Manuscript essd-2022-259, Second Revision

February 3, 2023

We thank the editors and reviewers for their helpful comments and appreciation. They have lead to substantial improvements in our study. Please find point-by-point responses below. *Our responses look like this*, quotes from the changed text look like this.

Since the last revision, newer aerial imagery covering London became available to us. We therefore updated our work to use the newer imagery, and use the older imagery as a point of comparison for unseen imagery. Furthermore, in response to the reviewers we have implemented four-fold cross-validation. Together these are substantial changes to the dataset, and the results tables have changed completely; we include the new results tables in this letter.

By using four-fold cross-validation, as well as testing against imagery from a different year, we have fully addressed the referees' questions about generalisation of the model to unseen imagery.

1 Referee 1

I thank the authors for providing detailed responses to my original comments. I still have concerns about the inability for the trained CNN to generalise to new images. I believe the major corrections I recommended last time, relating to demonstrating the ability for the tool to generalise to detect green roofs in previously unseen imagery, is still necessary to accept this publication.

I agree with the authors that it is not expected that this tool would be able to detect green roofs in different types of imagery e.g., satellite imagery with a coarser spatial resolution. However, to demonstrate that this paper and tool provides a significant advance, further evidence is required to demonstrate that the tool can at least generalise to detect green roof area in aerial imagery captured by the same or similar sensors at similar times of the year. If this is not possible, it is not clear what the purpose and benefit of training a CNN over other methods and datasets, such as those cited in this manuscript (e.g., LRW2019 and GLA).

Since the previous revision, new aerial imagery of London became available to us, collected in 2021. We therefore updated our study to use the 2021 imagery as its primary dataset, and used imagery from 2019 as an alternative test dataset. When the model trained on the 2021 dataset was tested against the 2019 dataset, precision was lower and recall higher than when tested against the 2021 dataset. However, we also tested performance of a model trained on the 2019 dataset, and found that it performed worse than the model trained on the 2021 dataset even when tested against the 2019 dataset. This demonstrates that the model was able to generalise to unseen imagery.

As a result of the change of dataset, all of the performance statistics have changed. Furthermore, in response to Referee 2 we applied four-fold cross-validation. The results shown in the tables are now the average across folds, whereas before we used a single split. Test data is still completely unseen during training. We therefore include the updated results tables in this letter.

By showing that performance is consistent between training splits through cross-validation, and by testing on completely unseen imagery from a different year, we demonstrate the ability of the

model to generalise.

In Section 2.1 we describe the two imagery sets.

The imagery used for segmentation comprised of raster images with red, green and blue bands from cloud-free mosaics of aerial imagery at 25 cm horizontal resolution (from [Getmapping Plc.(2020),] accessed under an academic license). Two sets of imagery were used, from 2019 and 2021. The imagery from 2021 was used as the primary dataset, with the imagery from 2019 providing an alternative dataset to test generalization. The collection dates for the imagery mosaic covering Greater London are shown in Figure 2. The 2021 imagery covers 1706 km^2 of which 1558 km^2 was inside the Greater London boundary, while the 2019 imagery covers 1527 km^2 of which 1422 km^2 was inside the Greater London boundary.

In Section 2.3 we describe the labels for the two imagery sets.

Labels were initially produced with reference to the 2019 imagery, and then were modified with reference to the 2021 imagery; the labels are different for the two datasets. In total, sample areas covered 7.8% of Inner London, resulting in 4.9×10^4 m^2 (in 2019) and 5.7×10^4 m^2 (in 2021) of green roofs labelled inside the CAZ, and 2.3×10^4 m^2 (in 2019) and 3.3×10^4 m^2 (in 2021) outside the CAZ.

In Section 2.4 we describe the four-fold cross validation.

Four-fold cross-validation was performed; as required computational resources grows with the number of folds, we decided four was a good compromise between testing performance thoroughly and limiting resource usage. The hand-labelled tiles were split into five sets, of which one was reserved as the test dataset. The random split was performed separately for positive and fully negative tiles to ensure all splits contained both positive and negative examples. For each fold, training was performed with 3 of these sets, and validation with one set. This is to demonstrate that good performance is not unique to a particular random split of training and validation data, and therefore tests the ability of the model to generalise. To reduce resource requirements, optimisation of the training method was performed by maximising validation F-score using the first fold only, with only the final selected configuration being cross-validated. The test dataset remained unseen to all models during training, and was not used for choosing the optimal configuration, allowing for a good estimate of out-of-sample performance.

