nature sustainability

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41893-019-0246-x

In the format provided by the authors and unedited.

Deep learning to map concentrated animal feeding operations

Cassandra Handan-Nader[1,2] and Daniel E. Ho[1,2,3]*

[1]Stanford Law School, Stanford University, Stanford, CA, USA. [2]Department of Political Science, Stanford University, Stanford, CA, USA. [3]Stanford Institute for Economic Policy Research, Stanford, CA, USA. *e-mail: dho@law.stanford.edu

Deep Learning to Map Concentrated Animal Feeding Operations

Cassandra Handan-Nader Daniel E. Ho

Supplementary Notes

Image data. We describe our raw data sources in more detail here. The first data source is the National Agriculture Imagery Program (NAIP) (`https://www.fsa.usda.gov/programs-and-services/aerial-photography/imagery-programs/naip-imagery/`). NAIP is administered by the U.S. Department of Agriculture's Farm Service Agency. It acquires aerial imagery annually during the agricultural growing season on a three-year cycle staggered across states. Images are orthorectified and disseminated publicly. In North Carolina, images are available at up to a 2 meter per pixel resolution for 2004 and 2005. In 2006, 2008, 2009-10, 2012, 2014, and 2016, images are available at up to a 1 meter per pixel resolution. We access the 2014-16 NAIP imagery through Descartes Labs, a geospatial analytics company that provides an API interface for satellite imagery (`http://www.descarteslabs.com/`). Of particular use is that Descartes overlays geographic metadata, such as state, county, and UTM grid. For our longitudinal analysis, we access NAIP imagery using Google Earth Engine, which provides more years of historical NAIP imagery (`https://earthengine.google.com/datasets/`).

Image Weighting. We over-sampled images from three categories in the hand validation process – images containing locations that EIGs tagged as CAFOs, their neighbouring images, and images matching a Google Places API search for common false positive categories. We did this to maximise the information in our training set given limited tagging resources. As a 25% random sample of the hand-validated images, the test set necessarily had a lower ratio of non-CAFO images to CAFO images than in the population of images across the state. Based on the number of images containing an EIG location, the expected class imbalance in the population of images was approximately 269 control images for every 1 CAFO image, compared to approximately 4 control images for every CAFO image in the test set.

A more extreme class imbalance would alter the expected precision, as even a very low false positive rate can correspond to a large number of false positives when applied to a large number of negative examples. However, precision and sensitivity on the statewide distribution of images may not apply to the primary use case for the model. If the use case is to detect CAFOs that may have been missed after a manual enumeration, then the "natural distribution" of all possible images in North Carolina may be inappropriate. In that setting, regulators and interest groups may opt to manually inspect facilities with high predicted probabilities of being CAFOs that were missed by the manual enumeration. To conserve resources, they might prioritise images with high probabilities and continue inspecting until the signal-to-noise ratio is too low.

If the model is intended as a complete substitute for manual enumeration, the natural distribution may consist of all images in a state. Here, sensitivity and precision will be fundamentally affected by the baseline rate of CAFOs. To provide a sense of this, we develop a weighting protocol to weight images in the test set to match the population distribution from which it was sampled (all images in North Carolina excluding landscape images, see Supplementary Figure 4 for the score distribution of excluded images). Supplementary Table 3 shows the distribution of images in the population and the test sample for each model. The sample weight for each image category was calculated as $\frac{\text{Pop. Image } n}{\text{Sample Image } n}$, so that sum of the weights added up to the total number of images in the population. We confirmed the representativeness of the weighted sample by comparing the weighted average model scores in the sample to the unweighted average scores in the population. The unweighted average swine and poultry scores in the population were 0.0135 and 0.0236 respectively, while the sample weighted averages were 0.0134 (weighted SE = 0.0009) and 0.0225 (weighted SE = 0.0010), respectively.

The downside to the population weighting approach is that, because the sample size is small

relative to the population size, sampling variability may cause feature distortion. Each ordinary control image in the sample, for example, is weighted up to represent nearly 400 such images in the population. The weighting approach also puts more weight on occluded CAFO images, since by definition there are more CAFO neighbour images than CAFO images in the statewide distribution, and neighbour images are more likely to depict only edges of CAFOs. It is not possible to correct for this by calculating weights after dropping occluded images, as occlusion was not known at the time of image sampling. The weighted figures should therefore be interpreted only directionally.

Supplementary Figure 8 shows that the area under the ROC curve (AUC) remains high for both the test set (unweighted) and statewide (weighted) distribution of images. The weighted AUC was 0.94 for the poultry model and 0.99 for the swine model, compared to unweighted results of 0.97 and 0.99, respectively. As expected, the precision-recall curve is most heavily influenced by the increase in class imbalance [1]. Average precision drops to 50% for the poultry model and 67% for the swine model. However, it of course matters greatly what the baseline is. The resource savings curve shown in Supplementary Figure 9 suggests that the model can nonetheless aid substantially compared to reviewing every image in the state. Reaching 95% sensitivity requires tagging fewer than 10% of images compared to the baseline, which remains a substantial saving of resources. In addition, the facility consolidation algorithm reduces the number of returned results by nearly half compared to raw image results, further improving efficient use of manual resources by reducing the number of items to review.

