ESTIMATING THE ELECTRICITY GENERATION CAPACITY OF SOLAR PHOTOVOLTAIC ARRAYS USING ONLY COLOR AERIAL IMAGERY

Brenda So[1], *Cory Nezin*[1], *Vishnu Kaimal*[1], *Sam Keene*[1], *Leslie Collins*[2], *Kyle Bradbury*[3], *Jordan M. Malof*[2]

[1]Department of Electrical & Computer Engineering, The Cooper Union, New York, NY 10003
[2]Department of Electrical & Computer Engineering, Duke University, Durham, NC 27708
[3]Energy Initiative, Duke University, Durham, NC 27708

ABSTRACT

In this work, the problem of developing algorithms that automatically infer information about small-scale solar photovoltaic (PV) arrays in high resolution aerial imagery is considered. Such algorithms potentially offer a faster and cheaper solution to collecting small-scale PV information, such as their location and capacity. Existing work on this topic has focused on the automatic identification and annotation of panels in the aerial imagery. We extend this work by showing that we can reliably infer the capacity of PV arrays given only (i) color aerial imagery and (ii) a precise annotation of the array location. First we demonstrate that accurate capacity estimates can be obtained simply by estimating the visible surface area of a solar array, regardless of tilt. We then build a more sophisticated model where we use additional image information related to the properties of the solar array to further improve the capacity predictions. We use a dataset of 362 manually annotated Google Earth images of solar arrays with known electricity generation capacity in North Carolina to measure the predictive performance of our models.

Index Terms—solar energy, detection, satellite imagery, photovoltaic, machine learning

1. INTRODUCTION

In this work the problem of developing algorithms that can automatically infer information about small-scale solar photovoltaic (PV) arrays, such as their capacity and energy generation, only using very high resolution (e.g., 0.3m per pixel) aerial imagery is considered. At such a high resolution it is possible to visually identify PV arrays in the imagery as well as their color, shape, and size. Fig. 1 shows an example of a Google Earth image where the PV array has been manually annotated with a red border. Automatic identification techniques could potentially yield PV information at much greater geo-spatial resolutions than are currently available (e.g., neighborhood- or city-level instead of state- or national-level) [1], [2].

Estimating solar PV array generation capacity from aerial imagery can be separated into two problems: (1) the automatic detection and annotation of the solar arrays in the imagery (e.g., the red polygon in Fig. 1), and (2) inferring capacity using the identified array imagery. Existing research has focused primarily on the automatic detection and annotation of PV arrays [1], [2]. In this work, we investigate the second problem of estimating electricity generation based on the imagery. Specifically, we investigate models to infer the capacity of individual PV panels using only (i) color aerial imagery of the solar array and (ii) a precise (polygonal) annotation of the array in the imagery.

In order to develop predictive models, we created a dataset consisting of 362 manually annotated PV arrays in high resolution Google Earth images in the US state of North Carolina. The ground truth capacity of each PV array was self-reported by the array owners. Using this data, we first show that capacity is well estimated using a simple regression on the area of each panel. We then show that these estimates can be substantially improved using a more sophisticated model incorporating simple image intensity and texture statistics.

Fig. 1. Example of a solar PV array in a high resolution color aerial image. The manual annotation of the panel in the image is shown in red.

The remainder of this paper is organized as follows: Section 2 describes the aerial imagery dataset; Section 3 describes the regression models used for predicting solar

array capacity; Section 4 describes the variable selection process for our regression models; Section 5 presents our experimental design and results; and Section 6 presents our conclusions and ideas for future work.

2. THE SOLAR ARRAY IMAGERY DATASET

For our ground truth data on solar array capacity, we used a dataset with both individual solar array electricity generation capacity and their precise geospatial coordinates. This dataset contains over 4,500 solar arrays throughout the state of North Carolina, based on data from the North Carolina Utility Commission and curated and compiled by the North Carolina Sustainable Energy Association.

Using the locations of these arrays, we used a stratified sampling approach to select 500 solar arrays distributed over the total range of generation capacities in the dataset. We then collected Google Earth imagery for the selected arrays. Each array was manually annotated and the size (in square meters) was estimated from those annotations. We then eliminated those arrays that were larger than a standard residential solar array insolation by only selecting those arrays that were less than 50 square meters. The resulting dataset contains 362 solar array aerial images with known electricity generation capacity.