In Section 2.7 we describe testing of the model using the 2019 versus the 2021 imagery.

As an additional test of out-of-sample performance, we included a second imagery dataset from a different year; the primary imagery was for the year 2021, the alternative dataset for the year 2019. First, training was performed using the 2021 imagery and labels, using k-fold cross validation to test the sensitivity of the performance to the train-test split. This model was tested against imagery and labels from the same year (2021) but also from an earlier year (2019). We compared the pixel-value distributions of the roof selected between these datasets. Further, we trained a single model using the 2019 imagery and labels, with exactly the same data split (i.e. the same geographic locations of tiles) as the first fold of the primary model; this model was used to provide a benchmark for the performance of the primary model by testing against both 2021 and 2019 test data. Model design optimisation was performed only with the 2021 imagery and labels.

In Section 3.1 we present the results of the four-fold cross validation.

The performance statistics averaged across folds for green roof identification are given in Table 4; performance statistics for all folds are given in Table A4, and the full confusion matrix in A3. Table 5 gives the same statistics calculated in terms of building counts rather than area; with performance statistics for all folds in Table A3 and the full confusion matrix in A5 Table 6 compares the performance of models trained on 2019 and 2021 imagery and labels, and tested against both 2019 and 2021 imagery and labels.

Results of the hyperparameter search are shown in Table A1. The best performance was found