Supplementary Methods

Oversampling of difficult control images. We randomly sampled high-scoring false positive images from early rounds of model training and categorised them according to Google's place categories (see `https://developers.google.com/places/web-service/supported_types` for a complete taxonomy.). We obtained the following list of common false positive categories: store, campground, airport, RV park, lodging, school, church, park, parking, shopping mall, roofing contractor, supermarket, car dealer, cemetery, department store, and restaurant. Because 50,000 metres is the maximum search radius for the Google Places API, we then randomly seeded 5 point locations at least 50,000 metres apart in each county and downloaded all images returned in a Google Places API search for the above-listed categories. These images were added to subsequent training rounds as control images. We note that this makes the classification task on the test sample more difficult than in the natural setting.

Spatial algorithm for facility detection. We designed the following algorithm to consolidate image-level predictions into unique facility locations.

1. Compute the class activation map for each pixel position (x, y) as $M(x, y) = \sum_{g=1}^{2048} w_g f_g(x, y)$, where $f_g(x, y)$ is the feature map in the global average pooling layer at pixel position (x,y) and w_g is the weight for the CAFO class corresponding to feature g.

2. Identify activated pixel positions $A(x, y)$ where $M(x, y) \geq 0.5 \max\{M(x, y)\}$.

3. Using a standard k-means implementation, find k possible clusters of pixel activations $A(x, y)$ that are at least 150m apart, with a maximum of $k = 5$, and record their centroids $C_k(x, y)$.

4. For each centroid $C_k(x, y)$ found in step 3:

(a) Centre image on position $C_k(x, y)$ within the UTM grid and crop the image at 299x299 pixels.

(b) Repeat steps 1-3 on centred image.

(c) For each cluster of activated pixels found in step 3 representing an object of interest, draw a polygon around the outer rim of the activated pixels belonging to that object cluster.

(d) Record the centroid of each object cluster in UTM coordinates.

(e) Record the area of each object cluster using the shoelace algorithm, $A = \frac{1}{2} \mid \sum_{i=1}^{n-1} X_i Y_{i+1} + X_n Y_1 - \sum_{i=1}^{n-1} X_{i+1} Y_i - X_1 Y_n \mid$, where X is a vector of x coordinates for the object cluster, and Y is the corresponding vector of y coordinates. Convert metres to feet in order to obtain square footage.

5. Re-score all re-centred images to obtain new swine and poultry image predictions. Discard images no longer classified as poultry CAFOs.

6. Within each UTM zone, compute the Euclidean distance between the UTM coordinates of all CAFO objects. Combine CAFO objects that are within 250 metres of each other.

7. Record the facility coordinates as the average of the object centroids. Record the maximum square footage across combined objects as the square footage estimate for the facility.

8. Repeat 6-7 until the minimum distance between all consolidated CAFO objects is 250 metres.

Facility validation. We hand validated each predicted facility location across the state to determine whether it truly identified a poultry CAFO. For the subset of facilities in Duplin and Cumberland counties, we performed a more detailed hand validation in order to obtain ground truth size estimates for each facility. For each true positive predicted poultry facility, we counted the number of barns visible in the satellite image as ground truth for the facility size. In this process, we also manually verified the 250 meter threshold that we used to define the overlap between the EIG facility list and our modelled list across the state. We first matched each model-identified facility to the nearest EIG poultry location. We then plotted EIG poultry locations side-by-side with the modelled locations on an interactive Google map to determine whether the closest location was indeed the same facility.

Supplementary Figure 11 shows the distribution of distances between the EIG location and the modelled location for facilities manually confirmed to be the same location in the modelled list and the EIG list. Only two out of 187 facilities were located more than 250 metres from their EIG location, yet determined to be the same facility by manual verification. This confirmed the appropriateness of the 250 meter threshold, which matches the definition that the algorithm used to split facilities. Supplementary Figure 12 illustrates edge cases that we found during manual inspection that would introduce a small amount of error into automated matching between EIG and the modelled list. In a few cases, the EIG definition of a facility conflicted with the model's definition of separation by 250 metres, such that the model would find two facilities where EIGs would only find one, or vice versa. For swine facilities, this exercise would likely increase in complexity because permit addresses can be located away from the actual CAFO site. Fortunately, EIG satellite-based tagging for poultry was largely consistent with our model's identification of facilities.

