

BENV0096: MSc ESDA Dissertation

Identifying solar photovoltaic arrays from aerial imagery

Laurence Watson

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Energy Systems and Data Analytics
of
University College London.

Energy Institute
University College London

Word count: 9976

August 28, 2019

Coursework Submission

Identifying solar photovoltaic arrays from aerial imagery

Title

Student name**Laurence Watson**

Student number (This can be found on your student identity card)**18163921**

Module title**MSc ESDA Dissertation**

Module Code**BENV0096**

Submitted to**UCL Energy Institute, Bartlett School of Energy, Environment and Resources**

Statement of authorship:

I certify that the attached coursework exercise has been completed by me and that each and every quotation, diagram or other piece of exposition which is copied from or based upon the work of other has its source clearly acknowledged in the text at the place where it appears; all field work and/or laboratory work has been carried out by me with no more assistance from the members of the department than has been specified; and all additional assistance which I have received is indicated and referenced in the report .

Signature of Student**Date Submitted:****28/08/2019**

You are reminded that the normal ESDA regulations apply when applying penalties for late submission – see the ESDA General Moodle page for details.

I, Laurence Watson, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Acknowledgements

A large number of people have positively impacted this project. Thanks go firstly to my supervisor Professor Paul Ruyssevelt for his wisdom and advice and secondly to Aidan O'Sullivan, the creator and director of this course, for his efforts and encouragement throughout the inaugural year of ESDA at the Energy Institute.

Thanks and cheers to my friends and course-mates who have supported, encouraged and inspired me during this work, in particular Clem Attwood, Simon Perez Arango, Angus Campbell, Ayrton Bourn, Daoquan Zhang, Farid Sarwari, and Xinrui Deng. Additional thanks to Damien Tanner and Jack Kelly of Open Climate Fix for their motivation and vision the journey is just beginning. I'd also like to recognise the efforts of the fine people at Azavea Labs who share their software and answer questions for the public good.

Finally, special thanks to Caitlin O'Kelly, Zach Eaton-Rosen, Patrick de Mars and Connor Galbraith for their advice, expertise and support.

Contents

1	Introduction	8
1.1	Research Question	10
2	Literature Review	12
2.1	Drivers of UK solar PV deployment	13
2.2	PV Forecasting and Nowcasting	14
2.3	Motivation for Solar PV datasets	16
2.4	Deep Learning approaches	16
2.5	Deep Learning and Solar	17
3	Methodology	19
3.1	Outline of methodology	19
3.2	Model choices	22
3.3	Evaluation metrics	23
3.4	Model Creation and Deployment	23
4	Data	25
4.1	Existing data	25
4.2	Aerial Imagery and labelled PV image datasets	29
4.3	UK Imagery	31
4.4	Data Collection and Processing	31
5	Results	33
5.1	Results using USA Distributed Solar PV Dataset	33

<i>Contents</i>	5
5.2 Results using UK dataset	34
6 Discussion	38
6.1 Implication of the results	38
6.2 Limitations of the approach	39
6.3 Recommendations	41
7 Conclusions	42
Appendices	44
A Images and Data Gathering	44
A.1 Query used for OpenStreetMap	44
A.2 UK Image Tile Locations and Ages	45
B Predictions	46
C Ethics	54
C.1 Ethics form	54
Bibliography	56

List of Figures

3.1	Methodology workflow	20
4.1	UK insolation taken from Burnett et al. (2014)	26
4.2	Distribution of PV Installations by postcode district	27
4.3	Commissioning date and size of 840,777 UK Solar PV Arrays (13.13GW)	28
4.4	Capacity of UK PV Installations	29
4.5	Counts of UK PV Installations	29
4.6	Histogram of Open Street Map solar PV label sizes	30
4.7	OpenStreetMap labels overlaid onto UK aerial images	31
4.8	Two examples of OpenStreetMap label difficulty	32
5.1	Model 1 predictions in Knowsley, UK	36
A.1	UK aerial imagery age and distribution	45
B.1	Model 1 Large and small successful UK predictions, Bristol	47
B.2	Model 1 Unsuccessful predictions at Bristol docks	48
B.3	Model 1 Mixed but many successful predictions over urban Cambridge, UK	49
B.4	Model 1 PV Predictions in urban area, Knowsley, UK	50
B.5	Detail of Model 1 PV Predictions, Knowsley, UK	51
B.6	Model 1 PV predictions after different training epochs, UK. PV in lower images never successfully identified	52
B.7	Unsuccessful prediction outputs from Model 2, UK	53
C.1	UCL Ethics Declaration	55

List of Tables

5.1	Accuracy of Model 1 on Validation Set of USA imagery	33
5.2	Confusion Matrix for Model 1 for USA PV and Background classes	34
5.3	Accuracy of Model 1 classifying UK images	34
5.4	Confusion Matrix for Model 1 classifying UK images	35
5.5	Model 1 Predictions and deployment statistics by Local Authority	36

Chapter 1

Introduction

Knowing the location of solar PV arrays is essential for several reasons: for efficient and stable grid operation, to enable better forecasting techniques, and to understand the drivers of solar deployment. This paper will investigate the extent to which solar PV locations are known already in the UK and assesses a methodology to identify and precisely map PV using aerial imagery when locations are unknown or inaccurate.

PV locations are important when systems are not directly measured. Large amounts of solar generation is embedded, meaning it is connected directly to the distribution networks rather than the transmission network. Distribution Network Operators (DNOs) are responsible for these distribution networks while the main transmission network is controlled by National Grid. Only larger PV arrays like solar farms are directly metered, so the power output of embedded solar must be modelled. This creates a challenge for the system operator National Grid additional to the variability of output due to weather. Network capacity and stress must also be monitored and the system kept in balance. While installations that wish to be connected to the grid must apply for permission and register, increasingly large private wire and “off grid” solar arrays still impact demand. As renewables market share increases, system integration costs are estimated to be £10- 25/MWh for annual penetrations of up to 50-65% ([Committee on Climate Change 2019](#)). More accurate forecasting will reduce integration costs ([Inman et al. 2013](#)), which can be informed by location data.

Increasing the accuracy of PV generation forecasts has been discussed for as long as there has been solar PV ([Colantuono et al. 2014](#), [Taylor, Leloux, Hall, Everard, Briggs &](#)

Buckley 2015, IEA 2013). One approach that can utilise advances in machine learning and statistical methods is called “nowcasting” (Saint-Drenan et al. 2011, Kuhn et al. 2018, IEA 2013). Simply, more accurate short-term forecasts could reduce the cost of infrequent high-impact mismatches in predictions and actual generation, both in cash paid to standby generators or demand response and in terms of carbon emissions with gas generators that make up two-thirds of the balancing market and a small number of very polluting and rarely deployed open cycle gas turbines and reciprocating diesel engines.

A second motivation for high spatial-resolution data is to understand drivers of deployment and geographic variation. This is of particular interest at domestic level where the majority of installations have been carried out. Understanding deployment rates in urban, suburban and rural areas may be useful for future planning, and to examine who has benefited the most from policy-driven incentives.

The context for this paper is that as of mid-2019, the United Kingdom (UK) is home to over one million solar photovoltaic (PV) arrays, making up over 13GW of installed capacity. The over 90% of these arrays are domestic and smaller than 4kW in size, although most capacity comes from around 54,000 larger scale commercial and industrial arrays with almost half of total capacity from the 457 solar farms larger than 5MW. Thanks to a generous subsidy regime, solar deployment in the UK grew rapidly from almost nothing in 2010 to 12GW in 2017 when growth flatlined as subsidies fell. Of the two main subsidies, the direct payment for generation Feed-in-Tariff (FIT) scheme ran for 9 years from April 2010, while the Renewable Obligation (RO), an obligation on suppliers to buy renewable energy, continues having reached a level of 19% for large suppliers across all renewables (Ofgem 2019b). The UK follows in the footsteps of several other countries in the adoption and subsequent removal of PV subsidies after large initial deployment rates. Germany, Spain, Czech Republic, Italy and Canada have all displayed the same pattern over different time periods (Wilson 2018).

The UK’s system operator National Grid’s lowest to highest future solar scenarios see between 2 and 17GW of additional solar by 2030 and 13-40GW by 2050 (Grid 2019), while the Committee on Climate Change sees 4GW/year of solar in their Further Ambition scenario (Committee on Climate Change 2019), matching peak rates during 2013-16,

for a “net zero” scenario now legislated in principle. Solar remains the UK’s most popular energy source, with 89% positive respondents in a UK-wide attitude tracker ([BEIS 2019a](#)). More solar power therefore seems inevitable, despite the bumpy transition for solar PV’s route to market from guaranteed 20 year FITs to new models like long-term private wire and sleeved¹ power purchase agreements and government-backed contracts for difference, a multi-technology auction to support clean power.

The central issue for the UK addressed by this paper is that the FIT database is the most comprehensive data-collection for smaller-scale PV installations, with the Renewable Energy Planning Database (REPD) focused on larger commercial projects and the Renewable Obligation database containing sparse location data. The impetus for the FIT database registry has ended with the subsidy. New PV projects currently have no motivation or obligation to register their information. The fraction of larger projects that require grid connections will be known to National Grid but their details may not be published. Other countries and regions may also have no centralised database of PV locations and this is also a significant motivator to develop this approach. This paper proposes to address this by identifying PV arrays directly.

1.1 Research Question

The following research question is therefore proposed:

Can UK solar PV arrays be identified accurately using aerial imagery and machine learning?

To answer this question the currently available solar PV datasets for the UK are explored and approaches for algorithmic identification of PV arrays from aerial imagery are examined using techniques proposed by [Malof et al. \(2019\)](#). A machine-learning model is developed and trained in a supervised approach with imagery from California, USA, and applied to UK imagery. This paper then presents and compares prediction datasets for several Local Authority regions.

¹Private wire or “behind the meter” refers to when the electricity user is located adjacent to the power generator and do not use the distribution network. Sleeved PPAs are purchase agreements via the distribution grid through an intermediary like a utility.

This subject is useful and topical given the post-subsidy post-tracking solar regime in the UK, and the expected global growth of solar worldwide. Geospatial solar PV datasets are of use to policymakers, grid operators and investors making decisions which require an understanding of the geographical distribution of solar PV, all of whom could make use of a robust methodology to produce those datasets.

The remainder of this paper is structured as follows: Chapter 2 examines previous work on spatial solar deployment trends, aerial imagery analysis, deep learning for solar detection and nowcasting. Chapter 3 presents the machine learning framework for image processing, model creation and the evaluation metrics, while Chapter 4 details the data and necessary pre-processing steps. Chapters 5 and 6 present the results and discusses the strengths and weaknesses of the analysis. Finally, Chapter 7 concludes the paper with a summary and suggestions for future work to accurately identify solar PV arrays.

Chapter 2

Literature Review

This chapter restates the major motivation for this paper and provides a brief overview of existing analysis of UK solar, drivers of deployment, and solar initiatives. This is followed by a discussion of motivations and current work on solar forecasting and nowcasting. Nowcasting is the forecasting of generation output in real-time, or near real-time, with high spatio-temporal accuracy, the better to deal with the problems arising from variable generation. Finally, deep learning approaches for image identification are considered including efforts focused on identifying solar PV with computer vision and overhead imagery.

As mentioned in Chapter 1, embedded generation connected directly to distribution networks can increase uncertainty for generation predictions. This is because a significant amount of embedded generation including PV arrays are not directly measurable by the System Operator, and must be modelled instead. The UK's System Operator is National Grid Electricity System Operator (NGESO), whose System Operability Framework ([ESO 2019](#)) notes that increased frequency response services will be needed to maintain grid frequency as renewable penetration increases. System shocks often focus attention on problems, for example a 7GW positive imbalance caused by solar PV in Germany in 2010 led to renewed interest in short-term forecasting (Saint-Drenan, 2011).

Solar met 3.4% of UK electricity demand in 2018, ([Solar 2019](#)), but due to difficulties in measuring embedded generation, is not reported specifically in government statistics, instead bundled together with other renewables. The technical performance of PV in the UK has been improving at around 1% per year ([Taylor, Leloux, Hall, Everard,](#)

Briggs & Buckley 2015), over 10% in total over 2002-2013+, so assumptions must be constantly refreshed to ensure accuracy.

