Measurement & spatial analysis of London's tree canopy cover: 2018 methodology report

This report was prepared for the Greater London Authority by

Breadboard Labs

2018

Executive summary

The data presented is a high resolution, 25 cm per pixel, map of tree canopy cover for the Greater London area and a number of related regional summaries of that data. It is accompanied by a document describing the machine learning and image processing techniques used to generate that map as well as the series of evaluations that were carried out to determine its accuracy and error characteristics.

This work was carried out under Breadboard Lab's European Space Agency funded project, Curio Canopy, in collaboration with the Greater London Authority who were a support partner of that project. All the algorithms and techniques developed were implemented using Google Earth Engine which was also leveraged to perform the processing needed to create a London wide high resolution tree canopy map.

1. Introduction

The urban forest plays a key role in making cities habitable environments for people. Trees remove carbon dioxide from the atmosphere, filter air pollution and produce oxygen. They also play a role in combating climate change and extreme weather events by providing shade, cooling the air and absorbing stormwater. Importantly, trees and green spaces also provide a range of of social, cultural and health benefits and have a positive impact on people's physical and mental well being[1] .

The importance of the urban forest and the services it provides is increasingly being recognised and management policies are being developed to help promote, develop and manage this critical resource. The ability to effectively measure and monitor the resource is crucial if such policies are to be developed in an informed way and implemented efficiently.

[1] http://www.fao.org/forestry/urbanforestry/87029/en/

This document describes the generation of a high resolution map of tree canopy through the use of aerial imagery and machine learning techniques. Such canopy information can be used to help measure and monitor the urban forest and to guide policy and management decisions. The approach taken is extremely scalable and the data produced makes it possible to generate canopy cover estimates for an arbitrary area of interest while also, importantly, giving a sense of the distribution of canopy cover within that area of interest. These characteristics differentiate the approach taken from traditional survey based approaches such i-Tree canopy[2].

1.2 Background

Breadboard Labs was founded in 2014 to create people-centred solutions to environmental problems. In particular, our aim is to build tools that enrich people's view of the world around them and help them take an active part in the effort to protect their environment. Inspired by amazing initiatives like the LandSat[3] and Sentinel[4] satellite imagery programmes and also the Open Data and citizen science movements we began investigating how satellite imagery could be used to give an overview of the urban forest and to help guide a crowdsourced effort to gather data on the ground. We received funding from the European Space Agency to investigate our ideas which led

[2] https://canopy.itreetools.org/
[3] https://landsat.usgs.gov/
[4] http://www.esa.int/Our_Activities/Observing_the_Earth/Copernicus/Sentinel-2

to the Curio Canopy [5] and ultimately to the development of our environmental education and outreach platform Curio[6].

The Greater London Authority acted as a support partner in the Curio Canopy project. The core focus of that collaboration was an investigation of the use of Sentinel 2 data and aerial imagery for canopy and green infrastructure measurement monitoring. Preliminary analysis was carried out on the efficacy of Sentinel 2 imagery for urban green infrastructure analysis and our existing high resolution imagery based canopy analysis solution was developed further and refined. This document is solely focused on the high resolution canopy analysis work carried out.

1.3 Structure of the report

A brief overview of imagery and technology used to carry out the work will be presented. This will be followed by a detailed description of the approach taken and evaluations carried out. The high resolution canopy map was used to produce a number of canopy analysis outputs including borough and ward level canopy cover analysis. These and outputs will be presented and discussed. Finally some concluding remarks and plans for future work will be presented.

2. Imagery & Technology

A high-level description of the imagery as well as the technologies used follows.

2.1 Data used

High resolution colour infrared (CIR) imagery was used to produce the tree canopy cover layer. Colour Infrared Imagery consists of red, green and infrared bands. The inclusion of an infrared band at the expense of a blue band is significant as infrared light can be used to help identify vegetation. The imagery was collected in September 2016 and purchased by the GLA from the GeoInformation Group[7].

The CIR imagery provided is available at a 10 cm per pixel resolution. There are, however, considerable storage and computational implications in processing imagery at that resolution and an early phase evaluation found that there was little advantage to using the 10 cm imagery over the resampled 25 cm alternative. The CIR imagery used was made up of 467 image tiles covering the entire Greater London area.

[5] https://business.esa.int/projects/curio-canopy
[6] https://www.curio.xyz
[7] http://www.geoinformationgroup.co.uk/

2.2 Software used

Google Earth Engine is a cloud-based platform for planetary-scale geospatial analysis that brings Google's massive computational capabilities to bear on remote sensing tasks.[8] It is designed to be accessible to and empower not just remote-sensing experts but a broad range of users with varying levels of technical expertise. Significantly, as well as many useful imagery and remote sensing tools, Google Earth Engine provides a data catalogue consisting of imagery, geophysical, climate & weather and demographic data. The imagery catalogue includes continuously updated LANDSAT[9] and Sentinel imagery[10]. The ease of access to data and the algorithms and computational resources needed to extract value from that data make Google Earth Engine an extremely powerful resource. It is not surprising that it has already had big impact on the scientific community and on a number of environmentally focused projects[11].

The platform includes an IDE[12], through which algorithms can be created to perform various data processing and analysis tasks. A more comprehensive overview of the platform can be viewed online, with introductory videos and more comprehensive documentation available through the Earth Engine home page. The CIR imagery was uploaded to Earth Engine and the IDE was used to develop, evaluate and apply the canopy mapping algorithms to the full set of CIR imagery. More details on the algorithms and evaluations carried out will be provided in the sections that follow.

[8] *Google Earth Engine: Planetary-scale geospatial analysis for everyone (Gorelick et al, 2017)* https://www.sciencedirect.com/science/article/pii/S0034425717302900
[9] LANDSAT is a US government programme of earth observation satellites that has been active since the 1970s. It provides the longest temporal record of moderate resolution multispectral data of the Earth's surface on a global basis. More information available at https://landsat.usgs.gov/
[10] Sentinel is part of the European Space Agency's Copernicus satellite programme https://sentinel.esa.int/web/sentinel/home
[11] Google Earth Engine case studies https://earthengine.google.com/case_studies/
[12] An IDE (Integrated Development Environment) is a computer software programme that provide facilities to programmers to develop new programmes. The Google Earth Engine IDE is available at https://code.earthengine.google.com/

3. Approach

The approach taken to building a canopy map involved a combination of data processing techniques. In particular, machine learning and image processing techniques were employed.

Machine learning is a particular branch of Artificial Intelligence concerned with models that extract patterns from data in an automated way. Provided with some labelled training data, a machine learning algorithm is capable of creating a model that can then be applied to unseen data.

In this particular case, we wanted to label each and every pixel in the CIR imagery set as being either a canopy pixel or not. A reasonably small set of image pixels that have been labelled as canopy or not can be used to build a machine learning model capable of performing such as task. This model can then be applied to each of the nearly 30 billion pixels contained in the 9 gigabits of CIR imagery that cover the Greater London area, a task far beyond human capabilities.

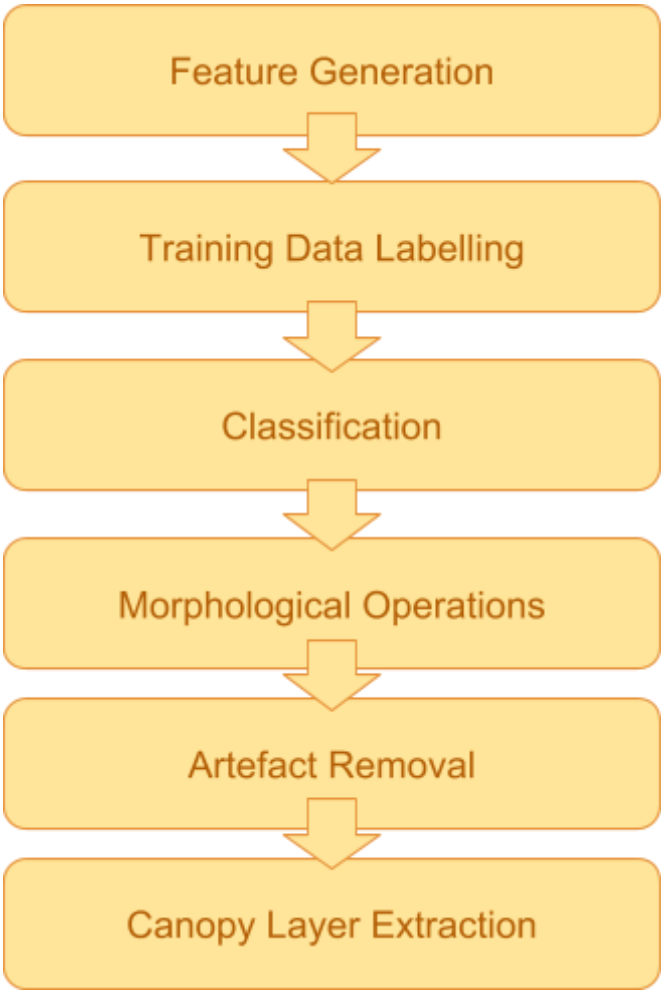
Image processing techniques take an image as an input and the output may be an altered image or characteristics/features associated with that image. This covers a broad range of techniques including reducing noise[13] in an image or extracting information about what we would, intuitively, recognise as the textures present in an image. Information about things like the texture in the neighbourhood (pixels immediately surrounding) around a pixel can be a useful input to machine learning techniques in the case of this particular task. Image smoothing techniques, which are a set of techniques for removing noise from an image, are also useful in order to clean up the raw output produced by the machine learning techniques applied.

The techniques discussed were applied as part of a series of distinct processing steps that were referred to as the Curio Canopy Data Pipeline. Each of the steps in the pipeline and their purpose will now be outlined in greater detail.

3.1 Curio Canopy Data Pipeline

Curio's canopy layer extraction algorithm has been developed using the Google Earth Engine[14] (Earth Engine) platform discussed previously and involves a number of distinct phases, each of which can be seen in the process chart below. Each step has a distinct purpose, a set of inputs and a set of outputs. The outputs from one step form the inputs to the steps further down the pipeline. Each step will now be discussed in turn.

[13] Image noise is any undesirable by-product of image capture that obscures the desired information. In the context of earth observation imagery, this may include errors introduced by a satellite sensor or the presence of clouds.
[14] https://earthengine.google.com/



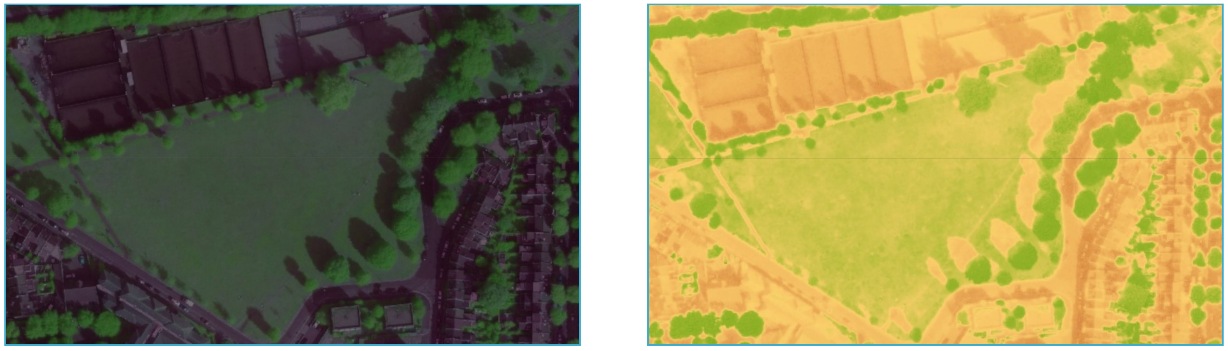
Feature Generation

Earth Engine's machine learning classification libraries are based on a pixel based approach whereby each pixel is examined and classified in isolation. Machine learning techniques rely on a description of the phenomena they are trying to model in order of discern patterns that can be used to classify objects. The set of attributes that make up that description are known as features. It is easy to appreciate that expanding the description of each pixel beyond its basic band information to include neighbourhood, texture and normalisation based information as well as other data transformations might make it easier to uncover patterns that could be used to classify the pixel. The creation and evaluation of features for use in machine learning forms the first stage in the canopy analysis pipeline. In all, more than 30 features were generated. A brief summary of some of the key classes of feature generated follows.

