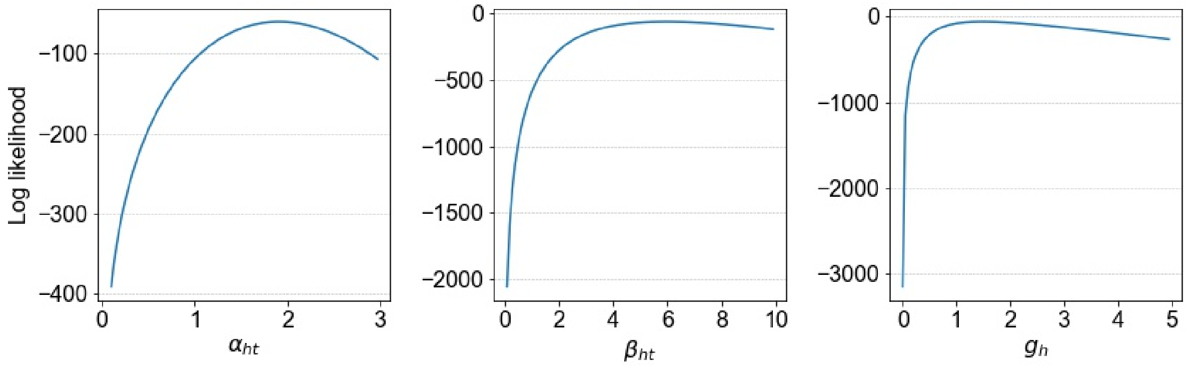

Figure 2: The log-likelihood function is uni-modal with respect to α_{ht}, β_{ht} and c_h, with the other parameters kept fixed.

The log-likelihood is maximised by finding the α_{ht} and β_{ht} parameters for each gamma distribution that best represent the energy consumption at each time step. Then for each house we find the value of the PV capacity, c_h, that maximises the likelihood of the inferred energy consumption values being from their respective gamma distributions. This is then repeated until the log-likelihood converges (see Algorithm 1, set h = 1 and follow the instructional comments).

5. PV Capacity Inference with Only Clearsky Solar Irradiance

We extend the above approach to the inference of PV capacity only using clearsky irradiance (as opposed to a solar proxy for each panel), under the assumption we have a group of houses in close proximity (e.g., sharing the same postcode). The clearsky (solar) irradiance, s_{td}, is the solar irradiance that would be incident on the ground if there was no atmospheric or weather interference. It is calculated using existing physical models, as it only depends on the location on Earth relative to the Sun. To identify how much of the solar irradiance is actually incident on the ground, the clearsky solar irradiance is multiplied by the clearsky irradiance factor, $f_{td} \leq s_{td} \leq 1$, defined as the solar irradiance at a particular time compared to the maximum recorded solar irradiance of the year.

Assuming all houses receive the same solar irradiance, an additional step is added in the previous algorithm to find an estimate of the solar proxy factor, namely f_{td}, for each time step across all houses, so that we maximise the likelihood that the inferred energy consumption values come from their corresponding distributions. Consider the new likelihood function which factorises the probability distribution across houses and is no longer conditioned on f_{td} as it is now being inferred:

$$\begin{aligned}&\mathcal{L}_{\text{irr}}(f_{td}|\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t, \boldsymbol{C}, \boldsymbol{R}_{td}, \mathbb{I}_{td}, \tau) = \\ &\quad \prod_{h=1}^{H} \phi(x_{htd}|\alpha_{ht}, \beta_{ht})^{\mathbb{I}_{htd}} \Phi(\tau c_h f_{td}|\alpha_{ht}, \beta_{ht})^{(1-\mathbb{I}_{htd})}.\end{aligned} \tag{7}$$

This likelihood function is factorised over houses (H) in close proximity, as opposed to Equation (3) that factorises over days (N) for a single house. To infer the solar proxy, the likelihood is conditioned on the predicted gamma distribution parameters. This is due to the close houses sharing the same solar proxy, f_{htd}, but having unique PV capacity, c_h, and unique gamma distribution parameters representing energy consumption, α_{ht} and β_{ht}. The value of f_{td} maximising this likelihood is given by:

$$f_{td}^* = \underset{f_{td}}{\operatorname{argmax}} \log \mathcal{L}_{\text{irr}}(f_{td}|\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t, \boldsymbol{C}, \boldsymbol{R}_{td}, \mathbb{I}_{td}, \tau). \tag{8}$$

The approach finds the solar proxy factor that gives either the most likely energy consumption, or an upper bound on the energy consumption in the case of censoring. As illustrated in Algorithm 1, this becomes an extra step in the algorithm as the most likely gamma distribution parameters for each house, α_{ht} and β_{ht}, the most likely PV capacity for each house, c_h and the most likely solar proxy values, f_{td}, are inferred iteratively and until convergence.