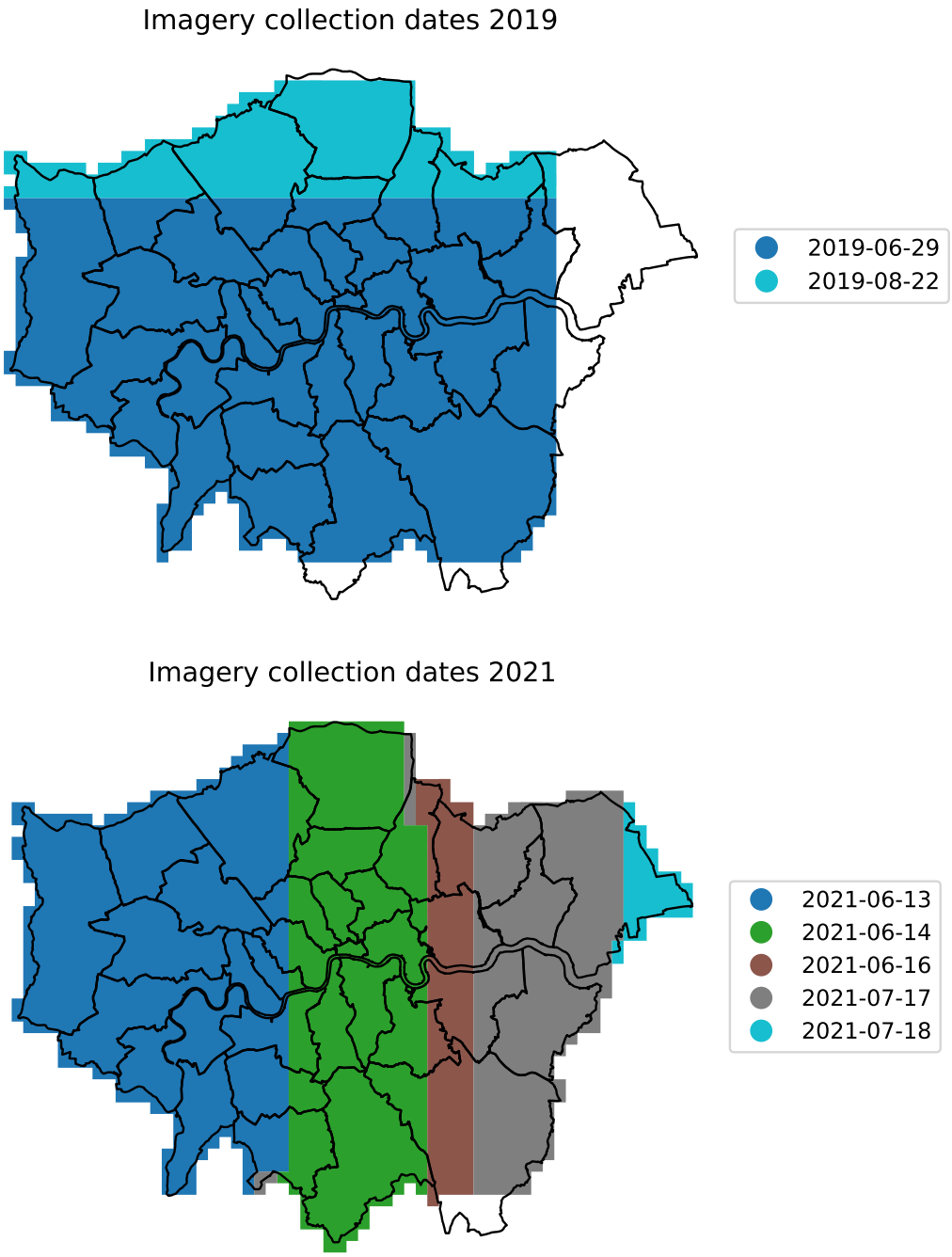

Figure 2: Collection dates for the aerial imagery mosaic covering Greater London. The primary imagery dataset used was that from 2021, while the 2019 data were used for comparison.

Table 4: Performance metrics for the green roof identification method, calculated based on area. For the full set of statistics for all folds see Supplementary Table A4

Dataset	Accuracy	IoU	Precision	Recall	F-score
Testing 2021	0.9925	0.3970	0.6505	0.5046	0.5683
Testing 2019	0.9921	0.3707	0.5072	0.5793	0.5408
Validation	0.9923	0.4666	0.7148	0.5733	0.6363
Training	0.9941	0.5773	0.7374	0.7266	0.7320

Table 5: Performance metrics for the green roof identification method, calculated based on building counts. For the full set of statistics for all folds see Supplementary Table A6

Dataset	Accuracy	IoU	Precision	Recall	F-score
Testing 2021	0.9924	0.4128	0.5721	0.5972	0.5844
Testing 2019	0.9892	0.3256	0.3934	0.6536	0.4912
Validation	0.9920	0.4516	0.6024	0.6433	0.6222
Training	0.9925	0.4444	0.5404	0.7146	0.6154

Table 6: Comparison of test dataset performance model trained on 2019 imagery and labels with the model trained on 2021 imagery and labels.

Dataset	Accuracy	IoU	Precision	Recall	F-score	
Trained on 2019						
Testing 2021	1	0.9921	0.3682	0.6379	0.4655	0.5382
Testing 2019	1	0.9909	0.3580	0.4580	0.6212	0.5273
Trained on 2021						
Testing 2021	1	0.9929	0.4134	0.6801	0.5131	0.5849
Testing 2019	1	0.9923	0.3782	0.5216	0.5790	0.5488

Table A1: Table listing the hyperparameters that were tuned, which values were tested, and the final value used for classification.

Parameter	Tested values	Final value
Loss function	Cross entropy, Lovasz, Focal, F-measure	F-measure
Learning rate	5.e-3, 5.e-4, 5.e-5, 5.e-6	5.e-5
Random augmentations	None; flips and 90^o rotations; crops and flips and 90 ^{o} rotations; flips and fully random rotations; 90% crops and flips and 90^o rotations; flips and 90^o rotations and sharpness; flips and 90^o rotations and sharpness; elastic distortion; alterations to gamma and colour	Flips, rotations, elastic distortions, alterations to gamma and colour
Max epochs	100	100
Pretrained model frozen	True, False	True

Table A3: Full confusion matrix for the green roof identification method, calculated based on area. TP, TN, FP, FN are as a proportion of total building footprint area in the hand-labelled areas.