2.1 Drivers of UK solar PV deployment

A government survey from 2012 (DECC 2012) found domestic PV installations are typically located in more affluent, higher energy consuming households. Areas with higher proportions of detached housing, low proportion of social or low value housing tend to have higher amount of installations. Rural areas have greater PV density per household than urban areas. Balta-Ozkan et al. (2015) found similar results with income, education level, electricity sales, irradiation and share of detached houses all having a positive impact in a thorough spatial econometric analysis of domestic installations smaller than 10kW.

An empirical analysis of 807,969 German residential solar PV systems also found socioeconomic status had an impact on PV adoption, as well as spatial spill-over effects (Dharshing 2017). Effects from urban vs suburban densities and settlement types were ambiguous. In agreement with DECC (2012) and Balta-Ozkan et al. (2015), higher income and education levels were positively correlated with PV density, along with employment rates. In the USA, residential solar deployment was also found to increase with annual household income (Yu et al. 2018). Peer effects have been found in California and in Austin, Texas, whereby higher solar PV deployment in a region resulted in higher probability of further deployment (Bollinger & Gillingham 2012, Robinson et al. 2013). This implies the possibility that neighbourly envy of visible solar PV is a motivating factor.

Many UK city authorities have set targets for PV deployment; the Mayor of London has pledged to reach 0.85GW of solar PV by 2030 and 2GW by 2050, from around 100MW in 2019 (Mayor of London 2018), although only 100MW by 2030 will come from Greater London Authority programmes under the Mayor's authority with the rest expected to come through national policies. Tools to aid deployment through evaluation of site-specific solar potential could aid the achievement of such targets, like the London Solar Opportunity Map being developed by UCL's Energy Institute (UCL 2019).

2.2 PV Forecasting and Nowcasting

The research group Sheffield Solar of the University of Sheffield, have extensively studied the variability of UK solar PV, offering a live generation estimate and a production forecast that is used by National Grid. A Sheffield Solar study identified the maximum distance between two solar PV arrays under various conditions where the assumption that their generation is correlated still holds. They found under “clear-sky” conditions PV generation could be assumed correlated at distances beyond 350km, while under “variable sky” conditions the distance was just 15km ([Taylor, Leloux, Everard, Briggs & Buckley 2015](#)).

This is important both for simulating PV on a grid and in the context of increasing solar capacity ([Taylor, Leloux, Everard, Briggs & Buckley 2015](#)). The researchers note that PV generation simulations in low and medium-voltage networks can be considered to be synchronised with the addition of some noise, due to their smaller geographical spread. One conclusion is that variable sky conditions, for example varying cloud cover and sunshine, means PV generation with a higher spatial variance and this provides some justification for “nowcasting” approaches which are discussed later in this chapter.

Due to uncertainties around PV output fluctuations, Distribution Network Operators (DNOs) oversize their networks or restrict PV integration onto the network. The UK's largest DNO, Western Power Distribution, has paused new solar PV developments in Cornwall since around 2013 due to constraints on the transmission network ([BBC 2013, RegenSW 2016](#)). The challenge of integrating PV output on power systems due to large uncertainties in generation forecasts is now widely appreciated.

Nowcasting, as defined in the opening to this chapter, has been developed for solar PV in Germany that improves on methods interpolating between metered PV installations by using satellite-based radiation assessment methods. Errors in satellite derived irradiation, which drives production, were found to have a higher variance when the diffuse fraction (the ratio of diffuse irradiation to global horizontal irradiation) was higher. That is to say, the accuracy was better in clear-sky conditions than on variable-sky ([Saint-Drenan et al. 2011](#)).

Hour-ahead and Day-ahead forecasts for PV generation are commonly used, for ex-

ample by System Operators in the USA like CAISO and MISO, and higher spatial resolution methods are being implemented or considered in many regions. Persistence forecasts are used for the smallest time steps, with whole sky imagery-based methods useful for 10s of minute forecasts, geostationary satellite imagery for 5 hour forecasts, and numerical weather predictions for up to 10 day forecasts ([IEA 2013](#)). While stochastic learning techniques without exogenous inputs i.e. learning solely from past outputs can provide good accuracy, this can be improved by including meteorological data and sky imagery. Several case studies have found forecasts over a given area have reduced error compared to a point forecast, with three showing an approximate two-thirds reduction in error rate at a country-sized area (USA, Japan), and one finding a 22% reduction in error at a distance of 100km ([IEA 2013](#)). No approach can give perfect accuracy but any improvement in forecasting is positive as it reduces all the risks outlined in Chapter 1.

Swift changes in solar irradiance at solar PV arrays caused by passing clouds can result in 50% of higher changes in generation output in seconds ([Mills et al. 2009](#)), and hourly sampling can also underestimate actual power output due to a fast response in power output to a change in irradiation, but a slow temperature response which has a limiting effect on efficiency ([Paulescu et al. 2013](#)). Nowcasting passing clouds to improve forecast accuracy is suggested with ARIMA models ([Paulescu et al. 2013](#)), and including cloud speeds into persistence models has shown to be beneficial ([Lipperheide et al. 2015](#)).

A comparison of machine learning algorithms for short-term solar forecasting found hybrid approaches to perform better than traditional numerical weather prediction models, with Artificial Neural Networks and Support Vector Machines producing the largest error reductions, beating a cloud-movement forecasting technique ([Dobbs et al. 2017](#)). Another hybrid approach successfully combined measured reference PV arrays and satellite-data-derived nowcasting methods, improving on the status quo in Australian PV systems ([Bright et al. 2018](#)). Models that can learn location-specific variances also appear promising, with strong results in precipitation nowcasting ([Shi et al. 2017](#)). Commercial PV sites are already using state-of-the art nowcasting techniques like allsky imagerbased nowcasting ([Kuhn et al. 2018](#)).

2.3 Motivation for Solar PV datasets

Detailed geo-located PV datasets are crucial for some promising approaches to PV generation forecasting. Existing methods like surveys and grid connection information are often limited in availability and spatial resolution. Accurate PV databases can also be used as inputs to algorithms to find solar PV installations in other regions, or newly built projects. A scalable and automated approach to PV detection offers an appealing route to the creation of accurate and near-complete datasets wherever overhead imagery can be sourced. Satellite imagery offers increasingly high-resolution imagery and imagery and satellites like Ikonos, WorldView-2 and 3 ([Bradbury et al. 2016](#)) and newer commercial offerings have <1m resolution and high revisit rates, opening this methodology up worldwide. Easily available PV data could have a huge impact on network planning decisions, ease of forecasting and analysis for capital deployment.

Aerial imagery has long been used to examine urban infrastructure, social structures and socio-economic conditions ([Green 1957](#), [Henderson & Utano 1975](#), [Jensen & Cowen 1999](#)). More recently, satellite imagery has also been used to study and identify human and natural features. This imagery alongside computer vision techniques could be used to identify and create a PV database to aid effective forecasting. However, it first requires models to learn to identify PV arrays. The following paragraphs outline the background for improvements in deep learning for object identification before discussing the applicability to PV.

2.4 Deep Learning approaches

Deep Learning approaches have shown significant improvement on a wide variety of problems including object classification ([Krizhevsky et al. 2012](#), [LeCun et al. 2015](#)). Deep learning approaches with aerial or remote sensing imagery has been used for building feature extraction ([Lari & Ebadi 2007](#)), including at scale using convolutional neural networks (CNNs) ([Huang et al. 2016](#), [Maggiori et al. 2017](#), [Yang et al. 2018](#)), and object detection ([Cheng et al. 2016](#)).

Domain adaptation allows knowledge gained about a source domain to be applied to a target domain, even when the domains have different distributions ([Pan et al. 2010](#)).

Transfer learning is commonly implemented to address this problem ([Pan et al. 2010](#), [Shin et al. 2016](#)). In the case of aerial imagery this applies to the problem of models learning patterns of data in one geographic region and subsequent evaluation of a different region. Differences in building and terrain type, materials and vegetation affect the performance of models deployed in different regions.

A central and recurring problem is that aerial imagery classification approaches have been found to generalise poorly across geographies ([Wang et al, 2017](#)). Transfer learning and “fine-tuning” ([Yosinski et al. 2014](#), [Maggiori et al. 2017](#)), a form of transfer learning that uses a small amount of training data from a new region to adjust the parameters of the model that has been trained on a large amount of data from an original region, have been found to be useful in addressing this issue ([Malof et al. 2019](#)) as well as careful experimental design that considers cross-site and cross-region evaluations of performance ([Wang et al. 2017](#)). This topic will be returned to in Chapter 5.

2.5 Deep Learning and Solar

Several groups have employed these techniques to identify solar PV arrays in US cities using satellite imagery ([Malof et al. 2015](#)) and high resolution aerial imagery ([Malof et al. 2016](#)), providing several deep learning frameworks for solar PV classification like “SolarMapper” ([Malof et al. 2019](#)) and “DeepSolar” ([Yu et al. 2018](#)). Both these models use convolutional neural networks (CNNs), a class of artificial neural network commonly used to analyse imagery. They also both implement transfer learning, whereby the last few layers of a pre-trained network are removed and then retrained on images similar to the desired deployment area for the model.

The SolarMapper model used a training set of 16,000 labelled solar arrays over 400km² of images over three California cities. From two-fold cross validation on the Distributed Solar Array Dataset (see Chapter 4) it yielded a precision and recall of 0.76 and 0.77 respectively. For comparison, the winning entry of a competition to classify satellite imagery of the Amazon rainforest achieved an F score of 0.93.

The DeepSolar model is a semi-supervised deep learning framework which used a training set of 366,467 images from across the USA with 46,090 images labelled as

containing solar panels, along with 12,986 validation images and a test set of 93,500 images. To obtain the USA-wide predictions, over 1 billion images were processed. It used “greedy layer-wise training” by adding another convolutional neural network to the intermediate layers of the classifier to obtain solar array outlines without supervision. It yielded a precision and recall of 0.93 and 0.89 for residential areas.

To be more explicit as these methods will be referred to in Chapter 3, SolarMapper used precise pixel-wise labels for fully supervised learning, then producing pixel-wise predictions. In contrast, DeepSolar used images only labelled as containing or not-containing solar PV as input data to train a classifier before producing pixel-wise labels in a semi-supervised fashion. Estimating the capacity of solar PV arrays has been shown to be feasible from annotated outlines of arrays in aerial imagery ([So et al. 2017](#)), and both SolarMapper and DeepSolar calculate detected PV array size estimates from their predictions. In both cases, PV technologies, degradation due to age and shape distortion due to the angle of the PV are not assessed and a constant conversion value is used. A conversion factor is discussed in Chapter 4.

Deep learning approaches often rely on supervised learning which requires large amounts of training data. Existing labelled image datasets include a solar PV array dataset for four cities in California, USA ([Bradbury et al. 2016](#)), which formed the basis of the SolarMapper model. This image dataset is publicly available and will be discussed in Chapter 4. The training set used by DeepSolar is not publicly available. Other notable PV datasets that do not include imagery are the Global Power Plants Dataset ([Institute 2019](#)), and the Open PV Project ([DOE 2019](#)). Proposed steps to construct a UK-specific training set can be found in the following chapter.

Chapter 3

Methodology

In this section an approach that uses image analysis tools is set out. The use of computer vision and overhead imagery is an appropriate methodology for solar PV detection because of its proven success across multiple domains including infrastructure detection ([Camilo et al. 2017](#), [Malof et al. 2016](#), [Huang et al. 2016](#)), and the relative ease of scaling the technique to large areas. An alternative method would be to scrutinise images by hand, either as a discrete task or by appealing to crowdsourcing approaches like Open Street Map. Hand-labelling has been used to create datasets like the Distributed PV Dataset ([Bradbury et al. 2016](#)), at a significant time cost. Using exhaustively hand-labelled image datasets as the starting point for a computer vision approach greatly enhances the value of that effort and this was the stated motivation of the datasets creators. A trained and calibrated model could then produce new datasets quickly and cheaply.

