
Article:
Lambert, J.P.T. orcid.org/0000-0001-7034-7219, Childs, D.Z. orcid.org/0000-0002-0675-4933 and Freckleton, R.P. orcid.org/0000-0002-8338-864X (2019) Testing the ability of Unmanned Aerial Systems and machine learning to map weeds at subfield scales: a test with the weed Alopecurus myosuroides (Huds). Pest Management Science. ISSN 1526-498X

https://doi.org/10.1002/ps.5444

Testing the ability of Unmanned Aerial Systems and machine learning to map weeds at subfield scales: a test with the weed Alopecurus myosuroides (Huds).

J P T Lambert, D Z Childs & R P Freckleton.

Correspondence:

JPT Lambert, Department of Animal & Plant Science, University of Sheffield, Western Bank, Sheffield, S10 2TN.

E-mail: Jptlambert1@sheffield.ac.uk

U.K. Tel: (+44) 114 222 0017

Abstract

BACKGROUND

It is important to map agricultural weed populations in order to improve management and maintain future food security. Advances in data collection and statistical methodology have created new opportunities to aid in the mapping of weed populations. We set out to apply these new methodologies (Unmanned Aerial Systems - UAS) and statistical techniques (Convolutional Neural Networks – CNN) for the mapping of black-grass, a highly impactful weed in wheat fields in the UK. We tested this by undertaking an extensive UAS and field-based mapping over the course of two years, in total collecting multispectral image data from 102 fields, with 76 providing informative data. We used these data to construct a Vegetation Index (VI), that we used to train a custom CNN model from scratch. We undertook a suite of data engineering techniques, such as balancing and cleaning to optimize performance of our metrics. We also investigate the transferability of the models from one field to another.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/ps.5444

RESULTS

The results show that our data collection methodology and implementation of CNN outperform pervious approaches in the literature. We show that data engineering to account for “artefacts” in the image data increases our metrics significantly. We are not able to identify any traits that are shared between fields that result in high scores from our novel leave one field our cross validation (LOFO-CV) tests.

CONCLUSION

We conclude that this evaluation procedure is a better estimation of real-world predictive value when compared to past studies. We conclude that by engineering the image data set into discrete classes of data quality we increase the prediction accuracy from the baseline model by 5% to an AUC of 0.825. We find that the temporal effects studied here have no effect on our ability to model weed densities.

Keywords:

Unmanned Aerial Systems, weed mapping, Convolutional Neural Networks, black-grass, management

1.1 Introduction

The core objective of plant population ecology is to understand changes in numbers of individuals/organisms across time and space (1). Achieving this depends on methods that permit plants to be mapped and monitored at informative scales (2-4). Surveys of plant populations have been undertaken using a variety of different methods such as transect sampling, quadrat sampling and with Unmanned Ariel Systems (UAS) (5-7). Each of these methods has an inherent trade-off between the area that can be surveyed and the intensity at which the subjects can be studied in that area (8). Transect and quadrat sampling can be either used for small area, high intensity studies or large area, low intensity studies, but typically not both (9).

UAS present a unique opportunity for ecological monitoring because, potentially, they can yield data across both large spatial areas and at high survey intensity. This bridges the gap between local scales at which interactions matter, and larger landscape scales at which environmental variation is important (10). UAS have been applied in a range of ecological scenarios including mapping communities (11), population monitoring (12) and mapping individuals in small areas (13). However, few studies have focused on mapping populations at differing times and places, or the challenges of the homogeneous of the environment.

An economically important agricultural crop such as winter wheat (Triticum aestivum L.) may be significantly impacted by competition from weeds (14). Weed species add additional costs to the production of crops by increasing the need for agricultural inputs: e.g. in one national-scale audit, it was estimated that weeds cost the Australian economy A$3.5B a year (15). Monitoring data can reduce costs by facilitating precision application of inputs such as herbicides, or better-informed cultural management (16). Ecological monitoring depends on being able to locate and enumerate individuals or species within a given

environment (17). Patches of weeds have shown to be persistent over 10 years, therefore mapping in one year represents a potential predictor of future occurrence (18). There are many challenges in the mapping of weeds such as their fast growth rates, and highly variable spatial and temporal distributions (19). Given the potential value of monitoring data, and the possibility of rapid large-scale acquisition of data using UAS, there is clear interest by researchers and farmers in applying this technology to measure weed populations (20).

Despite the potential for data derived from UAS to improve weed management, previous research has highlighted significant issues in the use of them to monitor weed populations (6). Specifically, images and models calibrated to measure weeds in one environment appear to perform poorly when transferred to another. There are several reasons for this limited transferability, for example, variation in weather conditions or different growth stages of the weed or crop. As crop plants grow over the field season their phenology changes, as does that of the weeds (21). This results in changes in the spectral properties of the crop and weed species, both in the visible spectrum and beyond (22, 23). Moreover, common crops are grown in many different varieties, each with their own unique phenology and physiology (24-26). The statistical methodology of random forests (RF) and a dataset of mean pixel values from UAS image plots, as used in our previous study of weed monitoring does not fully capture the extent of these variations, thus failing to generate highly transferable models (6).

Supervised machine learning is a statistical method that generates a classification output after being presented with an unclassified input, having previously been trained on data consisting of known inputs and outputs (27). All such models are trained using "features". A feature is a numeric representation of the unclassified input. In the case of an image input, these can be engineered by researchers i.e. texture, colour, shape or they can be

abstractly and randomly defined by the model and adapted over iterations. Here we highlight key network methods that are used in supervised machine learning.

Neural Networks (NN) conceptually mimic biological neurons in their node-like structure. Each node is interconnected to others and sends a "signal" if threshold values are passed. Threshold values are tuneable at each node and are adjusted automatically over the course of fitting the model. An important advantage of NN is that they can bypass the need for domain knowledge of the dataset (feature engineering), allowing more abstract and potentially useful features to be used. This does, however, make the model less interpretable, as the features that are used are selected without logical justification. As with most statistical methods, NN perform better when trained on more data.

CNN are a type of NN specifically applied to image data sets. Convolutional Neural Networks (CNN) have emerged as the most common, and frequently best performing, model for image classification tasks in the machine learning literature (28). CNN learn a sparser connection between regions of an image than traditional NN models by imposing spatial dependencies upon the pixels in the image (29). This may be of use when analysing weed distributions because these are spatially dependant (30-32). CNN do not use user defined features such as colour, shape or texture to learn from the data. Instead CNN create abstract feature maps and then through training/iterations, assigns importance to different feature maps (33) representing different states in the image. These components of a CNN make them a well-suited method for mapping weed populations, but the underpinning model correspondingly harder to interpret. Spatial information is retained, and automated abstract feature identification can identify common aspects among the classes of data that human feature selection would otherwise miss (34).

Here we investigate how images collected from UAS can be classified using CNN to predict weed densities in unseen images. We explore how data engineering can be undertaken to improve the results and account for the heterogenous nature of the environment. We also investigate the seasonal effects of mapping on our ability to correctly predict weed densities by comparing our models between years and the week of survey, thus addressing key limitations from past literature. Finally we assess true out of sample predictions of CNN models to assess their transferability across populations.

2.1 Materials & Methodology

2.1.1 Description of dataset

We studied Alopecurus myosuroides (black-grass) in populations of Triticum aestivum L. (winter wheat). 1.9 million hectares of wheat is cultivated per year in the UK, making it the most widely grown crop, with A. myosuroides becoming a significant problem throughout the UK (35).

Our field sites were part of an ongoing study by the Black Grass Resistance Initiative (BGRI) into herbicide resistance levels in the weed nationally. We surveyed 102 new fields across the arable regions of the UK. Late season monitoring (13rd June – 12th August in 2016 and 2017) was chosen as previous work shows that the weeds are distinguishable from the surrounding wheat crops at this time (6). This represents a BBCH weed growth stage of 87-89 (36).

Fields were subject to a range of differing management practices, across farms from 80 to 3000 ha. The populations of black-grass had previously been measured in fields using the methodology developed by (3, 35) to estimate plant density states in a plot. Plots of 20x20m were chosen as this allowed large amounts of contiguous ground-truthed data on the

densities of black-grass in a field to be collected. The average field was 8ha with 110 plots per field, depending on the varying extents of the field. Five ordinal density states of black-grass were denoted: absent, low, medium, high, very high,(0, 1-160, 161-450, 451-1450 and 1451+, plants per $20m^2$ respectively). This method allows for multiple observers to be used, enabling large spatial scales to be covered with minimal misclassification error between observers.