Vegetation Index Features

Healthy vegetation exhibits a very distinctive signal across the light spectrum and in the red and infrared bands in particular. In remote sensing that signal is often characterised and measured using what are known as vegetation indexes. The use of vegetation indexes or features that help detect and measure vegetation levels are useful for canopy analysis. The Normalised Vegetation Index (NDVI) along with the two-band version of Enhanced Vegetation Index (EVI) and the Transformed Difference Vegetation Index (TDVI) were generated. The absence of a blue band in the available CIR imagery meant that the traditional EVI was not option.

Fig 1. A false-colour visualisation of the base bands and NDVI visualisation of the same scene



Texture Based Features

Intuitively taking texture into account can be expected to be useful in distinguishing between canopy and other green vegetation. Earth Engine provides a number of methods for identifying texture by providing entropy and gray-level co-occurrence matrix (GLCM) based features.

Fig 2. A false-colour visualisation of the base bands and an EVI entropy based visualisation of the same scene

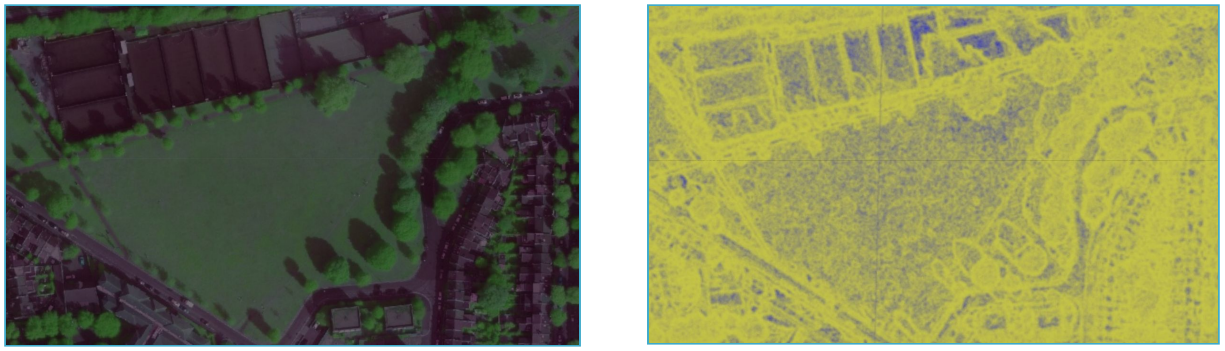
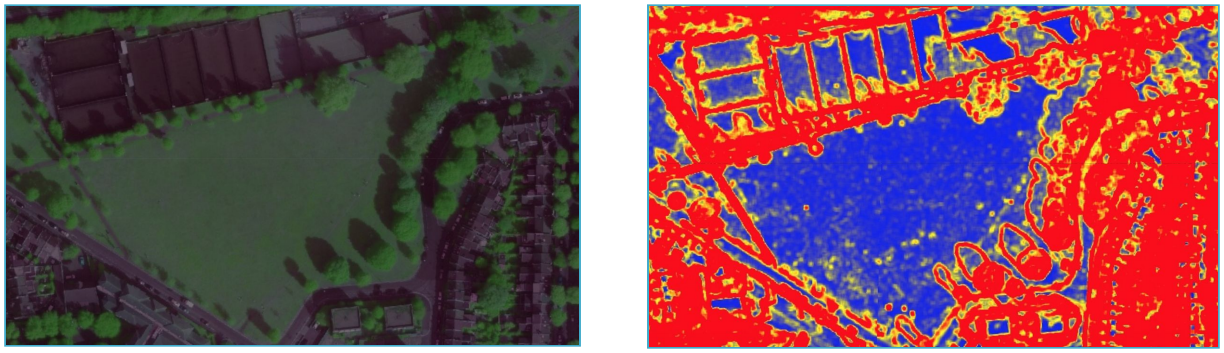


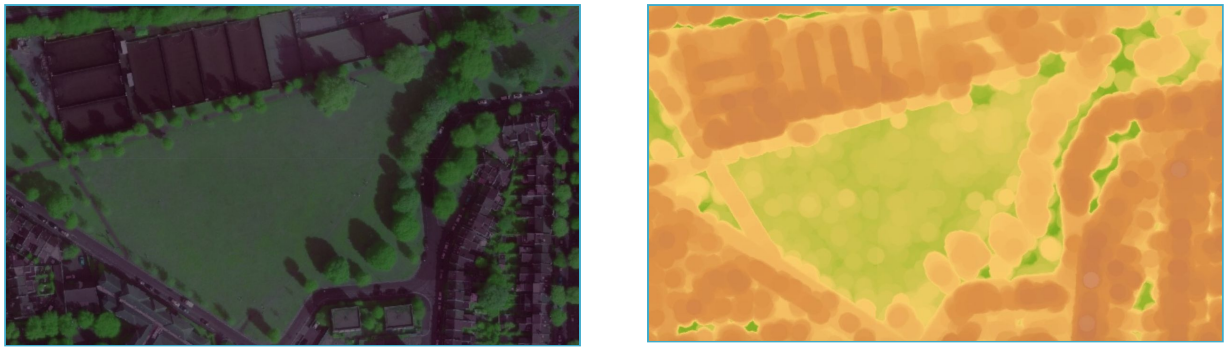
Fig 3. A false-colour visualisation of the base bands and a GLCM Inertia based visualisation of the same scene



Neighbourhood and Smoothing Based Features

Taking into account the characteristics of the neighbourhood surrounding a pixel might also be useful. For instance, the minimum or median EVI value within a certain radius. Generating such features can also have a smoothing effect on the image, especially when repeated. Figure 6 displays an example of such smoothing and a number of such features were added to the pixel features vector used as the basis for our analysis

Fig 4. A false-colour visualisation of the base bands and a visualisation of the same scene after a minimum EVI value kernel has been applied twice.



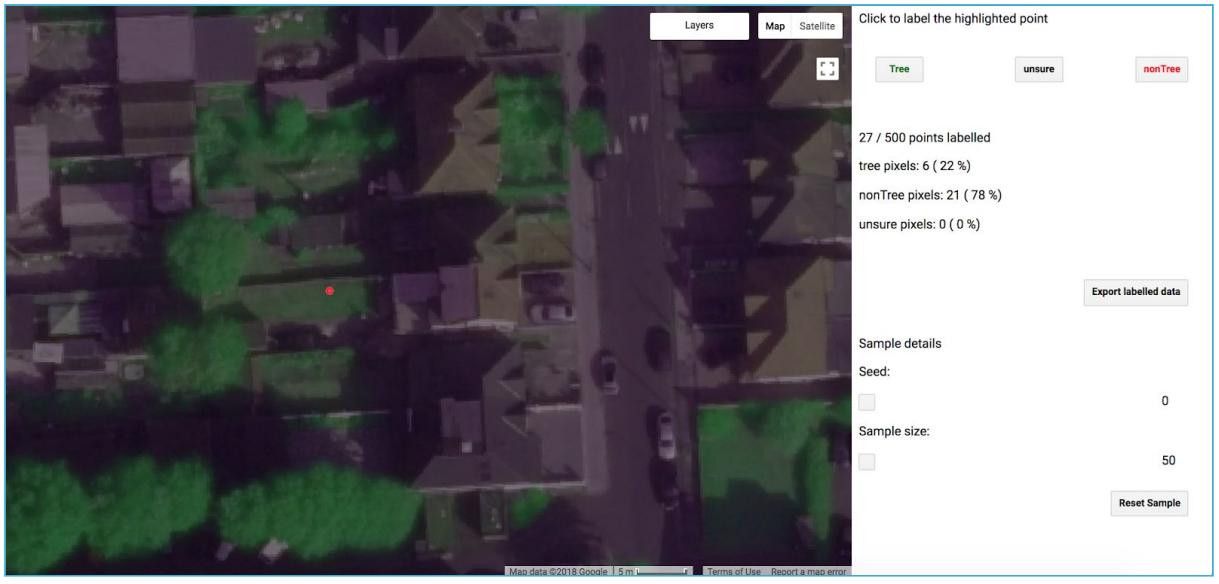
Training and Validation Data Generation

Machine learning algorithms require labelled data in order to build a model that can then be applied to unseen data. A separate set of labelled data, unseen by the model, is also needed to properly assess the model's performance. The details of such evaluations will be discussed in later sections, however, a mechanism of labelling the CIR imagery was needed.

Google Earth Engine's web interface makes it extremely easy to hand label imagery and to build classifiers by placing markers or drawing polygons on the aerial imagery in order to identify particular features. With this approach it is usually possible to quickly develop a reasonably effective classifier and also to respond to weaknesses in the model by labelling more data to correct errors. However, this approach proved problematic when trying to build a model that would generalise beyond a small set of tiles and when trying to determine the likely real world performance of the model. The data created in this manner inevitably suffers from biases and it is unlikely to be balanced and reflective of real world distribution.

An alternative approach is to generate a set random points on a plane and to label them. It solves all the issues highlighted with the previous approach and the generation of such data also creates an additional source of canopy cover data. This distribution of labelled points in a given area can be used as an additional estimate of canopy cover. In order to facilitate the generation of such data a simple user interface and app was developed using the Earth Engine User Interface library. The interface developed can be seen below

Fig 5. Random points on a plane app in Earth Engine



The interface is a useful tool for generating a large amount of balanced data. The target pixel is highlighted in red and the user has three options; 'tree', 'notTree' and 'unsure'. The 'unsure' option is not intended to be used frequently but exists to help prevent ambiguous pixels being labelled

arbitrarily. These pixels are later removed. Once a pixel has been labelled the user is presented with another random pixel from the experiment area.

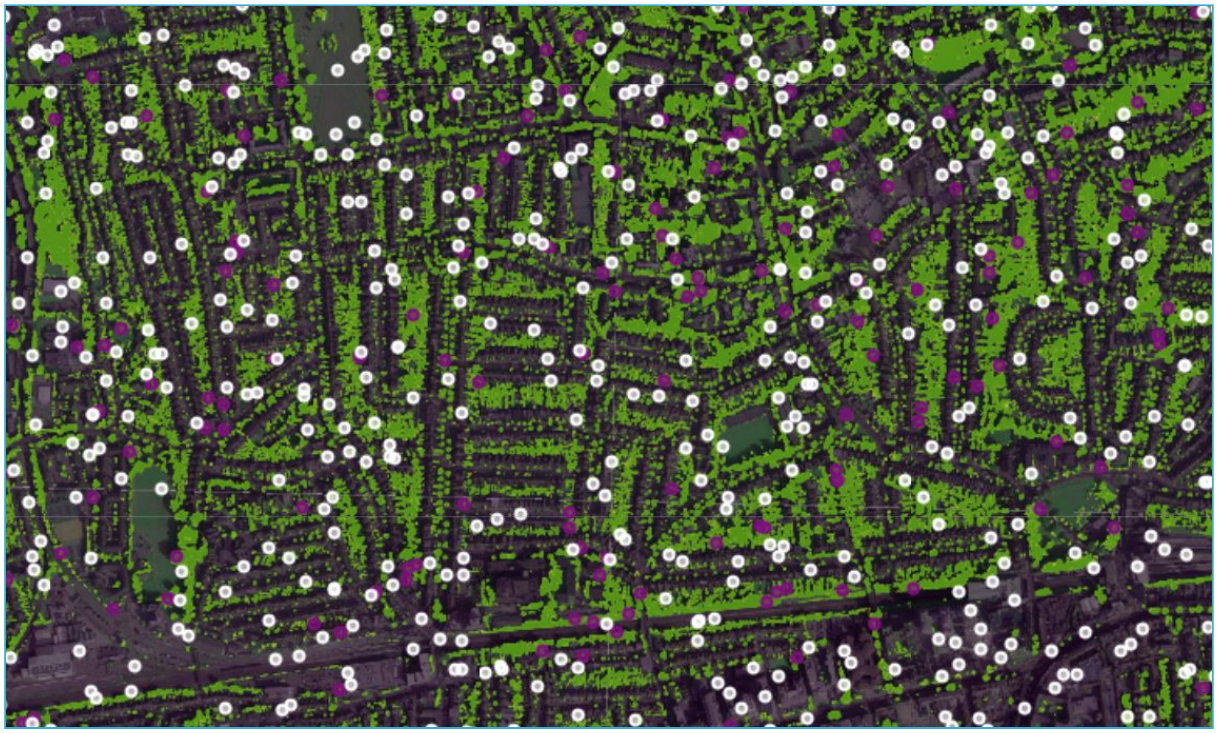


Fig 6. A sample of the random points on the plane that were generated in a given area overlaid on the canopy layer produced. Pixels that were labelled as being canopy are purple and non-canopy pixels are white.

