SOLAR CAPACITY ESTIMATION FROM AERIAL IMAGERY

A study of the correlation between areas in satellite images and capacity for solar panels

JULY 25, 2016

DUKE UNIVERSITY
Händestam Jacob, jachan@kth.se
Lundblad Therese, tlundbla@kth.se

Abstract

The amount of solar panels is increasing rapidly. Current solar power capacity estimation relies on solar panel owners reporting capacity, which is a slow process. This report presents a method to predict the power capacity of solar panels using the area from two dimensional satellite images. The goal is to show the correlation between these two variables and to quantify it.

A data set from North Carolina Sustainable Energy Association, *NCSEA*, was used were solar panels were reported with a capacity and an approximate location for solar cells in North Carolina, USA. A polygon was created for the solar panels' boundaries using satellite images and from these polygons an area was calculated using a Matlab script. After creating a dataset with both capacity and area, a linear correlation was calculated. A linear regression analysis was run in order to calculate confidence and prediction intervals.

The results show a strong correlation between solar panel area in satellite images and solar panel capacity. A general linear correlation was quantified, alongside with a refined correlation for solar panels with an area less than or equal to 200 m^2.

Foreword

We would like to start this report with expressing our gratitude towards the people and institutions that gave us the opportunity to perform this study. First towards Dr. Kyle Bradbury, who gave us support and guidance throughout the research project. We would also like to thank Dean Jim Gaston, who helped us arrange our stay at Duke University.

In addition to this we would like to thank the Department of Energy Technology at the Royal Institute of Technology, Stockholm, Sweden for the scholarships and funding that enabled this research.

Table of Contents

1. Introduction

The human population is increasing and so are the greenhouse gas emissions associated with everyday life. In the U.S. electricity generation accounts for a third of the total greenhouse gas emissions [1]. Many of the basic human needs, such as access to clean water, nourishing food, a controlled climate and sanitation are dependent on electricity access [2]. In order to reach good living standards for all in a sustainable way without the negative consequences associated with global warming, an increase in clean and renewable electricity generation is needed. This electricity can come from a wide variety of sources, with solar photovoltaics, solar PV, being one of them. Increasing the amount of renewable energy also decreases fossil fuel dependency [3].

Solar PV is a method for direct conversion of solar energy to electricity using a semiconductor, contrasting solar thermal power that first converts the solar energy to heat, which is then converted into electricity at a later stage [4]. During the last decade the prices of solar panels have dropped with more than 70 % due to upscaling in manufacturing and technological improvements. During the same time period a rapid increase in solar panel installations has been observed [5]. Modern solar panels have an average capacity of 0.08 to 0.15 kW/m^2 [6, 7].

1.1. Objectives and limitations

Currently electricity generation capacity from private solar cells is monitored through reports from solar panel owners. Gathering this data is a slow process, not adapted to the rapid increase of solar cell installations that has been seen recently. In addition to the process being slow it also allows for human errors to occur. Removing the human factor from the capacity monitoring could possibly increase the processing time and reliability. In order to determine if the process can be automated or not the correlation between solar panel area and capacity is to be assessed. Focusing on small scale solar power, defined as solar panel arrays with capacities under 1 MW, a prediction equation for long range solar panel capacity estimation is to be calculated. This means that solar fields are not included in the scope of the study. Furthermore, confidence and prediction intervals for this equation are to be quantified.

It is assumed that the reported solar panel information is correct. Solar panels not in the data set are not considered. In addition to this, it is assumed that solar panel areas can be predicted from two dimensional, *2D*, satellite images.

2. Method

The study was based on a data set from North Carolina Sustainable Energy Association, *NCSEA*, including a reported ID, capacity and an approximate location for solar cells in North Carolina, USA. In order to create a representative data sample, the data set from *NCSEA* was divided into eleven categories based on solar panel capacity. Since the main purpose was to study private solar cells, the panels with larger capacity than 1 MW were not included. Within each category 70 IDs were randomized using Matlab's RANDSAMPLE function [8].

The solar panels corresponding to the selected IDs were found in Google Earth using their approximate location. In the majority of cases, multiple solar panels existed for each ID. Area approximations were performed through manually identifying each panel's corners, creating one or more polygons that could be exported as a .kml-file containing coordinates for polygon corners.