Machine learning in computer vision has benefited from a myriad of shared and open neural network architectures for many different purposes. While there is no firm consensus on best practice, empirical results and competitions have directed attention towards convolutional neural nets (CNNs), a type of artificial neural network which have shown very positive results in object detection in images ([Long et al. 2015](#)). CNNs will therefore be used in this methodology.

3.1 Outline of methodology

The methodology follows the best practice used in the RasterVision geospatial analysis framework ([Azavea 2019b](#)) and is shown in Figure 3.1. It begins with a dataset of la-

belled and georeferenced overhead imagery and divides it into smaller images of a fixed dimension, typically 300 pixels by 300 pixels, along with any labels applicable to the image. This small image segment is commonly called a “chip” and will be referred to as such throughout. The label can be treated as an array with the same dimensions as the image. For two categories of label, i.e. PV or not-PV, the array values are binary, but multiple categories of labels and prediction are possible. The labels identify whether the chip contains any PV arrays or parts of an array and specifies the pixels exactly. Vector labels are rasterised and combined with images into scenes, then randomly divided into training and validation sets with an 80:20 split. The validation set was also used as the test set for model evaluation.

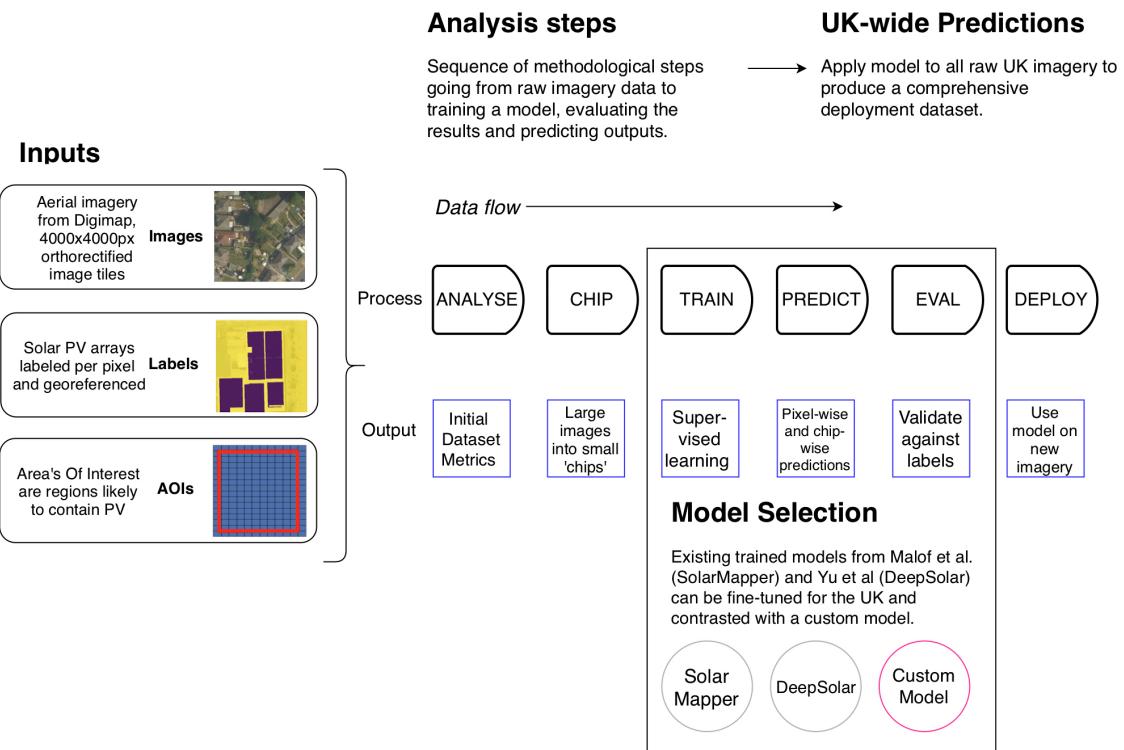


Figure 3.1: Methodology workflow

Several detection tasks are now possible. First and most simply, these positive (containing PV) and negative (no PV) chips can be used to train a classifier with a binary output for a given chip, described as chip classification. Secondly, precisely labelled training data with pixel-level classes can be used to do semantic segmentation, that is labelling every pixel in the image, where the result is a pixel-level prediction of desired

classes. This is sometimes referred to as mapping ([Malof et al. 2019](#)).

Two sets of image chips were created, one using the Distributed Solar PV Dataset images and labels, and one created from OpenStreetMap labels ([OpenStreetMap contributors 2019](#)) and corresponding UK image tiles. Additionally, for evaluation purposes, a collection of 372 UK imagery chips were hand selected and labelled as containing PV or not, with one third rural, one third urban with no PV and one third urban containing PV. These two sets of image chips are used to train two models, Model 1 and Model 2, trained on USA and UK data respectively.

Both models used the same neural network design. U-Net, a type of CNN ([Ronneberger et al. 2015](#)), that has been used successfully for overhead image segmentation ([Dong et al. 2019](#)) was used as a base model using the FastAi plugin for RasterVision ([Azavea 2019a](#)) with PyTorch ([Paszke et al. 2017](#)), a computer vision package, and using model architecture from Resnet18 ([He et al. 2015](#)). Model 2, a UK-specific model was trained using overhead imagery and 9,508 labels taken from Open Street map, with all matched image tiles being more recent than 2012. This resulted in 1763 image tiles comprising 4000 by 4000 pixels being selected. These were then processed into labelled image chips and divided into a training and validation set for semantic segmentation in the manner described in Chapter [5](#).

PV areas are calculated using a simple coefficient to convert from pixel count to installed capacity. PV panels in the UK usual come in a standard size and power output. A 300W PV panel in a standard size measures 1m by 1.65m. This means a square metre of PV has a power rating of 182W. Area distortion due to roof angle, PV technology, panel degradation due to age and panel orientation are not considered. This yields a simple formulation for PV array capacity C in terms PV array area A , the PV area calculated from the pixel count and scaling of aerial imagery used, with coefficient γ (following notation in ([Malof et al. 2019](#)):

$$C = \gamma A \tag{3.1}$$

Where γ is $181.82W/m^2$ from the values used above.

3.2 Model choices

The batch size is the number of training instances (that is, images with associated labels) seen by the model during a weighting update step. Training is done through stochastic gradient descent, which calculates a parameter update by computing the loss function over gradient of the batch. A small batch size means updates will be sufficiently noisy to avoid saddle-points and local minima in the gradient descent, while not being too small to be unrepresentative, and with a fraction of the computation effort. A batch size of 8 is used in this analysis to balance these considerations.

An epoch is complete when the model has seen every training instance. The number of epochs essentially mean the amount of training undertaken. “Early stopping” is used, which means training is halted when the validation loss stops decreasing. This helps to prevent over-fitting. The main model used for predictions in this analysis trained for 500 epochs and results are compared with the same model trained for 20 and 100 epochs in Chapter 5. All training was carried out using a Titan Xp 12GB single GPU, and training and experiments balanced overall compute time necessary against model rigour. 500 epochs seemed a good balance with the levelling-off of validation loss and compute time.

Solar PV is a small fraction of area on the ground and of the pixels in aerial imagery. This is known an imbalanced class, with far more background than PV. This is a problem as if the training and validation data contains a similarly unbalanced proportion of solar PV, high accuracy can be obtained simply by predicting no PV in any sample. This can be addressed by balancing the training and validation sets either by over-sampling PV data or under-sampling the background class, the latter being used for this method. Training data is augmented with vertical and horizontal flips and random crops of the images. Images are normalised to the same mean and variance in pixel intensity values as ImageNet, which the base model is trained on.

Techniques considered but not used were semi-supervised methods which have been shown to deliver improvements across many problems ([Berthelot et al. 2019](#)), including on solar PV identification ([Yu et al. 2018](#)), however all things being equal a fully-supervised approach should perform better if data is available. Semi-supervised approaches can be useful to deal with uncertain datasets and this is discussed more fully

in Chapter 5. Two models are created to compare results and assess the viability of using OpenStreetMap labels to create training data for the UK.

3.3 Evaluation metrics

Evaluation of the outputs is done by calculating the precision (the ratio of correct predictions to positive predictions) and recall (the ratio of correct predictions to positive samples) of predictions. These two metrics and their weighted combination into the F1¹ score is a common measure for image-based object detection ([Cheng et al. 2016](#), [Malof et al. 2015](#)). This can be done per-chip for classification, or per-pixel depending on the task. Typically research has balanced the trade-offs between precision and recall and sought to maximise the F1 score ([Malof et al. 2019](#), [Yu et al. 2018](#)).

The impact of a PV false positive and false negative in this study is the same. If the count and capacity of PV predictions are normalised against a known upper limit, for example official statistics, this will just change the spatial distribution of predicted arrays. Therefore maximising F1 score is the right approach here. Other experiments should consider this when deciding on evaluation metrics, for example when the underlying statistics are unknown. State of the art could be considered to be precision of 0.93 and recall of 0.89 in residential areas with a semi-supervised method ([Yu et al. 2018](#)), although the training data used is not available. A precision of 0.77 and recall of 0.76 was obtained using the same dataset Model 1 will use for training ([Malof et al. 2019](#)).

3.4 Model Creation and Deployment

Once Model 1 and Model 2 are created and calibrated, they are tasked to identify PV in imagery previously unseen by the models. Seven UK Local Authority regions are selected on the basis of a high concentration of PV installations, a high number of labels on Open Street Map, and large proportion of urban development, and image tiles corresponding to each region are selected from the UK imagery dataset ([Digimap 2019](#)) and input to the model.

Predictions for each region are obtained in the form of pixel masks. These can be

¹The balanced F_1 score here is given by $F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$

seen in Chapter 5 and in Appendix B. Data from the predictions can be obtained from the pixel masks, which are numerical arrays of binary values. If firm cluster boundaries are not present, image processing methods can be used for thresholding (Otsu 1979). To turn pixel-level predictions into PV array counts and sizes, the number of connected-components, that is, adjacent pixels of the same class, are then counted and their area stored. A formula to convert pixel sizes to PV array capacity is shown in the following chapter. From this a count of PV array predictions and the corresponding capacities are created. The resulting predictions are compared to database statistics in Chapter 5.

Chapter 4

Data

The datasets used in this paper fall broadly into two categories. The first is data relating to UK solar deployment from UK government sources like the Department for Business, Energy and Industrial Strategy (BEIS). The second consist of geo-referenced imagery data and image labels for the purpose of training and refining a machine learning model for solar PV identification. Once a model has been created, it can be tasked with identifying new solar PV arrays that are not precisely mapped in existing datasets. However this relies on availability of overhead imagery from which to make predictions.

An aside on imagery definitions. Aerial imagery is used throughout this paper and usually refers to images created by planes, drones or other surveying craft. Overhead imagery ([Tanner et al. 2009](#)) incorporates imagery from aerial sources and satellites. Satellite imagery has historically been of much lower resolution and of less use for computer vision techniques but has improved in the last few decades as discussed in Chapter [2](#).

Briefly before the datasets are examined, it is useful to note the spatial distribution of solar insolation on the UK, that is, where the sunlight provides the most power over the year, as shown in Figure [4.1](#). [Burnett et al. \(2014\)](#) show climate change will increase southern solar insolation while decreasing in the northern parts of the UK.