Examples of false positives. Supplementary Figure 13 provides examples of false positive images. The top left panel depicts a swine CAFO whose lagoon is washed out in the image, making it appear closer to a poultry CAFO. The top right panel depicts the warehouse of a flooring manufacturer. Interestingly, this is an example that the EIG manual review classified as a poultry CAFO, showing that manual review does not always represent ground truth. The bottom left panel depicts an airplane hangar, and the bottom right panel depicts feed storage facilities.

We note that because even the human eye may not be able to positively distinguish these facilities from CAFOs that higher-resolution imagery will be particularly promising for reducing this source of noise.

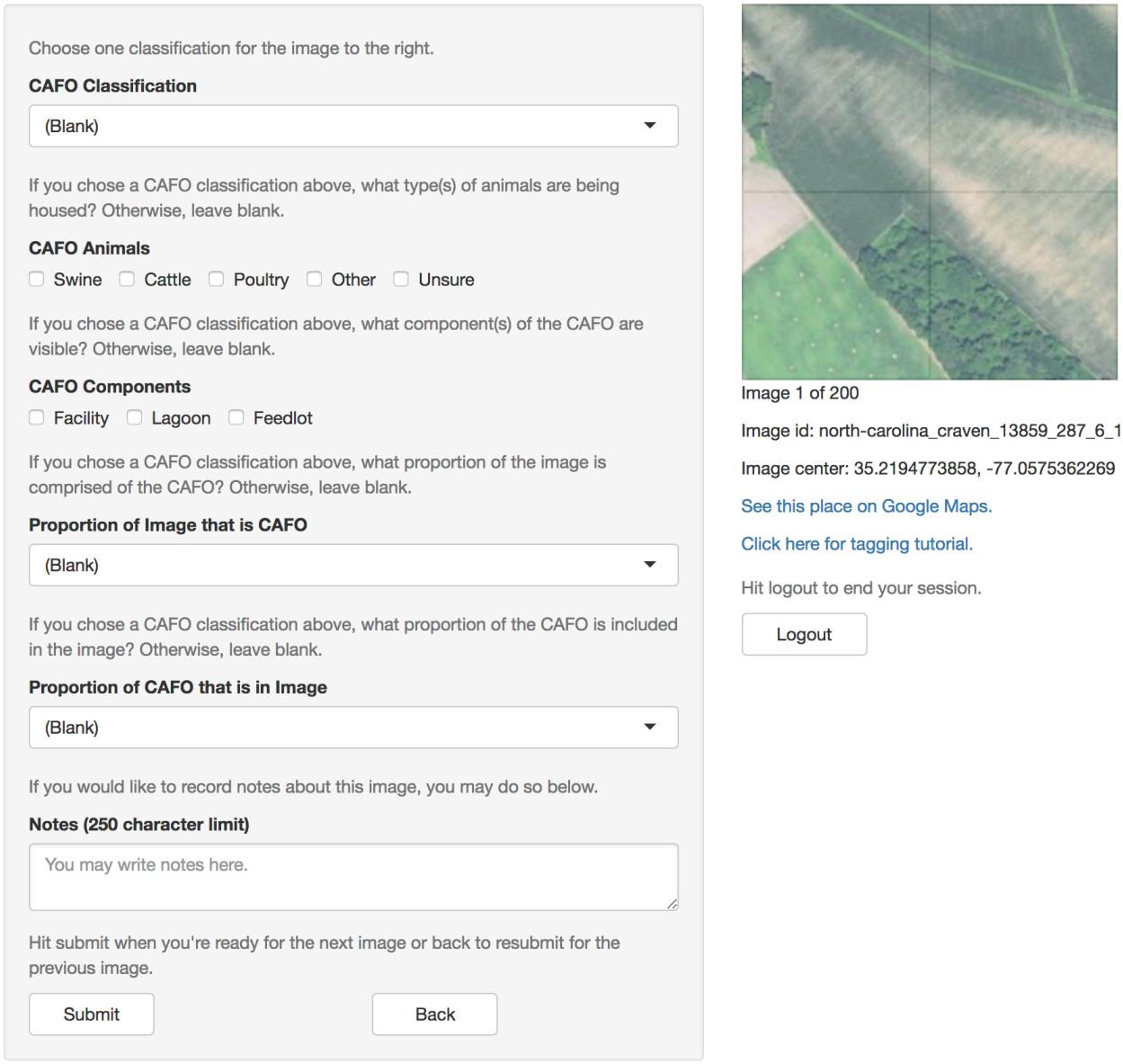
Supplementary Discussion

Ecological Impact. The left panel of Supplementary Figure 14 displays a CAFO that is surrounded on three sides by the Old Millpond tributary, with potential "land application" areas (where manure is applied) abutting the tributary. The right panel of Supplementary Figure 14 displays abandoned facilities detected by our model, with substantial waste surrounding the once-operational CAFO.