2.1.2 UAS platform

A widely available commercial UAS platform was chosen to allow for low entry costs and high repeatability. We used the 3DR solo UAS[1] as it permits third party imaging systems to be attached and operated. The Parrot Sequoia[2] was chosen as the imaging sensor as this sensor has been specifically designed for use with UAS. This sensor records images in four discrete calibrated spectral channels: Green 550nm (f_g), Red 660nm (f_r), Red-Edge 735nm (f_{re}) and Near Infrared 790nm (f_n) at 1.2Mp. The sensor possesses a "sunshine sensor" that standardised against variable lighting conditions over the course of a flight by continuously recording the light conditions in each spectral channel and then automatically calibrating the outputs to the absolute values.

All flights were carried out following UK rules and regulations controlling the use of UAS for scientific research. Flights were conducted within 2 hours either side of solar noon to reduce the effect of sun angle. The optimum flight parameters to cover each field in the minimal amount of time were a flight height of 100m and an image overlap of 60% (37). Each flight generated thousands of subfield scale images that are stitched together to create a single orthomosaic image, encompassing an entire field using relatively few ground control points. For this Agisoft Photoscan was used. This software also creates Vegetation Indices

[1] "Solo - The Smart Drone | Commercial Drone Platform." https://3dr.com/solo-drone/. Accessed 11 Jan. 2018.

[2] "Sequoia - MicaSense." https://www.micasense.com/sequoia/. Accessed 11 Jan. 2018.

(VIs) from the individual bands of the sequoia. The average ground sample distance (GSD) of all the flights was 8.27cm pixel^{-1}.

Of the 102 fields that were flown, 76 generated data of high enough quality to analyse. The fields that were not suitable to be analysed were discarded for the following reasons: poor image quality, significant image stitching artefacts and sensor failure.

The calibrated spectral channels of the sequoia sensor allow for VIs to be calculated for each pixel. VIs are used as they reduce multiband observations to a single numerical index (38). We used Green Normalized Differential Vegetation Index GNDVI (equation 1) to classify images:

$$GNDVI = \frac{f_n - f_g}{f_n + f_g} \quad (1)$$

All subsequent references to the data, refers to the GNDVI dataset See appendix Table 5, for statistical measurements of the GNDVI dataset.

Our choice to base our analysis on GNDVI is because high biomass crops such as wheat cause saturation of chlorophyll levels in the red wavelength, resulting in poor performance when using Normalized Differential Vegetation Index NDVI (equation 2) (39).

$$NDVI = \frac{f_r - f_n}{f_r + f_n} \quad (2)$$

Previous studies have focused on the NDVI owing to its correlation with plant vigour and growth (40). However, when needing to discriminate between invasive populations, vigour and growth rates with NDVI has shown to be uninformative in cases of high saturation of a spectral channel (41). Analysis based on UAS imagery has often overlooked this feature of NDVI , but is recognised in satellite remote sensing work (42-44).

The ground-truthed density data were overlaid on each georectified orthomosaic using GIS packages in R. Then the orthomosaic maps were split into 20x20m subplots, each geographically relating to the ground-truthed observations. This creates a dataset of images at the 20x20m scale, which our subsequent analysis area is based on. The resulting image dataset consists of 12,313 unique measurements of black-grass at 20x20m scale covering the full range of black-grass densities. The densities are however not evenly distributed. The breakdown as follows: Absent = 14.5% Low = 53.1% Med = 17.3% High = 8.2% Very High = 6.9%.

2.1.3 Modelling approach and metrics

We used a CNN to train a classifier on our black-grass image data. The model structure was taken from one of the top performing methods on the industry standard image database, ImageNet (45), called GoogLeNet (34). Whilst we use the structure of GoogLeNet, it is important to note that we do not use the pretrained model weights and biases that allowed the model to score so highly on ImageNet. Here we highlight four common components of our chosen model framework, that are then stacked together with other components such as batch normalisation and dropout to create a variety of different network structures:

(1) Convolution: The convolutional step involves extracting features from an image whilst maintaining their spatial context, by using a filter to pass over an image and computing the dot product to create a generalised feature map.

(2) Addition of Non-Linearity: Non-Linearity is introduced to the feature maps by applying a Rectified Linear Unit (ReLU), this speeds up the training process when compared to tanh/sigmoid activation functions. This means that model convergence will occur with a lower computational cost (46).

(3) Pooling: Pooling of the feature map is used to reduce dimensionality. This reduces the parameter number in the network, a key stage in preventing overfitting. Pooling also makes the network more stable to distortions in the training images (47).

(4) Fully connected final layer: This combines all the neurons of the previous layer and applies an activation function to determine the final classification of an image. The most common form of activation function is SoftMax and the predictions always sum to 1 (48).

CNN have been successfully applied to many datasets similar to ImageNet through a process known as transfer learning, whereby only the weights of the connected final layer of a pretrained model are altered (49). We do not use the process of transfer learning as our proposed dataset is significantly different from that of ImageNet. Instead, we use the GoogleLeNet structure and independently train all layers of our model.

To model a CNN three data sets are needed: training, validation and test sets. Each dataset comprises pairs of input images and target vectors. Target vectors act as a labelling method and are what the model tries to predict when given a new image. In our example the input image is a 20x20m image plot and the target vector represents the five different ordinal density states. CNN are trained using a variety of parameters. From our initial exploration of the modelling we settled on using the following as our standards: a decaying momentum beginning at 0.1 and halving every 32000 steps as our optimizer, categorical cross entropy as our loss function and a batch size of 128.

We report, where appropriate, three metrics for our models. These are: Multiclass AUC, Cohen's kappa and weighted Cohen's kappa. AUC refers to the Area Under the Receiver Operating Characteristic (ROC) curve, that is the true positive rate (Sensitivity) against the true negative rate (Specificity). AUC is used for its ability to differentiate between two groups, and is equal to the probability that the classifier will rank a randomly chosen

positive example higher than a randomly chosen negative example (50). AUC values range from 0 – 1. We plot a diagonal line from (x=0, y=1) to (x=1, y=0) known as the line of equality or the random chance line (51). Points that fall below this line represent non-informative models where random classification would perform better. For the x-axis in our AUC plots we use 1 – Specificity.

The categorical predictions of a model and ground-truthed observations can be viewed as different raters. This allows us to assess the degree to which they agree or disagree and utilise Cohen's kappa statistic (52) (equation 3):

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{3}$$

Where ρ_o is the observed agreement and ρ_e is agreement due to chance. This results in a range from 1 indicating complete agreement between raters, through 0 indicating that agreement is only due to random allocation and -1 indicating complete disagreement.

AUC and kappa do not consider the ordinal structure of our data, with observations ranging from Absent to Very High in incrementing ordered categories. Therefore, an observation of Absent and a prediction of Low is closer to agreeing than if the prediction were Very High. We therefore used weighted Cohen's kappa (equation 4):

$$\kappa_w = 1 - \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} \omega_{ij} x_{ij}}{\sum_{i=1}^{k} \sum_{j=1}^{k} \omega_{ij} m_{ij}} \tag{4}$$

Where κ is the number of categories, ω_{ij}, χ_{ij} and m_{ij} represent the weight from the matrix. This allows us to count disagreements differently (53). The weighted kappa is on the same scale and distribution as the base Cohen's kappa. We use a squared weighting matrix of 1, 4, 9, 16 and 25 ranging from agreement to significant disagreement, to penalise significantly wrong agreements.

2.1.4 Model refinement: data balancing

We checked the performance of the model in several respects. First, we analysed the effect of balancing the data in terms of the distribution of observations among density states. This is important because the dataset is heavily weighted towards the Low-density state, comprising over 50% of the dataset. Such imbalanced distributions can lead to lazy or biased classifiers, whereby the model can default to predicting the majority class but will nevertheless still score well in many metrics such as error or accuracy rate. To investigate this, we created balanced datasets and use metrics as outlined above. In our dataset the Very High class had the smallest representation with only 565 examples in the training set. We therefore randomly sampled 565 of each remaining density states, to create a balanced training set of 2825 images. The same balancing process was repeated for the validation and testing data sets resulting in 800 and 575 images respectively.

2.1.5 Model refinement: data cleaning

It is important to consider the quality of imaging data. Specifically, many of our 20x20m aerial plots contain "artefacts" that were not accounted for in our ground observations. Figure 1 shows examples of three such types of artefacts. In Figure 1 an overhanging tree, the tramline and the field hedgerow in the top right hand corner are introducing significant noise into the image that does not represent either wheat or black-grass. It is this excess noise/uncategorised data we aimed to remove.