The polygon area was calculated in Matlab, see Appendix I, and the sum of the polygon areas associated with an ID was added to the original data set. After acquiring an adequate data set containing both capacity and area, a linear regression analysis was performed in order to examine the correlation between capacity and area. To obtain a realistic relation for small areas, the intercept of the trendline was forced to zero. This seemed reasonable as it insinuated that no area means no capacity. The trendline was calculated by using Excel's LINEST function [9]. Subsequently, confidence and prediction intervals were calculated using [10]:

$$y_i \pm t_{n-2} s_y \sqrt{\frac{1}{n} + \frac{\left(x_i - \bar{x}\right)^2}{SS_x}} \tag{1}$$

$$y_i \pm t_{n-2} s_y \sqrt{1 + \frac{1}{n} + \frac{\left(x_i - \bar{x}\right)^2}{SS_x}} \tag{2}$$

where t_{n-2} is the critical value for a two-sided test, as a function of the probability of error and number of observations. The probability of error was set to 5 % and the number of observations, *n*, was 554. Using Excel’s T.INV.2T function the critical value was calculated. [11]. The residual standard error, s_y, was calculated as:

$$s_y = \sqrt{\frac{SSE}{n-2}} \tag{3}$$

where *SSE* is the sum of squared errors in the linear model. The residual standard error can also be calculated using Excel’s LINEST function.

The sum of squares of deviations of data points from their sample mean, SS_x, was calculated as:

$$SS_x = \sum_{i=1}^{n} \left(x_i - \overline{x} \right)^2 . \tag{4}$$

This can also be calculated using Excel’s DEVSQ function [12].

The coefficient of determination, R^2, is defined as the explained variation over the total variation, and was calculated as [13]:

$$R^2 = \frac{SS_{reg}}{SS_{tot}} \tag{5}$$

where SS_{reg} is the regression sum of squares, which measures the variability in the dependent variable which is due to the regression model. The total sum of squares, SS_{tot}, measures the total variability in the dependent variable. The coefficient of determination can also be calculated using Excel's LINEST function.

In order to ensure that the results were not obtained by chance, a P-value and a significance F value was calculated. These values were calculated using Excel's regression tool. The P-value describes the likelihood that the slope of the trendline is correct and didn't occur by chance. The significance F value indicates whether the regression output might have been created by chance [14].

In order to establish an improved correlation for smaller areas, the procedure was repeated for areas between 0 and 200 m^2, which included 486 observations. This limit was set after analyzing the change of residual standard error for different intervals. Of the total polygon area for each ID merely 12.3 % were over 200 m^2 and only 6.66 % of all polygons areas were larger than 200 m^2.

3. Result

The data sample and the corresponding trendline is shown in Figure 1 and Figure 2. The trendline shows capacity as a function of the solar panel area.