Notice that this approach is realistically not feasible over a single house, as the inferred solar proxy factor would likely overfit, being dependent on a single data point. On the contrary, finding the solar proxy factor across a batch of houses (assumed to be adjacent) regularises the most likely solar proxy factor and allows for a prediction of its true value (see Algorithm 1). Let us emphasise that this is not intended to be an improvement of the above algorithm, but rather an approach that works with more limited data and an exploratory result towards alternative methods to using weather to predict solar irradiance for inferring PV generation.

6. Datasets

To evaluate the approach a subset of the US Pecan Street dataset and a constructed UK dataset are used.

The Pecan Street dataset provides energy data for houses with PV panels [28]. We have cleaned the data by creating censored smart meter readings for each house, as obtained from their energy consumption and generation. From this dataset, 30 houses have been identified to have solar panels and selected and the PV Capacity is in the range 2.5 kW to 10.2 kW for the Pecan Street dataset.

Alongside this, we have constructed a dataset representing smart meter readings from the UK, comprising solar energy generation from locations across the UK from the Sheffield Solar microgen dataset [29], and smart meter readings from London households[3]. Curation of this dataset is required as there are no large scale public UK datasets that record energy usage, PV energy generation and local solar irradiance. To create the dataset a selection of smart meter readings for 260 houses are taken from the London smart meter dataset and treated as energy consumption. A manual check was conducted to ensure that there is no PV energy generation on the selected houses. Each house has been randomly assigned one of the 50 available PV generation profiles from the microgen dataset. The maximum power output for each PV system in the microgen dataset has been re-scaled to demonstrate how the algorithm works across a range of PV capacity values in the range 0.5 kW - 6.5 kW with 20 values drawn randomly in each 0.5kW band, additionally 20 houses were included with no PV panel (0 kW). Then, the half-hourly smart meter readings for the combined dataset are calculated using the formula $r_{htd} = \max(0, x_{htd} - \tau f_{td} c_h)$. Only the daylight hours of each day have been used. As the assignment of the PV panels to the smart meters is random this can be considered as a worst-case scenario as there is typically correlation between the energy usage of a home and the PV capacity.

The solar proxy has been implemented as seen in previous papers by taking the known proportion of the PV energy generation through the day [5]. There is potential for transpo-

[3] data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households (Accessed 16/04/20)

sition errors, where there is a difference between the plane of incidence for the solar proxy to the PV system that is being evaluated. However, it has been shown that the geometry of the proxy system can be recovered and mapped to the geometry of the PV system being evaluated removing this source of error [4]. Furthermore, it is shown that the factors other than orientation only affect the maximum power output, hence they can be absorbed into the PV capacity term which represents the true power input to the house from the PV system.

The clearsky solar irradiance is calculated using existing techniques such as the Bird model (as seen in the package PVLIB) [30, 31].

7. Empirical Evaluation

To evaluate the performance of the algorithm four metrics are used. The root mean square error (RMSE) measures the accuracy of the inferred values by taking its distance from the true value, regardless of the true PV capacity. RMSE is a useful metric in a situation when we are not interested in the contributions of individual houses and instead focus on the total prediction of PV generation,

$$\text{RMSE} = \sqrt{\frac{\sum_{h=1}^{H}(\hat{c}_h - c_h)^2}{H}}.$$

Note that, $\hat{c}_h$ denotes the inferred PV capacity and c_h is the real PV capacity. To measure the precision we also calculate the mean absolute percentage error (MAPE). The MAPE calculates the absolute error as a percentage of the real PV capacity,

$$\text{MAPE} = \frac{1}{H}\sum_{h=1}^{H}\left|\frac{\hat{c}_h - c_h}{c_h}\right|.$$