Dataset	K-fold	Land area (km^2)	Built area (km^2)	TP	TN	FP	FN
Testing 2021	1	3.9	1.282	0.0050	0.9879	0.0024	0.0048
	2	3.9	1.282	0.0047	0.9876	0.0026	0.0051
	3	3.9	1.282	0.0047	0.9872	0.0030	0.0051
	4	3.9	1.282	0.0054	0.9876	0.0026	0.0044
	average	3.9	1.282	0.0049	0.9876	0.0027	0.0048
Testing 2019	1	3.9	1.282	0.0047	0.9877	0.0043	0.0034
	2	3.9	1.282	0.0041	0.9878	0.0042	0.0040
	3	3.9	1.282	0.0048	0.9861	0.0058	0.0033
	4	3.9	1.282	0.0052	0.9880	0.0039	0.0029
	average	3.9	1.282	0.0047	0.9874	0.0045	0.0034
Validation	1	4.0	1.333	0.0060	0.9887	0.0021	0.0033
	2	4.0	1.278	0.0070	0.9851	0.0032	0.0047
	3	3.9	1.255	0.0060	0.9839	0.0030	0.0071
	4	4.0	1.261	0.0079	0.9844	0.0026	0.0052
	average	4.0	1.282	0.0067	0.9856	0.0027	0.0050
Training	1	12.0	3.798	0.0089	0.9851	0.0029	0.0031
	2	12.0	3.853	0.0075	0.9863	0.0025	0.0036
	3	12.1	3.875	0.0077	0.9860	0.0033	0.0030
	4	12.0	3.870	0.0084	0.9863	0.0028	0.0025
	average	12.0	3.849	0.0081	0.9859	0.0029	0.0031

Table A4: Full performance metrics for the green roof identification method, calculated based on area.

Dataset	K-fold	Accuracy	IoU	Precision	Recall	F-score
Testing 2021	1	0.9929	0.4134	0.6801	0.5131	0.5849
	2	0.9923	0.3774	0.6423	0.4778	0.5480
	3	0.9919	0.3659	0.6075	0.4792	0.5358
	4	0.9930	0.4325	0.6720	0.5482	0.6038
	average	0.9925	0.3970	0.6505	0.5046	0.5683
Testing 2019	1	0.9923	0.3782	0.5216	0.5790	0.5488
	2	0.9918	0.3329	0.4946	0.5044	0.4995
	3	0.9909	0.3436	0.4509	0.5907	0.5114
	4	0.9932	0.4328	0.5697	0.6431	0.6042
	average	0.9921	0.3707	0.5072	0.5793	0.5408
Validation	1	0.9947	0.5298	0.7454	0.6468	0.6926
	2	0.9922	0.4735	0.6893	0.6020	0.6427
	3	0.9899	0.3754	0.6696	0.4607	0.5459
	4	0.9923	0.5048	0.7538	0.6045	0.6709
	average	0.9923	0.4666	0.7148	0.5733	0.6363
Training	1	0.9940	0.5957	0.7514	0.7419	0.7466
	2	0.9938	0.5497	0.7487	0.6741	0.7094
	3	0.9937	0.5510	0.7018	0.7194	0.7105
	4	0.9947	0.6120	0.7479	0.7710	0.7593
	average	0.9941	0.5773	0.7374	0.7266	0.7320

Table A5: Full confusion matrix for the green roof identification method, calculated based on counts of buildings. TP, TN, FP, FN are as a proportion of total building footprint area.

Dataset	K-fold	Building count	TP	TN	FP	FN
Testing 2021 1	12026	0.0063	0.9835	0.0075	0.0027	
	2	12026	0.0067	0.9742	0.0168	0.0022
	3	12026	0.0068	0.9727	0.0183	0.0022
	4	12026	0.0070	0.9791	0.0119	0.0020
	average	12026	0.0067	0.9774	0.0136	0.0023
Testing 2019	1	12026	0.0067	0.9739	0.0181	0.0013
	2	12026	0.0064	0.9617	0.0304	0.0016
	3	12026	0.0067	0.9543	0.0378	0.0012
	4	12026	0.0068	0.9618	0.0302	0.0012
	average	12026	0.0067	0.9629	0.0291	0.0013
Validation	1	11505	0.0077	0.9832	0.0066	0.0024
	2	11724	0.0078	0.9701	0.0195	0.0026
	3	11167	0.0081	0.9684	0.0215	0.0021
	4	11820	0.0085	0.9767	0.0132	0.0016
	average	11554	0.0080	0.9746	0.0152	0.0022
Training	1	30932	0.0065	0.9820	0.0100	0.0016
	2	30717	0.0072	0.9705	0.0209	0.0015
	3	31069	0.0070	0.9674	0.0242	0.0014
	4	30731	0.0071	0.9758	0.0159	0.0012
	average	30862	0.0069	0.9739	0.0177	0.0014