3. REGRESSION MODELS

In this section, we present two general regression models that we use for estimating the capacity of a PV array based on information extracted from (i) color aerial imagery and (ii) annotations of the PV arrays. In the subsequent discussion, we will use c to denote the total capacity of a PV array, γ to denote the capacity per square meter (of surface area) of a PV array, and α to denote the surface area of the PV array. Both of the proposed models estimate capacity using a linear relationship between the area of the array and the capacity of the array [3]:

$$c = \gamma\alpha + \gamma_0 \tag{1}$$

Here γ_0 is a bias parameter, which is included because equation (1) is an approximation, and may not actually have an intercept at zero. The main difference between the two proposed models in this work is that our first model assumes a global, fixed, γ parameter for each panel, and the second model attempts to infer a unique γ for each PV array based on the imagery data of the array.

3.1. Baseline regression using a global estimate of capacity per unit area, γ

This approach relies on estimating a single γ and γ_0 parameter for all PV arrays. This is achieved by performing a linear regression of the solar array capacity, c, onto the array surface area, α, as in equation (1). Fig. 2 shows the result of applying this regression model using the entire PV dataset. It illustrates that the estimated annotation areas are indeed very correlated with c, yielding a correlation coefficient of 0.83, with greater than 1% statistical significance. In Section 5 we extend our analysis and measure the predictive ability of the model in equation (1) using cross-validation.

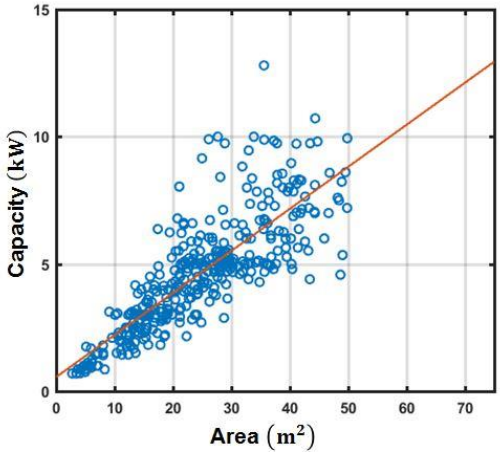

Fig. 2. Demonstration of the relationship between the area and generation capacity of solar PV arrays in the dataset. Each point corresponds to a single solar array. The line is the simple regression fit in which a single γ is learned for all PV arrays. The correlation coefficient of this fit is 0.83.

3.2. A regression model using a unique estimate of capacity per unit area for each solar array, γ_i

The approach described in this section assumes that each PV array has a unique γ, which can be estimated using statistics computed from its imagery and/or annotations. Our hypothesis here is that γ may vary depending upon the underlying physical properties of the solar cells, such as the chemistry of the solar cells, or the manufacturer's design. These physical differences may, in turn, exhibit different colors, reflectance, or texture in the imagery. We can compute statistics from the imagery that encode these relevant qualities, and then use them to estimate a unique γ for each panel. To examine this hypothesis we use the following model:

$$\gamma_i = \boldsymbol{\beta}^T \boldsymbol{v}_i + \beta_0 \,. \tag{2}$$

Here $\boldsymbol{v}_i$ denotes a vector of imagery statistics for the ith PV array (which are described in Section 3.3), $\boldsymbol{\beta}$ denotes a vector of regression weights shared by all solar arrays, γ_i is the capacity per unit area for the ith array, and β_0 is a bias term. Note that bolded symbols refer to vector quantities. The measurements in $\boldsymbol{v}_i$ consist of statistics that can include the means and variances of colors that are computed using pixels within the PV annotations.