(Figure 1 near here)

To achieve this, we subsampled each individual 20x20m plot into 16 smaller images. Figure 1 demonstrates the outline of this subsampling grid. This yielded a dataset of 197 008 images. We then manually examined this dataset and set aside all subsamples that we determined to contain artefacts. In the case of Figure 1 only two subplots of "pure wheat" remained ((1, 2), (1, 3)), that were subsequently used in what we will refer to as the Clean dataset. This created a clean data set of 101,907 images and Artefact dataset of 95101 images. The training and test sets were the same as the previous experiments, but now "cleaned". We use the Clean and Artefact datasets to build models and predict on the test data of the other dataset e.g. clean model on artefact test data, and vice versa. This allows us to test the influence of data cleaning.

To make a comparison with our ground observations, we must upscale the subplot predictions back up to the 20x20m scale at which ground observations were recorded. There is often variation in density within each plot, but this is not recorded. In a hypothetical situation this could mean that the model is perfectly fitting the subplot test data, but then being penalised as we are unable to ascertain the observed level of black-grass in that specific subplot, only the entire 20x20m plot. We therefore take the median prediction from each subplot of one 20x20m plot as the model observation. This gives us a prediction of only the areas of the image with wheat and/or black-grass in them, at a scale that allows for comparison to our ground truthed data.

2.1.6 Model transferability: Field level Cross validation

To test out-of-sample/new field performance we conducted leave-one-field-out cross validation (LOFO-CV) trails and created 76 models, i.e one per field. Each model was trained using the baseline model parameters and cleaned upscaled subplots from all the fields. One

field was withheld from the training dataset to become the test set in each new model. We report back metrics at field level (i.e. not 20x20m plot level) as not all fields have the full five density states present.

2.1.7 Modelling workflow – baseline model

Having created the relevant datasets for each question we trained a model using our standard parameters. We began the analysis with a simple baseline test of how the models perform when 10% of the entire data is randomly selected as the test set. The model was then used to predict the ground-truthed observations of the relevant test set. We then calculated all relevant metrics and plot a ROC curve where appropriate. This assessed the performance of the CNN and established a baseline against which further analysis could be benchmarked. We investigated the effect of data balancing, data engineering and LOFO-CV against the baseline model.

To account for possible differences owing to variation in the date or survey or between years, we grouped the LOFO-CV models by years with 38 and 43 fields in 2016 and 2017 respectively and took the mean values of the AUC for each year. Each field season lasted 6 weeks and averaged the same number of fields each week. Consequently, we grouped the LOFO-CV models by week and took the mean values of AUC. Owing to the design of our field season we begin in the south and move north over the course of the season, so latitudinal effects will also be present but are not accounted for.

3.1 Results

3.1.1 Baseline Model

We find that the baseline model gives an AUC of 0.78, a weighted kappa of 0.59 and an average misclassification rate across all states of 17.8% as seen in Figure 2. We see that the Very High and Absent density states show the AUCs closest to x=1, y=1. This means that these density states are easier to distinguish for the model than the states in between.

(Figure 2 near here)

3.1.2 Data Balancing

The same training and evaluation parameters were used to train a model for the data in which the proportions of the density states were balanced. We see that by balancing the data set we slightly reduced the AUC and Cohen's kappa of the model (see Figure 3 for the ROC plot), whilst slightly increasing the weighted kappa and increasing the misclassification rate to 22.4%. This is most likely a consequence of the reduced number of training samples, leading to a poorer ability of the model to generalise features unique to each class. Table A1-A4 present statistical analysis on the differences between curves. (54). The results in Table A1 show that when the curves from Figure 2 (baseline model) are compared to those of Figure 3 (data balanced) that all but the Low density state curve are statistically non-significantly different. Balancing the dataset or not therefore does not affect the predictive performance of the models. We therefore continue to use the unbalanced dataset for the rest of our analysis.

(Figure 3 near here)

3.1.3 Data Cleaning

To examine how the data cleaning process (Figure 1) affects our models a new model was trained using the same parameters as the baseline model, but using the unbalanced, Clean

dataset. Figure 4 shows us that the AUC increased by 4.6%, a significant improvement with a similar misclassification rate to the baseline of 17.5%. Table A2 presents the statistical breakdown of the individual comparisons of AUC to the baseline.

(Figure 4 near here)

The images vary greatly in quality, with some having a large amount of high quality coverage, whilst in other cases only a small amount of the image is of good quality. We therefore divided the dataset according only to the percentage cover of good quality data of the original 20x20m plots remaining after the cleaning, regardless of black-grass level. Five equal categories of coverage of the 16 subplots, ranging from <20% (~3 subplots) to >80% (13-16 subplots) were established. Looking at the Multi-class AUC values for each plot in Figure 5, we see there is a ~6% difference in the lowest (0.67, <20%) and highest values (0.73, 60%-80%). We highlight the statistical differences between the categories with the highest and lowest AUCs in table A3. Showing that whilst the individual density states lines are not significantly different, the overall graphs are significant in conjunction with Figure 5.

(Figure 5 near here)

3.1.4 Analysis artefact data

Having shown in Figure 4 that cleaning and upscaling the data results in improved metrics from the baseline we next investigated the predictive performance of models fitted to the "artefact" images. To do this we used the 95101 artefact images set aside from the training set, predicted on the artefact images from cleaning the test data and then upscaled. Figure 6 suggests that the artefact plots still have features within them that allow us to classify black-grass as accurately as the Clean model (Figure 4). It also shows that with a higher weighted kappa and lower misclassification rate of 15.5%, it does better at not making large ordinal disagreements e.g. Very High observation Vs Absent prediction, when compared to the Clean

model. The Clean model predicted Absent when a Very High was observed in 8.75% of cases, compared with the artefact model only predicting 6.3% of such cases.

(Figure 6 near here)

As shown in Figure 7, the clean model can predict the black-grass levels in the artefact dataset with some degree of accuracy, with an AUC of 0.61 and misclassification rate of 17.1%. However, the model for the artefact data is not able to predict the clean test dataset accurately, with an AUC of 0.463, a misclassification rate of 42.1% and the AUC for all density states were significantly different as shown in Table A4. This suggests that the features used by the artefact model are not conducive to black-grass identification. Therefore, the features in the model for Figure 6, must not be directly related to black-grass. This also suggests that our manual screening of the data may have been overly strict, and we are thereby missing data that could increase the ability of the model to generalise features for the identification of black-grass.

(Figure 7 near here)

3.1.5 Out of Sample predictions - LOFO-CV

Here we examine the true out of sample prediction for the dataset. In all our previous models we have used an initial random 10% as our test dataset as described in our initial test set. Therefore, the model has been trained on a large sample of each individual field, allowing it to generalise features specific to that field, making it more sensitive to outliers. Thus, our reported results so far are not truly out of sample and may have limited repeatability in further studies, even when using the standardised data collection methodology described here.

(Figure 8 near here)

Figure 8 shows the mean AUC of the fields is 0.54 with a range of 0.38 0.81. This means that LOFO-CV predictions for these models are frequently no better than random. The kappa metrics were not used here as most of our out of sample fields did not contain the full range of black-grass densities and so are penalised for lack of agreement when there are no observations of a level.

3.1.6 Temporal Effects

To investigate temporal effects on the results of our out of sample predictions, we studied whether the year or the week we visited the field had any effect on the AUC. Figure 9 shows the mean and standard errors of the AUC for each year and week. Neither year nor week has a significant effect on the model performance measured by the AUC of the model, with adjusted R^2 values of -0.011 and 0.008 respectively. This means that the temporal variation in the time surveying has not influenced our results.

(Figure 9 near here)

4.1 Discussion

We set out to predict distributions of weed densities using UAS imagery and CNN. We have devised a standardised and repeatable UAS data collection methodology, applied it over multiple years across the major arable areas of the UK and utilised data engineering techniques to increase the quality of our datasets. Whilst the weeds have been shown to be detectable, it is by no means a simple task, as both species are grasses with many similar traits. Our main conclusion is that data engineering increases the performance of our metrics the most, relative to other methods attempted when given a sample of known states in a field.

Increases in performance such as these are not common for CNN in the computer vision literature. There was no evidence that temporal factors such as year or time of sampling, affects the performance of the out of sample predictions.

However, when predicting on fields with no previous ground truthing (i.e. true out of sample data), the success as revealed by our metrics was highly variable. This may be due to the problem of dataset shift (55). Dataset or covariate shift occurs when there is a change in distribution of the classes between the training and test datasets. We know from our ground observations that on an individual field-by-field basis that it is rare to find fields with the full five density state distribution and there are no cases where all five are present in an equal distribution. One way of counteracting this issue in the literature is by constructing a density estimation of the labels in the test dataset and reweighting the training dataset accordingly (56). This approach is not applicable in a fully automated UAS system for the prediction of density states, as it is still dependant on ground-truthed observations from skilled observers.