Supplementary Figure 15 provides an additional example of the vivid ecological impacts of defunct CAFO facilities detected by the model. The poultry facilities open around 2008, with little evidence pointing to a liquid manure storage system. Recall that the use of liquid vs. dry manure storage systems is a critical distinction, with the latter largely exempted from permitting requirements in North Carolina, as they arguably pose greater risk discharge into the water system. Over time, operators appear to build more pit storage capacity, with the ones added by 2013 directly abutting the Upper Little River. As these pits are used, they appear to not be maintained, exhibiting algae growth and signs of non-maintenance. By 2013, at least part of the facility is defunct, with a roof collapsed.

Environmental justice comparison between poultry and swine sites. One of the most contentious questions about CAFOs involves implications for "environmental justice." In 2014, environmental interest groups filed a formal complaint with EPA alleging that North Carolina's swine permitting process discriminated against minorities in violation of the Civil Rights Act. The settlement arising out of that process requires the county to develop an "[environmental justice] geographical information tool" by 2019. We here illustrate how our methods – which may be particularly useful in the case of largely unpermitted poultry facilities – can provide the geographic distribution of facilities that is up-to-date within three years (the cycle length of NAIP updates). We merge 2010 census block demographic information to examine the demographic correlates of CAFO siting. Supplementary Figure 16 plots densities for census blocks with and without poultry CAFOs in red and gray, respectively, for educational attainment, income, and race. The presence of poultry CAFOs is defined by the true positive facilities returned by the model at a threshold of 0.5. We observe that CAFOs are disproportionately sited in communities with lower educational attainment and at the lower income range. We observe that while both poultry and swine CAFO presence are similarly associated with lower proportions of college graduates and lower household incomes, the presence of swine CAFOs is more strongly associated with higher proportions of non-white residents in accordance with previous work [2]. These figures show that while both swine and poultry CAFO siting decisions have environmental justice dimensions, the racial dimension appears distinct.

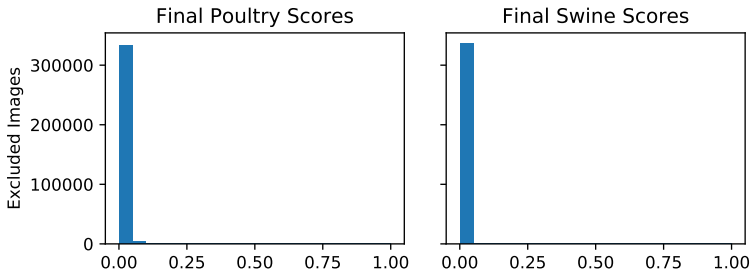
Supplementary Figures



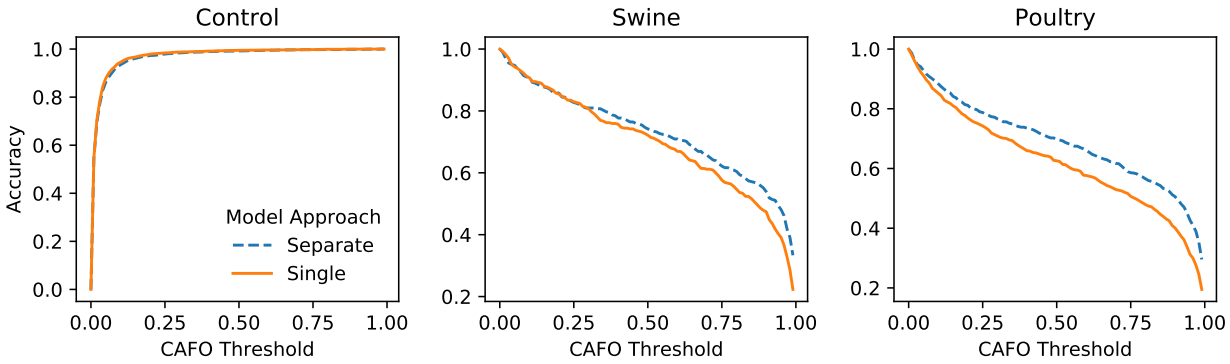
Supplementary Figure 1: Tagging tool for image validation. Screen short of web app for tagging images, developed using a PostgreSQL database for image meta-information and R's Shiny package. Research assistants were instructed to examine the NAIP tile and navigate to the Google Maps or Google Earth for historical imagery. Upon examination of features, they coded (a) whether the image was of a CAFO (CAFO, not CAFO, unsure), (b) the animal type (swine, cattle, poultry, other, unsure), (c) components observed (facility, lagoon, feedlot), (d) proportion of the image that is comprised of the CAFO (very little of the image $< 10\%$, some of the image 10-25%, much of the image 25-50%, majority of the image 50-75%, almost all of the image $> 75\%$), and (e) proportion of the CAFO that is in the image (some of the CAFO $<33\%$, much of the CAFO 33-67%, majority of the CAFO $>67\%$). To assist taggers with the image proportion thresholds, the app superimposed a 2x2 square grid on top of the image.