Using the labelling app 24,000 pixels were labelled. The motivation for labelling so many points was to ensure a level of sampling was approached that would result in estimates with a standard error of 2% or less on any of the London electoral wards that were targeted for examination as part of the evaluations of the model. These estimates could then be used to assess the canopy model's performance and as a reference point when examining the results of other canopy studies carried out in these areas.

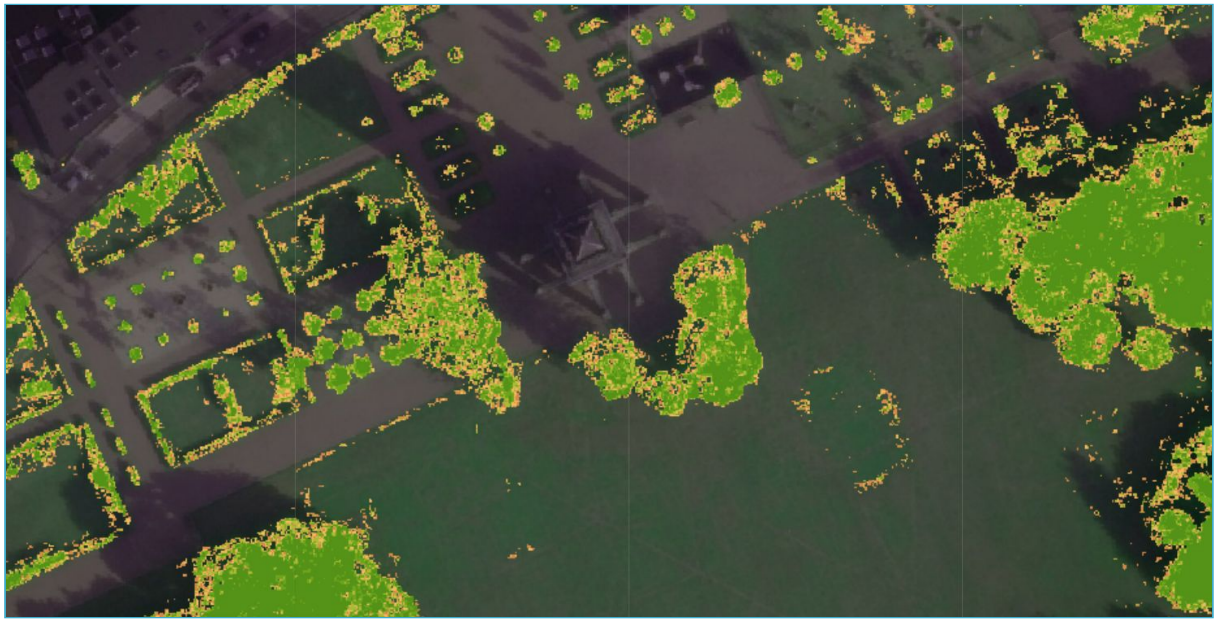
It is worth noting that the effort to generate that amount of data was about 6 person-days. The labelling app directly mimics the i-Tree Canopy approach which recommends at least 500-1000 points per study area. This advice stems from the standard error characteristic of estimating a binomial distribution and it is unavoidable that such levels of data labelling will be required to produce reasonable estimates. Such requirements might make a London wide ward level or sub ward level i-Tree Canopy report a prohibitive task and highlight the value of an automated or semi-automated approach. The i-Tree Canopy approach is extremely elegant and robust, however, it is limited to point data and can only give a broad estimate of the canopy cover. It does not provide information about the underlying distribution of canopy cover without ever increasing levels of sampling being required.

Classification & Morphological Operations

Once a set of features have been generated, a classifier can be built using the labelled data or a proportion of it. Earth Engine provides a number of machine learning classifiers. A number were investigated but it quickly became apparent that of the options available the Random Forest classifier performed best.

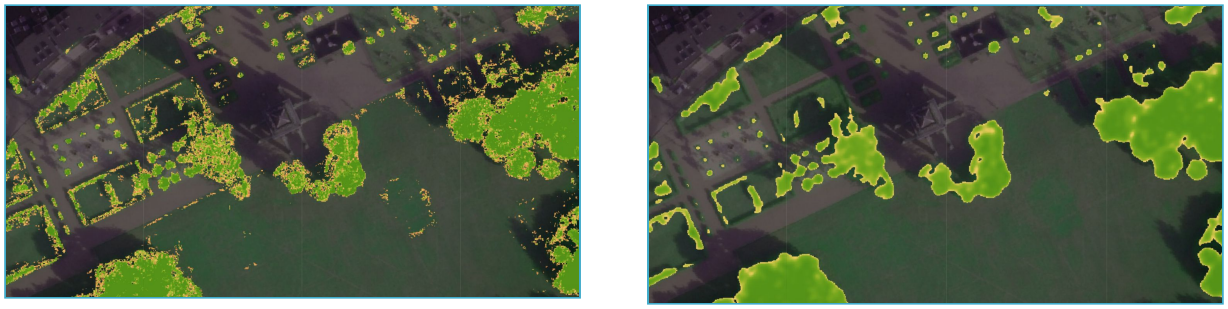
Earth Engine supports a pixel based approach to classification and each pixel in an image is classified. In the image below just pixels with a classification value greater than 0.5 are displayed. It is possible to generate a probabilistic output from Earth Engine's implementation of the Random Forest algorithm and this is used in the visualisation below and in later stages of our canopy layer pipeline. Pixels with lower probability values are displayed in yellow and orange.

Fig 7. All 'canopy' pixels with a probability greater than 0.5



It doesn't make sense to look at each pixel in isolation when considering objects such as trees and the output generated is a little noisy as a result. Earth Engine provides a number of morphological operations that can be used to take the probabilities of a pixel's neighbours into account in determining its probability and to smooth out the raw pixel classifications. The results of applying a number of these operations can be seen below and it is this output that forms the basis our canopy solution.

Fig 8. Morphological operations applied to canopy layer



Artefact Reduction & Exporting Canopy

The individual canopy objects can be identified and converted from raster format to a vectorised object using Earth Engine. The identification of these canopy objects opens up the possibility of generating object wide statistics that capture the characteristics of a unit of canopy such as its area and also the average probability score of the pixels contained. The canopy layer was exported in both formats, however, it is advised that where possible a raster version should be used.

It is possible to use the probability statistics or other qualities of each vectorised section of canopy to select and remove artefacts. However, although applying artefact removal can improve the model accuracy and also the visual appearance of the layer it can also lead to a loss in accuracy. Ideally it should help reduce the false positive rate but if the removal process is too severe it can also impact on the overall accuracy. The effect of the process can depend on the nature of the layer produced and so the effects of artefact removal and the exact parameters used should be evaluated. The canopy layer delivered has not had artefact removal applied but this option is available to users of the data.

3.2 Model Evaluation and Refinement

To begin with, the canopy model was evaluated using a standard machine learning methodology, 10-fold cross-validation, which is explained in greater detail in the section that follows. The results of that evaluation technique were used to refine a number of parameters used in the various stages of the Curio Canopy data pipeline. Parameters such as the set of features used and size of the kernel widths used in morphological operations were iteratively examined and refined in order to maximise the model performance.

The labelled data used in these evaluations was predominantly from the Ealing area. To ensure that the model's performance over the Greater London area was understood, a further 'generalised performance' evaluation was carried out. Our collaborators in the GLA selected 11 tiles that were representative of a range of different urban or semi-urban terrain. Labelled data for each of the 11

tiles was generated and the model's performance on each tile was assessed. The results of this evaluation were also used to iteratively refine our model and ensure that the parameters selected led to a good generalised performance.

The existence of a separate i-Tree canopy study data recently completed for Ealing by Trees for Cities as part of a broader i-Tree Eco study also opened up an opportunity to compare the ward level results produced using the canopy model with such studies. Although these studies were carried out using different imagery, such comparative evaluations were useful in validating the model and understanding its characteristics.

Each of the evaluations carried out will now be discussed in turn.

Cross-validation Accuracy Analysis

Using the 24,500 labelled pixels generated through the data labelling app, it was possible to determine the canopy model's likely real world accuracy using 10-fold cross-validation. All 'unsure' data was removed and the remainder was split into ten. Nine of the subsets were then combined to form a set of training data with the the tenth being used as validation data. This process was repeated 10 times with each of the ten sections being used as a validation data set.

The outcome of this validation process is a set of accuracy measurements for the model. These accuracy measurements help describe the accuracy of the model in producing the desired outputs, in this case 'canopy' vs. 'non canopy'.

Overall Accuracy	93.24% (+-0.45)
Consumers Accuracy (Canopy)	87.06% (+- 1.7)
Producer's Accuracy (Canopy)	77.24% (+-2.2)
Consumers Accuracy (Non Canopy)	95.55% (+-0.53)
Producer's Accuracy (Non Canopy)	97.17% (+-0.44)

The Producer's Accuracy is the map accuracy from the point of view of the map maker (the producer). This assesses how accurately real features on the ground are correctly picked up and classified. It is the probability that a certain land cover of an area on the ground has been accurately classified by the model. In this case, it is the proportion of the land cover classes correctly labelled as 'canopy' and 'non-canopy'. The Producer's Accuracy is complement of the Omission Error, and can be understood to represent how thoroughly the model captures the desired classification without omission (i.e. how well is the model capturing all the canopy cover without missing canopy pixels?).

The Consumer's Accuracy is a measure of reliability and it is the accuracy from the point of view of the map user. Consumer accuracy tells us how often, when a prediction of a particular class is made, that we can expect the prediction to be correct in reality. The Consumer's Accuracy is complement of the Commission Error, and can be understood to represent how well the model captures the desired classification accurately without including incorrect areas (i.e. how much non-canopy cover is being incorrectly captured/ 'false positives').

The model's performance and the statistics above reflect the fact that this is a minority class problem[15]. The distribution of canopy to non-canopy pixels is heavily skewed towards non-canopy pixels. In fact, 19.8% of the pixels labelled were canopy pixels and this can be confidently determined as the canopy coverage for the 27 tiles covered in this analysis (the number of sample points means the standard error is 0.25%). The Random Forest algorithm will try to maximise its overall accuracy rather than maximising predictions in relation to any one class. Although the ability to predict the presence of as much canopy as possible is desirable, not at such a cost as to devalue the need to predict the areas that are not canopy, especially as it would lead to an overestimation of canopy. Although resampling was investigated it was rejected and naturally balanced training data was used for all other analysis carried out.