Our study is the first to use repeated UAS surveys and deep learning statistical methodology to assess the impact of the significant heterogeneity in conditions across time and space on automated monitoring of weed densities. Anderson & Gaston (57) outline many areas in which UAS can be used in ecology and emphasise the need for temporally resolved studies, allowing for scale appropriate measurements using UAS that can be at user defined times and locations. This is a change in precedent from remote sensing work using satellite data, where data was only available at set times, resolutions and spectral frequencies. However, many previous studies using UAS have focused on repeated visits to one single site over time (58) or multiple sites at one time point (59). The use of trial plots in some studies does allow for a more detailed assessment of certain variables (60). However, in real world applications of methodologies and management decisions developed under these controlled settings, much more spatial and temporal variability when applied in agronomic use cases

will be encountered, thus reducing the transferability and scope of the studies (61). Therefore, our focus of only using "live" uncontrolled agronomic scenarios, does result in reduced reported metrics but allows our work to be applied with a more realistic understanding of the results that would be seen in the field.

Neural networks have previously been used and compared to other statistical methods, to classify the state of weed populations at a range of spatial scales (62-64). Barrero (13), trained a NN with a user defined texture feature derived from NDVI to identify a weed species amongst a single rice paddy. They reported a 99% precision on test data, with no reported recall score. This is most likely an overstatement of the model performance and approach. However, this study only focused on the binary classification issue of presence/absence of a weed, a much simpler and less informative on-farm metric, and only considered predictions from a single field at a single time point, suggesting that the performance is being overstated with no LOFO-CV being attempted. It is to be expected that our metrics (AUC, Cohen's kappa and weighted Cohen's kappa) are lower than the equivalent ones reported in the NN study, due to our focus on multiple fields spanning a wide variety crop conditions and for the more advanced use of density state predictions. Therefore, our results are more representative and transferable than these studies due to our LOFO-CV analysis, for methodologies involving UAS and machine learning to map weed populations going forward. However, our results indicate a more extensive and controlled analysis of the transferability of models is still needed.

The process of manually screening the datasets for artefacts is a slow and non-reproducible or scalable task. In the future we propose to train a classifier to automatically partition an entire dataset into clean and artefact sections. This approach is comparable to work that quantifies the data quality of video using a CNN (65). This would allow us to

expand our analysis into other Vegetation Indices by improving and standardising the data processing pipeline.

With the artefact dataset predicting to the same if not higher standards in our metrics than the clean dataset, it stands to reason then that a composite modelling approach could be undertaken to channel the clean and artefact subplots to their respective models and then recombined at the upscaling stage. This is a concept similar to ensemble based classifiers, where multiple differing model types are trained on the same data set and aggregate their predictions for the test set (66). Our approach described here would use this concept but instead of differing model types on the same dataset, we propose the same model on differing datasets and aggregating their predictions. This would reduce the amount of data loss and combine the differing feature sets of the models to aid in the detection of arable weeds.

4.1.2 Concluding remarks

We have demonstrated here how data engineering of UAS imagery and use of CNN can be used to classify weed densities. We highlight the methodological improvements resulting in increased prediction accuracy compared to past research using a variety of metrics, statistics and data collection procedures that provide a more detailed assessment of true model performance. All our models apart from the LOFO-CV are composed of a random 10% of individual subplots for the test set. This means that the models will have most likely been exposed to some in-field examples of the test set, and therefore can generate features that are specific and not generalised to the detection of the weed. We can conclude that when only considering the images of a new field and no other data, we cannot be highly confident in the ability of most of our models to map the black-grass in the field. Whilst we don't show a significant improvement in LOFO-CV testing with no apparent factors that make an

individual field be predicted well or poorly. We believe that the robustness of this evaluation procedure is a greater estimation of real-world predictive value when compared to past studies, which consequently overestimate their applicability. Therefore, the methodology set out in this paper represents a new standard in the area of weed mapping with UAS due to the expanded capabilities of data collection, statistical methods and evaluation procedures.

Acknowledgments

JL was funded by a studentship from the Grantham Centre for Sustainable Futures. Collection of field density data was funded by the BBSRC (BB/L001489/).

Appendix

Table 1 Non-Equal dataset AUC's compared to the Equal datasets. Used (54) to test the statistical difference of the AUC of each Density state.

Density State	AUC 1	AUC 2	D	p-value
Abs	0.75	0.77	-0.94	0.34354
Low	0.66	0.58	2.87	0.004032
Med	0.59	0.55	1.48	0.138393
High	0.56	0.58	-0.46	0.643204
V High	0.71	0.74	-1.00	0.313908

Table 2 Non-Equal dataset AUC compared to the Cleaned dataset.

Density State	AUC 1	AUC 2	D	p-value
Abs	0.75	0.73	0.76	0.445291
Low	0.66	0.66	-0.34	0.727911
Med	0.59	0.52	3.72	0.000193
High	0.56	0.53	1.81	0.069353
V High	0.71	0.67	1.34	0.177991

Table 3 Worst performing bracket AUC from data quality testing (20% <)(AUC 1) compared to the best performing AUC bracket (60% - 80%) (AUC 2).

Density State	AUC 1	AUC 2	D	p-value
Abs	0.61	0.74	-2.25	0.023968
Low	0.64	0.71	-1.47	0.140146
Med	0.56	0.53	0.23	0.813781

High	0.72	0.56	1.12	0.261717
V High	0.73	0.86	-0.70	0.479862

Table 4 Clean model AUC's from predicting on the artefact dataset compared to the Artefact model AUC's from predicting on the clean dataset.

Density State	AUC 1	AUC 2	D	p-value
Abs	0.66	0.51	12.22	2.43E-34
Low	0.62	0.53	11.62	2.94E-31
Med	0.52	0.5	4.70	2.56E-06
High	0.51	0.5	2.88	0.003877
V High	0.69	0.5	9.26	1.92E-20

Table 5 Statistical measurements of the GNDVI pixel values for each vegetation group.

GNDVI	Mean	Standard Deviation
Black-grass	0.336	0.007
Winter wheat	0.304	0.011

References

1. Gibson DJ. Methods in comparative plant population ecology: Oxford University Press; 2014.
2. Harper JL. Population biology of plants. Population biology of plants. 1977.
3. Queenborough SA, Burnet KM, Sutherland WJ, Watkinson AR, Freckleton RP. From meso-to macroscale population dynamics: a new density-structured approach. Methods in Ecology and Evolution. 2011;2(3):289-302.
4. Symonides E. On the ecology and evolution of annual plants in disturbed environments. Vegetatio. 1988;77(1-3):21-31.
5. Burnham KP, Anderson DR, Laake JL. Estimation of density from line transect sampling of biological populations. Wildlife monographs. 1980(72):3-202.
6. Lambert J, Hicks H, Childs D, Freckleton R. Evaluating the potential of Unmanned Aerial Systems for mapping weeds at field scales: a case study with Alopecurus myosuroides. Weed research. 2018;58(1):35-45.
7. McIntyre G. A method for unbiased selective sampling, using ranked sets. Australian Journal of Agricultural Research. 1952;3(4):385-90.
8. Rondinini C, Wilson KA, Boitani L, Grantham H, Possingham HP. Tradeoffs of different types of species occurrence data for use in systematic conservation planning. Ecology letters. 2006;9(10):1136-45.
9. Braunisch V, Suchant R. Predicting species distributions based on incomplete survey data: the trade-off between precision and scale. Ecography. 2010;33(5):826-40.
10. Brown JH. On the relationship between abundance and distribution of species. The american naturalist. 1984;124(2):255-79.
11. Laliberte AS, Herrick JE, Rango A, Winters C. Acquisition, orthorectification, and object-based classification of unmanned aerial vehicle (UAV) imagery for rangeland monitoring. Photogrammetric Engineering & Remote Sensing. 2010;76(6):661-72.
12. Hardin PJ, Jackson MW. An unmanned aerial vehicle for rangeland photography. Rangeland Ecology & Management. 2005;58(4):439-42.