4.1 Existing data

The most significant and comprehensive publicly available datasets of UK solar PV are the Feed-In-Tariff (FIT) database ([Ofgem 2019a](#)) and the Renewable Energy Planning Database (REPD) ([BEIS 2019b](#)), along with the official Solar Photovoltaics Deployment

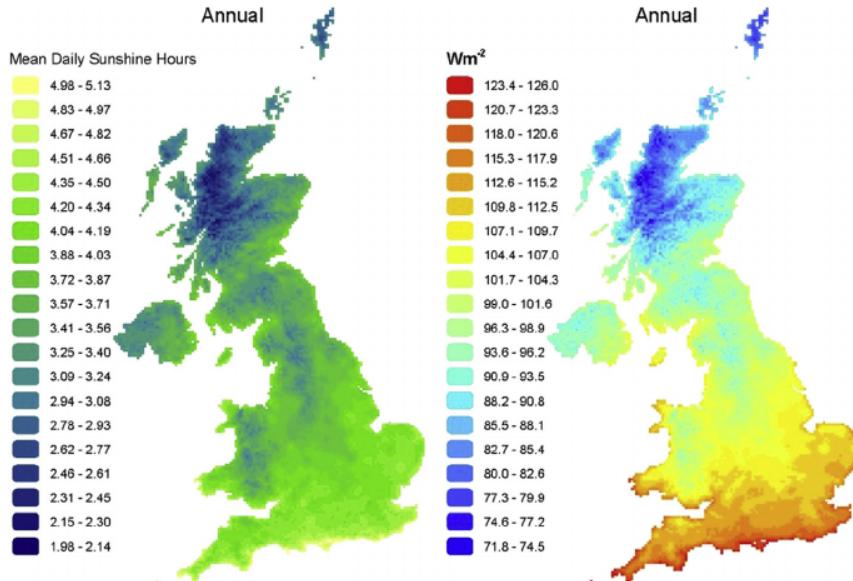


Figure 4.1: UK insolation taken from [Burnett et al. \(2014\)](#)

monthly figures ([BEIS 2019c](#)). The Renewables Obligation (RO) database contains PV arrays the others do not, but provides very little location information and where it does, often supplies the address of company owners of the sites rather than the PV arrays themselves. The Solar Photovoltaics Deployment statistics make use of the RO registry and can be treated as the best source of aggregated ground truth information, but does not give any information at the level of the individual installation. Therefore for this study the FIT and REPD databases will be used as they provide installation-level data like postcode, capacity and commissioning date.

The location of around 60% of solar capacity is known accurately in the REPD, mostly comprising larger utility-scale solar farms and industrial rooftops. Point coordinates supplied in the REPD are typically accurate to a hundred meters from a small-sample visual inspection. This study then will focus on the remainder of UK PV which makes up the long tail of smaller capacity arrays. The location of the installations in the FIT database is generally given at the level of postcode district, of which there are around 2,800 in the UK, with an average area of 85km^2 . These two sources can be combined to the coarser spatial level of postcode district and is shown in Figure 4.2.

FIT tariff data specified in the REPD was used to identify and remove 56 installations making up 224.9MW that were duplicated between the FIT database. The Renew-

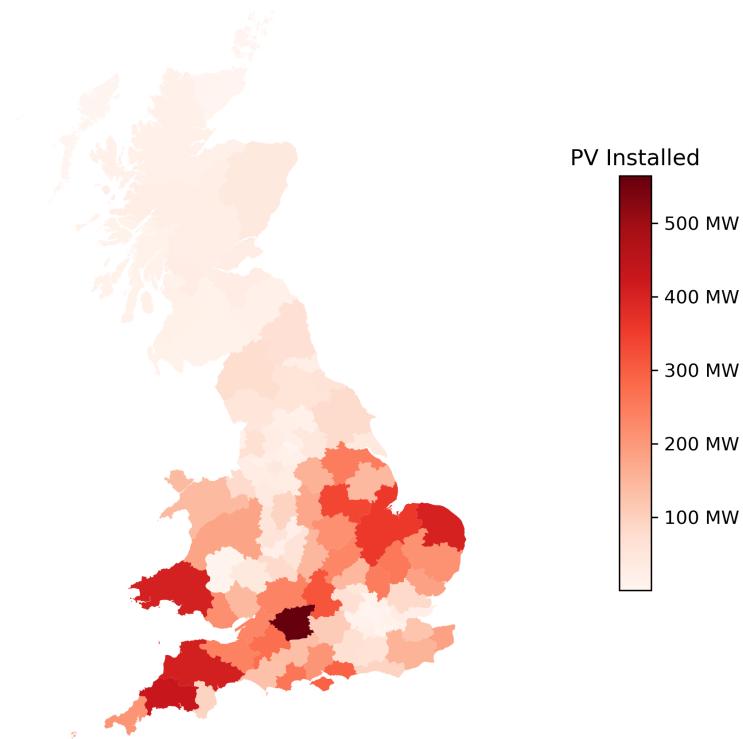


Figure 4.2: Distribution of PV Installations by postcode district

able Energy Planning Database (REPD) provides coordinate locations for around 8GW of solar PV capacity. The total number of operational installations in the two databases was 840,777, with a total capacity of 13.13GW. Considering the Solar Photovoltaics Deployment statistics, this leaves approximately 164,021 installations totalling 116MW un-accredited and left out from this paper, or 2% of the UKs total capacity and 16% of total installations.

Both FIT and REPDB databases contain commissioning dates for PV installations, which allows an examination of deployment over time as shown in Figure 4.3. Strategic deployment sizing can be seen both in array size to hit upper bounds of FIT qualification levels, and in time to ensure arrays received the highest FIT levels before each reduction. The banding at 30kW in Figure 4.3, marked with *, could be because below that array size, the FIT is paid by a deemed percentage of generation, rather than directly measured. This could provide a financial incentive to install arrays up to this size if the deemed amount was more than the alternative. The reason for banding in the chart at 100kW, marked †, is less certain, but above that threshold export meters must be directly approved ([Ofgem](#)

2018) and some utilities use the 100kW level as a threshold for differentiating supply contracts ([Energy 2016](#)). A distinction should be drawn between domestic, commercial

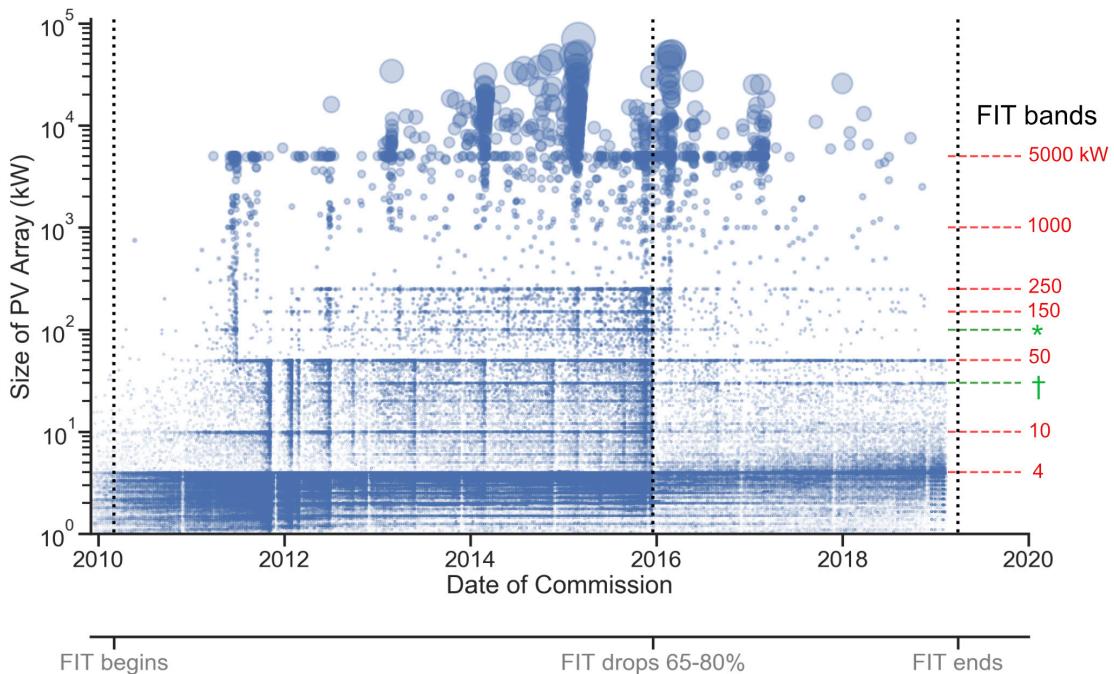
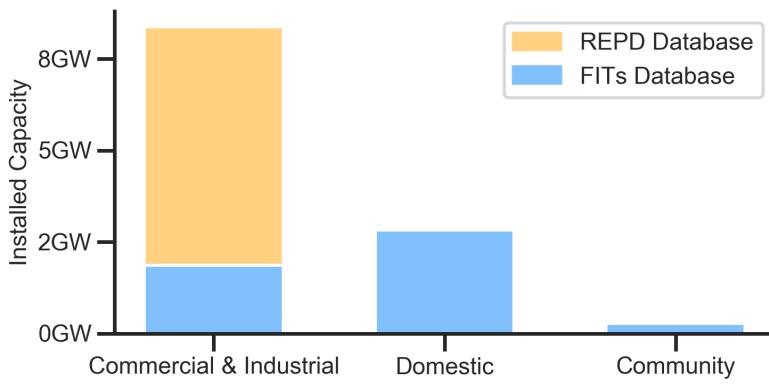
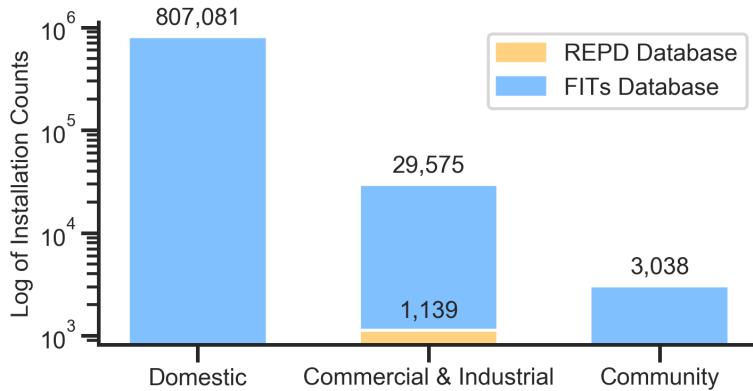


Figure 4.3: Commissioning date and size of 840,777 UK Solar PV Arrays (13.13GW)

and industrial, and community PV installations. Domestic installations tend to be of small size, less than 50kW, while commercial arrays are larger and make up all the very large sizes. Community installations are between these two but make up a small number of installations overall. The motivations and access to capital and space of these groups may differ substantially, and drivers of PV deployment could reasonably be expected to differ substantially. When investigating spatial and econometric trends it is therefore reasonable to treat them separately. It could be naively assumed commercial installations would be sited in the South of the UK to benefit from maximum solar insolation, taking into account planning and grid constraints, as commercial developers can explore more locations than homeowners wishing to develop on their own land. This will be explored in Chapter 5. As seen in Figures 4.4 and 4.5 and outlined in Chapter 1, the majority of installations are small and domestic, and the minority and largest source of capacity are commercial and industrial.

**Figure 4.4:** Capacity of UK PV Installations**Figure 4.5:** Counts of UK PV Installations

4.2 Aerial Imagery and labelled PV image datasets

Two approaches are considered for the supervised training of a machine learning model to identify solar PV. Firstly using existing work and secondly by constructing a new UK dataset.

The Distributed Solar Photovoltaic Array Location and Extent Data Set for Remote Sensing Object Identification, referred to as the Distributed Solar PV Dataset ([Bradbury et al. 2016](#)), provides a thorough catalogue of aerial imagery and exhaustive PV array labels, exhaustive meaning all existing PV arrays are labelled as far as the authors are aware. The dataset version used in this paper contains 19,863 PV labels. The dataset was created by hand-labelling 601 high resolution images covering four Californian cities. Two annotators were used for each label and their data entries merged to ensure accuracy. The three main sources of error were missed arrays, false positives and incorrectly drawn

labels, and these errors will be applicable to a UK dataset as detailed in Chapter 5.

No such labelled dataset currently exists for the UK, but by combining overhead aerial imagery of the UK ([Digimap 2019](#)) with crowd-sourced solar PV generator labels ([OpenStreetMap contributors 2019](#)), this paper will explore the possibility of creating one. Aerial imagery also provides the means to geolocate the bulk of UK PV, by using it as an input to a trained PV detection model and examining the predictions made on it.

OpenStreetMap (OSM) has 9,511 labelled solar PV installation geometries (that is, defined pixel labels, known as “ways” by OSM), and over 30,000 point locations. The labels will be used in this study, and the points left out. As seen in Figure 4.6, the modal array size is between $50\text{-}75m^2$. Using the conversion coefficient γ described in Equation 3.1, the modal array capacity is therefore 9-14kW. This is two to three and a half times larger than the modal domestic installation, and not perfectly representative. However, some of this discrepancy may come from oversized label areas.