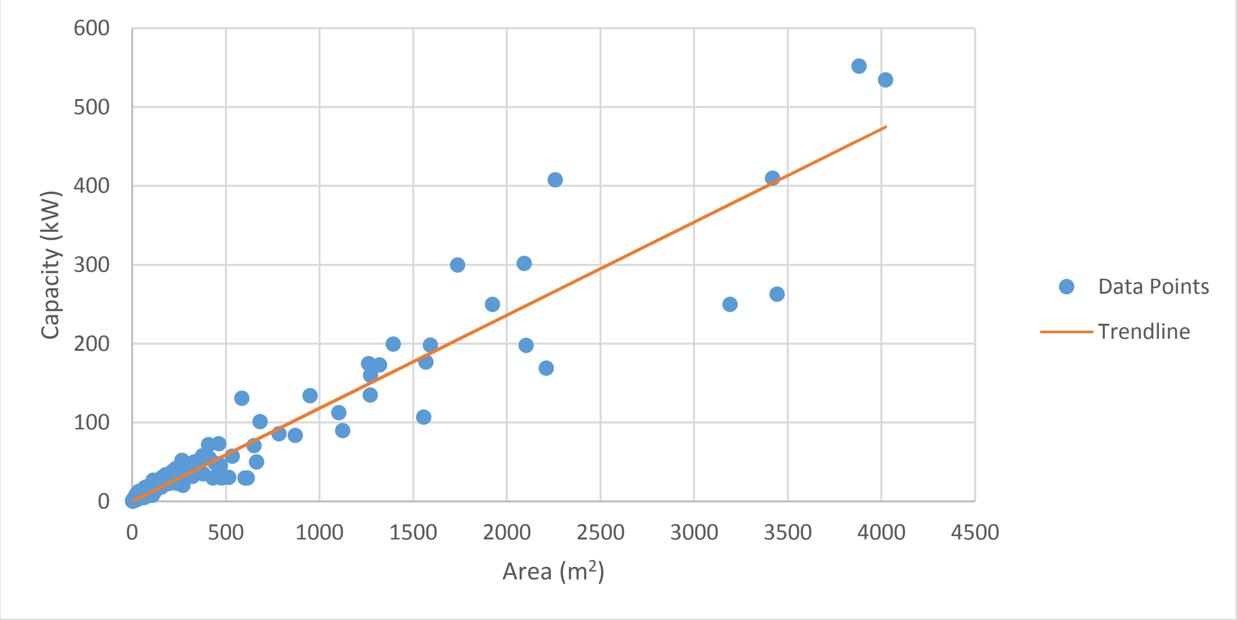

Figure 1. Data points and trendline for area as a function of capacity for all areas.

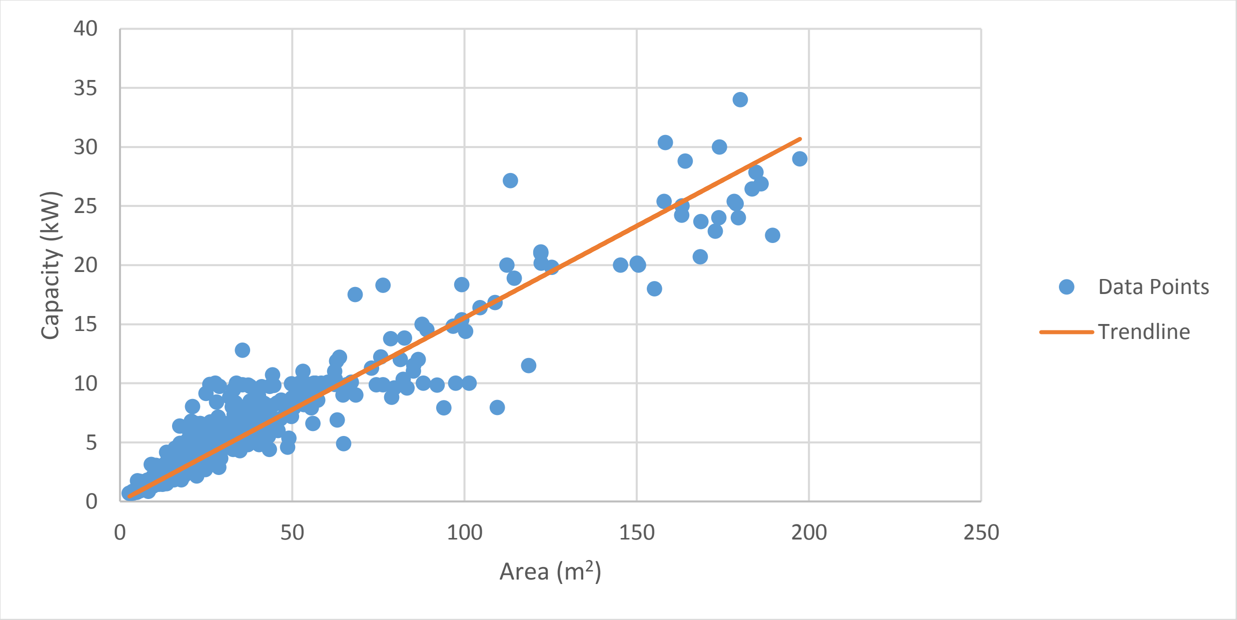

Figure 2. Data points and trendline for area as a function of capacity for areas less than or equal to 200 m^2.

The slope and intercept of the linear correlation and the regression variables used for when all areas are investigated, when the intercept was forced to zero is presented in Table 1.

Table 1. Values from linear correlation and –regression for case of all areas, when intercept was forced to zero.

Variable	Value	Unit
Slope	0.118	kW/m^2
Intercept	0	kW
$t_{\text{n-2}}$	1.9643	-
s_{y}	14.8	kW
$\bar{x}$	147.1	m^2
SS_{x}	1.07E+08	(m^2)^2
R^2	0.901	-
P-value	0	-
Significance F	0	-

The results for the capacity-area correlation for areas less than or equal to 200 m^2 with an intercept forced to zero is presented in Table 2. It contains the slope and intercept for the linear correlation and the regression variables.

Table 2. Values from linear correlation and –regression for case areas less than or equal to 200 m^2, when intercept was forced to zero.