13. Barrero O, Rojas D, Gonzalez C, Perdomo S, editors. Weed detection in rice fields using aerial images and neural networks. Signal Processing, Images and Artificial Vision (STSIVA), 2016 XXI Symposium on; 2016: IEEE.
14. Lemerle D, Verbeek B, Cousens R, Coombes N. The potential for selecting wheat varieties strongly competitive against weeds. Weed Research. 1996;36(6):505-13.
15. Sinden J, Jones R, Hester S, Odom D, Kalisch C, James R, et al. The economic impact of weeds in Australia. Technical Series. 2004;8.
16. Pedersen SM, Fountas S, Have H, Blackmore B. Agricultural robots—system analysis and economic feasibility. Precision agriculture. 2006;7(4):295-308.
17. Kremen C, Merenlender AM, Murphy DD. Ecological monitoring: a vital need for integrated conservation and development programs in the tropics. Conservation biology. 1994;8(2):388-97.
18. Gerhards R, Christensen S. Real-time weed detection, decision making and patch spraying in maize, sugarbeet, winter wheat and winter barley. Weed research. 2003;43(6):385-92.
19. Freckleton RP, Sutherland WJ, Watkinson AR, Stephens PA. Modelling the effects of management on population dynamics: some lessons from annual weeds. Journal of Applied Ecology. 2008;45(4):1050-8.
20. Pena JM, Torres-Sánchez J, de Castro AI, Kelly M, López-Granados F. Weed mapping in early-season maize fields using object-based analysis of unmanned aerial vehicle (UAV) images. PloS one. 2013;8(10):e77151.
21. Xiao D, Tao F, Liu Y, Shi W, Wang M, Liu F, et al. Observed changes in winter wheat phenology in the North China Plain for 1981–2009. International Journal of Biometeorology. 2013;57(2):275-85.
22. Steven M, Malthus T, Demetriades-Shah T, Danson F, Clark J. High-spectral resolution indices for crop stress. High-spectral resolution indices for crop stress. 1990:209-27.
23. Wang N, Zhang N, Dowell FE, Sun Y, Peterson DE. Design of an optical weed sensor usingplant spectral characteristics. Transactions of the ASAE. 2001;44(2):409.
24. Lawless C, Semenov M, Jamieson P. A wheat canopy model linking leaf area and phenology. European Journal of Agronomy. 2005;22(1):19-32.
25. Sakamoto T, Van Nguyen N, Ohno H, Ishitsuka N, Yokozawa M. Spatio–temporal distribution of rice phenology and cropping systems in the Mekong Delta with special reference to the seasonal water flow of the Mekong and Bassac rivers. Remote Sensing of Environment. 2006;100(1):1-16.
26. Vina A, Gitelson AA, Rundquist DC, Keydan G, Leavitt B, Schepers J. Monitoring maize (Zea mays L.) phenology with remote sensing. Agronomy Journal. 2004;96(4):1139-47.
27. Friedman J, Hastie T, Tibshirani R. The elements of statistical learning: Springer series in statistics Springer. Berlin; 2001.
28. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, et al. Backpropagation applied to handwritten zip code recognition. Neural computation. 1989;1(4):541-51.
29. Sünderhauf N, McCool C, Upcroft B, Perez T, editors. Fine-Grained Plant Classification Using Convolutional Neural Networks for Feature Extraction. CLEF (Working Notes); 2014.
30. Freckleton RP, Hicks HL, Comont D, Crook L, Hull R, Neve P, et al. Measuring the effectiveness of management interventions at regional scales by integrating ecological monitoring and modelling. Pest management science. 2017.
31. Mortensen D, Dieleman JA, Johnson G. Weed spatial variation and weed management. 1998.
32. Perry N, Hull R, Lutman P, editors. Stability of weed patches. 12th European Weed Research Symposium, Wageningen, The Netherlands; 2002.
33. Zeiler MD, Fergus R, editors. Visualizing and understanding convolutional networks. European conference on computer vision; 2014: Springer.
34. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al., editors. Going deeper with convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition; 2015.
35. Hicks HL, Comont D, Coutts SR, Crook L, Hull R, Norris K, et al. The factors driving evolved herbicide resistance at a national scale. Nature ecology & evolution. 2018;2(3):529.

36. Lancashire PD, Bleiholder H, Boom T, Langelüddeke P, Stauss R, WEBER E, et al. A uniform decimal code for growth stages of crops and weeds. Annals of applied Biology. 1991;119(3):561-601.
37. Ballesteros R, Ortega J, Hernández D, Moreno M. Applications of georeferenced high-resolution images obtained with unmanned aerial vehicles. Part I: Description of image acquisition and processing. Precision Agriculture. 2014;15(6):579-92.
38. Wiegand C, Richardson A, Escobar D, Gerbermann A. Vegetation indices in crop assessments. Remote Sensing of Environment. 1991;35(2-3):105-19.
39. Gitelson AA, Kaufman YJ, Merzlyak MN. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. Remote sensing of Environment. 1996;58(3):289-98.
40. Sripada RP, Heiniger RW, White JG, Weisz R. Aerial color infrared photography for determining late-season nitrogen requirements in corn. Agronomy Journal. 2005;97(5):1443-51.
41. Underwood E, Ustin S, DiPietro D. Mapping nonnative plants using hyperspectral imagery. Remote Sensing of Environment. 2003;86(2):150-61.
42. da Silva Junior CA, Nanni MR, Teodoro PE, Silva GFC, de Lima MG, Eri M. Comparison of mapping soybean areas in Brazil through perceptron neural networks and vegetation indices. African Journal of Agricultural Research. 2016;11(43):4413-24.
43. Peña JM, Torres-Sánchez J, Serrano-Pérez A, de Castro AI, López-Granados F. Quantifying efficacy and limits of unmanned aerial vehicle (UAV) technology for weed seedling detection as affected by sensor resolution. Sensors. 2015;15(3):5609-26.
44. Torres-Sánchez J, López-Granados F, Peña JM. An automatic object-based method for optimal thresholding in UAV images: Application for vegetation detection in herbaceous crops. Computers and Electronics in Agriculture. 2015;114:43-52.
45. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. International Journal of Computer Vision. 2015;115(3):211-52.
46. Krizhevsky A, Sutskever I, Hinton GE, editors. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems; 2012.
47. Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov RR. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:12070580. 2012.
48. Simard PY, Steinkraus D, Platt JC, editors. Best practices for convolutional neural networks applied to visual document analysis. null; 2003: IEEE.
49. Shin H-C, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE transactions on medical imaging. 2016;35(5):1285-98.
50. Fawcett T. An introduction to ROC analysis. Pattern recognition letters. 2006;27(8):861-74.
51. Carter JV, Pan J, Rai SN, Galandiuk S. ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves. Surgery. 2016;159(6):1638-45.
52. Fleiss JL, Cohen J, Everitt B. Large sample standard errors of kappa and weighted kappa. Psychological bulletin. 1969;72(5):323.
53. Cohen J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological bulletin. 1968;70(4):213.
54. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988:837-45.
55. Moreno-Torres JG, Raeder T, Alaiz-RodríGuez R, Chawla NV, Herrera F. A unifying view on dataset shift in classification. Pattern Recognition. 2012;45(1):521-30.
56. Gretton A, Smola AJ, Huang J, Schmittfull M, Borgwardt KM, Schölkopf B. Covariate shift by kernel mean matching. 2009.
57. Anderson K, Gaston KJ. Lightweight unmanned aerial vehicles will revolutionize spatial ecology. Frontiers in Ecology and the Environment. 2013;11(3):138-46.
58. JONES IV GP, Pearlstine LG, Percival HF. An assessment of small unmanned aerial vehicles for wildlife research. Wildlife Society Bulletin. 2006;34(3):750-8.

59. Getzin S, Wiegand K, Schöning I. Assessing biodiversity in forests using very high-resolution images and unmanned aerial vehicles. Methods in Ecology and Evolution. 2012;3(2):397-404.
60. Holman FH, Riche AB, Michalski A, Castle M, Wooster MJ, Hawkesford MJ. High throughput field phenotyping of wheat plant height and growth rate in field plot trials using UAV based remote sensing. Remote Sensing. 2016;8(12):1031.
61. Concepción ED, Díaz M, Baquero RA. Effects of landscape complexity on the ecological effectiveness of agri-environment schemes. Landscape Ecology. 2008;23(2):135-48.
62. Irmak A, Jones J, Batchelor W, Irmak S, Boote K, Paz J. Artificial neural network model as a data analysis tool in precision farming. Transactions of the ASABE. 2006;49(6):2027-37.
63. LÓPEZ-GRANADOS F, PEÑA-BARRAGÁN JM, JURADO-EXPÓSITO M, Francisco-FERNÁNDEZ M, Cao R, ALONSO-BETANZOS A, et al. Multispectral classification of grass weeds and wheat (Triticum durum) using linear and nonparametric functional discriminant analysis and neural networks. Weed Research. 2008;48(1):28-37.
64. Mansourian S, Darbandi EI, Mohassel MHR, Rastgoo M, Kanouni H. Comparison of artificial neural networks and logistic regression as potential methods for predicting weed populations on dryland chickpea and winter wheat fields of Kurdistan province, Iran. Crop Protection. 2017;93:43-51.
65. Le Callet P, Viard-Gaudin C, Barba D. A convolutional neural network approach for objective video quality assessment. IEEE Transactions on Neural Networks. 2006;17(5):1316-27.
66. Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews). 2012;42(4):463-84.