Supplementary Figure 2: CAFO image proportions. Examples of poultry and swine CAFO images at each image proportion category determined by the research assistants using the tagging tool in Supplementary Figure 1.

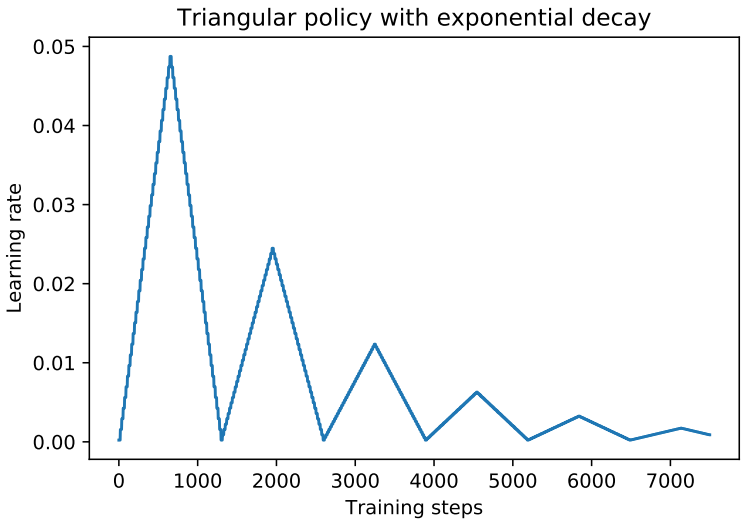
Supplementary Figure 3: Low propensity CAFO images. Representative examples of control images excluded from hand validation and from subsequent model training, validation, and test sets. The prototype model used to exclude images was based on permit locations for swine CAFOs, as these are required to have permits in North Carolina, increasing the reliability of the training data locations. We hand validated images at all swine permit locations in Duplin County, a high-swine-producing jurisdiction. For control images, we took a random 10% sample of images in the county and hand-validated that they did not include any type of CAFO. We trained a CNN on the Duplin images and obtained 98% accuracy when testing on completely unseen images from Sampson County, a neighbouring jurisdiction with a similarly high volume of swine permits. We use a CAFO probability threshold below 0.007 to exclude low propensity images from subsequent hand validation based on manual inspection of a random sample of images.



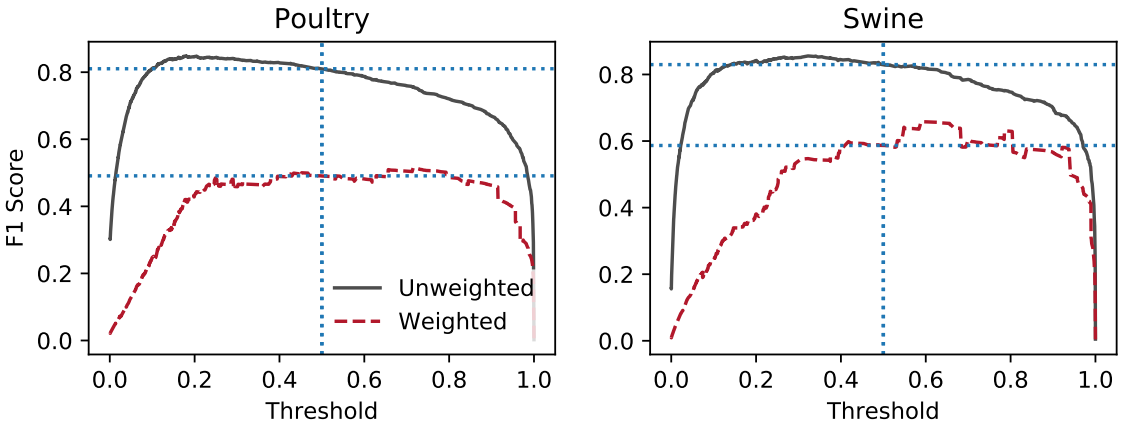
Supplementary Figure 4: Score distribution of excluded images. Poultry (left) and swine (right) score distributions from the final models for images excluded from hand validation with the 0.007 threshold from the prototype model. The excluded images (about 20% of the 1,684,879 images in North Carolina) also had scores concentrated below 0.01 in the final swine and poultry models.