Table A6: Full Performance metrics for the green roof identification method, calculated based on building counts.

Dataset	K-fold	Accuracy	IoU	Precision	Recall	F-score
Testing 2021	1	0.9932	0.4266	0.6354	0.5648	0.5980
	2	0.9919	0.3938	0.5478	0.5833	0.5650
	3	0.9916	0.3952	0.5280	0.6111	0.5665
	4	0.9928	0.4387	0.5913	0.6296	0.6099
	average	0.9924	0.4128	0.5721	0.5972	0.5844
Testing 2019	1	0.9919	0.3718	0.4915	0.6042	0.5421
	2	0.9899	0.3260	0.4097	0.6146	0.4917
	3	0.9854	0.2798	0.3163	0.7083	0.4373
	4	0.9896	0.3455	0.4099	0.6875	0.5136
	average	0.9892	0.3256	0.3934	0.6536	0.4912
Validation	1	0.9937	0.5000	0.7157	0.6239	0.6667
	2	0.9902	0.3883	0.5252	0.5984	0.5594
	3	0.9909	0.4171	0.5407	0.6460	0.5887
	4	0.9934	0.5185	0.6614	0.7059	0.6829
	average	0.9920	0.4516	0.6024	0.6433	0.6222
Training	1	0.9937	0.4771	0.5920	0.7108	0.6460
	2	0.9923	0.4381	0.5428	0.6943	0.6093
	3	0.9914	0.4145	0.4922	0.7241	0.5860
	4	0.9927	0.4548	0.5471	0.7294	0.6252
	average	0.9925	0.4444	0.5404	0.7146	0.6154

Table A7: Standard deviation of performance metrics between folds, calculated using area.

Dataset	Accuracy	IoU	Precision	Recall	F-score
Testing 2021	0.0005	0.0310	0.0330	0.0333	0.0317
Testing 2019	0.0010	0.0450	0.0497	0.0572	0.0471
Validation	0.0019	0.0677	0.0414	0.0812	0.0647
Training	0.0004	0.0316	0.0238	0.0409	0.0254

with the F-measure loss (F-score improvement 0.3), which may be because the class imbalance is large. We found that the augmentation which provided the greatest improvements in performance were the non-destructive transformations (flips and rotations) which provided an F-score improvement of 0.10 versus no augmentations, the effect of the other augmentations (elastic tranformation, colour shift, random gamma adjustment) were smaller, improving F-score by only a further 0.03. We found that over-sampling the positive tiles was more effective than not including any negative-only tiles, including all tiles without resampling, or under-sampling negative tiles. We experimented with the proportion of positive tiles to be achieved by resampling, and found the best results when 50% of tiles contained positive pixels. Training was roughly two times faster per epoch with the pre-trained part of the model frozen, so augmentation experiments were performed with it frozen, when the best combination was found training was repeated with the model un-frozen but this did not lead to an increase in F-score. We found that the building-intersection step increased testing precision by 0.05 on average across the folds for the 2021 testing dataset and 0.11 for the 2019 testing dataset with no effect on recall, showing that across the building-intersection step plays an important role in suppressing false positives. Figure 6 shows the distribution of colours in the predictions for the two imagery sets: generally true positives, false positives, and false negatives have strongly overlapping colour distributions which are similar between the two imagery sets.