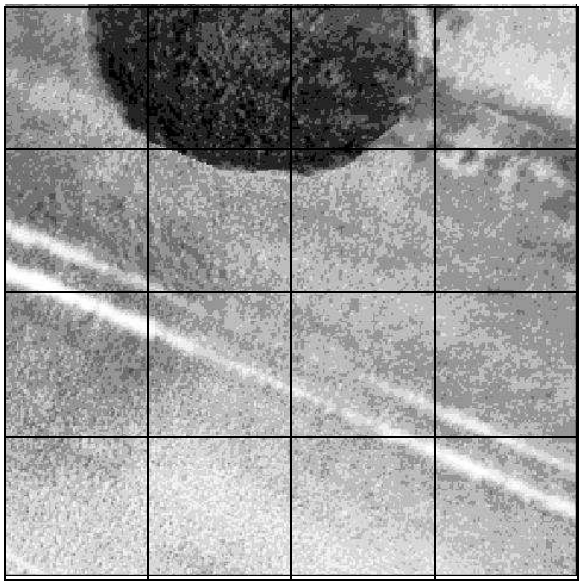

Figure 1 Example of a Very High, 20x20m plot with significant non-black-grass "artefacts", reducing the signal in the image coming from the Very High level of black-grass that was observed on the ground in this plot. The grid overlay represents the subsampling methodology used to break each image into 16 smaller representations of the entire plot. The subplots are referenced by their position relative to the bottom left hand corner (1,1) and top right corner (4,4).

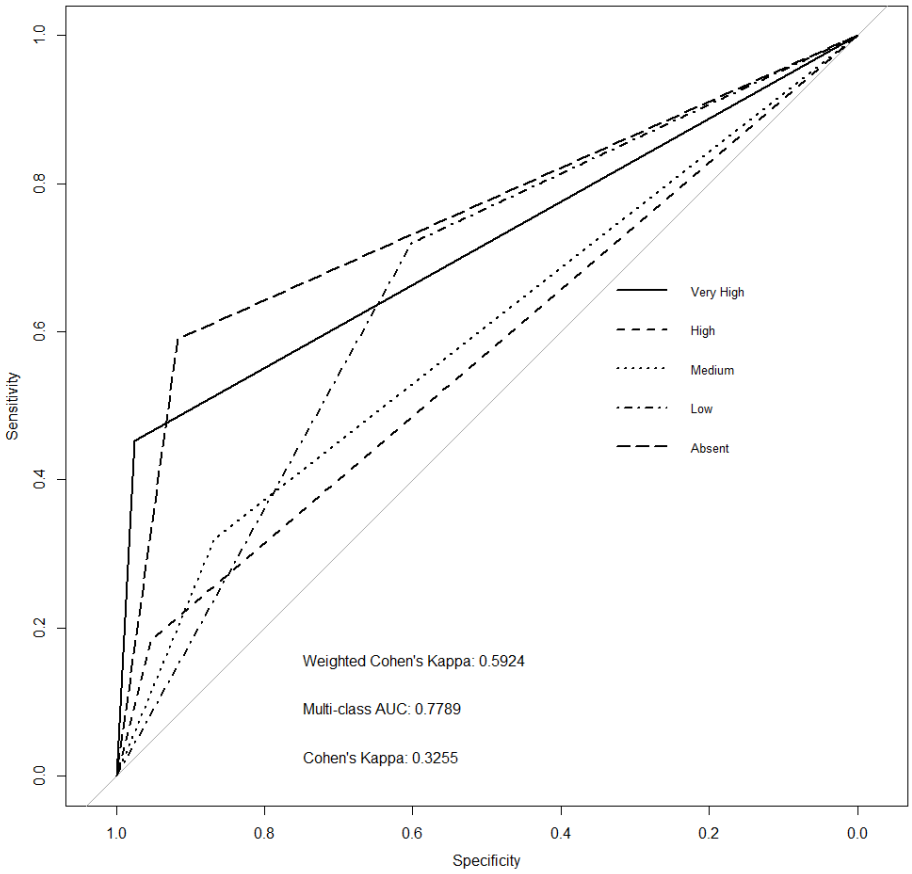

Figure 2 Baseline, ROC plot of a CNN trained using 90% of the dataset and used to predict the multiclass black-grass density state of the completely withheld random 10% of data.

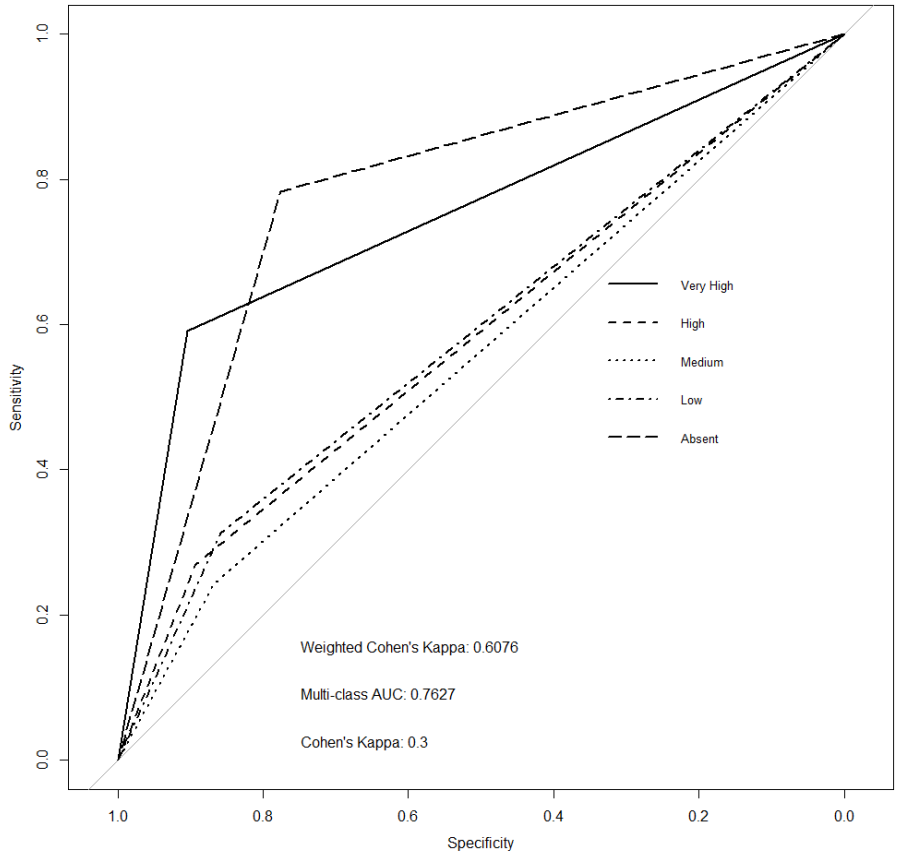

Figure 3 ROC plot of a CNN trained using 90% of the balanced dataset used to predict the multiclass black-grass density state of the completely withheld random 10% of balanced data.

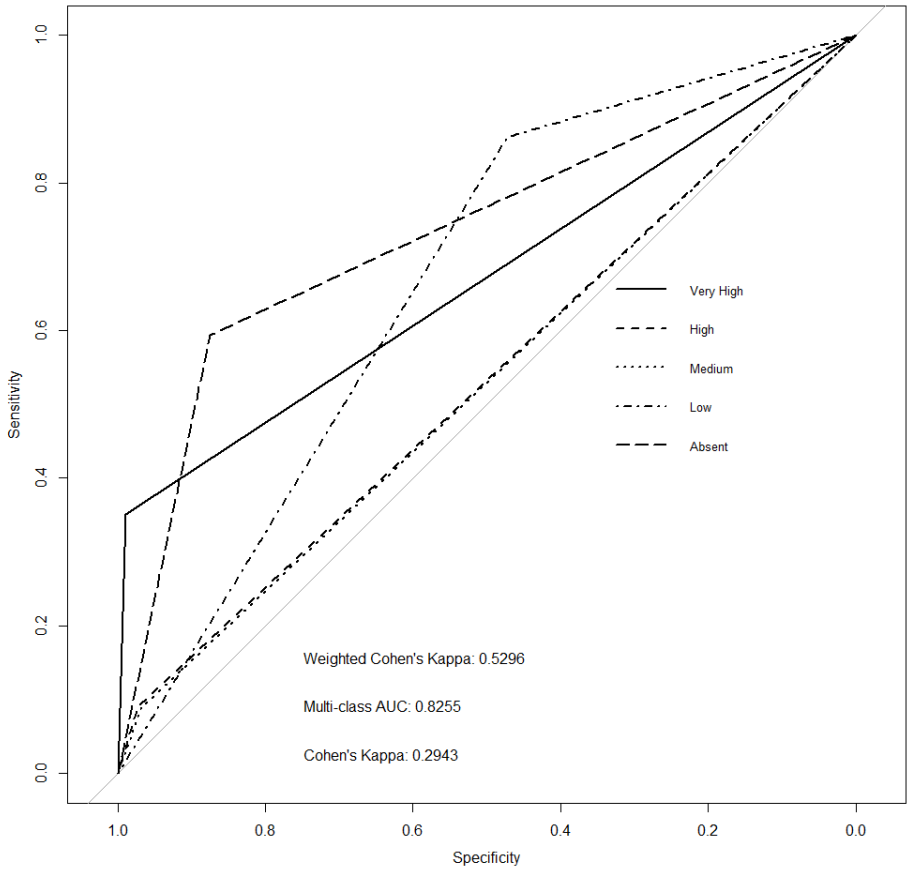

Figure 4 ROC plot of a CNN trained using 90% of the entire Clean subplot dataset used to predict the multiclass black-grass density state of the completely withheld random 10% of Clean data. The subplot predictions are then scaled back up to 20x20m plots for comparisons to our ground observations.