Variable	Value	Unit
Slope	0.155	kW/m^2
Intercept	0	kW
$t_{\text{n-2}}$	1.9649	
s_{y}	1.91	kW
$\bar{x}$	39.2	m^2
SS_{x}	671934.8	(m^2)^2
R^2	0.951	-
P-value	0	-
Significance F	0	-

The results from the linear correlation and –regression, when all areas are investigated, without forcing the intercept to zero, is shown in Table 3.

Table 3. Values from linear correlation and –regression for case of all areas, when intercept was not forced to zero.

Variable	**Value**	**Unit**
Slope	0.117	kW/m^2
Intercept	1.93	kW
t_{n-2}	1.9643	-
s_y	14.7	kW
$\bar{x}$	147.1	m^2
SS_x	1.07E+08	(m^2)^2
R^2	0.924	-
P-value	0	-
Significance F	0	-

The confidence- and prediction intervals, when all areas are investigated, and the intercept was forced to zero is shown in Figure 3.

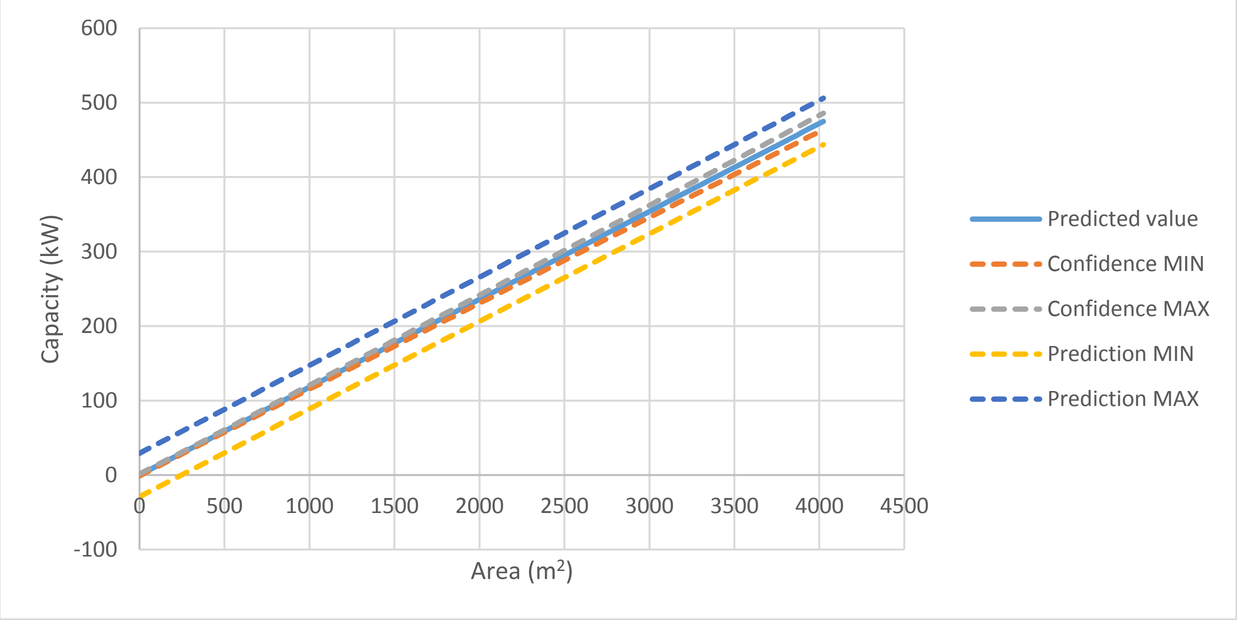

Figure 3. Confidence- and prediction intervals for all areas when the intercept was forced to zero.

Out of all collected data points, 96.9 % are within the prediction interval.

The confidence- and prediction intervals are shown in Figure 4, when areas less than or equal to 200 m^2 were investigated, and the intercept was forced to zero.