Generalisation Performance

To help estimate the model's performance across London, labelled data was used to generate validation data across the 11 tiles selected by the GLA team. This validation data allowed a parallel estimation of overall canopy cover using a similar method to i-Tree canopy. Based on this validation data, the overall model accuracy across the 11 tiles was found to be 94.87% with a standard deviation of 3.7%. It's notable that in the case of tile 102, the accuracy dropped to 86% while the results in many of the other tiles were in the very high 90's. Further investigation revealed that tile 102 contains a number of fields with shrub like features that exhibit a high NDVI signature. The full set of results of this analysis can be viewed in the following google spreadsheet.

Tile Name	Overall accuracy	Consumer's accuracy nonTree	Consumer's Accuracy Tree	Producer's Accuracy nonTree	Producer's Accuracy Tree
101	96.89%	100.00%	76.00%	96.55%	100.00%
102	86.31%	98.58%	22.22%	86.88%	75.00%
228	97.04%	96.95%	97.37%	99.22%	90.24%
224	94.87%	94.79%	95.00%	96.81%	91.94%
317	91.49%	90.00%	94.12%	96.43%	84.21%
453	91.92%	91.95%	91.67%	98.77%	61.11%
451	95.24%	100.00%	85.45%	93.39%	100.00%
309	98.80%	100.00%	88.24%	98.68%	100.00%

[15] For more detail see https://towardsdatascience.com/dealing-with-imbalanced-classes-in-machine-learning-d43d6fa19d2

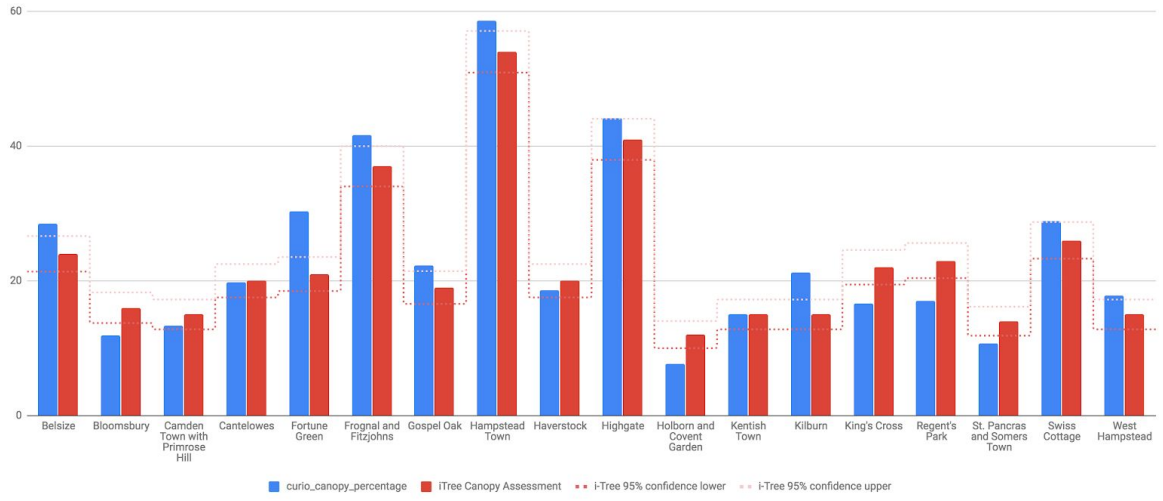
95	98.80%	98.68%	100.00%	100.00%	88.24%
655	95.03%	94.93%	95.65%	99.24%	75.86%
401	97.13%	98.60%	90.32%	97.92%	93.33%
Average Accuracy	94.87%				
Standard Deviation	3.72%				

Ward Level and Model Comparison Analysis

The provision of data from i-Tree canopy studies that were recently conducted in the Boroughs of Ealing and Camden meant that it was possible to perform comparative analysis on a ward by ward basis in both boroughs. In the case of Ealing, the large amount of labelled validation data generated could be used to generate ward level estimates of canopy cover using a method similar to i-Tree canopy. We refer to it as Curio Sampler in the sections that follow.

Camden i-Tree Study Comparison

An i-Tree Canopy study was conducted for the Borough of Camden in 2016. A thousand sample points were labelled in each ward. A comparison of the i-Tree study estimates and those produced using our canopy model can be seen in the graph below.



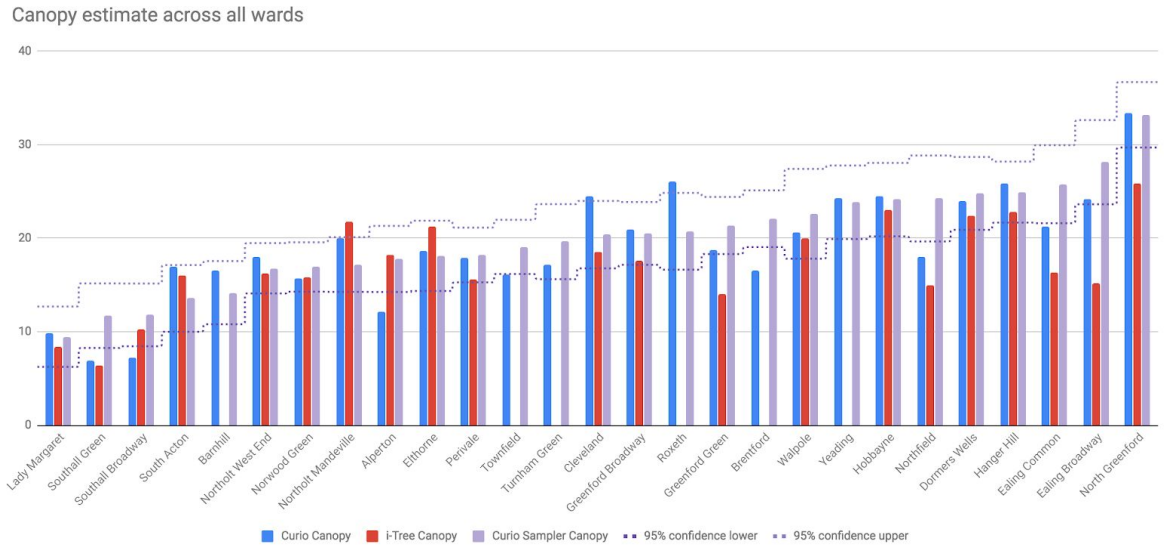
The raw data and chart can be found in the following Google Spreadsheet. The two estimates are reasonably well aligned with the exception of Fortune Green. Across all wards, the mean error is 0.8 with a standard deviation of 4.41. The mean absolute error is 3.77. These figures suggest that the model is very marginally biased towards overestimating canopy relative to the i-Tree study figures. There is, however, some variation at a ward level and a ±3-4% difference in the estimates of canopy cover at that level can be expected. It is worth noting that the i-Tree study was conducted using a different set of imagery, captured during a different time period, and so some variation in results is to be expected.

Ealing i-Tree Study & Curio Sampler Comparison

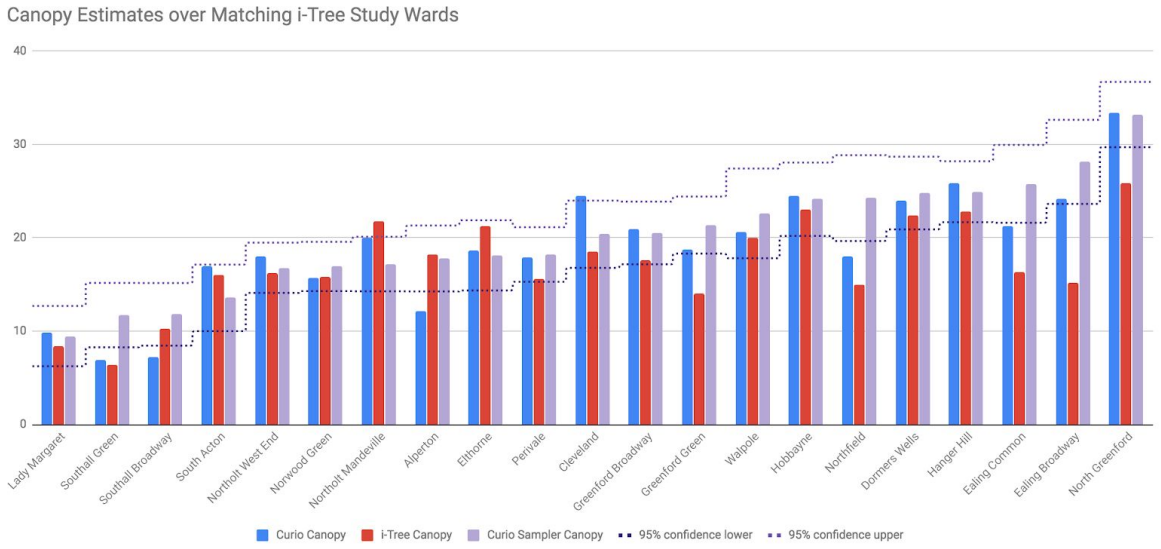
An i-Tree Canopy study was conducted in Ealing by Trees for Cities. The application developed for labelling data, was applied extensively to the Ealing area. The resulting labelled data are referred to as Curio Sampler. This provided a means of comparing the canopy model results with ward level

canopy estimates similar to i-Tree canopy produced using exactly the same imagery. It also made it possible to examine Curio Canopy's performance against a broader number of wards than were covered in the Ealing i-Tree Canopy study.

In the graphs that follow the data was ordered based on Curio Sampler Canopy percentage in order to make the graph easier to read. It is also worth noting that the Curio Sampler Canopy trend line is surrounded by 95% confidence interval lines. These bounds were determined based on binomial distribution standard error which is a function of the mean value and the sample size



The raw data and chart can be found in the following Google Spreadsheet. Based on this work, Trees for Cities reran their evaluation of Ealing Broadway which led to the canopy cover estimate for that ward being updated from 12.5 to 15.2. The updated i-Tree results are included in this analysis.



Overall, the average error against the i-Tree study is 1.95 with a standard deviation of 3.86. The mean absolute error against the i-Tree study is 2.75. The various combinations of comparative results are summarised in the table below.

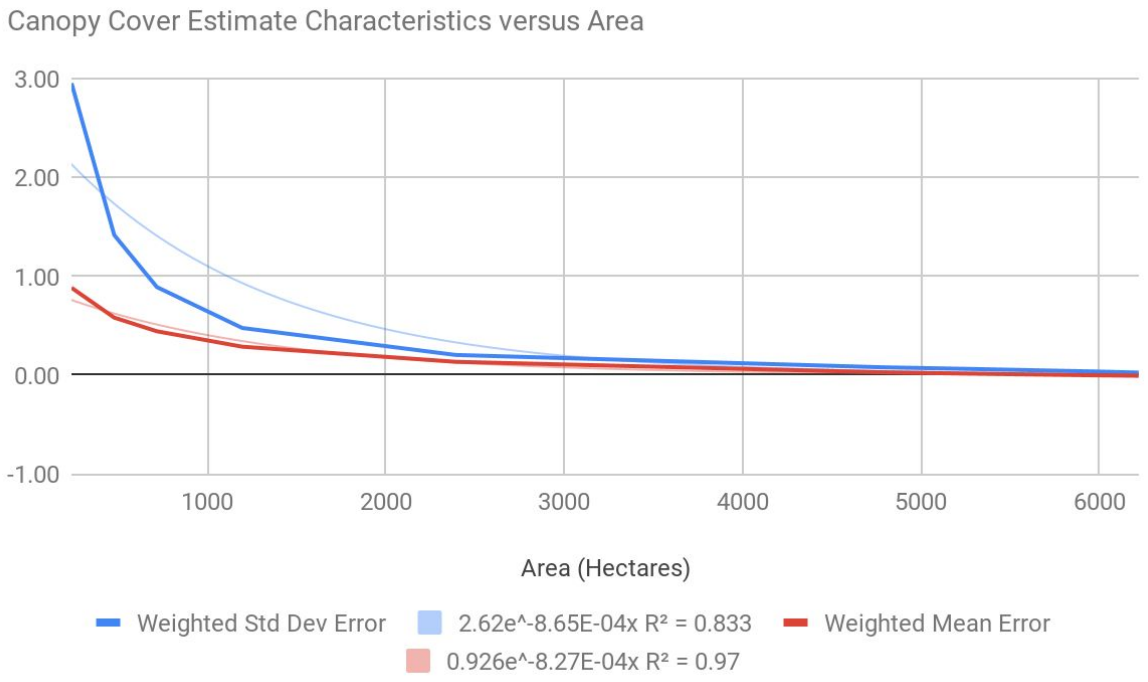
	Curio Canopy versus Curio Sampler	i-Tree versus Curio Sampler	Curio Canopy versus i-Tree
Mean Absolute Error	2.48	3.44	2.75
Average Error	-1.08	-3.03	1.95
Standard Deviation	3.00	4.70	3.86