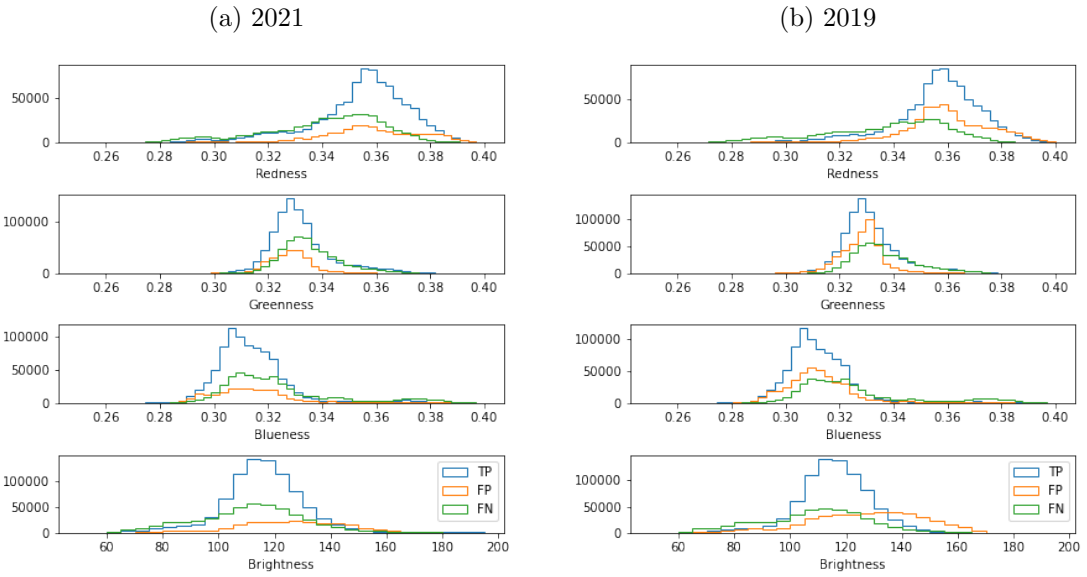

Figure 6: Colour and brightness of pixels in false positive (FP), true positive (TP), and false negative (FN) groups for (A) 2021 imagery and labels, (B) 2019 imagery and labels. The model was trained on 2021 imagery and labels.

In Section 4.1 we discuss the performance of the model, including on unseen imagery.

The segmentation model achieves a high level of accuracy (0.99). Precision and recall based on area for the 2021 testing dataset are 0.65 and 0.50 respectively, with an F-score of 0.57 (Table 4). Based on counts of buildings instead, precision is lower (0.57) and recall is higher (0.60). This indicates that the model is effective at identifying green roofs, and that many of the false positives are small areas on buildings with no green roof.

Given that the survey covers such a large and diverse area, and the green roof fraction is low in many areas, it is important to consider the false positive rate. Tables A3 and A5 suggest that we expect 0.3% of the built area to be incorrectly identified as green roof, which is comparable to the

green roof area in some districts that have very little green roof e.g. Waltham Forest, but small in areas with more green roof.

Comparing the performance of the same model (trained on 2021 imagery and labels) for the two testing datasets (2021 versus 2019), precision was lower for the 2019 dataset (from 0.57 to 0.39) but recall was higher (from 0.59 to 0.65) (Table 4). This means that with the alternative imagery dataset the model tends to include more a higher proportion of spurious green roofs. The difference in precision is greater when calculated in terms of building counts rather than area (0.59 to 0.39) (Table 5), suggesting that the additional false positives take the form of small areas on buildings without real green roofs. Imagery in the alternative set was completely unseen during training and optimisation. However, as Table 6 shows, performance is just as good or better than a model trained on the 2019 images and labels. This demonstrates model can generalise to unseen imagery, although with some loss of precision.

It is recognised as a strength to this manuscript that the dataset is being made publicly available and that it covers the entire Greater London area, but if these are the primary novel factors in the paper, then it is at odds with the text which focusses primarily on the development and training of the CNN. Further emphasis has to be placed in the discussion section on what the repercussions of deriving such a dataset is. What is the benefit of having this information over and above what was provided in the previous datasets? Section 4.3. discusses the differences between the figures contained within the different datasets, but little information is provided on the repercussions of this.

In order to emphasise the additional value provided by this dataset, we have added some detail to the discussion section covering the implications of the differing estimates. We have also added a new Section 4.5 which discusses the additional utility of this dataset through providing data at the level of single buildings. The previous studies do not make available their detailed geospatial results, only district-level averages.