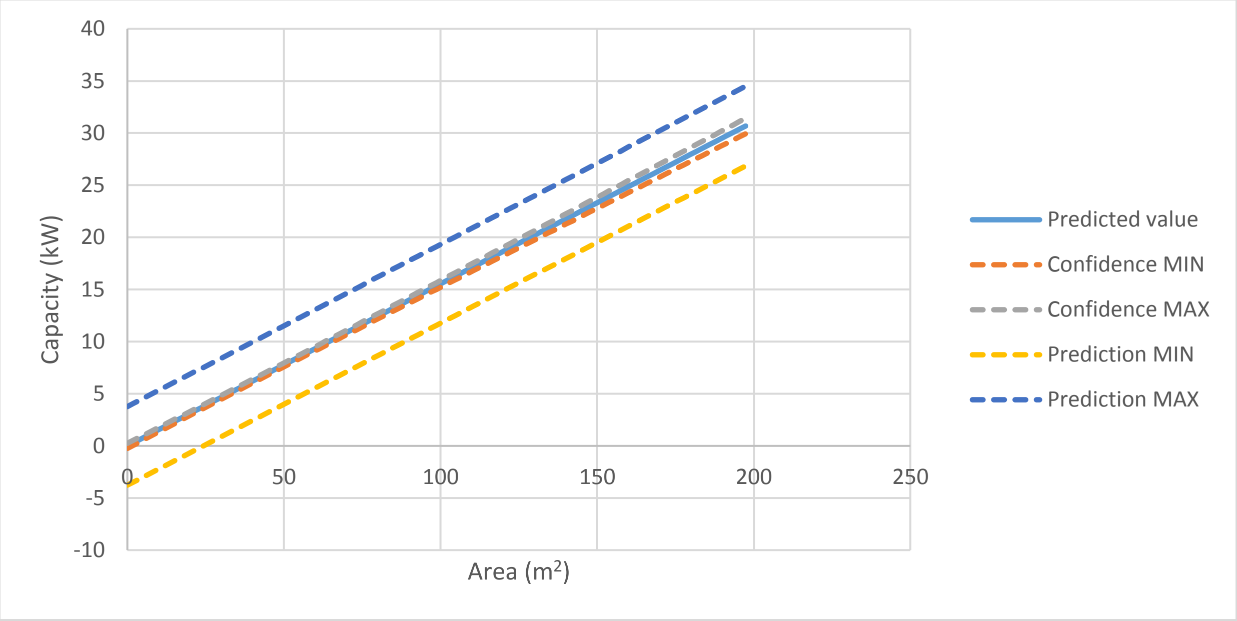

Figure 4. Confidence- and prediction intervals for areas less than or equal to 200 m^2 when the intercept was forced to zero.

Out of these collected data points, 93.6 % are within the prediction interval.

The average capacity per area unit between year 2007 and 2015 is plotted in Figure 5. During the studied time period an average increase of 1.64 % per year was detected.