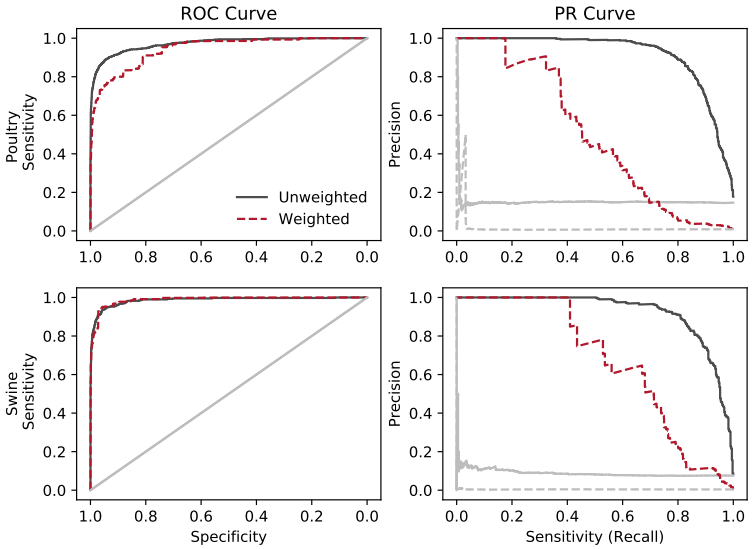
Supplementary Figure 5: Model accuracy comparison. Proportion of images correctly classified (accuracy) for control images, swine images, and poultry images at score thresholds for CAFOs varying from 0 to 1, using two modelling approaches – a single model that generates probability scores for the swine and poultry classes at once, or two separate models that generate scores for the swine and poultry classes, respectively, as reported in the paper. While the two approaches produce similar accuracy for control images, the separate model approach was more successful at distinguishing between swine and poultry images.



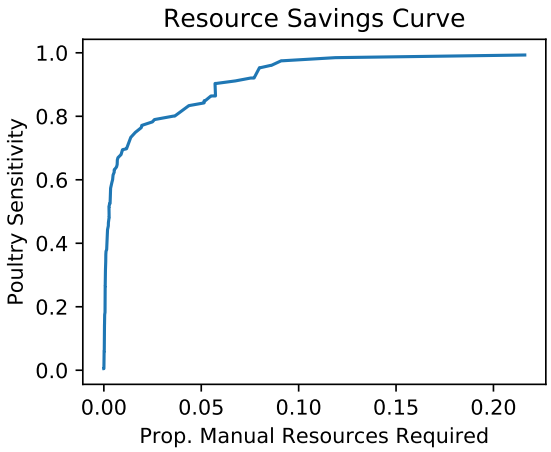
Supplementary Figure 6: Triangular learning rate policy. Visual depiction of the triangular learning rate policy algorithm using a minimum bound of 0.0002, a maximum bound of 0.05, and a cycle length of 1250 training steps. The learning rate α of the CNN controls the rate at which parameters of the loss function are updated in gradient descent. The triangular learning rate policy linearly cycles through learning rates between minimum and maximum bounds within a set number of training steps, with the maximum bound declining by an exponential factor at the end of each cycle. The algorithm updates the learning rate α at each training step i with the procedure described in [3], using a step size of 2 times the epoch size. We arrived at the maximum learning rate bounds for each model by calculating average classification error on the validation set with a range of learning rates, and determining the minimum learning rate where the first difference of the average classification error ceased to decrease. This occurred at $\alpha = 0.05$ for the swine model and $\alpha = 0.04$ for the poultry model. We obtained the minimum bound for the learning rate by finding the minimum learning rate at which the average classification error decreased by at least half a percent. This occurred at $\alpha = 0.0002$ for the swine model and $\alpha = 0.0003$ for the poultry model.



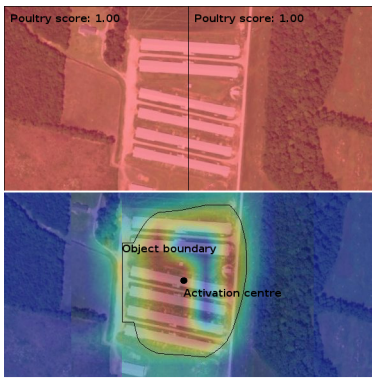
Supplementary Figure 7: F1 score curves. Harmonic mean of precision and sensitivity (recall), or "F1 Scores," at each classification threshold from 0 to 1. The F1 score is calculated as $2 \cdot \frac{\text{precision} \cdot \text{sensitivity}}{\text{precision} + \text{sensitivity}}$ at each threshold. As a visual aide, the dotted horizontal blue lines represent the F1 values at a threshold of 0.5, while the dotted blue vertical line demarcates the 0.5 threshold. The solid black series represents the unweighted test image set presented in the paper, while the dashed red series represents the weighted test image set described in the Supplemental Notes. The latter attempts to capture the more extreme class imbalance in the population of North Carolina images. The threshold of 0.5 represents a stabilisation point of the harmonic mean for both the weighted and unweighted test samples.