In order to incorporate the relationship in equation (2) into our original model, equation (1), we can substitute equation (2) into equation (1) and simplify:

$$c_i = (\boldsymbol{\beta}^T \boldsymbol{v}_i + \beta_0)\alpha_i + \gamma_0 \ ,$$
$$c_i = [\boldsymbol{\beta}^T,\ \beta_0]\ [\alpha_i \boldsymbol{v}_i^T,\ \alpha_i]^T + \gamma_0 \ ,$$
$$c_i = \boldsymbol{\beta}'^T \boldsymbol{v}_i' + \beta_0' \ , \qquad (3)$$

In equation (3), $\boldsymbol{v}_i' = [\alpha_i \boldsymbol{v}_i,\ \alpha_i]$ and $\boldsymbol{\beta}' = [\boldsymbol{\beta}^T, \beta_0]^T$. For consistency, we use β_0' to replace γ_0 as the bias parameter. Equation (3) is a regression model that implicitly models each array as having its own γ_i parameter that is determined by the solar array image statistics, $\boldsymbol{v}_i$. The parameters $\boldsymbol{\beta}'$ and β_0' can be estimated using standard linear regression.

3.3. Selecting solar array image descriptors for inclusion in the regression model

In this section we discuss the motivation for each of the imaging statistics that we included in our model for γ_i, given in equation (2). Our first imaging feature relates to the brightness, or intensity, of each PV array, which may be related to how well it absorbs light. For this purpose we use the HSV colorspace, in which the intensity of the color is encoded in the 'Value' component. We use the median of the Value component across all pixels in the array annotation to capture this quality. We also included the standard deviation of the Value component in order to encode simple texture information, which may be related to the type of solar array (e.g. thin film vs monocrystalline silicon vs polycrystalline silicon [4]).

The size of the array itself may indicate more than just the total generation capacity, but also the efficiency of the solar arrays. For example, it is less expensive to cover a rooftop with thin film solar than to cover it with crystalline solar. Therefore, the area of the polygon is included as a feature for estimating the generation capacity per square meter as well.

To help further justify the choice of these model variables, Fig. 3 shows 2-dimensional histograms of each variable and the target variable, capacity per square meter, γ_i. To produce these plots, a γ_i for each PV array was estimated by setting $\hat{\gamma}_i = c/\alpha$, where α was measured from the manual array annotations. It is clear from the plots that the three predictor variables are anti-correlated with γ_i.

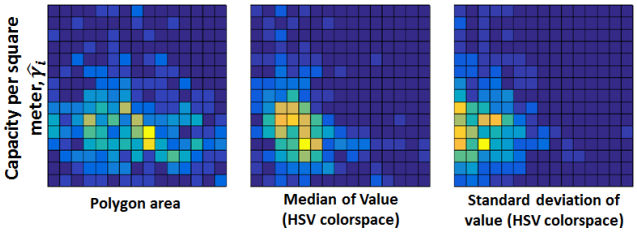

Fig. 3. Two-dimensional histograms of capacity per square meter, γ_i versus solar array image features of the PV arrays. Brighter squares indicate higher frequencies of occurrence. Each of these three features show weak anti-correlation with γ_i, and were included in the γ_i prediction model.

4. THE PREDICTIVE PERFORMANCE OF THE REGRESSION MODELS

In this section we estimate the predictive performance of the two proposed regression models using K-fold cross-validation. Here we use $K = 30$ to balance estimation accuracy and computational speed. The predicted capacity values given by each regression model are compared to the true values using mean-squared-error (MSE). Lower MSEs indicate better performance, and this will be the primary metric that is used to compare the two proposed models.

For both regression models employed in this work we use ridge regression [5] to infer the model coefficients. Ridge regression requires that the user to specify a parameter, λ, for regularizing the model. We optimize the value for this parameter, denoted λ^*, using a second K-fold cross-validation scheme performed within each training fold. In this inner cross-validation, we set $K = 20$. The λ^* value is chosen so to be the λ that yields to the lowest MSE on the inner cross-validation. Once λ^* is attained, ridge regression is applied with λ^* to the current training fold data to infer the regression weights for each model (e.g., $\boldsymbol{\gamma}$, γ_0, $\boldsymbol{\beta}'$ and $\boldsymbol{v}_i$). These parameters are then used to predict capacity values on the corresponding testing set.

This procedure yields two capacity predictions, $\hat{c}_i$, for each panel: one from each of the two regression models. Because cross-validation is randomized, this can lead to different MSE results each time the cross validation is performed. To measure the variance on the cross-validation MSE estimates, the whole cross-validation procedure is repeated 30 times.