In Section 4.3 we compare our district-level estimates with the results of earlier studies.

Our estimate of green roof area in the CAZ in 2021 ($2.3 \times 10^5\ m^2$) is higher than the LRW2019 estimates and the AMR estimate for 2013 and 2015, but lower than the AMR estimates for 2017. For Greater London, the identified area is higher than the 2016 and 2017 estimated areas from LRW2019. While individual-building data from previous studies are not available for comparison, local-authority district (LAD) level data are available from for 2017 from LRW2019 [Livingroofs Enterprises Ltd(2019), European Federation of Green Roof and Green Wall Associations (EFB) and] In Figure 10, we compare our results for 2019 with the estimates for each LAD in 2017 from LRW2019: the results are strongly correlated, but some LADs have quite different results. According to this, most LADs have gained some green roof between 2017, with a few losing some. Newham (Nwm) and Hillingdon (Hdn) appear to have gained the most green roofs between 2017 and 2019. Our estimate for Havering (Hvg) is close to zero, because the 2019 imagery does not cover Havering (see Figure 2). Where estimates are differ by a small amount it may be due to differences in methodology or errors rather than a real change.

In Section 4.5 we discuss the use of the new dataset, and give an example of the kind of analysis that can be performed.

The dataset provides far greater detail than is available publicly from previous work in London. Green roof polygons are provided for individual areas of green roof, and are identifiable for individual buildings. This will enable new insights into the distribution of green roofs in London which were not possible before. For example, using the building use classifications given by the UKBuildings dataset, we can calculate the distribution of green roofs between building uses. As shown by Figure 11, non-residential buildings make up most of the buildings with green roofs (56%), with around 1.2% of non-residential building footprint area covered by green roofs compared to 0.3%

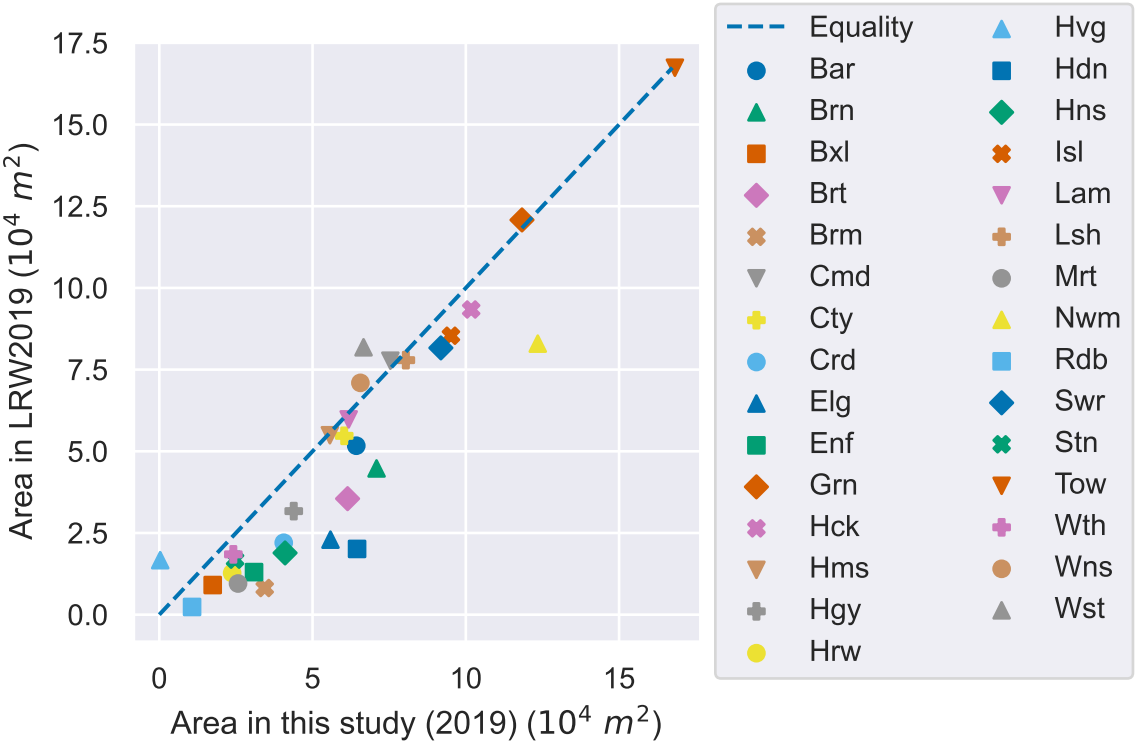

Figure 10: Scatter plot showing estimated green roof area in LADs of Greater London, comparing the estimates from [Livingroofs Enterprises Ltd(2019)] (2017) to our estimates (2019).