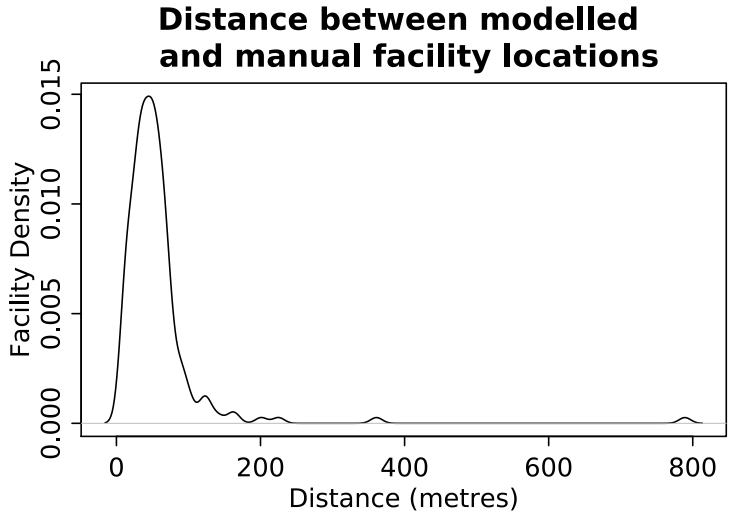
Supplementary Figure 8: Accuracy on weighted images. Population-weighted (red) and unweighted (black) receiver operating characteristic (ROC) curve (left) and precision-recall (PR) curve for the poultry (top) and swine (bottom) models on the test set of images. The weighted curves reflect the expected class distribution of images in North Carolina (approx. 269 control images for every 1 CAFO image), while the unweighted curves reflect the distribution of classes in the test set as originally sampled (approx. 4 control images for every 1 CAFO image). The grey lines represent the performance of a random classifier, or a model whose scores have no relationship to ground truth beyond chance (dashed for weighted images, solid for unweighted).



Supplementary Figure 9: Resource savings curve. Estimated manual resource savings at varying levels of sensitivity for the poultry model based on the weighted precision and recall curves presented in Supplementary Figure 8. The proportion of manual resources required is calculated as the number of predicted poultry images returned by the model divided by the total number of images in North Carolina. We estimate that 95% of poultry facilities could be detected using fewer than 10% of manual resources.



Supplementary Figure 10: Comparison of image-level predictions and facility localisation. Poultry CAFO split across two scored images in UTM grid system (top), and reconciliation into a single facility using class activation maps and k-means clustering (bottom). The model scores do not reflect the proportion of the facility that is within the image, leading to double counting of facilities when summing scores. With the localisation within the image provided by class activation maps, the algorithm is able to infer that these two images have activated on a single facility. Because the edges of the facility and its immediately surrounding area were most likely to activate, while some areas of the facility interior often did not activate, we calculate the CAFO object area as the area within the activation boundary.



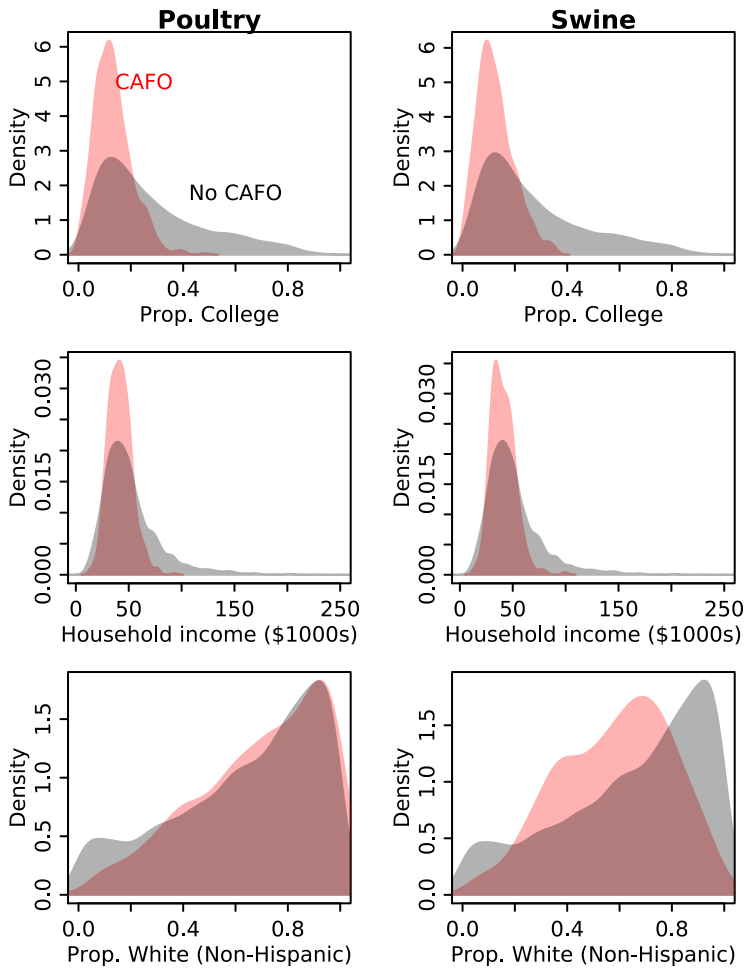
Supplementary Figure 11: Distance between modelled and manual locations. Distribution of distance in metres from modelled point location to the corresponding EIG facility location in Duplin and Cumberland, for locations manually validated to be the same facility in both datasets. The 250 meter threshold correctly matched modelled locations to EIG locations in 99% of cases.