Furthermore, to identify the bias in the estimations the mean normalised bias error (MNBE) is calculated,

$$\text{MNBE} = \frac{1}{H}\sum_{h=1}^{H}\frac{\hat{c}_h - c_h}{c_h}.$$

We also measure the classification rate, which is the number of houses correctly predicted to have a solar panel. In order to classify the presence of a PV panel, we have selected a PV

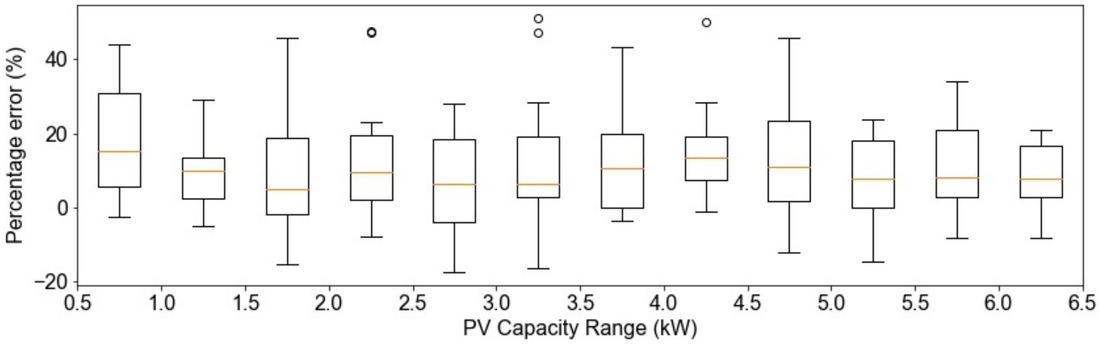

Figure 3: Each boxplot represents the distribution of percentage error for the 20 houses in each 0.5 kW band of PV capacity. The performance of the approach appears to be independent of the PV capacity, however there is a bias and the approach consistently over-estimates the PV capacity which the approach can be adjusted for.

capacity threshold, and assumed that if the inferred generation is above the given threshold, then a PV panel is present.

7.1. PV Capacity Inference with Solar Proxy

For our constructed UK dataset we have selected 260 houses from the smart meter readings from London households dataset that do not have a PV panel installed, and have treated the readings as energy consumption. They are then combined with PV generation as described in Section 6 to create the dataset used for the experiments. For the 30 houses with from the US Pecan Street dataset we have run the algorithm twice: once with the censored smart meter readings as described above, and once with the solar generation removed to emulate if the houses did not have solar panels.

7.1.1. Inference of PV Capacity

The first key takeaway from Table 1 is the 100% classification rate, meaning that the algorithm correctly identified every incident where there was a PV panel present and every incident where there was not. This was achieved by setting a threshold at 0.05 kW, and any value below this was reported as not having a PV panel. Furthermore, the results indicate that the performance of the algorithm does not vary with the size of the panel and there does not appear to be a range of values where the approach fails, showing the algorithm performs in the range of expected PV capacity values.

However, there is a consistent bias in the results as the MNBE, as Figure 3 shows. The algorithm typically infers the PV capacity to be larger than the true value: this is most likely caused by the distribution representing energy usage over-estimating the amount of energy used and hence leading to a larger PV capacity being inferred. This can potentially be addressed by finding a distribution that better represents the energy usage in the specific scenario, or if the bias is known for the dataset the inferred values can be corrected for the bias.

	MAPE (%)	MNBE (%)	RMSE	Classification (%)
UK dataset				
0 kW	-	-	0.01	100
0.5 - 1.0 kW	18	17	0.14	100
1.0 - 1.5 kW	10	9	0.16	100
1.5 - 2.0 kW	13	9	0.32	100
2.0 - 2.5 kW	14	12	0.44	100
2.5 - 3.0 kW	13	6	0.41	100
3.0 - 3.5 kW	14	11	0.62	100
3.5 - 4.0 kW	15	13	0.73	100
4.0 - 4.5 kW	15	15	0.81	100
4.5 - 5.0 kW	15	12	0.92	100
5.0 - 5.5 kW	12	7	0.71	100
5.5 - 6.0 kW	13	11	0.96	100
6.0 - 6.5 kW	10	8	0.76	100
Average	**13**	**11**	**0.64**	**100**
Pecan Street				
Average	29	-18	1.68	100

Table 1: The MAPE, MNBE, RMSE and classification metrics for each band of 20 PV capacity values on the UK dataset and across the Pecan street dataset. The performance is relatively consistent across all ranges of values with the RMSE increasing as expected due to the larger PV capacity values.