It is encouraging that both i-Tree Canopy and Curio Canopy generally follow a similar trend relative to the Curio Sampler trend line. The i-Tree Canopy score for Ealing Broadway, even after being updated, is still out of trend, the reason for which we are not able to explain. The Curio Canopy model overestimates the canopy coverage in both Cleveland and Roxeth compared to the i-Tree Canopy estimate by a margin of 4.1 & 5.3% respectively. An investigation into these wards uncovered that a considerable proportion of both consists of golf courses. The Curio canopy model sometimes struggles to distinguish rough grassland areas of golf courses with a very strong NDVI signal from tree canopy. This led to the overestimates in both these cases. Investigations were made into how to minimise such effects, however it is noted that this can be a problem with image recognition based approaches (Brennan, Michael, Gerald Mills, and Tine Ningal. "Dublin Tree Canopy Study." (2017) IR/Sample comparison: Annex 1, p. 33-39).

Canopy Cover Estimate Characteristics as a Function of Area

One of the advantages of the image processing based approach to canopy cover estimation used in this project is that the detailed canopy map it produces makes it possible and easy to estimate canopy cover for any arbitrary area of interest. However, it is important to understand the error characteristics of the model and how they may vary over different selected areas of interest.

The i-Tree canopy comparison evaluations carried out indicate that the canopy cover model's mean error across all tiles or wards tends towards zero or a very slight overestimate. This implies that as the area under investigation is increased the error and variation in that error will decrease. In order to investigate whether this is definitely the case and to characterise the model's performance, the Ealing wards were incrementally combined to generate a series of larger areas of interest. For each combined area of interest, the area was recorded along with the difference in the Curio Canopy estimate relative to the Curio Sampler estimate. As each combination of wards produced different total combined areas, the weighted mean error by area, as well as the mean error, were examined. A summary of data generated can be seen in the chart and table below.



Number of Wards Combined	Number of Combinations	Mean Area	Mean Error	Weighted Mean Error	Std Dev Error	Weighted Std Dev Error	Std Dev Area
1	27	239.	0.94	0.88	3.16	2.95	74.95

2	351	478.	0.60	0.58	1.47	1.42	102.15
3	2925	717	0.46	0.44	0.91	0.89	122.42
5	80730	1196	0.30	0.29	0.48	0.48	151.29
10	8436285	2392	0.14	0.13	0.20	0.20	188.08
20	888030	4784	0.03	0.03	0.08	0.08	170.68
22	80730	52637	0.02	0.01	0.06	0.06	151.29
25	351	5980	0.00	0.00	0.04	0.04	102.15

This evaluation was carried on the Ealing wards alone from which some of the model training data was derived. However, it is an investigation of a trend that was evident in all the evaluations carried out and is a general characteristic of the model. Ealing was selected as the main focus of the early work in this project as it encompasses a wide range of land use types. The high level of labelled data generated for the borough of Ealing made this evaluation possible. Given that the model's accuracy figures for both the Ealing and the generalised evaluations were largely the same, it is reasonable to expect that the same area based error curve applies across the Greater London area. Based on the trend evident in the data it can be conservatively proposed that estimates generated for areas greater than 2,500 hectares will have standard error of 0.2%. For areas below that threshold, the chart and table should be used as a guide to the standard error and the trend line fitted can be used if an exact value is needed.

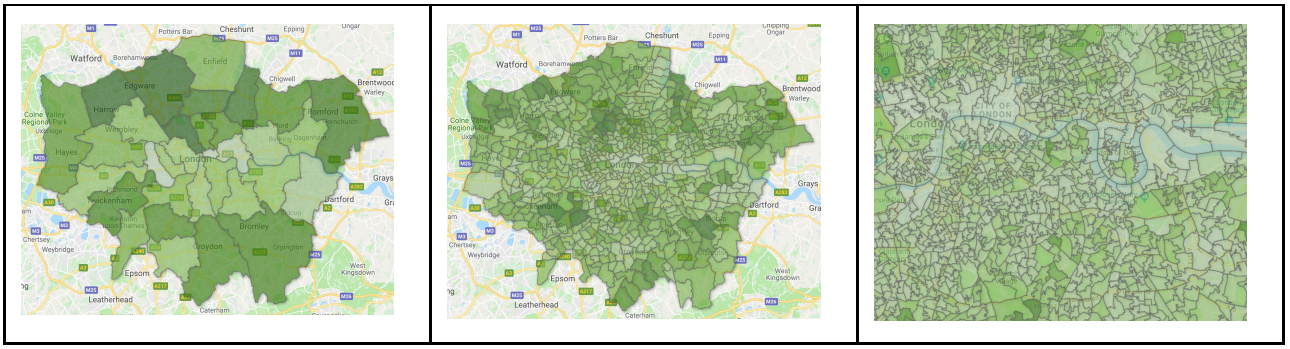
3.3 Summary of final selected approach

A machine learning and image processing approach was taken to generating a high resolution map of canopy cover across the Greater London area. Labelled data was used to train and fit a model that was then applied across the full set of 25 cm per pixel imagery provided to produce a detailed canopy map.

Fig 9. A close up view of the tree canopy map

That high resolution map can then be used to produce canopy cover estimates for any area of interest by simply counting the number of pixels marked as canopy as a proportion of the total number of pixels in that area. The error associated with the canopy estimate will depend on the size of the area of interest. As part of this study, borough, ward and Lower Super Output Area (LSOA) level canopy estimates were generated. Visualisations of the estimates produced can be seen in Fig 10 below

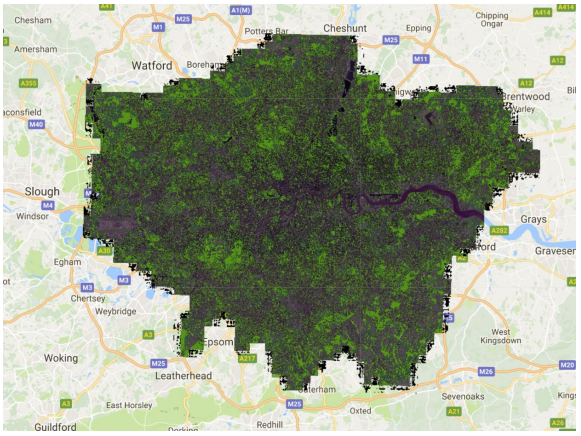
Fig 10. Visualisations of wider-area canopy estimates



4. Results & Outputs

The techniques described throughout this report were used to create a tree canopy prediction model, which then generated a high resolution, 25 cm per pixel, map of tree canopy in Greater London. Using that canopy map, the average canopy cover across Greater London is estimated to be 21.06 ± 0.2%.

Fig 11. High resolution canopy layer

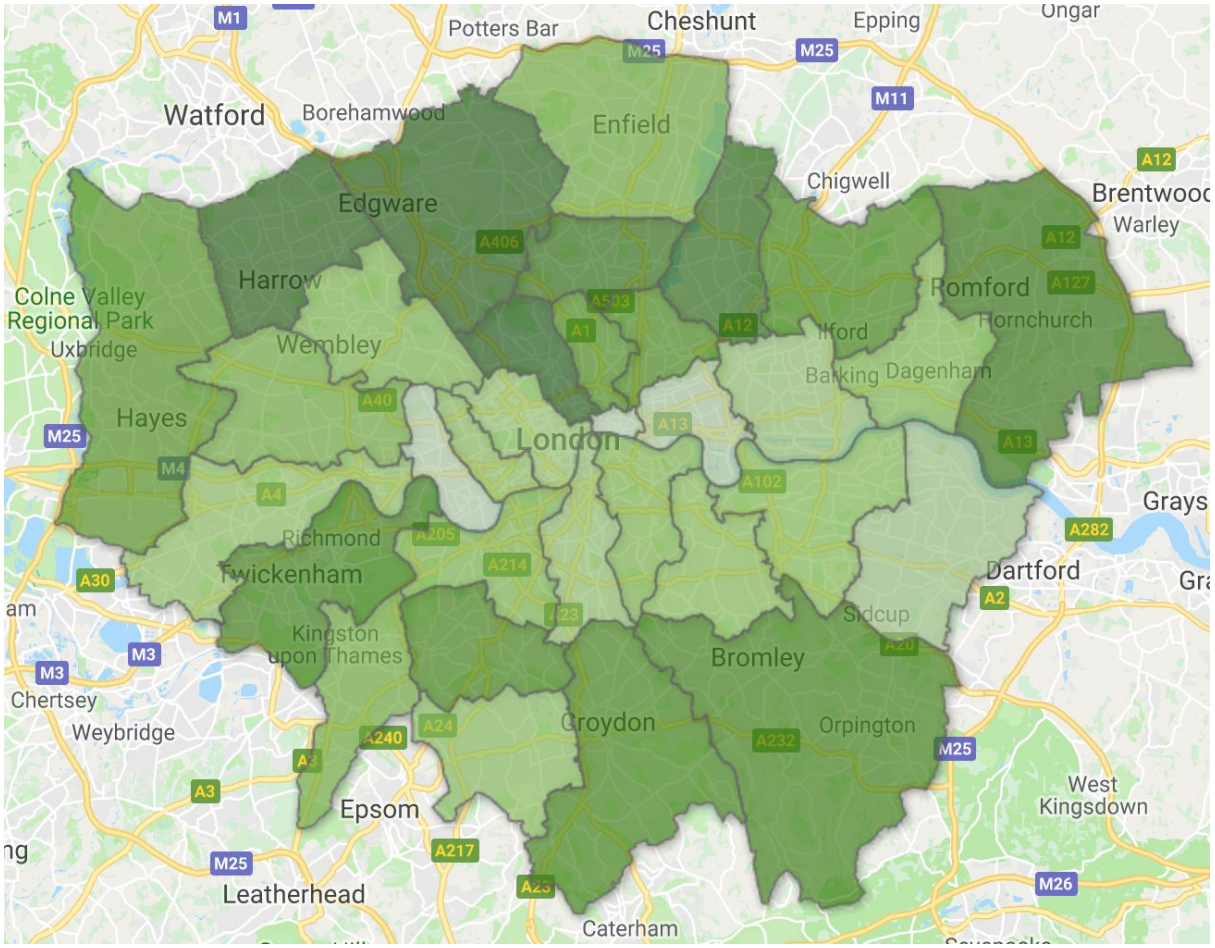


It is possible to use that canopy cover map to estimate canopy cover for particular areas of interest. As part of this study, borough, ward and Lower Super Output Area (LSOA) level canopy estimates were generated. Transport for London also produced a hexagon grid that is being used by the Greater London Authority and that grid was also used to produce a detailed summary of canopy cover. For each set of areas of interest a Google Fusion Table was produced which contains all of the related data. The borough level estimates will be presented in detail and followed by sections describing the ward, LSOA and hexagon resources produced.

4.1 Borough Level Canopy Figures

A visualisation of the canopy cover figures for all 33 London boroughs can be seen in Fig 12. The canopy cover estimate for each borough is accompanied by an estimate of the standard error attached to that figure. That error estimate is based on the area of the borough and the work described in the 'Canopy Cover Estimate Characteristics as a Function of Area' Section.