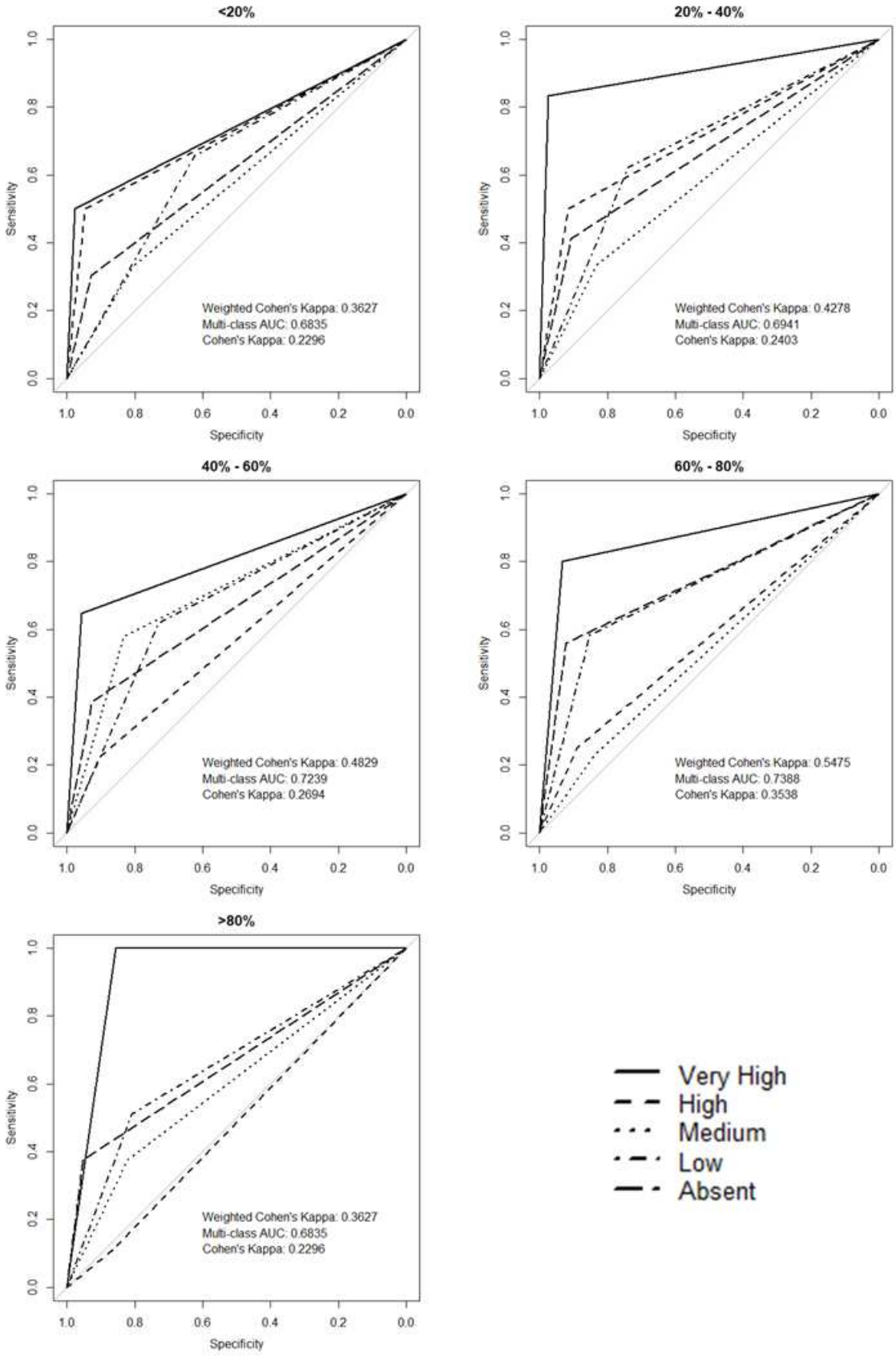

Figure 5 ROC plots showing how the percentage cover of the subplots in the Clean dataset affect performance (measured as AUC and kappa).

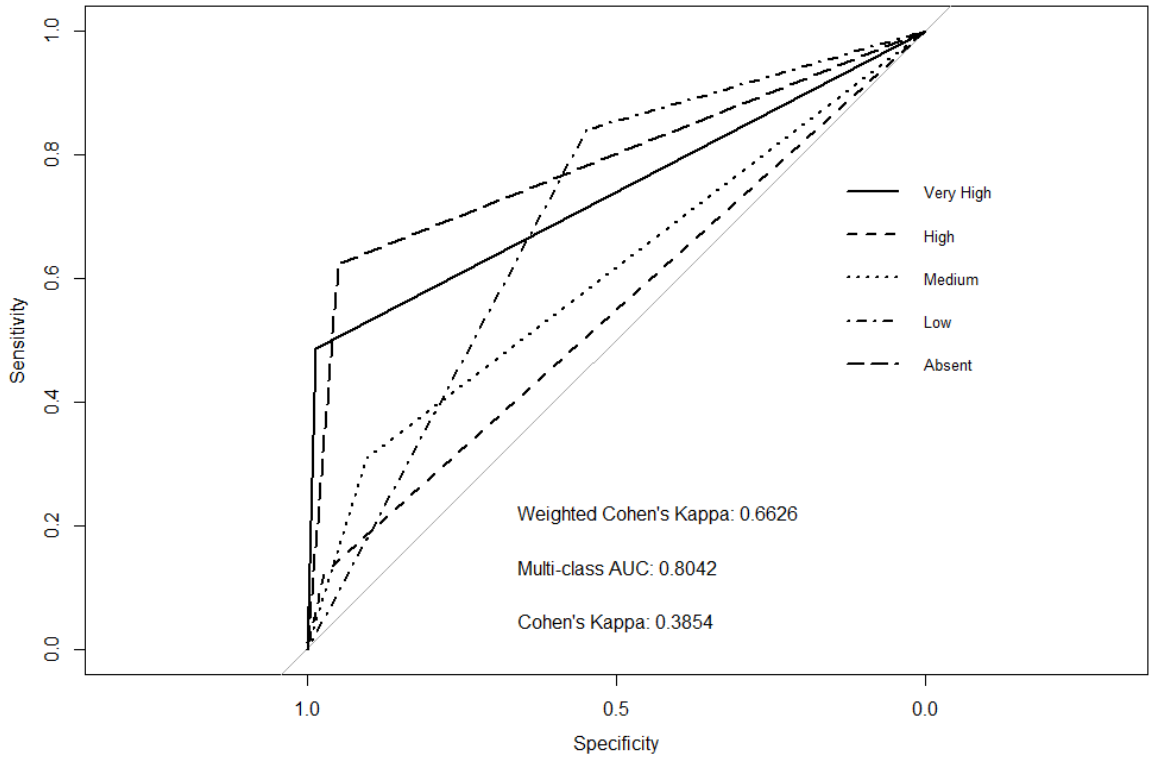

Figure 6 ROC plot of a CNN trained using 90% of the artefact subplot dataset used to predict the multiclass black-grass density state of the completely withheld random 10% of artefact data.