There is also a noticeable drop in performance from the constructed UK dataset to the

Pecan Street dataset, with the MAPE across the dataset going from 13% to 29%. The drop in performance on the Pecan Street dataset is expected as it typically has PV systems with larger PV capacity, and if the PV generation is large relative to the energy usage it can be difficult to accurately infer the PV capacity as more information is censored. This explanation is validated by the MNBE showing an 18% under-estimate of PV capacity in the Pecan Street dataset on average due to the high energy generation censoring a large proportion of the smart meter readings. This is a difficult issue to address with censored smart meters and providing a lower bound on the PV capacity for these households may have to suffice. Combining this approach with satellite imagery will provide an upper and lower bound on the potential PV generation.

7.1.2. Failure without Consideration of Censoring

To show the importance of correctly handling censored observations, we have demonstrated in Table 2 what happens if we treat the censored smart meter readings ($r_{htd} = 0$ kWh) as a net reading and ignore censoring. Following Algorithm 1, we have set $\mathbb{I}_{htd} = 1$ for all data points. The algorithm fails to detect the presence of PV panels or correctly infer the PV capacity, demonstrating the importance of using an appropriate likelihood function to deal with censored observations.

	UK Dataset		
	PV	No PV	All
MAPE	99.6%	-	-
RMSE (kW)	1.23	0.01	0.87
Classification	0.4%	98.2%	49.3%

Table 2: Error metrics and classification rate of PV capacity inference without censoring

7.2. PV Capacity Inference with Clearsky Solar Irradiance

For this experiment we have used the constructed UK dataset, and present the outcomes in Table 3. We have taken a sample of 250 houses with PV panels and 250 houses without PV panels. The smart meter readings have been calculated similarly to the previous experiment, and we have normalised the actual PV generation from one of the microgen dataset houses and randomly assigned a PV capacity between 1.5 kW and 3.0 kW to each house with a PV panel, whereas the houses without a PV panel have been left as is. The solar irradiance and PV capacity data have then been disregarded and we have implemented our algorithm to infer the PV capacity for each house. In our implementation, once the inference of a PV capacity has gone below 0.1 kW, we have classified the house as not having a solar panel and removed it from the next iteration as it would no longer contribute useful information to the problem.

	UK Dataset		
	PV	No PV	All
MAPE	33.0%	-	-
RMSE (kW)	0.55	0.01	0.39
Classification	100%	100%	100%

Table 3: Error metrics and classification rate of PV capacity inference with clearsky solar irradiance

Whilst the RMSE is similar to the previous results for the constructed UK dataset, the MAPE is larger (more than twice as large), indicating that the approach without solar irradiance incurs a larger error in houses with solar panels of smaller PV capacity. To the best of our knowledge, this is the first approach to infer PV capacity that does not require solar irradiance or weather data as an input. Whilst in practice if the location is known then weather data could be acquired, the purpose of this implementation and experiment was to show the feasibility of alternative approaches to solar dissagregation, in particular approaches that do not inherit errors related to inferring solar irradiance from weather data. Weather-based approaches to solar disaggregation to infer PV capacity are still considerably more accurate, however we believe this shows the feasibility of an alternative approach and could inspire related research efforts that use local clusters of households to infer PV generation [5].

7.3. Runtime and Scalability

The runtime on a single Intel 7th gen i7 CPU core to detect the presence and capacity of a PV panel averaged at 9.32 seconds per house (8.54 seconds per house excluding data loading) across the 240 houses with PV panels and at 5.74 seconds per house (5.06 seconds per house excluding data loading) for the 20 houses without PV panels. There are approximately 25 million homes in the UK, of which about 1 million homes have PV panels, which means to infer the PV capacity for all buildings across the UK it would require 16000 CPU-hours. As each house uses a separate instance of the algorithm, the workload could be parallelised across multiple CPUs and, based on current AWS EC2 T2 prices, it could be conducted for under $400. Further speed-ups are expected by employing GPU-based computations.