Fig 12. Visualisation of borough-level canopy figures



Name	Canopy Cover Estimate	Area (Hectare)	Standard Error Estimate (%)
Barking and Dagenham	17.99	3,765.71	0.20
Barnet	27.63	8,646.15	0.20
Bexley	14.33	6,405.38	0.20
Brent	18.18	4,309.97	0.20
Bromley	23.58	14,966.65	0.20
Camden	28.19	2,172.15	0.36
City of London	2.37	314.24	1.29
Croydon	23.24	8,624.53	0.20
Ealing	19.34	5,535.25	0.20
Enfield	19.27	8,193.29	0.20

Greenwich	17.35	5,022.96	0.20
Hackney	22.60	1,900.30	0.43
Hammersmith and Fulham	12.11	1,710.71	0.49
Haringey	25.45	2,949.33	0.21
Harrow	27.52	5,031.21	0.20
Havering	24.87	11,408.27	0.20
Hillingdon	21.72	11,534.03	0.20
Hounslow	16.59	5,641.88	0.20
Islington	22.28	1,481.22	0.58
Kensington and Chelsea	16.86	1,235.05	0.68
Kingston upon Thames	19.12	3,714.30	0.20
Lambeth	15.88	2,716.77	0.25
Lewisham	17.33	3,520.10	0.20
Merton	23.19	3,750.00	0.20
Newham	15.58	3,851.98	0.20
Redbridge	22.78	5,626.05	0.20
Richmond upon Thames	23.77	5,857.60	0.20
Southwark	17.95	2,980.17	0.20
Sutton	16.77	4,370.12	0.20
Tower Hamlets	12.88	2,149.67	0.36
Waltham Forest	26.28	3,868.87	0.20
Wandsworth	18.19	3,510.11	0.20
Westminster	16.17	2,196.42	0.20

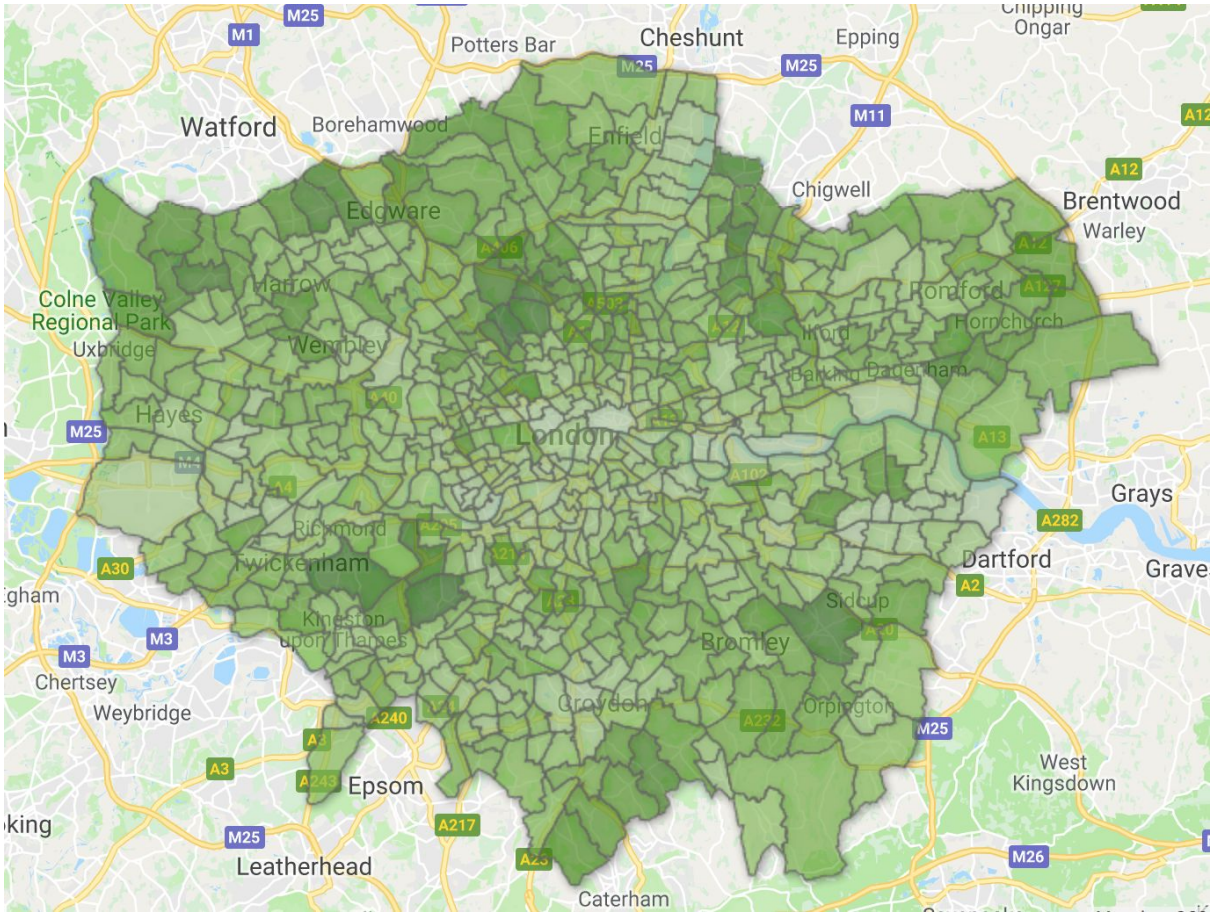
The raw borough level data and some other related visualisations can be found in the following Google Fusion Table. That data is also available as a Google spreadsheet which includes the calculation of the London wide canopy figure.

4.2 Ward Level Canopy Figures

Canopy cover figures for all 630 Greater London electoral wards were also generated and a visualisation of those figures can be seen in Fig 13 below. The data underpinning this visualisation

and some alternative visualisations can be found in the following Google Fusion Table. That data is also available as a Google spreadsheet which includes another separate calculation of the London wide canopy figure.

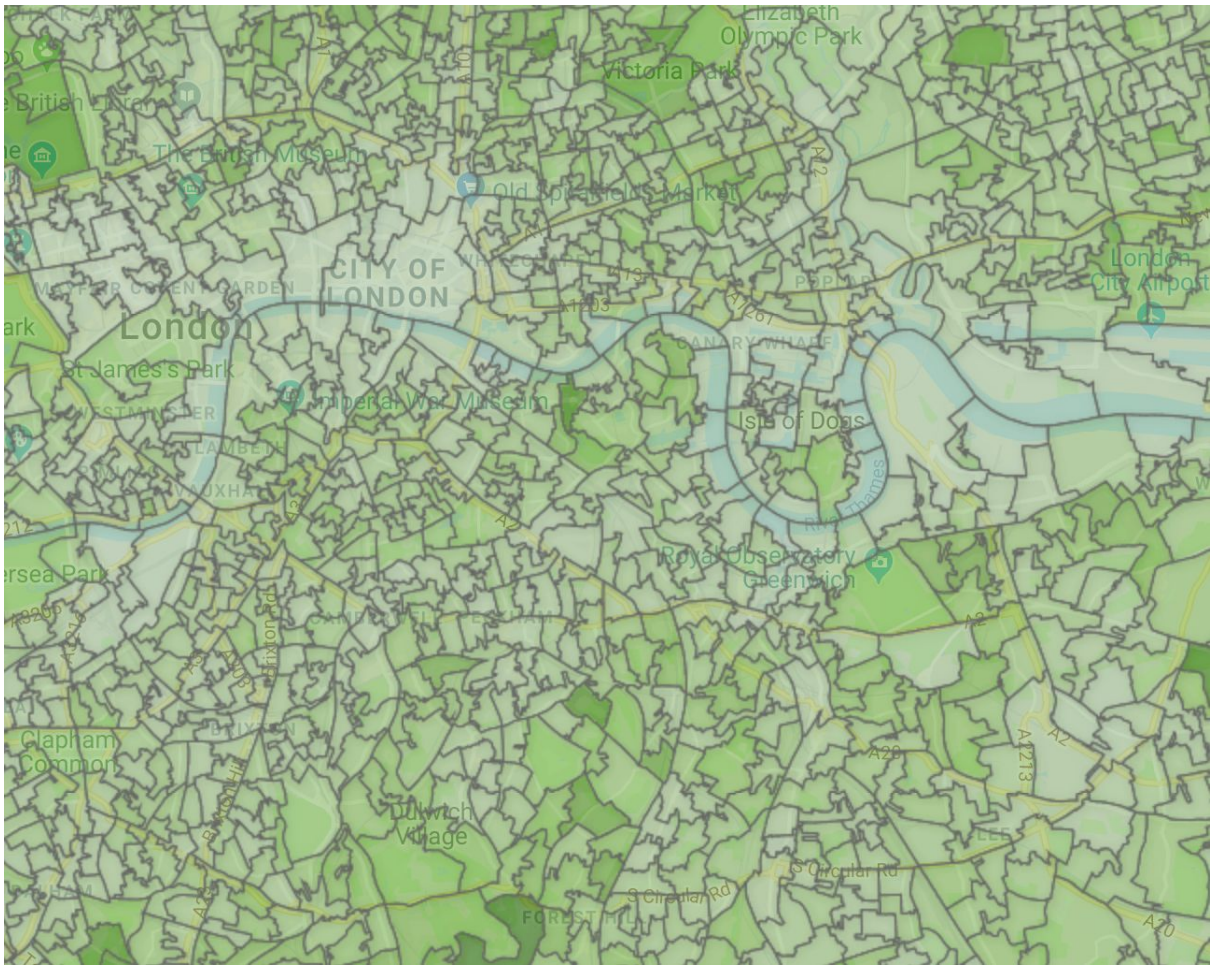
Fig 13. Visualisation of ward-level canopy figures



4.3 Lower Super Output Area Level Canopy Figures

A visualisation of the canopy cover figures for a proportion of the 4,835 Lower Super Output Areas processed can be seen in Fig 14 below. The data for the full set of LSOAs is available in the following Google Fusion Table

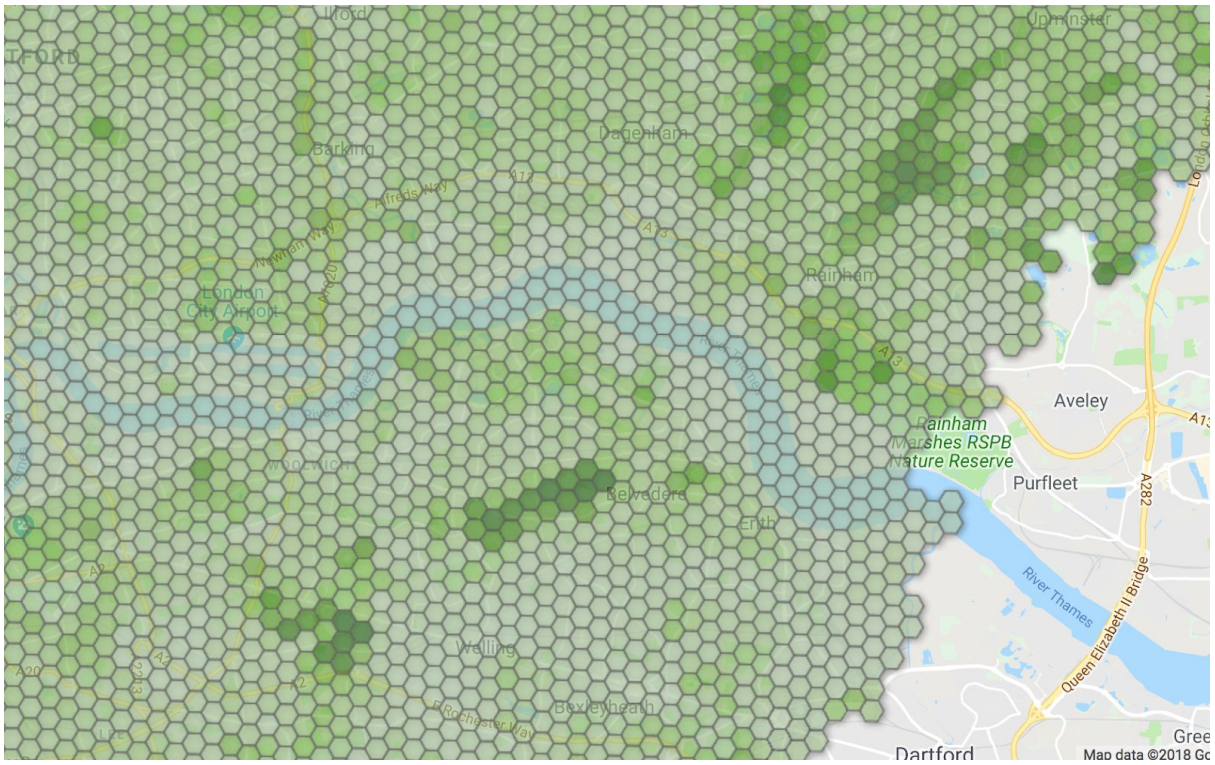
Fig 14. Visualisation of LSOA-level canopy figures



4.4 Hexagon Level Canopy Figures

The hexagonal grid produced by the Greater London Authority team was also used to produce canopy cover estimates. A section of that grid can be seen in Fig 15 below and the complete data is available in the following Google Fusion Table.