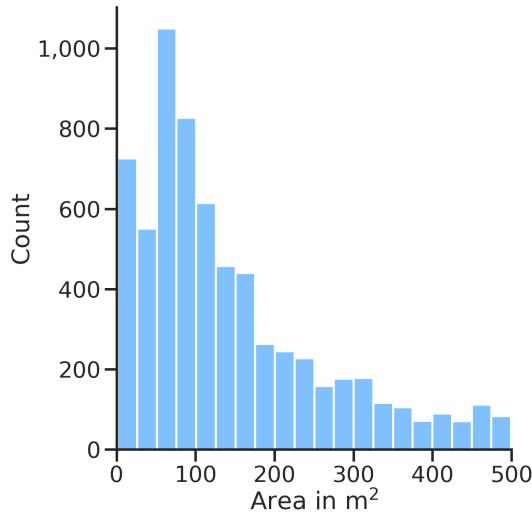


Figure 4.6: Histogram of Open Street Map solar PV label sizes

This resource is the best publicly available for PV geospatial data labels in the UK. Worldwide, OpenStreetMap contains over 240,000 solar labels. However, unlike the Distributed Solar PV Dataset, locations have not been covered exhaustively. OpenStreetMap data is known to be imperfect for the purpose of training machine learning datasets ([Maggiori et al. 2017](#)), with a proposed solution being to train on the whole set of imperfect labels, then fine-tune on a small set of manually corrected ones. This is similar to the

transfer learning approach of calibrating PV detection models on different geographies as mentioned in Chapter 2.

4.3 UK Imagery

Aerial imagery was obtained for the UK ([Digimap 2019](#)), comprising of 4000 by 4000 pixel georeferenced orthorectified images supplied in British National Grid coordinates (EPSG: 27700), with a resolution of 25cm per pixel. Image age varied over the UK as shown in Appendix A, Figure A.1. Images from 2015 and 2017 were preferred in this analysis, in order to maximise the chance of PV arrays being present on the imagery. However, this could not be guaranteed and presents a source of error. Labels in Figure 4.7 are successful with the exception of the second topmost label which stretches across two gaps between PV arrays and the top rightmost is misaligned. Labels in Figure 4.8 A show missing PV where images predate labels, while in B labels cover whole building footprints, not just the PV part of the rooftops.



Figure 4.7: OpenStreetMap labels overlaid onto UK aerial images

4.4 Data Collection and Processing

Geospatial shapes containing PV arrays were obtained from Open Street Map (See Appendix A for the query used). PV label shapes were combined with image tiles through a geospatial overlay, taking the intersection of the PV label shape and the image tile shape.



Figure 4.8: Two examples of OpenStreetMap label difficulty

Taking the intersection is necessary for labels that cross image tile edge boundaries. The label intersections were converted to geometry masks to enable visual assessment to check label accuracy.

The strength of the data sources assembled for this study lies in both the quantity of image data and labels, both of which are necessary for approaches using neural nets. Government-produced statistics also give a high level of confidence to regional deployment numbers. These both give a firm basis for the approach to identify solar PV. Limitations include the age of images, and the accuracy and completeness of labels. This will be corrected for by selecting more recent images and using accurate labels as far as is possible, but even given these limitations there seems clear value and potential in improving PV location accuracy.

Chapter 5

Results

5.1 Results using USA Distributed Solar PV Dataset

Model 1 was trained using the methodology in Chapter 3 on USA images from the Distributed Solar PV Dataset described in Chapter 4. It was first evaluated on a validation set of USA images from the same dataset. Secondly Model 1 was tested on a different dataset made up from data from across the UK. The accuracy results for this first test are shown in Table 5.1. The background class results are included for completeness but discussion will focus on the PV class results. A confusion matrix is a results table commonly used to show absolute values of classification outcomes. In a confusion matrix, the total of all cells sum to the number of samples. Correct predictions can be found on the diagonal of the table, and incorrect predictions found in off-diagonal cells. The confusion matrix for the first evaluation are shown in Table 5.2.

Class	f1	precision	recall
PV	0.6948	0.8273	0.6324
Background	0.9998	0.9997	0.9998

Table 5.1: Accuracy of Model 1 on Validation Set of USA imagery

As shown in the tables above, precision was similar to but not as good as the reported results of the SolarMapper algorithm, which achieved precision, recall and F1 of 0.76, 0.77 and 0.76 respectively (Malof et al. 2019), and weaker than those reported by Yu et al. (2018) for PV identification.

Precision was higher than recall, meaning the model rarely mislabelled other objects

	Actual PV	Actual Background
Predicted PV	1,082,916	629,606
Predicted Background	448,652	2,313,838,826

Table 5.2: Confusion Matrix for Model 1 for USA PV and Background classes

as PV arrays, but was less effective in identifying all PV arrays present. [Malof et al. \(2019\)](#) employed multi-fold cross-validation over each city in the dataset when training, an approach that was not repeated here, which could account for the lower accuracy results.

5.2 Results using UK dataset

Model 2, the UK-specific model, was not successful. While training loss and validation loss were observed to reduce during training, predictions when visually examined showed no learning of PV-specific features. Precision and variance never exceeded 2% for the PV class. A sample prediction showing this can be seen in Appendix B, Figure B.7. Two sources of error were identified that contributed to this result. Firstly background class training images were contaminated by large numbers of PV pixels which were missing labels and misclassified. Secondly, pixel labels were not precisely accurate, often including roof sections or suffering from misalignment, as shown in Chapter 4. This issue and potential solutions are discussed further in Chapter 6. Treating the labels as unreliable data to classify image chips rather than pixel-level labels could yield better results. Given inaccuracy, Model 2 was not used for further predictions, as it demonstrably would not address the questions of this dissertation.

Model 1 performed a classification task using image chips from UK aerial imagery selected by hand with and without PV, ensuring a mix of rural and urban images. The resulting accuracy is shown in Table 5.3 and the confusion matrix in Table 5.4.

Class	Precision	Recall	F1
PV	1.0	0.589	0.741

Table 5.3: Accuracy of Model 1 classifying UK images

All PV predictions are correct, likely due to two factors. Firstly, hand-selecting im-

	Actual PV	Actual No PV
Predicted PV	123	0
Predicted No PV	86	163

Table 5.4: Confusion Matrix for Model 1 classifying UK images

age chips that clearly contained PV strongly biased the sample towards easily identifiable PV samples, and secondly, even a small number of pixels predicted as PV in the image chip was counted as a positive identification. That is, even a prediction with low confidence in identifying a PV array from the model was counted positively. Recall is lower than in other reference models because of this.

Model 1 was tasked with a semantic segmentation task on the same image chips, producing pixel-wise labels to investigate predictions qualitatively. These can be seen in Appendix B, Figure B.6, and suggest more training could have improved results. Model 1 seems more success in predicting with solar farms with larger PV arrays than smaller rooftop arrays in these images.

Predictions were made on recent UK imagery covering sets of Local Authority areas, primarily city regions, selected on the basis number of PV installations according to the combined REPD and FIT databases, and date of imagery being at least as recent as 2015. Pixel-level segmentation map outputs were assessed using connected-component analysis with a simple area filter of 10 pixels to remove any pixel groups that could not reasonably represent a PV module, in contrast to the previous classification task. This enabled an object count, and a total area, to produce a count of PV arrays predicted and their total capacity within the area. These results are contrasted against the capacity and counts from the official statistics and shown in Table 5.5. Due to the lack of a labelled dataset to verify predictions, precise statistics could not be generated as in the previous tasks. Results must instead be assessed qualitatively by examining image predictions. Results are also compared to regional PV deployment data from the original combined dataset of the FIT and REPD.

These results clearly have low predictive power, with large over and under-estimates of both installation count and of installed capacity. Predicted mean and median array sizes vary less than predicted counts and capacities, but the inaccuracies in the latter mean little

Local Authority	Count		Capacity (MW)		Mean Size (kW)		Median Size (kW)	
	Predicted	Actual	Predicted	Actual	Predicted	Actual	Predicted	Actual
Bristol	2423	3131	27.33	11.14	11.28	3.56	1.97	2.76
Cambridge	3764	999	13.43	3.47	3.57	3.47	0.95	2.76
Canterbury	31	1499	0.47	5.82	15.26	3.89	6.35	3.38
Knowsley	518	2423	1.52	7.57	2.93	3.12	1.45	2.88
Liverpool	277	1752	1.2	6.57	4.34	3.75	1.39	3
Manchester	673	3132	2.62	10.33	3.89	3.3	1.91	2.7
Peterborough	906	5203	4.1	17.22	4.52	3.31	1.62	2.55

Table 5.5: Model 1 Predictions and deployment statistics by Local Authority

confidence can be had in these estimates. Median array size is underestimated with the exception of Canterbury, for which only a small number of arrays have been predicted.

Investigating these semantic predictions of UK imagery using Model 1 which was trained on USA images, it can be seen that the model does identify PV and is working in principle, as shown in Figure 5.1. The model has predicted two rooftop PV arrays in the centre, predicted a few pixels of an array in the lower part of the prediction (circled on the left), and avoided making predictions on several glass conservatories. However three PV arrays (circled on the right) are not identified by the model.



Figure 5.1: Model 1 predictions in Knowsley, UK

Exhaustive labelled ground-truth data for a given region does not currently exist for the UK. For example using OpenStreetMap labels for verification would not provide good

measures of accuracy. False positives could not be identified as the dataset is not complete. Instead, images with large numbers of predictions in Bristol, Cambridge and Knowsley were qualitatively assessed and are shown in Appendix B. An image of Bristol showed the entirely successful identification of a large solar farm and of some small installations (Figure B.1). One clear issue this highlights is that a solar farm may be counted as a large number of small arrays when compared to official statistics. An image with high prediction count from Cambridge showed a large number of predictions over an urban area including some apparent true results and a large number of false positives, including the misclassification of glass rooftops (Figure B.3). Most predictions were appeared reasonable, although a notable mistake was the misclassification of crates at a city docks (Figure B.2). Some representative prediction images are shown of Knowsley, UK in Figures B.4 and B.5.

While Model 1 performed accurately compared to other work on USA images similar to its training data, it cannot be said to have performed well on UK data. The significant errors found in a qualitative assessment strongly suggest reduced accuracy. This is likely to be due to differences in terrain and building composition of UK cities in relation to the USA training data, and in the areas of interest used for predictions. The method has nonetheless been demonstrated and shown to work in principle. Possible improvements to the model for greater accuracy and effective verification are discussed in the next chapter.

Chapter 6

Discussion

6.1 Implication of the results

The accuracy of results from Model 1 in identifying USA PV arrays confirms the capability of using machine learning and aerial imagery to perform this task. Results obtained from Model 1 on UK data suggests the approach does generalise to UK data, although further work is needed to improve model accuracy. Furthermore the experience with Model 2 trained with UK data and labels from OpenStreetMap show that crowd-sourced labels alone may not be suitable for a fully supervised learning approach. However there are several promising routes to improve the methods in this study which will be discussed further on in this section.

Regarding the specific performance of the model, Model 1 showed false positives on urban areas, particularly on rooftops with a grid-like pattern, certain roofs with large glass sections and of particular image features like stacked crates (Figure B.2). This is not to say all rooftop glass was miscategorised: false positives on conservatories were not observed. Model 1 was more successful in identifying some types of PV arrays, performing worse in urban environments from a qualitative analysis. This could be due to more varied rooftop features in urban areas. Some arrays missed were adjacent or very close to ones detected, which could indicate the potential for improvement with further training. Large solar farms were effectively identified, but this is of limited value for the UK as these installations are well-known.

Instances of predicted pixel clusters making up only part of a PV array were ob-

served. This could indicate insufficient training time. This work had to balance training the model with the time and computational resources needed but future work could invest more heavily in compute resources. Aerial imagery at 25cm resolution and a convolutional neural network approach was shown to be a successful combination to achieve results, reflecting existing research ([Yuan et al. 2016](#)), while other network architectures not used here due to time and compute constraints are left for further study.