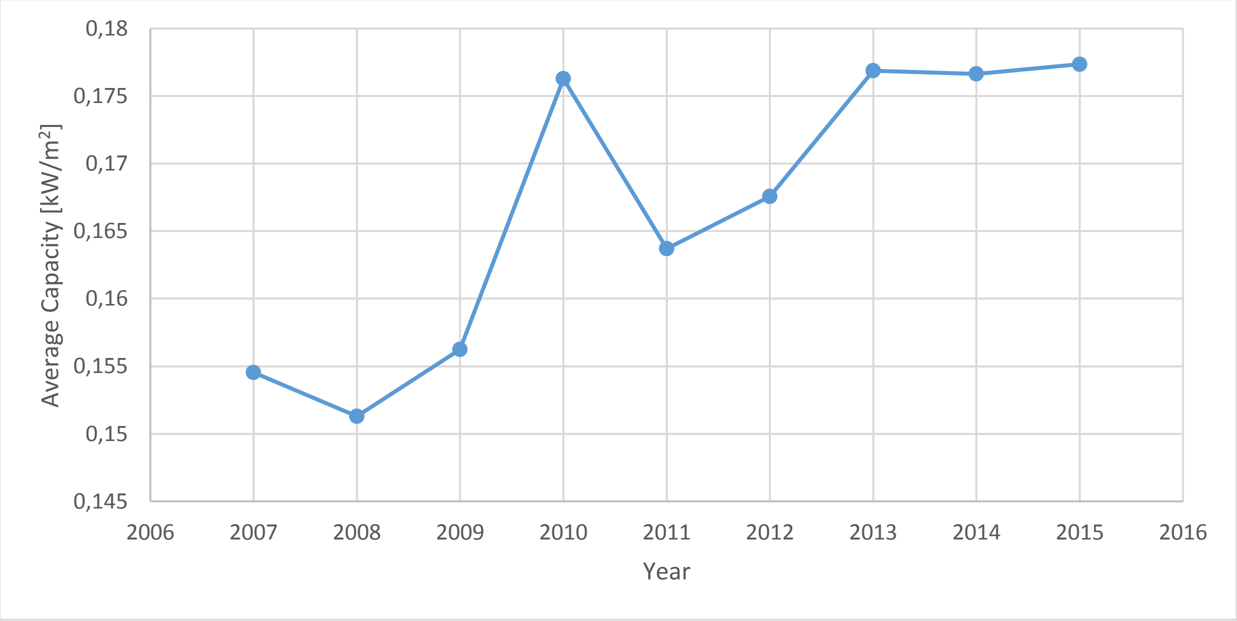

Figure 5. The average capacity per square meter for years 2007 to 2015.

4. Discussion

The result shows a strong correlation between the area and capacity of solar panels. This indicates that the capacity can be estimated when knowing the area of a solar panel. The prediction method has a large relative error when estimating the capacity of a single solar panel. It can however give an approximate capacity. The errors of predicted capacities compared to the real values are both positive and negative, and are evenly distributed. This means that although the method cannot predict the capacity of a single solar panel accurately, it is useful for capacity prediction from several solar panels as the errors cancel out to some extent, meaning that the aggregated relative error is lowered.

The R^2 values for the different cases are close to one, meaning that the majority of the variation around the mean can be explained by the model. In addition to this, for all cases the P-value and significance F value are both equal to zero which indicates that the slope of the trendline and regression output are real results, not created by chance.

The residual standard error for the cases when all areas are examined are rather large, around 14.7 to 14.8 kW. This can be derived from the fact that the data points deviate more from the trendline for areas over 200 m^2. The residual standard error is greatly reduced when only taking into account the solar panels that are smaller or equal to this value. Subsequently, the solar panel capacity estimations were more accurate.

The used input data depends on reports from solar panel owners. This could result in errors in several ways. Firstly the solar panel capacity could be entered incorrectly. In addition to this the data does not include information about how many arrays are included in one ID, which means that more, or fewer, solar panels than identified could be included in the reported capacity. It is also possible that the data is correct, but not representative for solar panels outside of North Carolina. This however seems improbable as it is likely that the solar panels within the U.S. have similar properties. In order to confirm that the correlation is valid for other locations, a similar process could be performed for a variety of different geographical areas.

Since *2D* images are used to estimate area, the result is a projection of the panel onto the ground beneath it. This can cause an error if the studied panels have different angles. An alternative could have been using three dimensional, *3D*, models. However, having an algorithm that is adapted to *2D* images increases the amount of IDs in which it can be used since *2D* images are more common and accessible than *3D* images. The third option is to estimate the slope of solar panels. If an average slope were to be used the results would not change, merely the slope of the equation would be changed so that it cancels out the average slope of solar panels. In order to use different slopes for different panels, more information is needed.

The solar panels were located from an approximate position. This might induce an error as there can be several panels close to the location. In cases where several panels were seen at the same distance from the given location, the ID was excluded from the sample. It is however possible that some of the recorded areas could belong to other panel IDs located nearby.

Estimations of solar panel area were performed manually which induces an error to the area approximations. Manual ocular corner identification could lead to different area approximations depending on the performer. This could be corrected through comparison to objects of known sizes. The error could also possibly be decreased through having multiple people performing the identification, therefore decreasing the dependence of a single person's performance. During the area estimation, the satellite image resolution could be a source of error. In addition to this the angle at which pictures are taken can also affect the area.

The capacity per area unit as a function of installation year increases during the studied time period. If the trend continues the correlation has to be changed accordingly. This could limit the time that the obtained correlation is valid. As technical improvements and material development continue this trend could change. For example, technical innovations could speed up the efficiency increase. Another possibility is that less effective, but significantly more cost efficient materials could be used in the future. This could decrease the solar panel capacity per area unit.

It would be of interest to collect all solar panel areas for the IDs of *NCSEA* data set to verify or improve the obtained correlation. This is however a time consuming process.

5. Conclusion

There is a strong correlation between solar power capacity and solar panel area. Therefore it is possible to estimate solar power capacity from satellite images. The method is better for assessing the capacity of all solar panels within an area, rather than estimating the capacity of a single solar panel. For areas under 200 m^2 the correlation is stronger, therefore the method estimates a power capacity with a greater accuracy within this range. Technical changes to solar panels affect the correlation, therefore it has to be updated as technology moves forward.