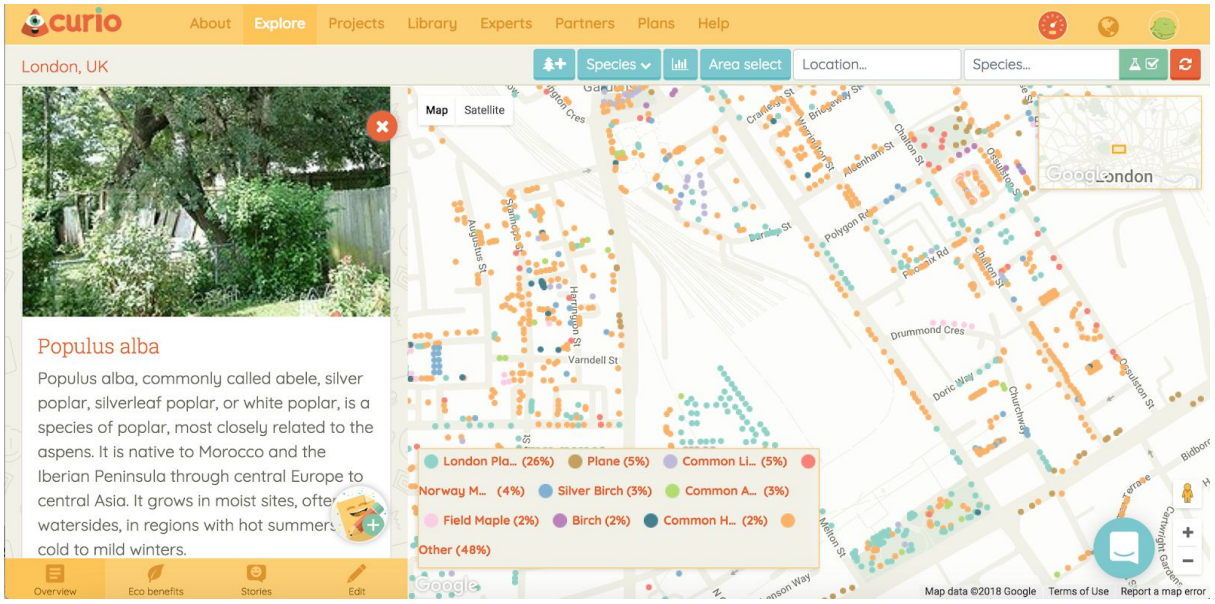
Fig 14. Visualisation of canopy figures broken down per hexagonal grid area



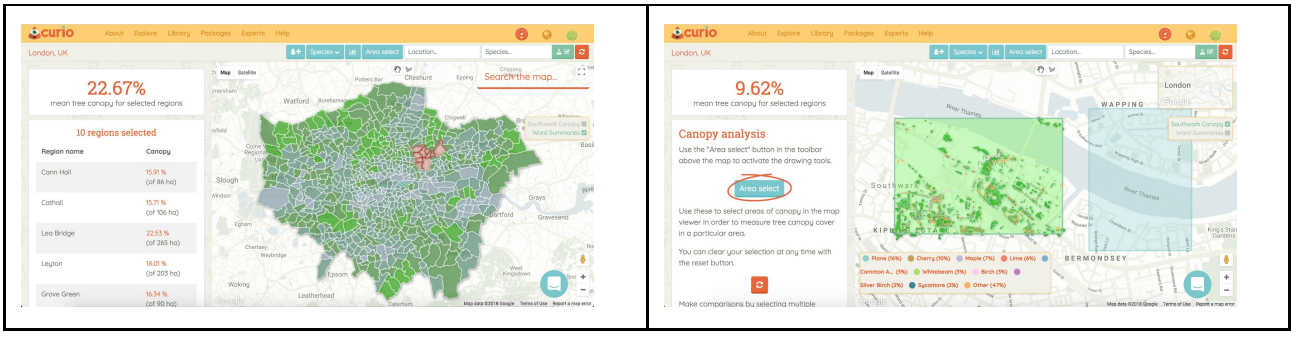
Integration with the Curio Platform

The Curio Canopy project led to the development of the environmentally focused education and outreach platform Curio[16]. The platform hosts information on over 2.1 million trees worldwide. Data on over 750,000 trees in the London area is available for view on the platform. This data comes from a number of sources including the citizen science efforts of Curio users but the vast majority comes from the efforts of the Greater London Authority to produce an open data inventory of London's trees.

[16] https://www.curio.xyz/



It is possible using the Curio platform to view and interact with ward and borough level canopy analysis and also overlay the canopy map on top of inventory data. Combining these data sources makes a lot of interest analysis possible and could also be useful in directing inventory and management activities more effectively. A demo of this is functionality is available to view through [Curio's eco data layers page](https://www.curio.xyz/eco-data-layers)[17].



5. Discussion

The canopy model's performance and the and advantages and disadvantages of the approach taken will now be discussed.

5.1 Model Accuracy Evaluation

A range of different evaluations of the model were carried out using a number of different data sources. In order to train and evaluate the model it was necessary to label the high resolution imagery provided. The mechanism for doing so was described in Section 3.1 and in total 24,000

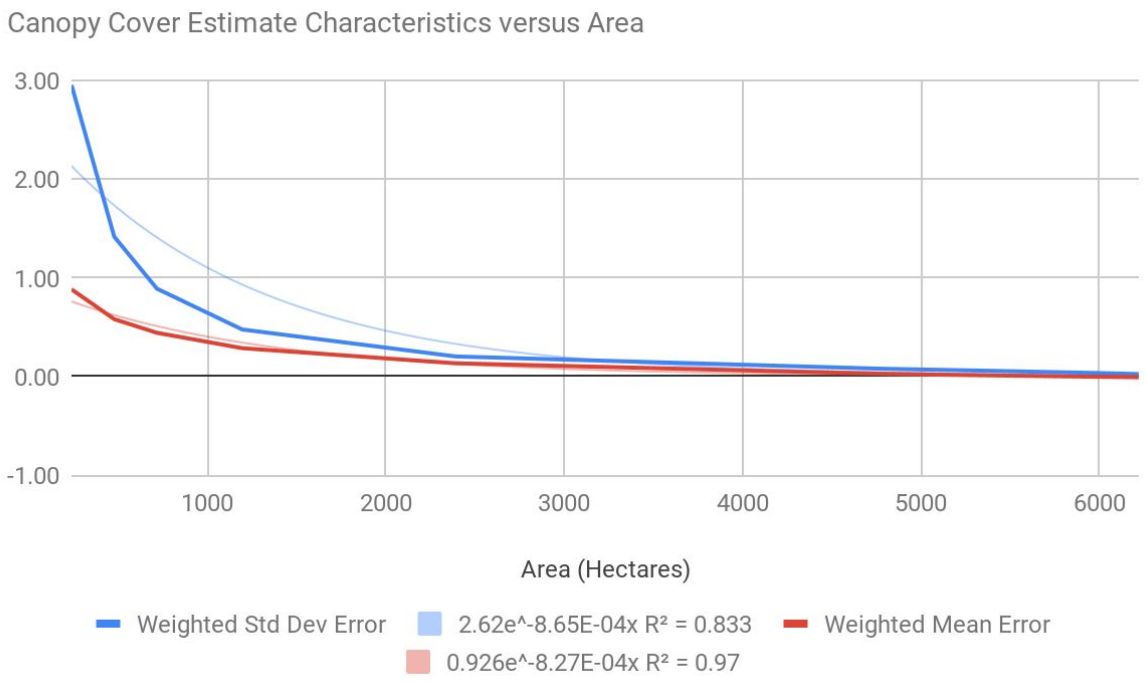
[17] https://www.curio.xyz/eco-data-layers

pixels were labelled. The primary focus of that labelling effort was Ealing with pockets of areas across London targeted as part of the Generalisation Evaluation described in Section 3.2.

Using the full set of data and the 10-fold cross-validation methodology, the model's per pixel accuracy on unseen data was estimated to be 93 ±0.47%. This evaluation involved data primarily from the Ealing area. One of the factors behind selecting Ealing as an evaluation target was that it is comprised of a wide range of urban landscapes, however, in order to verify if these results were reflective of what could be expected in other areas of London the Generalisation Performance evaluation described in Section 3.2 was carried out. This involved labelling data in 11 locations across London in order to determine the model's performance across those areas. The model's performance on this unseen data was found to be 94.87±3.7%. The variation in performance is much higher. This was interesting and this particular evaluation highlights a number of important characteristics of the model. Although overall, on average, the model tends to produce largely unbiased results, neither underestimating or overestimating canopy cover, it can perform poorly in certain circumstances. If the area of interest is quite small, less than 500 hectares for example, it risks targeting areas of terrain on which the model performs particularly strongly or weakly.

The relationship between the size of the area of interest and the standard error estimate should be taken into account when viewing ward, hex-grid or LSOA level estimates as these tend to be smaller and so can be expected to have higher error estimates. For example, the areas under investigation in the Generalised Performance evaluation were just 400 hectares and so the variation in performance is perhaps not surprising. One of the areas in which the model performed particularly badly was dominated by a field containing scrub-like bushes. Areas in which the model is known to perform poorly are discussed in Section 5.3. This same variation in performance was found in the ward level estimates provided which are described in Section 3.2.

An investigation of the model's performance as a function of area was carried out and a very clear pattern was uncovered. The variation in canopy cover error drops quickly as area increases and for areas of interest greater than 1000 hectares it can be expected to be less than 0.5%. It is possible that probability values produced by the model could be used to refine the error estimates for smaller areas however this was not investigated.



The model's performance relative to two i-Tree studies previously carried out in the Ealing and Camden areas was also investigated. The model's estimates generally followed the same trends across the wards investigated as those produced by the i-Tree studies, however, the variation in estimates was, in general, found to be ±3-4%. There are a number of important factors that must be taken into account when viewing the comparative figures. The i-Tree study figures are themselves estimates with margins of error and they have been produced using different imagery that may have been captured at different times of the year or even in different years altogether.

5.2 Benefits

There are a number of benefits to a machine learning and image processing based approach.