A nationwide scan only needs to be run once to detect panels and their associated PV generation. The algorithm can be re-run at set intervals at the discretion of the user to update their dataset of solar panel locations and capacity. Once the capacity of the solar panel is known - the inference of expected solar generation for the next time step is in the magnitude of seconds to infer the generation for all houses with solar panels, as it is a single calculation using the solar irradiance. The run times of the algorithm and inference show that this approach is fit for purpose at scale and affordable with cloud computing services.

8. Conclusions and Future Work

In this paper, we have proposed the first approach to dissagregate solar (PV) generation from energy consumption given censored smart meter readings and to infer the PV capacity. To evaluate our approach, we have used an appropriate subset of the Pecan Street dataset and a custom dataset that we have created by combining data supplied from the Sheffield Solar microgen dataset and the smart meter readings from London households. We have shown that if the solar irradiance is known as a proxy from a single panel, we can detect the presence of a PV panel successfully 100% of the time on our test data, and we can additionally infer the PV capacity. Using our approach, solar dissagregation can be performed on over 500,000 houses in the UK alone where alternatives that do not address the censoring of smart meters would fail. We have also shown that we can infer PV capacity with clearsky solar irradiance if we have a group of houses in a local area, and combined with the inferred values of solar irradiance using smart meter readings, we have presented the first approach to infer solar generation with clearsky solar irradiance.

In future work we plan to extend our approach to deal with daily censored smart meter readings. This has the added difficulty of not knowing at what time of day the censoring has occurred. We are also working with grid operators and energy suppliers to operationalise our research. Extensions to the work could look to combine this research with recent approaches to detecting solar panels using satellite imagery. Using both approaches in parallel can improve the confidence that a panel is correctly detected. Furthermore, satellite imagery can be used to detect the size of PV panels, and if our algorithm infers the real PV capacity to be below what is expected of a PV panel of the detected size, it could be used to identify faulty panels.

References

[1] A. Peruffo, E. Guiu, P. Panciatici, A. Abate, Aggregated Markov models of a heterogeneous population of photovoltaic panels, in: Proceedings of International Conference on Quantitative Evaluation of Systems, Springer, 2017, pp. 72–87.

[2] A. Peruffo, E. Guiu, P. Panciatici, A. Abate, Safety guarantees for the electricity grid with significant renewables generation, in: Proceedings of International Conference on Quantitative Evaluation of Systems, Springer, 2019, pp. 332–349.

[3] B. Schäfer, C. Beck, K. Aihara, D. Witthaut, M. Timme, Non-Gaussian power grid frequency fluctuations characterized by Lévy-stable laws and superstatistics, Nature Energy 3 (2) (2018) 119–126.

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

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

[6] R. Mohan, T. Cheng, A. Gupta, V. Garud, Y. He, Solar energy disaggregation using whole-house consumption signals, in: NILM Workshop, 2014.

[7] N. Sharma, P. Sharma, D. Irwin, P. Shenoy, Predicting solar generation from weather forecasts using machine learning, in: Proceedings of IEEE International Conference on Smart Grid Communications (SmartGridComm), IEEE, 2011, pp. 528–533.

[8] P. Chakraborty, M. Marwah, M. F. Arlitt, N. Ramakrishnan, Fine–grained photovoltaic output prediction using a Bayesian ensemble, in: Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012, pp. 274–280.

[9] E. C. Kara, C. M. Roberts, M. Tabone, L. Alvarez, D. S. Callaway, E. M. Stewart, Disaggregating

solar generation from feeder-level measurements, Sustainable Energy, Grids and Networks 13 (2018) 112–121.

[10] E. Vrettos, E. Kara, E. Stewart, C. Roberts, Estimating PV power from aggregate power measurements within the distribution grid, Journal of Renewable and Sustainable Energy 11 (2) (2019) 023707.

[11] F. Sossan, L. Nespoli, V. Medici, M. Paolone, Unsupervised disaggregation of photovoltaic production from composite power flow measurements of heterogeneous prosumers, IEEE Transactions on Industrial Informatics 14 (9) (2018) 3904–3913.

[12] H. Shaker, H. Zareipour, D. Wood, A data-driven approach for estimating the power generation of invisible solar sites, IEEE Transactions on Smart Grid 7 (5) (2015) 2466–2476.

[13] H. Shaker, H. Zareipour, D. Wood, Estimating power generation of invisible solar sites using publicly available data, IEEE Transactions on Smart Grid 7 (5) (2016) 2456–2465.