6. References

[1] EPA United States Environmental Protection Agency, "Sources of Greenhouse Gas Emissions," [Online]. Available: https://www3.epa.gov/climatechange/ghgemissions/sources.html. [Accessed July 2016].

[2] P. Driessen, "Electricity – A basic human right," Eco-Imperialism, 12 September 2003. [Online]. Available: http://www.eco-imperialism.com/electricity-a-basic-human-right/. [Accessed July 2016].

[3] European Commission, "Renewable energy - Moving towards a low carbon economy," [Online]. Available: https://ec.europa.eu/energy/en/topics/renewable-energy. [Accessed July 2016].

[4] SEIA - Solar Energy Industries Association, "Photovoltaic (Solar Electric)," [Online]. Available: http://www.seia.org/policy/solar-technology/photovoltaic-solar-electric. [Accessed July 2016].

[5] SEIA - Solar Energy Industries Association, "Solar Industry Data - Solar Industry Growing at a Record Pace," [Online]. Available: http://www.seia.org/research-resources/solar-industry-data. [Accessed July 2016].

[6] Swedish Solar Energy, "Facts About Solar Energy," [Online]. Available: http://www.svensksolenergi.se/fakta-om-solenergi. [Accessed July 2016].

[7] Solar Mango, "What is the capacity of the solar power system I require for my facility?," [Online]. Available: http://www.solarmango.com/faq/8. [Accessed July 2016].

[8] MathWorks, "RANDSAMPLE," [Online]. Available: http://se.mathworks.com/help/stats/randsample.html. [Accessed July 2016].

[9] Microsoft, "LINEST function," [Online]. Available: https://support.office.com/en-us/article/LINEST-function-84d7d0d9-6e50-4101-977a-fa7abf772b6d?ui=en-US&rs=en-US&ad=US&fromAR=1. [Accessed July 2016].

[10] T. Leininger, "Confidence and prediction intervals for SLR," 19 June 2013. [Online]. Available: http://www2.stat.duke.edu/~tjl13/s101/slides/unit6lec3H.pdf. [Accessed July 2016].

[11] Microsoft, "T.INV.2T function," [Online]. Available: https://support.office.com/en-gb/article/T-INV-2T-function-ce72ea19-ec6c-4be7-bed2-b9baf2264f17?ui=en-US&rs=en-GB&ad=GB. [Accessed July 2016].

[12] Microsoft, "DEVSQ function," [Online]. Available: https://support.office.com/en-us/article/DEVSQ-function-8b739616-8376-4df5-8bd0-cfe0a6caf444. [Accessed July 2016].

[13] X. Wang, "Linear Regression," 30 November 2009. [Online]. Available: http://www.stat.purdue.edu/~wangxiao/stat503/notes/Ch12.pdf. [Accessed July 2016].

[14] Excel STATISTICAL Master, "Understanding Regression Output in Excel," [Online]. Available: http://www.excelmasterseries.com/ClickBank/Thank_You_New_Manual_Order/ePUB_Files/Advanced_Regression/Text/Regression_Output.html. [Accessed July 2016].

Appendix I

The following Matlab script was used in order to go from a .kml-file containing Greenwich coordinates for the polygon corners of each solar panel to the respective area, being exported to an Excel file.

```
clear all
close all

R = 6371.008*10^3; %Mean radius of Earth

table = dir('Z:\data\solarcapacity\Polygons_GE'); %NCSEA_UID
ID = {table.name}';
ID = ID(3:end);

all_poly_area = [];
area_all = [];
for j=1:length(ID)
kmlStruct =
kml2struct(['Z:\data\solarcapacity\Polygons_GE\',num2str(ID{j}),'\poly.kml'])
; %Load Polygons from KML-file
Lat_all={kmlStruct.Lat};
Lon_all={kmlStruct.Lon};

area_tot = [];
for i=1:length(Lat_all)
    Lat_tot_i = Lat_all{i};
    Lon_tot_i = Lon_all{i};
    Lat_tot_i(isnan(Lat_tot_i(:,1)),:)=[];
    Lon_tot_i(isnan(Lon_tot_i(:,1)),:)=[];

    [x_i,y_i] = grn2eqa(Lat_tot_i,Lon_tot_i); %From Greenwich coord to
Cartesian coord
    x_i = x_i*R;
    y_i = y_i*R;

    area_i = polyarea(x_i,y_i); %Polygon area in ID
    area_tot = [area_tot,area_i]; %Area of each polygon in ID

    all_poly_area = [all_poly_area,area_i];
end
area = sum(area_tot) %Sum area of polygons in ID
area_all = [area_all;area]; %Sum of area of polygons for all IDs
end
all_poly_area = sort(all_poly_area);
max_poly_area = max(all_poly_area);

%Write to Polygon area to Excel-file
xlswrite('Z:\data\solarcapacity\MAIN - Duke_PV_01292015 - with Areas &
Calc.xlsx',ID,14,'A2');
xlswrite('Z:\data\solarcapacity\MAIN - Duke_PV_01292015 - with Areas &
Calc.xlsx',area_all,14,'B2');
```