Scalability and Level of Detail

One of the big advantages of the approach taken is that once a model has been built it can easily be applied across large areas without any great extra effort being required. This is not the case with sampling based approaches for which the effort required generally increases linearly with the number of areas to be sampled. For instance, to reach a standard error of about ±1.0% using the i-Tree methodology requires roughly 500 points to be generated and labelled per area of interest. For instance, to generate a canopy cover estimate for each of the 630 wards in London would require 315,000 sample points which is a considerable undertaking and would require about

200-300 person days. It's also important to note that the result of that effort would be just an overall estimate per ward with no spatial output indicating where within a ward the tree canopy exists.

Transparency & Reproducibility

The approach taken is easily reproducible and it is easy to visually inspect the the high resolution canopy map from which any canopy cover estimates are derived. The underlying technology used, Google Earth Engine, in particular makes it easy for the code and data used to be shared and for experiments to be reproduced using their infrastructure.

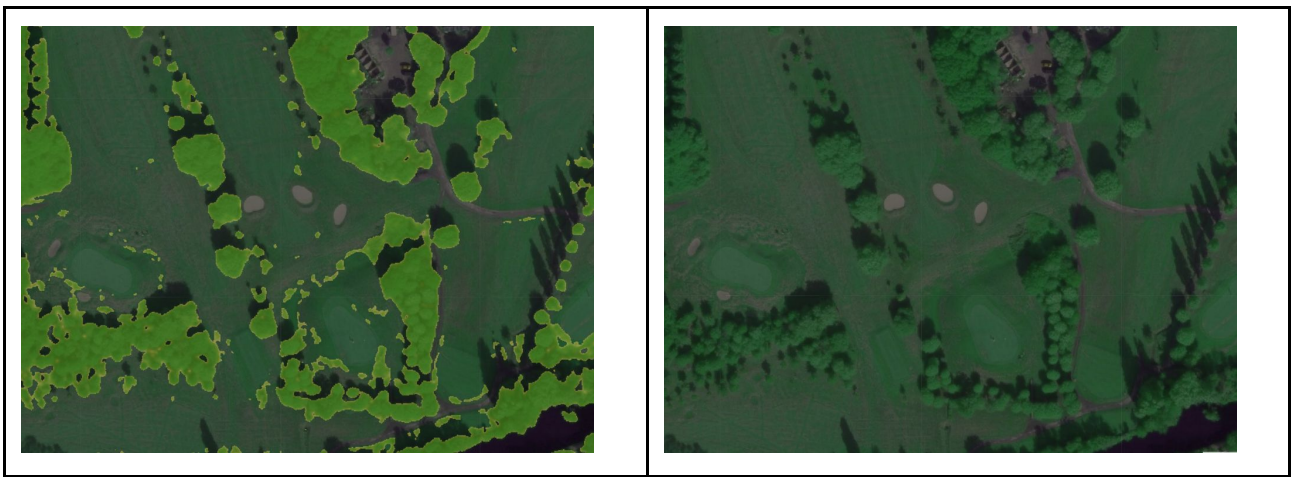
5.3 Limitations

An imperfect model

The machine learning approach taken is an attempt to build an algorithm capable of performing tasks that would normally require human expertise. Although human and superhuman like performance is sought, the approach taken is radically different and an imperfect statistical model whose real world view is limited by the set of features it is given is produced. Such a model can perform quite well but still make errors in ways that we find surprising, this is especially the case in the canopy cover detection model.

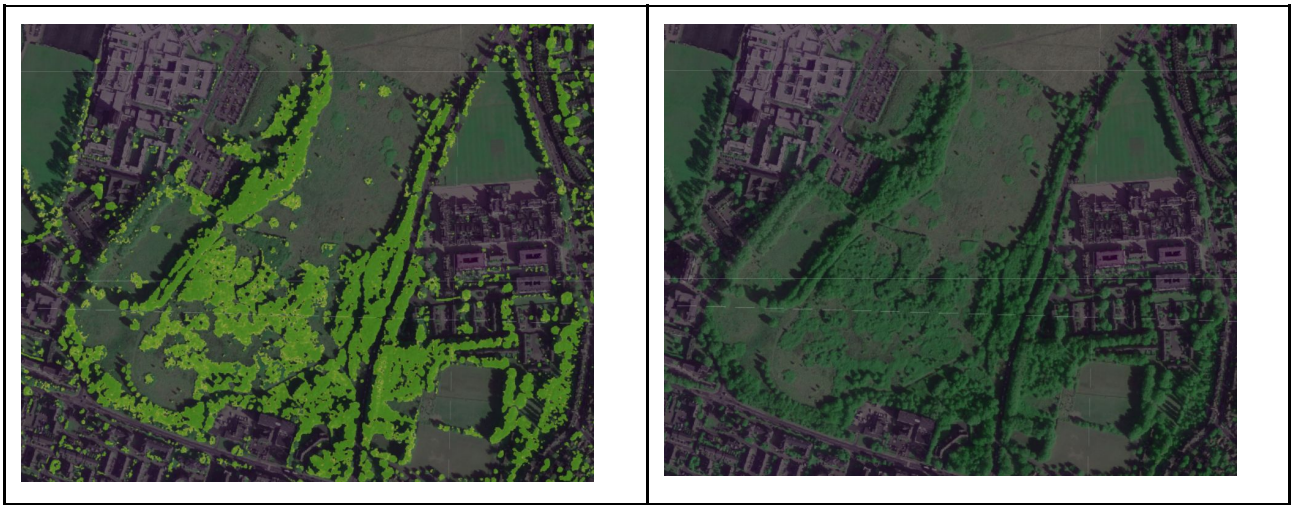
The model built uses the features described in Section 3.1. Naturally, the presence of Normalised Difference Vegetation Index features is extremely useful for detecting vegetation and texture based features are used to help distinguish canopy from grass and other vegetation. However, it was found in the case of golf courses that the model could sometimes perform poorly in surprising ways. It appears that the strong NDVI signals and textures found in fairways can confuse the model. An example of this can been seen in Fig 15 below

Fig 15. NDVI signatures in golf courses



Likewise unusual texture patterns that may not have been covered by the training data used to build the model such as football pitch markings or some grassy fringes can sometimes cause similar issues. The model can also sometimes struggle to distinguish scrub like bushes from tree canopy. The model operates at a per pixel level taking into account the characteristics of the area about 1 metre around that pixel via the features generated. That's a much smaller context than we would, subconsciously, use when making make judgements about objects being trees or bushes. Although when generating training data an effort was made to label scrub land correctly it is clear that despite that training data the model is not always successful in doing so. An example of this can be seen in Fig 16 below.

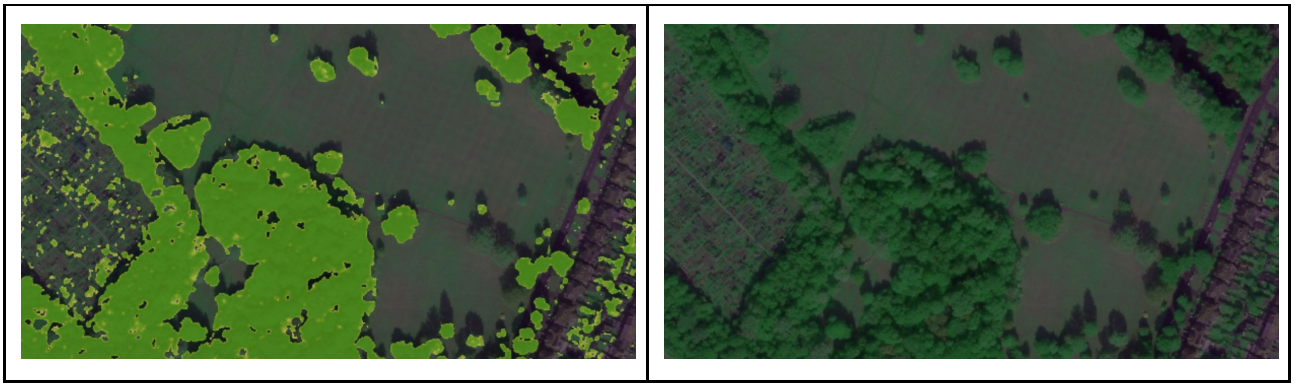
Fig 16. NDVI signatures in scrubland



The canopy visualisation reflects the probability values produced by the model. Lower probability values are yellower in colour. It is noticeable that many of the areas of error tend have lower probability values and be more marginal predictions

The model can also sometimes miss sections of canopy. Possible explanations for this are low NDVI values and certains types of canopy being under represented in the training data used. The presence of low NDVI values can possibly be explained by the fact that the imagery used in this project was captured in September. At that time of the year, although the leaves may still appear to be green, the leaf structure in some cases may have already begun to break down which affects the NDVI signature exhibited.

Fig 17. Examples of missed sections canopy



Having discussed some shortcomings of the model it is worth noting that when the machine learning algorithm is being trained if it isn't able to achieve perfect accuracy on the training data used it tends to balance out the errors it makes, such as the ones just discussed, in such a way as to maximise its overall performance. While the model occasionally behaves in ways that we may find surprising, over larger areas its performance tends to be quite good and we were able to achieve reasonably high accuracy rates.

Transparency

As stated earlier, it is possible to visually inspect the the high resolution canopy map from which any canopy cover estimates are derived. In fact, one of the advantages of the approach taken is the ability to inspect the distribution of canopy cover at such a high resolution. However, to the human eye the errors that can be seen in the high resolution map can affect confidence especially when it is viewed without a full appreciation of the model's characteristics over broader areas.

Aerial imagery

The approach taken requires the availability of high resolution multispectral imagery that ideally should include a near infrared band and be consistent across the area of interest. That imagery must also have been captured at an appropriate time of the year. For instance, as mentioned before, the imagery used in this project was captured in September and the leaf structure of some leaves that still appeared to be green may, importantly, have begun to break down and affected the NDVI signature they exhibited. The lack of height information available in aerial imagery is unfortunate as this would be very useful for distinguishing trees from bushes and for determining the quality and volume of vegetation.

Good Training Data

The model is entirely dependent on the data that is used to train it. The generation of this data can require some effort and it is important that this is done in a way that produces data representative of the intended area to which the model is to be applied. During the course of this work an application was developed to assist with that process. The labelling process is also entirely

dependent on a consistent definition or interpretation of the categories presented to the labeller being applied. It also worth noting that it can sometimes be difficult to accurately label a pixel based on the view provided by aerial imagery and for this reason an 'unsure' category was made available to labellers. These ambiguous points were then excluded from the training and evaluation data sets.

6. Conclusion

A novel and scalable approach to generating rich canopy cover information has been presented. The model developed performs quite well with an accuracy rate of 94% and can be used to produce ward and borough level canopy cover figures that are at least comparable in accuracy to those produced using traditional survey methods. The effort needed to produce such analyses using traditional methods would be so large as to make such similar London wide analysis intractable. It is clear that data and techniques presented make it possible to monitor, measure and manage the urban forest in ways that would be difficult to achieve using traditional means.

7. Future work

A number of related future projects are planned which look into improved data labelling and image recognition techniques. In collaboration with Niamh Donnelly and her supervisor Dr. Brian MacNamee of University College Dublin we have begun investigating the use of Convolution Neural Networks as an alternative canopy image recognition technique.

The importance of creating well balanced training and validation data was also highlighted in the course of this work. Given that canopy analysis is usually a minority class problem doing so in an efficient but balanced way while still capturing the variation in canopy type that exists is an interesting problem and one which will be investigated further.

This work focused on a single source of aerial imagery. An approach based on the use of multiple data sources could be expected to lead to improved results. In particular, data sources such as LIDAR data which can provide valuable height information should be investigated. The availability of height information should help in distinguishing low lying shrubs from trees and could also be used to derive green infrastructure volume information.

It is also planned to further investigate the use of Sentinel 2 imagery for green infrastructure analysis and, in particular, the possibility of deriving a proxy canopy cover scores from the 10 metre per pixel imagery. Such imagery is available world-wide and is freely accessible. Where high resolution imagery isn't available Sentinel 2 derived green infrastructure scores may be useful.