Fig. 4 shows the experimental results in the form a histogram of the MSEs that were obtained from each of the two models. The results show that the second model, employing a unique gamma for each PV array, outperforms the first model. The results show an improvement in the mean MSE between the two approaches of about 9%. The results are statistically significant at the 1% significance level. The effectiveness of the second regression model suggests that imagery data may be used to improve the efficiency estimates (in the form of the parameter) of individual panels.

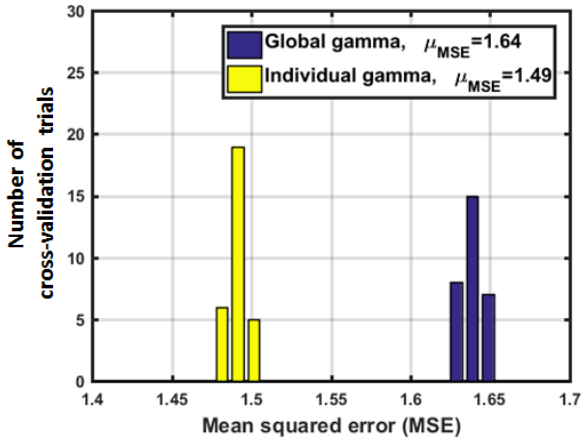

Fig. 4. Histogram of the mean squared error (MSE) of each of the two capacity prediction methods we propose, after each method was applied in 30 cross-validation trials. Estimating an individual capacity (per square meter), denoted γ_i, for each panel using image information provides better overall capacity estimates than using a single global estimate.

5. CONCLUSIONS AND FUTURE WORK

In this work we investigated two models for predicting small-scale PV array capacity using only (i) aerial imagery over the arrays and (ii) precise annotations of the arrays in the imagery. The baseline model assumes that all PV arrays share a common capacity-per-unit-area, denoted γ. It uses γ to predict PV capacity based only on the area of the PV array annotations. The second model assumes that each PV array has a unique γ parameter, which can be estimated using imagery information, such as intensity (brightness) and texture.

We measured the mean-squared-error (MSE) of the two models using cross-validation experiments. The results show that both models make accurate capacity predictions, but the second model yields a statistically significant improvement over the first. This suggests that imagery information can indeed predict the efficiency, in the form of a unique γ parameter, of PV arrays.

The results also suggest that PV array capacity can be estimated accurately based only on imagery information, further demonstrating the feasibility of a small-scale PV information collection approach that relies only on aerial imagery.

For future work, we propose expanding the scope of this analysis to incorporate more data samples of solar array capacity. In particular, acquiring imagery data of thin film, monocrystalline, and polycrystalline solar cells along with their capacity, will enable an analysis of how well these techniques could discriminate between the different solar array chemistries as well as balancing the classes of each type of panel to improve the regression models performance from this work.

6. ACKNOWLEDGEMENTS

The solar array location information was provided by the North Carolina Sustainable Energy Association (NCSEA). Many thanks to Therese Lundblad and Jacob Händestam from KTH Sweden, who assembled and manually annotated the dataset of solar array imagery in addition to implementing the initial regression model for this work.

7. REFERENCES

[1] J. M. Malof et al., "Automatic solar photovoltaic panel detection in satellite imagery," in *International Conference on Renewable Energy Research and Applications (ICRERA)*, 2015, pp. 1428–1431.

[2] J. M. Malof, K. Bradbury, L. M. Collins, and R. G. Newell, "Automatic detection of solar photovoltaic arrays in high resolution aerial imagery," *Appl. Energy*, vol. 183, pp. 229–240, 2016.

[3] L. K. Wiginton, H. T. Nguyen, and J. M. Pearce, "Quantifying rooftop solar photovoltaic potential for regional renewable energy policy," *Comput. Environ. Urban Syst.*, vol. 34, no. 4, pp. 345–357, 2010.

[4] B. Parida, S. Iniyan, and R. Goic, "A review of solar photovoltaic technologies," *Renew. Sustain. Energy Rev.*, vol. 15, no. 3, pp. 1625–1636, Apr. 2011.

[5] A. Hoerl and R. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, 1970.