Appendix II

The following describes the content in folder *"Z:\data\solarcapacity"* of the Virtual Drive, *"ei-edl-pap1.win.duke.edu"*. This folder was used in the project, *Solar Capacity Estimation from aerial Imagery*. Created by Jacob Händestam and Therese Lundblad, July of 2016.

- The folder *Distributed Solar Photovoltaic Array Location and Extent Data Set for Remote Sensing Object Identification* contains data from:

  https://figshare.com/articles/Distributed_Solar_Photovoltaic_Array_Location_and_Extent_Data_Set_for_Remote_Sensing_Object_Identification/3385780.

  Inside this folder, the *Capacity_Calculations.xlsx* file is a short first draft of the power capacity calculations for the data set. The other files are not modified.

- The *Image* folder contains the captured Google Earth images for the solar panels of each NCSEA_UID, before the polygons were placed. For each ID the image is saved as a .jpg- and .kml file, named after each ID.
- The *Polygons_GE* folder contains the captured Google Earth images including the polygons used for area calculation. The images are also saved as .jpg- and .kml files, named *poly*.
- The *Power categories (data collection)* folder contains an Excel file for each power category. These files were used to collect polygons from Google Earth, were it was noted if the polygon was collected or why the polygon could not be collected.
- The *coordinates_to_area.m* file was created to calculate the area of the collected polygons from Google Earth.
- The *Duke_PV_01292015.xlsx* file is the original data set that is untouched. This data set contains the NCSEA_UID, coordinates, capacity and much more for each reported solar panel set in North Carolina, USA.
- The *kml2struct.m* file reads .kml files and structures the data accordingly. This script is used as a function for the *coordinates_to_area.m* file in order to extract the coordinates from the .kml files.
- The *MAIN - Duke_PV_01292015 - with Areas & Calc.xlsx* file is the main file that was used throughout the project.
  - The *MAIN* sheet has the same structure as the original data set but with a new column added, the area for the polygons of the solar panels.
  - The *Capacity per Area for year* sheet calculates the average power capacity per area unit (kW/m^2) per year and illustrates this in a chart. Also, the average increase in kW/m^2 per year is calculated.
  - The *Analysis (all without intercep)* sheet contains the analysis when the intercept is forced to zero and all IDs are used. The first set of columns have the same structure as the *MAIN* sheet, except for the IDs without areas. The second set of columns contains a summary of the linear regression run on this data set. The built in regression tool of Excel is used. The third set of columns contains the calculations for the confidence and prediction intervals and illustrates these. The

fourth and last set of columns contains a test which calculates the predicted capacity and the confidence and prediction intervals for the collected data and calculates if the predicted capacity is within the confidence and prediction intervals.

  - The *Analysis (all with intercept)* sheet contains the analysis when the intercept is not forced to zero and all IDs are used. The structure is the same as the previous analysis sheet.
  - The *Analysis (0-200m2 without inte)* sheet contains the analysis when the intercept is forced to zero and IDs less than or equal to 200 m^2 are used. The structure is the same as the first analysis sheet.
  - The *Analysis (0-200m2 with interce)* sheet contains the analysis when the intercept is not forced to zero and IDs less than or equal to 200 m^2 are used. The structure is the same as the first analysis sheet.
  - The following sheets, #1, 0-0.002 MW to #11, 0.05-1.0 MW contain the IDs which are within these capacities, a regression analysis of the data is also shown.
  - The last sheet, *Area from MATLAB*, contains the ID and area which are imported from the *coordinates_to_area.m* script. These areas are exported to the *MAIN* sheet using the vlookup function.
- The *randomizer.m* file was created in order to randomly pick out IDs which were to be analysed.
- The *Solar Capacity Estimation from aerial Imagery.pdf* file is the written report of our project.