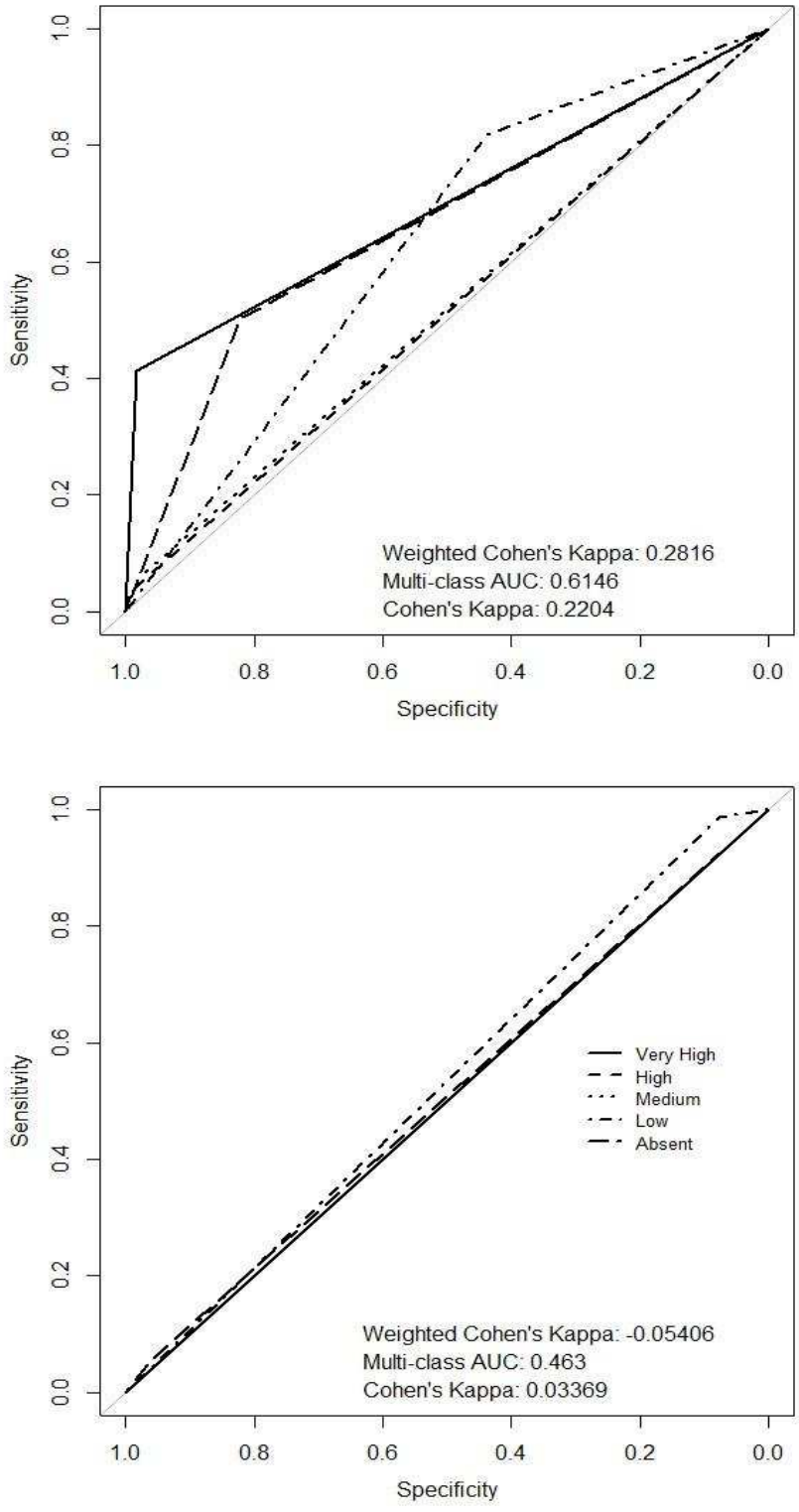

Figure 7 (a) ROC plot of a model trained using the Clean training set, then used to predict the five density level states in the artefact test set. (b) ROC plot of a model trained using the artefact training set, then used to predict the five density level states in the cleaned test set. The predictions are upscaled to plot level.

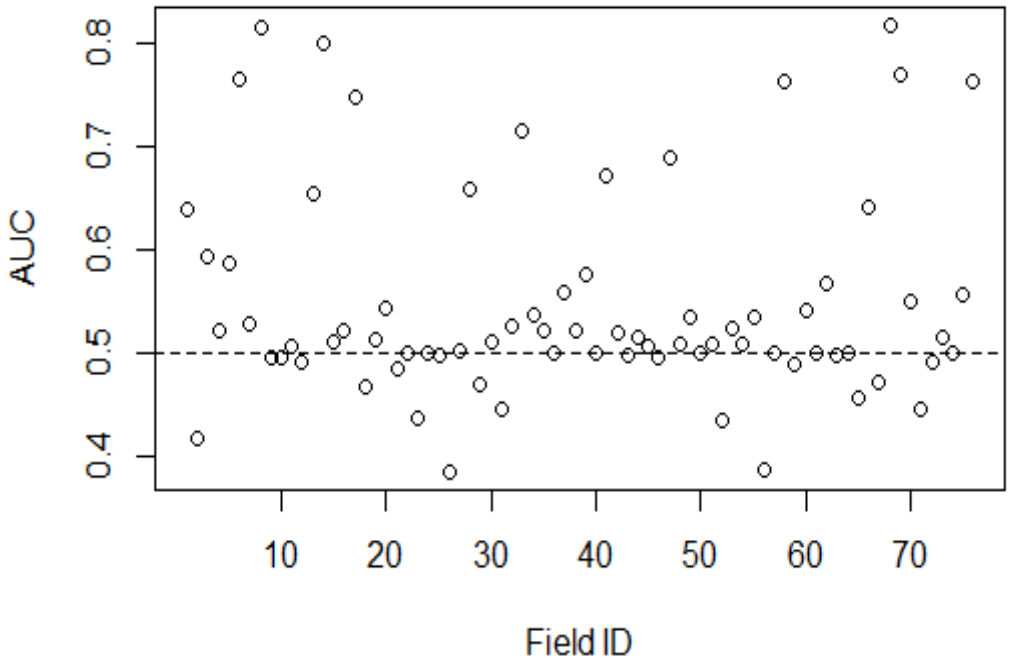

Figure 8 AUC of each field's out of sample prediction. Each point represents a separate model that was trained on all but the Field ID in question which is used as the test set. Field ID is a randomised ordering of the field names across both survey years.

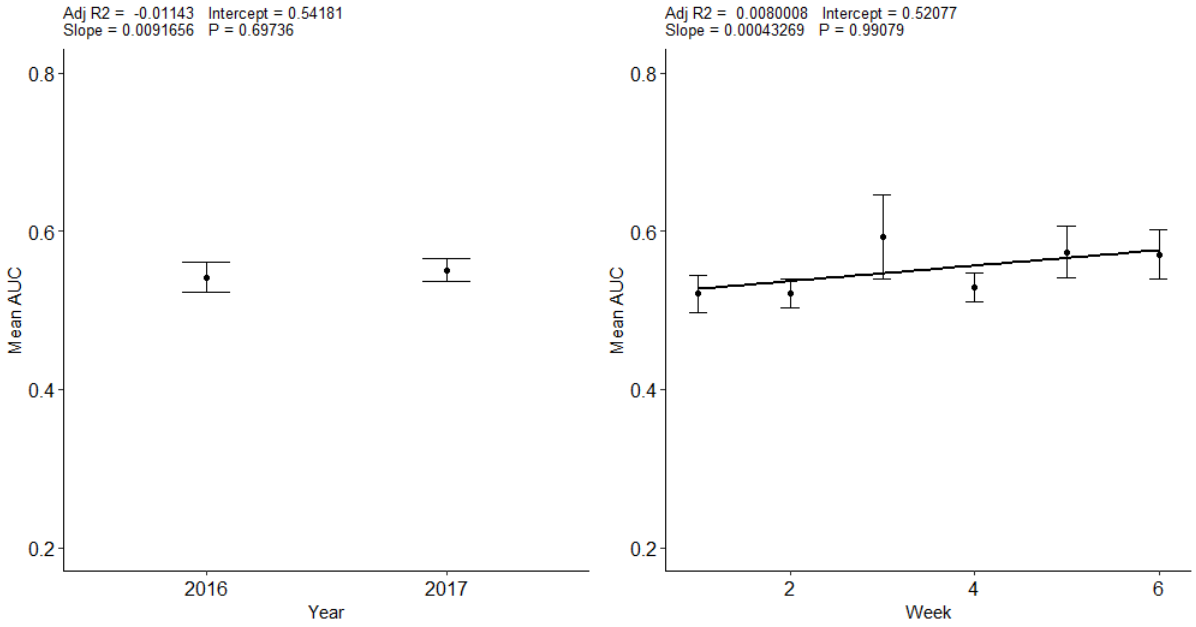

Figure 9 (left) Mean AUC for every model in each year. (right) Mean AUC for every model in each week.