The failure of Model 2 to produce reasonable results after training on the UK dataset suggests that these labels alone are insufficient within a fully supervised approach. [Maggiore et al. \(2017\)](#) suggests a two stage process to address this. Firstly, initialising a classifier on imperfect training data and then refined with precise labels. OpenStreetMap issues could be sidestepped if used as the basis for classifying an image chip in this way instead of labelling each pixel in the chip. [Yu et al. \(2018\)](#) also notably used this approach with an extremely large dataset, which could explain their improved accuracy. If image data were available, the 270,000+ PV labels globally on OpenStreetMap could be a powerful resource to follow their approach worldwide.

6.2 Limitations of the approach

The approach suffered from lack of a UK validation set. Predictions on new imagery could not easily generate evaluation statistics as ground-truth data was unknown. For the classification of UK chips the image classes were known but the process of choosing those chips did not use a rigorous process for selection, biasing the experiment. A better process could have been randomised selection of a mix of image types. Alternatively, several image tiles could have been chosen and then PV annotations applied rigorously by hand to create a small but accurate validation set from which to create image chips. However this could be prone to errors. This serves to highlight the value of work like the Distributed Solar PV Dataset, which is a wide portfolio covering and allowing training on varied landscapes with exhaustive labels. Several papers have found model accuracy reduces significantly when deployed on new geographic regions that contain differences from the training data ([Wang et al. 2017](#), [Malof et al. 2019](#)). This was also found in this paper, although the effect was not assessed rigorously due to the lack of a UK validation

set. However the Local Authority predictions suggest substantial reduction in accuracy. Neural network models are difficult to interrogate, but the literature suggests contextual differences in PV siting, building arrangement, materials and type, and vegetation and surface features could all play a part in the model learning to identify PV arrays. This could be investigated with imagery from very different-looking geographic regions.

This study made no effort to identify PV characteristics beyond size, which was done with a one-size-fits-all approach using a constant factor to convert from pixels to capacity. PV technology, for example mono- or poly-crystalline and array orientation and angle are not considered. PV orientation and angle could potentially be assessed by modelling the roof shape on which some arrays are situated. Overhead imagery containing a wider range of spectral bands could prove fruitful for assessing PV technology type. Counting and sizing arrays accurately from predicted data needs careful consideration. Numerous arrays in a field should be categorised as a single solar installation, but in an urban environment should be counted separately, for example with solar farms and a housing estate equipped with PV. OpenStreetMap data addresses this problem by grouping multiple labels together, called a “relation”. This is helpful when comparing predicted counts with official statistics. This could be a topic for a separate supervised learning approach on top of the work in this paper.

Avoiding large solar farms that are already known could be useful, for example to focus training on smaller urban PV. A ready solution is to appropriately use Areas Of Interest to select or mark for exclusion regions with known large solar arrays when selecting training data or when selecting images for predictions. Separating urban and rural PV could have produced more interesting accuracies and revealed the strengths and weaknesses of the model more clearly in this paper. Calibrating array size to pixel counts could yield better estimates of capacity. This could be done empirically against known array sizes and include PV characteristics mentioned above. Improved PV power output predictions, a motivation for this paper, could also be improved by data on array angle and orientation, adding further impetus to identify those characteristics.

6.3 Recommendations

As the approach was successfully demonstrated in principle, this paper recommends its continued development to create accurate PV location datasets. In particular to obtain satisfactory accuracy, two additions are suggested from existing studies. Firstly the use of fine-tuning of existing models on a small amount of regional data seems to improve accuracy. The cost of creating a whole new training dataset is minimised, instead using small set of labelled images. Secondly, a semi-supervised approach where crowd-sourced or otherwise imperfect imagery labels can be used to train a classifier, which can then be used to produce precise locations and sizes of PV arrays. In this way, the value of crowd-sourced databases like OpenStreetMap can be maximised for training data. Forecasting improvements augmented by PV locations could be deployed everywhere there is solar PV and the image data necessary. This lends weight to further development of this approach as it could have a significant impact worldwide.

Chapter 7

Conclusions

This paper set out a methodology to identify solar PV using aerial imagery for the UK using computer vision techniques. This approach has been demonstrated in principle. The research question should therefore be considered as answered positively. Existing work from object identification with computer vision and PV datasets has been brought together successfully. The development of this approach could yield the significant prize of worldwide solar PV identification, to the improvement of forecasting, network planning and policy design.

However, significant improvements in accuracy are needed to identify PV currently missed and avoid mis-categorisations. Variation in image source, resolution and quality could lead to poor generalisation of models. And geographic variability provides a barrier to easily deploying models on new regions, as demonstrated in this work. Crowd-sourced labels are a novel and significant source of data, if used appropriately. This work makes clear it would be wise not to treat them as precise ground-truth labels, instead to classify image regions into classes, as proposed in [Maggiori et al. \(2017\)](#), and implemented by [Yu et al. \(2018\)](#). The methods in this paper assess PV location and size. Further work is needed to infer other PV characteristics which may need different approaches.

This papers results and accuracies from the literature imply conclusions can be confidently drawn from prediction results when models are calibrated. Furthermore, the identification of other features from overhead imagery, whether for energy studies or the built environment seem feasible. Researchers should note that energy infrastructure like power stations, transmission lines or embedded renewables like onshore wind could be identified

with this approach, as well as smaller urban features like conservatories.

As data availability increases and some satellite imagery sources approach aerial images in resolution, the value and impact of this work is greatly enhanced. Aerial imagery is not widely available for most regions and infrequently updated due to its cost. Satellite imagery holds the possibility for rapid revisit rates and a global scope for identification methods. Computational methods as described in this paper scale easily to very large datasets, assuming adequate compute resources are deployed. To conclude, this paper has shown PV can be identified from UK aerial imagery. A method has been demonstrated, several avenues to increase model accuracy have been described, and the potential to scale the approach discussed, taking into account the limitations of the techniques and data used.

Appendix A

Images and Data Gathering

A.1 Query used for OpenStreetMap

OpenStreetmap was accessed at “<http://overpass-api.de/>” ([OpenStreetMap contributors 2019](#)) The query used to generate solar PV electricity generating arrays was as follows:

```
[out:json];
area["name"="United Kingdom"]->.searchArea;
(
node(area.searchArea)["generator:source"="solar"]\
['generator:output:electricity'='yes'];
way(area.searchArea)["generator:source"="solar"]\
['generator:output:electricity'='yes'];
relation(area.searchArea)["generator:source"="solar"]\
['generator:output:electricity'='yes'];
);
out body;
>;
out skel qt;
```

Query results were then filtered just for “ways” which are polygon shapes, that is, areas.

A.2 UK Image Tile Locations and Ages

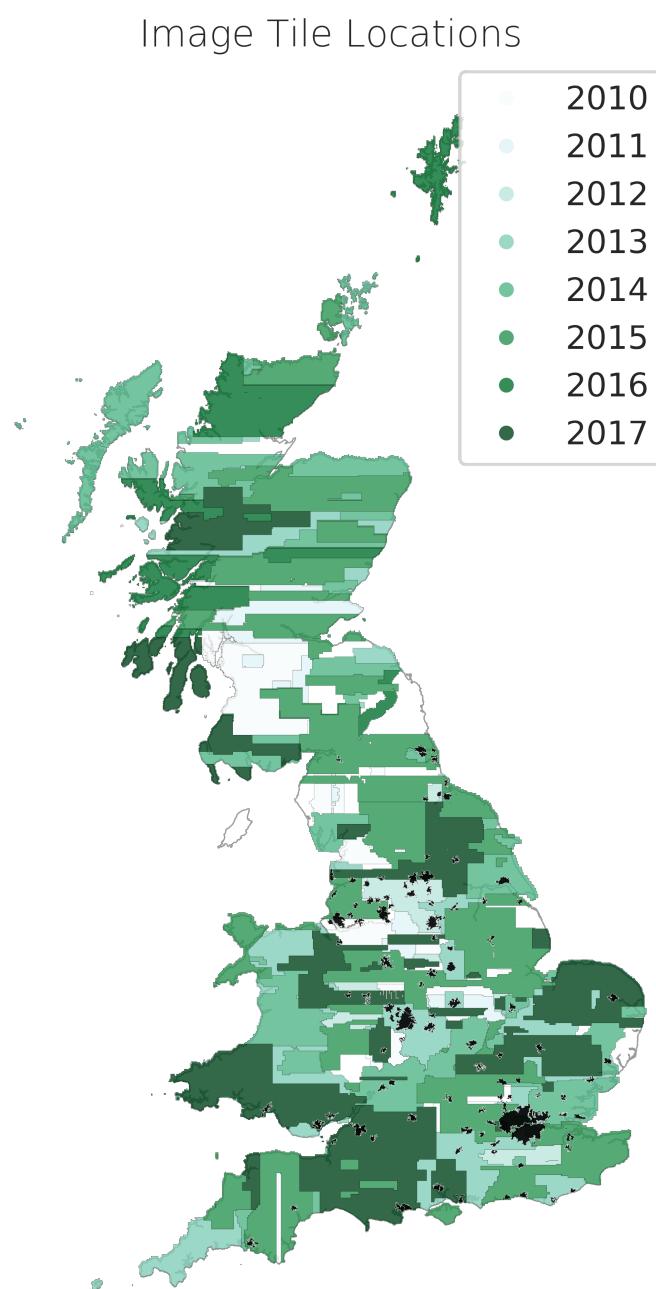


Figure A.1: UK aerial imagery age and distribution

Appendix B

Predictions

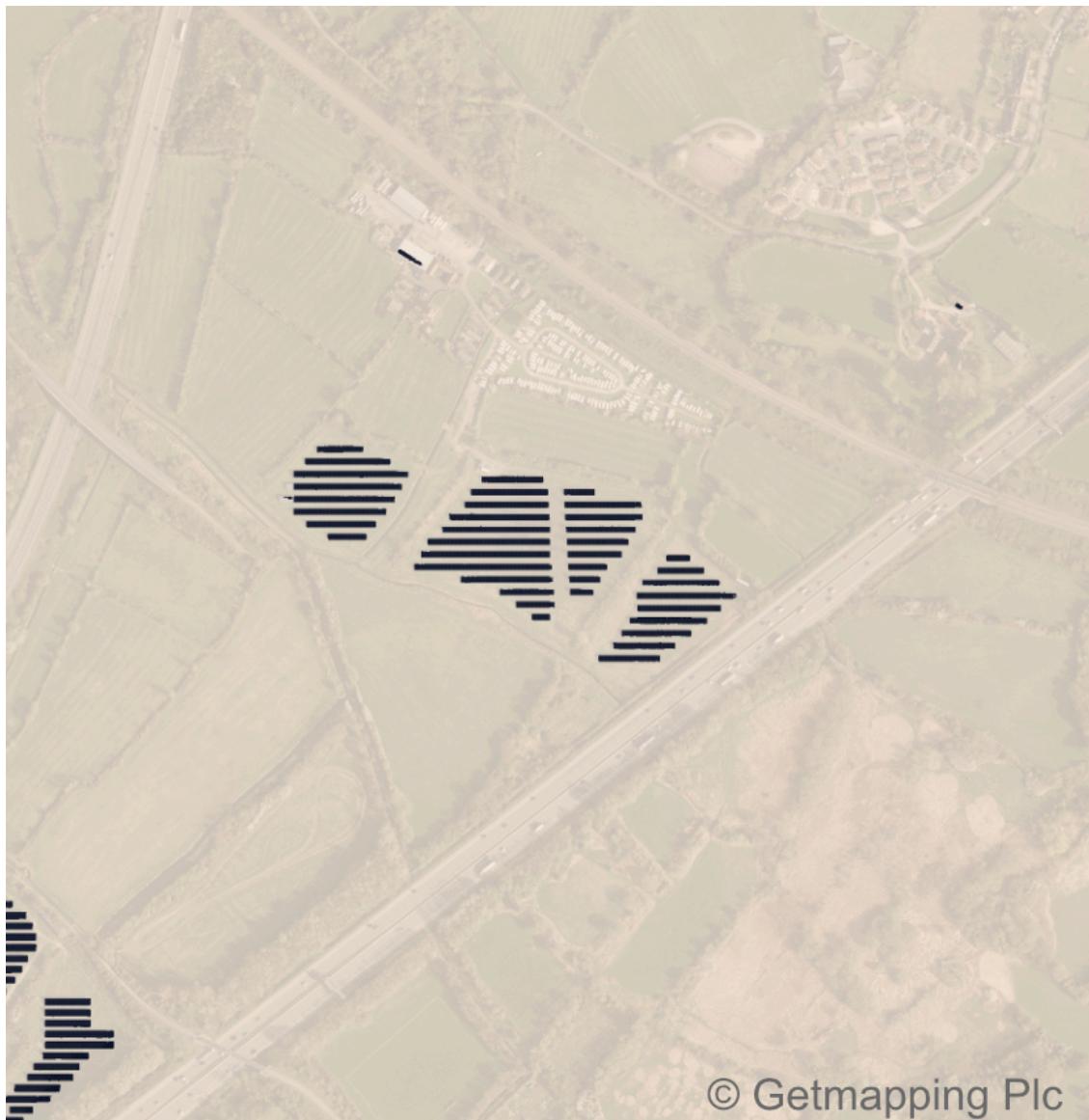


Figure B.1: Model 1 Large and small successful UK predictions, Bristol



Figure B.2: Model 1 Unsuccessful predictions at Bristol docks



Figure B.3: Model 1 Mixed but many successful predictions over urban Cambridge, UK