[14] Y. He, A. Gupta, H.-T. Cheng, V. Garud, Energy disaggregation techniques for whole-house energy consumption data, US Patent App. 14/543,824 (May 21 2015).

[15] C. M. Cheung, S. R. Kuppannagari, R. Kannan, V. K. Prasanna, Towards improved real-time observability of behind-meter photovoltaic systems: A data-driven approach, in: Proceedings of the Tenth ACM International Conference on Future Energy Systems, ACM, 2019, pp. 447–455.

[16] R. Perez, M. Beauharnois, K. Hemker, S. Kivalov, E. Lorenz, S. Pelland, J. Schlemmer, G. Van Knowe, Evaluation of numerical weather prediction solar irradiance forecasts in the US, 2011.

[17] S. Rehman, M. Mohandes, Artificial neural network estimation of global solar radiation using air temperature and relative humidity, Energy Policy 36 (2) (2008) 571–576.

[18] A. Mellit, A. M. Pavan, A 24-h forecast of solar irradiance using artificial neural network: Application for performance prediction of a grid-connected PV plant at trieste, italy, Solar Energy 84 (5) (2010) 807–821.

[19] F. Rodríguez, A. Fleetwood, A. Galarza, L. Fontán, Predicting solar energy generation through artificial neural networks using weather forecasts for microgrid control, Renewable Energy 126 (2018) 855–864.

[20] J. Yu, Z. Wang, A. Majumdar, R. Rajagopal, DeepSolar: A machine learning framework to efficiently construct a solar deployment database in the United States, Joule 2 (12) (2018) 2605–2617.

[21] J. Yuan, H.-H. L. Yang, O. A. Omitaomu, B. L. Bhaduri, Large-scale solar panel mapping from aerial images using deep convolutional networks, in: Proceedings of the Fourth IEEE International Conference on Big Data, IEEE, 2016, pp. 2703–2708.

[22] J. M. Malof, K. Bradbury, L. M. Collins, R. G. Newell, A. Serrano, H. Wu, S. Keene, Image features for pixel-wise detection of solar photovoltaic arrays in aerial imagery using a random forest classifier, in: Proceedings of the Fifth IEEE International Conference on Renewable Energy Research and Applications (ICRERA), IEEE, 2016, pp. 799–803.

[23] J. M. Malof, L. M. Collins, K. Bradbury, R. G. Newell, A deep convolutional neural network and a random forest classifier for solar photovoltaic array detection in aerial imagery, in: Renewable Energy Research and Applications (ICRERA), 2016 IEEE International Conference on, IEEE, 2016, pp. 650–654.

[24] G. Cheng, J. Han, X. Lu, Remote sensing image scene classification: benchmark and state of the art, in: Proceedings of the IEEE, Vol. 105, IEEE, 2017, pp. 1865–1883.

[25] Y. LeCun, Y. Bengio, et al., Convolutional networks for images, speech, and time series, The handbook of brain theory and neural networks 3361 (10) (1995) 255–258.

[26] J. M. Bright, S. Killinger, D. Lingfors, N. A. Engerer, Improved satellite-derived PV power nowcasting using real-time power data from reference PV systems, Solar Energy 168 (2018) 118–139.

[27] J. P. Klein, M. L. Moeschberger, Survival analysis: Techniques for censored and truncated data, Springer Science & Business Media, 2006.

[28] C. A. Smith, The Pecan Street Project: developing the electric utility system of the future, Ph.D. thesis, U. of Texas (2009).

[29] J. Taylor, J. Leloux, L. M. Hall, A. M. Everard, J. Briggs, A. Buckley, Performance of distributed PV in the UK: a statistical analysis of over 7000 systems, in: 31st European photovoltaic solar energy conference and exhibition, 2015.

[30] R. E. Bird, R. L. Hulstrom, Simplified clear sky model for direct and diffuse insolation on horizontal surfaces, Tech. rep., Solar Energy Research Inst., Golden, CO (USA) (1981).

[31] J. S. Stein, W. F. Holmgren, J. Forbess, C. W. Hansen, PVLIB: Open source photovoltaic performance modeling functions for MATLAB and Python, in: Proceedings of Forty-Fourth IEEE Photovoltaic Specialist Conference (PVSC), IEEE, 2017, pp. 1–6.