Supplementary Figure 12: Facility definition conflicts. Illustration of rare conflicts between EIG (blue) and modelled (red) definitions of a facility. On the left, the model counted both sets of buildings as belonging to the same facility, while EIG counted them as separate facilities. On the right, the opposite situation occurred. For the most part, however, the two methods of tagging agreed. Map data: Google, DigitalGlobe (2019).

Supplementary Figure 13: False positive examples due to poor image resolution. From top moving clockwise, (a) a swine CAFO whose lagoon is washed out in the image; (b) the warehouse of a flooring manufacturer, which the EIG process manually classified as a CAFO; (c) an airplane hangar; (d) feed storage facilities.

Supplementary Figure 14: Ecological impact. Two examples of the potential ecological impact of poultry CAFOs detected by the model: proximity to water (right) and abandonment (right). Map data: Google, DigitalGlobe (2019).

Supplementary Figure 15: Abandonment. Ecological impact of CAFO abandonment for a poultry facility with liquid manure storage near a river. From left to right, the liquid manure storage system is developed between 2008 and 2013, but the facility is abandoned by 2018 without cleanup of the system. Map data: Google, DigitalGlobe (2019).



Supplementary Figure 16: Environmental justice. Comparison of demographics in census blocks with CAFOs (pink) and without CAFOs (grey), limiting to poultry facilities (left) or swine facilities (right) only. Both poultry and swine CAFOs tend to be in areas with lower proportions of college graduates and lower household income, but swine CAFOs are more prevalent in areas with higher proportions of non-white residents.

Supplementary Tables

Oversampling	Sensitivity (Recall)	Precision
Naive	0.94	0.89
ADASYN	0.93	0.88
SMOTE	0.93	0.88
Undersampling		
Naive	0.95	0.77
Centroids	0.94	0.76
Near miss	0.92	0.86
No balancing		
Original	0.92	0.89

Note: Precision is calculated using the original ratio of control to CAFO images (7:1).

Supplementary Table 1: Class balancing techniques. Comparison of class balancing techniques from a prototype swine model. We compared six different types of class balancing methods: oversampling the minority class using naive random sampling, Synthetic Minority Oversampling Technique (SMOTE), and Adaptive Synthetic (ADASYN) sampling, and undersampling the majority class using naive random sampling, cluster centroid, and near miss sampling (for the specific implementation of these techniques, see the imbalanced-learn package documentation at `http://contrib.scikit-learn.org/imbalanced-learn/stable/introduction.html`.) We found that naive random oversampling of the CAFO class improved sensitivity by 2% while maintaining similar precision to the unbalanced model. Undersampling improved sensitivity but significantly reduced precision.

Criteria	**Swine**	**Poultry**
Min. prop. of image taken up by CAFO	25%	10%
Min. prop. of CAFO visible in image	Any	If image prop. under 25%, over 67%, otherwise any
Manure storage	Visible, liquid	Not visible, dry

Supplementary Table 2: CAFO image training criteria. Criteria for CAFO image inclusion into the training sample. The criteria reflect that the main visual distinction between poultry and swine CAFOs is the presence of a lagoon.

	Swine Test Set			Poultry Test Set		
Category	Pop. n	Sample n	Image weight	Pop. n	Sample n	Image weight
CAFO images	1756	414	4.24	3228	821	3.93
CAFO neighbour images	18146	559	32.46	28974	378	76.65
Google Places images	2894	841	3.44	2894	841	3.44
Other control images	1323063	3456	382.83	1310763	3664	357.74

Supplementary Table 3: Image weight distribution. Distribution of images in the population and the test set for each model. The sample weight for each image category was calculated as $\frac{\text{Pop. Image } n}{\text{Sample Image } n}$, so that sum of the weights added up to the total number of images in the population.

Supplementary References

[1] Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE* **10**, 1–21 (2015). URL `https://doi.org/10.1371/journal.pone.0118432`.

[2] Nicole, W. CAFOs and environmental justice: The case of North Carolina. *Environmental health perspectives* **121**, a182 (2013).

[3] Smith, L. N. Cyclical learning rates for training neural networks. *arXiv:1506.01186* (2017). URL `https://arxiv.org/abs/1506.01186`.