Prediction



Figure B.4: Model 1 PV Predictions in urban area, Knowsley, UK



Figure B.5: Detail of Model 1 PV Predictions, Knowsley, UK

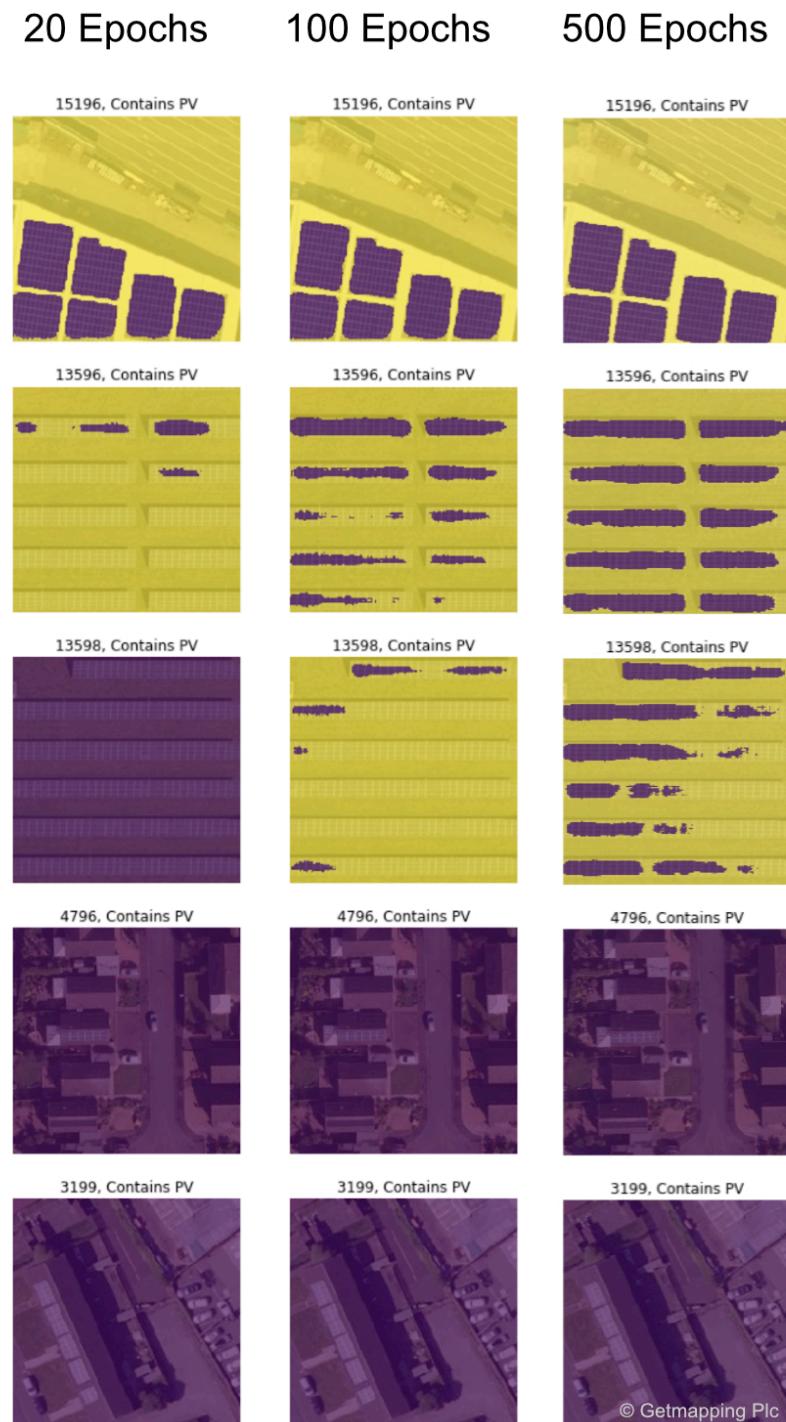


Figure B.6: Model 1 PV predictions after different training epochs, UK. PV in lower images never successfully identified

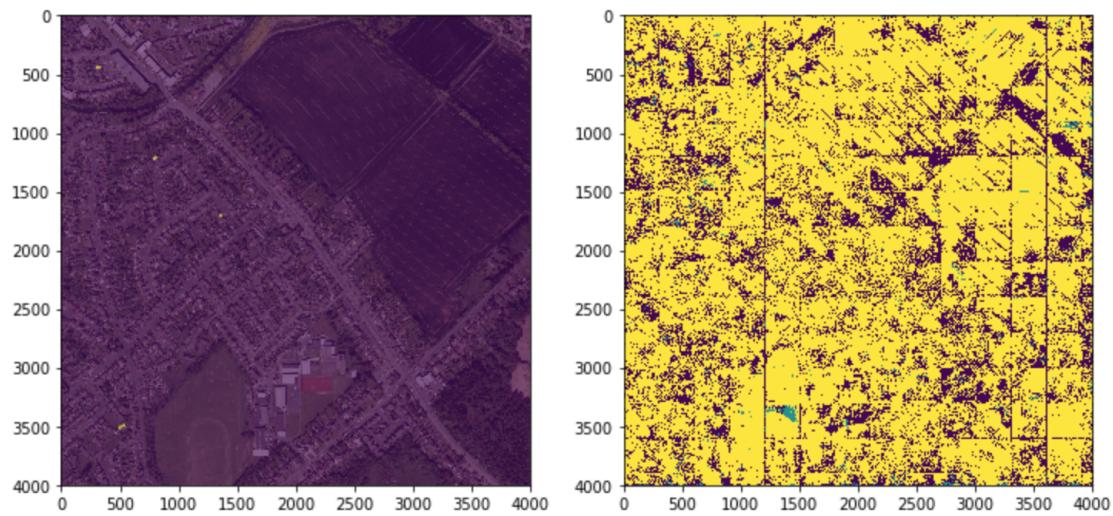


Figure B.7: Unsuccessful prediction outputs from Model 2, UK

Appendix C

Ethics

C.1 Ethics form

Step A6 – MSC ESDA Dissertation Ethics Declaration

Dissertation Research Proposal [COMPLETE THE TABLE BELOW]	
Title / Topic:	Where are solar photovoltaic arrays in the UK and what impact are they having?
Research Question(s) / Aims & Objectives:	Create framework for accurately locate solar PV arrays and use it to assess distribution of existing arrays.
Data & source (specify all data to be used; if none, explain why):	BEIS Renewable Energy Planning Database, Ofgem FIT Database, Aerial Imagery from EDINA Digimap University Services
Method(s) (specify all methods to be used):	Deep Learning frameworks (CNNs), geospatial analysis

Confirmation [TICK ALL APPLICABLE]:	
I have read and understood Step A1 – Does the research require a Risk Assessment?	<input type="checkbox"/>
EITHER TICK This planned research DOES require a risk assessment and appropriate approval will be secured before data collection starts.	<input type="checkbox"/>
OR TICK This planned research does NOT require a risk assessment.	<input checked="" type="checkbox"/>
I have read and understood Step A2 – Does the research require External research ethics approval?	<input type="checkbox"/>
EITHER TICK This planned research DOES require external ethics review and appropriate external ethics approval will be secured before the data collection starts.	<input type="checkbox"/>
OR TICK This planned research does NOT require external ethics review.	<input checked="" type="checkbox"/>
External ethics approval is not required and:	
I have read and understood Step A3 – Is the research Exempt from the need for ethics approval?	<input type="checkbox"/>
EITHER TICK This planned research IS EXEMPT from the need for research ethics approval.	<input checked="" type="checkbox"/>
OR TICK This planned research is NOT EXEMPT from the need for research ethics approval.	<input type="checkbox"/>
The research is not exempt from the need for ethics approval and:	
I have read and understood Step A4 – Does the research require High Risk ethics approval?	<input type="checkbox"/>
EITHER TICK This planned research IS deemed high risk and approval from the UCL Research Ethics Committee will be secured before the research starts.	<input type="checkbox"/>
OR TICK This planned research is NOT deemed high risk.	<input type="checkbox"/>
The research is not exempt from the need for ethics approval, does not require high risk ethics approval and:	
I have read and understood Step A5 – Does the research require ESDA low risk ethics review for questions-based methods OR BSEER low risk ethics review for other methods?	<input type="checkbox"/>
EITHER TICK This planned research requires MSc ESDA Low Risk Ethics approval for questions-based methods and approval will be secured before data collection starts.	<input type="checkbox"/>
OR TICK This planned research requires BSEER low risk ethics approval (for other methods), which will be secured before data collection starts.	<input type="checkbox"/>

Declaration by Students & Supervisors [COMPLETE THE TABLE BELOW]	
Student	UCL Supervisor
Name: Laurence Watson	Name: Prof Paul Ruyssenveit
UCL ID: ucbqlw0	UCL Position: Chair of Energy & Building Performance
UCL Email: Laurence.watson.18@ucl.ac.uk	UCL Email: p.ruyssenveit@ucl.ac.uk
I confirm that the information in this form is accurate to the best of my knowledge.	
I confirm that if the answers to any of these questions changes, I will go through this protocol again.	
Signature / E-signature: 	Signature / E-signature: Paul Ruyssenveit
Date: 4/06/2019	Date: 4/06/2019
Students	
<ul style="list-style-type: none"> Submit this signed page to the Dissertation Module Convenor following instructions from them. Include this signed form as a Dissertation Appendix. The Dissertation mark sheet asks the second marker whether this form was filled out correctly and, if not, what % deduction of the Dissertation Mark they recommend. 	

Figure C.1: UCL Ethics Declaration

Bibliography

Azavea (2019a), ‘PyTorch/fastai backend plugin for Raster Vision. Contribute to azavea/raster-vision-fastai-plugin development by creating an account on GitHub’, Azavea.

Azavea (2019b), ‘Raster Vision’, <https://rastervision.io/>.

Balta-Ozkan, N., Yildirim, J. & Connor, P. M. (2015), ‘Regional distribution of photovoltaic deployment in the UK and its determinants: A spatial econometric approach’, *Energy Economics* **51**, 417–429.

BBC (2013), ‘Renewables ’strain electricity grid”, *BBC News* .

BEIS (2019a), Public Attitudes Tracker - Wave 29, Technical report.

BEIS (2019b), ‘Renewable Energy Planning Database quarterly extract’, <https://www.gov.uk/government/publications/renewable-energy-planning-database-monthly-extract>.

BEIS (2019c), ‘Solar photovoltaics deployment’, <https://www.gov.uk/government/statistics/solar-photovoltaics-deployment>.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A. & Raffel, C. (2019), ‘MixMatch: A Holistic Approach to Semi-Supervised Learning’, *arXiv:1905.02249 [cs, stat]* .

Bollinger, B. & Gillingham, K. (2012), ‘Peer effects in the diffusion of solar photovoltaic panels’, *Marketing Science* **31**(6), 900–912.