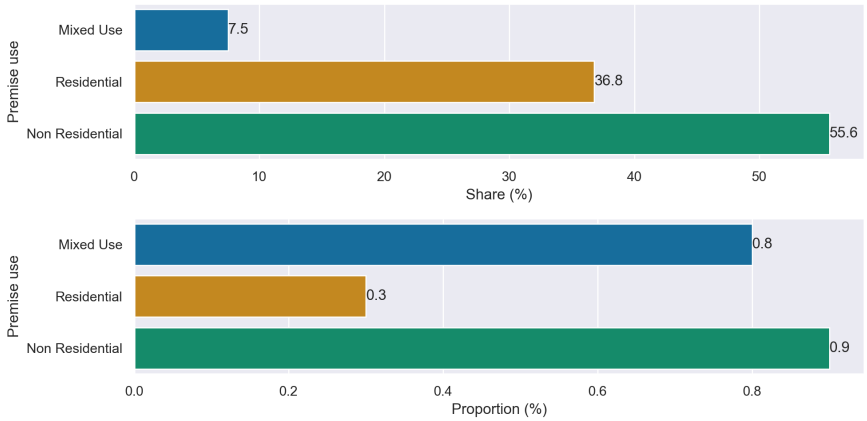

Figure 11: Share and proportion of green roofs for different building uses. Most green roofs are on non-residential buildings. Mixed use refers to buildings comprising both residential and non-residential uses.

of residential buildings. While a large fraction of green roofs occur on residential buildings, only a small proportion of residential buildings have a green roof. This illustrates the utility that this level of detail brings. Future work will extend this analysis to look in detail at the characteristics of buildings that have green roofs in London.

2 Referee 2

Summary: In the revised manuscript, the authors have clarified many things. My concern, however, remains the potential issues arising from the use a single training/validation split.

Major comment:

1. The major concern I have is the lack of cross-validation across training sets. The current model is constrained to work for the chosen training data and it is unclear how the feature selection would be impacted by a different training set. Some kind of cross-validation (Monte Carlo or k-folds) should at least be performed post hoc to confirm the ability of the model to generalize across training splits.

In response to this comment we have applied four-fold cross-validation. Models were trained on four different train-test splits. Generalisation was also tested against unseen imagery collected in a different year. Please see the answer to Referee 1 for details.

Minor comment:

1. colour (red, green, blue) raster images sounds strange. Maybe say rasters with red, green, and blue bands.
2. ‘ green roofs impose an structural loads and additional costs,’ no ‘an’

We have corrected these mistakes.

References

[European Federation of Green Roof and Green Wall Associations (EFB) and Livingroofs.org on behalf of the Gr
European Federation of Green Roof and Green Wall Associations (EFB) and Livingroofs.org on behalf of the Greater London Authority: Living Roofs and Walls: From Policy to Practice, https://livingroofs.org/wp-content/uploads/2019/05/LONDON-LIVING-ROOFS-WALLS-REPORT$_MAY-$2019.*pdf*, *accessed* : $2022-02-11, 2019$.

[Getmapping Plc.(2020)] Getmapping Plc.: High Resolution (25cm) Vertical Aerial Imagery, EDINA Aerial Digimap Service, https://digimap.edina.ac.uk, updated: 2020-09-27, Downloaded: 2021-10-14, 2020.

[Livingroofs Enterprises Ltd(2019)] Livingroofs Enterprises Ltd: London borough green roof infographics and maps, https://livingroofs.org/borough-green-roof-infographics-maps-london-green-r
accessed: 2022-02-11, 2019.