- Bradbury, K., Saboo, R., Johnson, T. L., Malof, J. M., Devarajan, A., Zhang, W., Collins, L. M. & Newell, R. G. (2016), 'Distributed solar photovoltaic array location and extent dataset for remote sensing object identification', *Scientific data* **3**, 160106.
- Bright, J. M., Killinger, S., Lingfors, D. & Engerer, N. A. (2018), 'Improved satellite-derived PV power nowcasting using real-time power data from reference PV systems', *Solar Energy* **168**, 118–139.
- Burnett, D., Barbour, E. & Harrison, G. P. (2014), 'The UK solar energy resource and the impact of climate change', *Renewable Energy* **71**, 333–343.
- Camilo, J., Wang, R., Collins, L. M., Bradbury, K. & Malof, J. M. (2017), 'Application of a semantic segmentation convolutional neural network for accurate automatic detection and mapping of solar photovoltaic arrays in aerial imagery', *IEEE Applied Imagery Pattern Recognition (AIPR) Workshop*.
- Cheng, G., Zhou, P. & Han, J. (2016), 'Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images', *IEEE Transactions on Geoscience and Remote Sensing* **54**(12), 7405–7415.
- Colantuono, G., Everard, A., Hall, L. M. H. & Buckley, A. R. (2014), 'Monitoring nationwide ensembles of PV generators: Limitations and uncertainties. The case of the UK', *Solar Energy* **108**, 252–263.
- Committee on Climate Change (2019), Net Zero - Technical Report, Technical report.
- DECC (2012), Identifying trends in the deployment of domestic solar PV under the Feed-in Tariff scheme, Technical report.
- Dharshing, S. (2017), 'Household dynamics of technology adoption: A spatial econometric analysis of residential solar photovoltaic (PV) systems in Germany', *Energy Research & Social Science* **23**, 113–124.
- Digimap, E. (2019), 'Aerial Imagery 1:250, (25cm resolution), Coverage: UK, © Getmapping Plc', <https://digimap.edina.ac.uk/roam/download/aerial>.

- Dobbs, A., Elgindy, T., Hodge, B.-M. & Florita, A. (2017), ‘Short-Term Solar Forecasting Performance of Popular Machine Learning Algorithms: Preprint’, p. 8.
- DOE, U. (2019), ‘The Open PV Project - Data.gov’, <https://catalog.data.gov/dataset/the-open-pv-project>.
- Dong, R., Pan, X. & Li, F. (2019), ‘DenseU-Net-based Semantic Segmentation of Small Objects in Urban Remote Sensing Images’, *IEEE Access* pp. 1–1.
- Energy, G. (2016), ‘Power Purchase Agreements’, <https://www.goodenergy.co.uk/business/our-generators/power-purchase-agreements/>.
- ESO, N. G. (2019), System Operability Framework (SOF) — National Grid ESO, Technical report.
- Green, N. E. (1957), *Aerial Photographic Interpretation and the Social Structure of the City*, American Society of Photogrammetry.
- Grid, N. (2019), Future Energy Scenarios, Technical report.
- He, K., Zhang, X., Ren, S. & Sun, J. (2015), ‘Deep Residual Learning for Image Recognition’, *arXiv:1512.03385 [cs]*.
- Henderson, F. M. & Utano, J. J. (1975), ‘Assessing general urban socio-economic conditions with conventional air photography’, *Photogrammetria* **31**(3), 81–89.
- Huang, Z., Cheng, G., Wang, H., Li, H., Shi, L. & Pan, C. (2016), Building extraction from multi-source remote sensing images via deep deconvolution neural networks, in ‘2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)’, IEEE, pp. 1835–1838.
- IEA (2013), ‘Photovoltaic and Solar Forecasting: State of the Art’, <https://www.nrcan.gc.ca/energy/renewables/solar-photovoltaic/publications/11930>.
- Inman, R. H., Pedro, H. T. C. & Coimbra, C. F. M. (2013), ‘Solar forecasting methods for renewable energy integration’, *Progress in Energy and Combustion Science* **39**(6), 535–576.

- Institute, W. R. (2019), ‘Global Power Plant Database — Earth Engine Data Catalog’, https://developers.google.com/earth-engine/datasets/catalog/WRI_GPPD_power_plants.
- Jensen, J. R. & Cowen, D. C. (1999), ‘Remote sensing of urban/suburban infrastructure and socio-economic attributes’, *Photogrammetric engineering and remote sensing* **65**, 611–622.
- Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012), Imagenet classification with deep convolutional neural networks, in ‘Advances in Neural Information Processing Systems’, pp. 1097–1105.
- Kuhn, P., Nouri, B., Wilbert, S., Prahl, C., Kozonek, N., Schmidt, T., Yasser, Z., Ramirez, L., Zarzalejo, L., Meyer, A., Vuilleumier, L., Heinemann, D., Blanc, P. & Pitz-Paal, R. (2018), ‘Validation of an all-sky imager–based nowcasting system for industrial PV plants’, *Progress in Photovoltaics: Research and Applications* **26**(8), 608–621.
- Lari, Z. & Ebadi, H. (2007), ‘Automatic extraction of building features from high resolution satellite images using artificial neural networks’, *Proceedings of the ISPRS Working Group I/5 and IV/3: “High Resolution Earth Imaging for Geospatial Information* p. 6.
- LeCun, Y., Bengio, Y. & Hinton, G. (2015), ‘Deep learning’, *nature* **521**(7553), 436.
- Lipperheide, M., Bosch, J. L. & Kleissl, J. (2015), ‘Embedded nowcasting method using cloud speed persistence for a photovoltaic power plant’, *Solar Energy* **112**, 232–238.
- Long, M., Cao, Y., Wang, J. & Jordan, M. I. (2015), ‘Learning transferable features with deep adaptation networks’, *arXiv preprint arXiv:1502.02791* .
- Maggiori, E., Tarabalka, Y., Charpiat, G. & Alliez, P. (2017), Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark, in ‘2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)’, IEEE, pp. 3226–3229.

- Malof, J. M., Bradbury, K., Collins, L. M. & Newell, R. G. (2016), ‘Automatic detection of solar photovoltaic arrays in high resolution aerial imagery’, *Applied energy* **183**, 229–240.
- Malof, J. M., Hou, R., Collins, L. M., Bradbury, K. & Newell, R. (2015), Automatic solar photovoltaic panel detection in satellite imagery, in ‘2015 International Conference on Renewable Energy Research and Applications (ICRERA)’, IEEE, pp. 1428–1431.
- Malof, J. M., Li, B., Huang, B., Bradbury, K. & Stretslav, A. (2019), ‘Mapping solar array location, size, and capacity using deep learning and overhead imagery’, *arXiv:1902.10895 [cs]*.
- Mayor of London (2018), Solar Action Plan For London, Technical report.
- Mills, A., Ahlstrom, M., Brower, M., Ellis, A., George, R., Hoff, T., Kroposki, B., Lenox, C., Miller, N., Stein, J. & Wan, Y.-h. (2009), ‘Understanding Variability and Uncertainty of Photovoltaics for Integration with the Electric Power System’, *The Electricity Journal* .
- Ofgem (2018), Guidance for generators, Technical report, Ofgem.
- Ofgem (2019a), ‘Feed-in Tariff Installation Report’, <https://www.ofgem.gov.uk/publications-and-updates/feed-tariff-installation-report-31-march-2019>.
- Ofgem (2019b), ‘Renewables Obligation (RO) Annual Report 2017-18’, <https://www.ofgem.gov.uk/publications-and-updates/renewables-obligation-ro-annual-report-2017-18>.
- OpenStreetMap contributors (2019), ‘Solar PV Generator Labels’, <https://overpass-turbo.eu/>.
- Otsu, N. (1979), ‘A Threshold Selection Method from Gray-Level Histograms’, <https://ieeexplore.ieee.org/document/4310076>.

- Pan, S. J., Tsang, I. W., Kwok, J. T. & Yang, Q. (2010), ‘Domain adaptation via transfer component analysis’, *IEEE Transactions on Neural Networks* **22**(2), 199–210.
- Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L. & Lerer, A. (2017), ‘Automatic differentiation in PyTorch’.
- Paulescu, M., Badescu, V. & Brabec, M. (2013), ‘Tools for PV (photovoltaic) plant operators: Nowcasting of passing clouds’, *Energy* **54**, 104–112.
- RegenSW (2016), Grid Constraints In the South West: Options For Connection, Technical report.
- Robinson, S. A., Stringer, M., Rai, V. & Tondon, A. (2013), GIS-integrated agent-based model of residential solar PV diffusion, in ‘32nd USAEE/IAEE North American Conference’, pp. 28–31.
- Ronneberger, O., Fischer, P. & Brox, T. (2015), ‘U-Net: Convolutional Networks for Biomedical Image Segmentation’, *arXiv:1505.04597 [cs]*.
- Saint-Drenan, Y.-M., Bofinger, S., Ernst, B., Landgraf, T. & Rohrig, K. (2011), Regional nowcasting of the solar power production with PV-plant measurements and satellite images, in ‘Proc. ISES Solar World Congress’.
- Shi, X., Gao, Z., Lausen, L., Wang, H., Yeung, D.-Y., Wong, W.-k. & Woo, W.-c. (2017), ‘Deep Learning for Precipitation Nowcasting: A Benchmark and A New Model’, *arXiv:1706.03458 [cs]*.
- Shin, H., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D. & Summers, R. M. (2016), ‘Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning’, *IEEE Transactions on Medical Imaging* **35**(5), 1285–1298.
- So, B., Nezin, C., Kaimal, V., Keene, S., Collins, L., Bradbury, K. & Malof, J. M. (2017), Estimating the electricity generation capacity of solar photovoltaic arrays using only color aerial imagery, in ‘2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)’, IEEE, pp. 1603–1606.

- Solar, S. (2019), ‘PV Live’, <https://www.solar.sheffield.ac.uk/pvlive/>.
- Tanner, F., Colder, B., Pullen, C., Heagy, D., Eppolito, M., Carlan, V., Oertel, C. & Sallee, P. (2009), Overhead imagery research data set—An annotated data library & tools to aid in the development of computer vision algorithms, *in* ‘2009 IEEE Applied Imagery Pattern Recognition Workshop (AIPR 2009)’, IEEE, pp. 1–8.
- Taylor, J., Leloux, J., Everard, A., Briggs, J. & Buckley, A. (2015), Quantifying PV fleet output variability in the UK: Consequences for Distribution Network Operators.
- Taylor, J., Leloux, J., Hall, L., Everard, A., Briggs, J. & Buckley, A. (2015), Performance of Distributed PV in the UK: A Statistical Analysis of Over 7000 Systems.
- UCL (2019), ‘UCL Energy Institute wins project to develop London Solar Opportunity Map’, <https://www.ucl.ac.uk/bartlett/energy/news/2019/jan/ucl-energy-institute-wins-project-develop-london-solar-opportunity-map>.
- Wang, R., Camilo, J., Collins, L. M., Bradbury, K. & Malof, J. M. (2017), The poor generalization of deep convolutional networks to aerial imagery from new geographic locations: An empirical study with solar array detection, *in* ‘2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)’, IEEE, pp. 1–8.
- Wilson, R. (2018), ‘European Solar PV By Country, based on BP Statistical Review of World Energy 2018’.
- Yang, H. L., Yuan, J., Lunga, D., Laverdiere, M., Rose, A. & Bhaduri, B. (2018), ‘Building extraction at scale using convolutional neural network: Mapping of the united states’, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **11**(8), 2600–2614.
- Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. (2014), How transferable are features in deep neural networks?, *in* ‘Advances in Neural Information Processing Systems’, pp. 3320–3328.

- Yu, J., Wang, Z., Majumdar, A. & Rajagopal, R. (2018), ‘DeepSolar: A Machine Learning Framework to Efficiently Construct a Solar Deployment Database in the United States’, *Joule* 2(12), 2605–2617.
- Yuan, J., Yang, H.-H. L., Omitaomu, O. A. & Bhaduri, B. L. (2016), Large-scale solar panel mapping from aerial images using deep convolutional networks, in ‘2016 IEEE International Conference on Big Data (Big Data)’, IEEE, pp. 2703–2708.