

# Economical crowdsourcing for camera trap image classification

Pen-Yuan Hsing\*<sup>1</sup>, Steven Bradley<sup>2</sup>, Vivien T. Kent<sup>3,4</sup>, Russell A. Hill<sup>4</sup>, Graham C. Smith<sup>5</sup>, Mark J. Whittingham<sup>6</sup>, Jim Cokill<sup>3</sup>, Derek Crawley<sup>7</sup>, MammalWeb volunteers, Philip Stephens<sup>1</sup>

<sup>1</sup>Conservation Ecology Group, Department of Biosciences, Durham University, South Road, Durham, United Kingdom DH1 3LE

<sup>2</sup>Department of Computer Science, South Road, Durham, United Kingdom DH1 3LE

<sup>3</sup>Durham Wildlife Trust, Chilton Moor, Houghton-le-Spring, Tyne and Wear, United Kingdom DH4 6PU

<sup>4</sup>Department of Anthropology, Durham University, South Road, Durham, United Kingdom DH1 3LE

<sup>5</sup>National Wildlife Management Centre, Animal and Plant Health Agency, Sand Hutton Campus, Sand Hutton, York, United Kingdom YO41 1LZ

<sup>6</sup>Biology, School of Natural and Environmental Sciences, Newcastle University, Newcastle-Upon-Tyne, United Kingdom NE1 7RU

<sup>7</sup>The Mammal Society, 18 St John's Church Road, London, United Kingdom E9 6EJ

\*email: penyuanhsing@posteo.is; phone: +44 (0) 1913341252

## Running title

Economical camera trap image classifications

## Word count

Summary (title, author information, running title, word count, number of tables and figures, abstract, and keywords): 442

Main text: 6,003

Acknowledgements: 60

References: 1,500

Figure captions: 628

## Number of tables and figures

Five figures, five supplementary figures, one table, three supplementary tables.

## Abstract

Camera trapping is widely used to monitor mammalian wildlife but creates large image datasets that must be classified. In response, there is a trend towards crowdsourcing image classification. For high-profile studies of charismatic faunas, many classifications can be obtained per image, enabling consensus assessments of the image contents. For more local-scale or less charismatic communities, however, demand may outstrip the supply of crowdsourced classifications. Here, we consider MammalWeb, a local-scale project in North East England, which involves citizen scientists in both the capture and classification of sequences of camera trap images. We show that, for our global pool of image sequences, the probability of correct classification exceeds 99% with eight concordant crowdsourced classifications per sequence. However, there is high variation among species. For highly-recognisable species, species-specific consensus algorithms could be even more efficient; for difficult to spot or easily-confused taxa, expert classifications might be preferable. We show that two types

of incorrect classifications – misidentification of species and overlooking the presence of animals – have different impacts on the confidence of consensus classifications, depending on the true species pictured. Our results have implications for data capture and classification in increasingly numerous, local-scale citizen science projects. The species-specific nature of our findings suggests that the performance of crowdsourcing projects is likely to be highly sensitive to the local fauna and context. The generality of consensus algorithms will, thus, be an important consideration for ecologists interested in harnessing the power of the crowd to assist with camera trapping studies.

**Keywords:** data science, data classification, camera traps, crowdsourcing, citizen science, MammalWeb

## 45 Introduction

For several centuries (Greenwood, 2007; Ratcliff, 2008), citizen science projects have engaged non-professionals in the scientific process (Bonney et al., 2014). While ecological research has spearheaded the development of citizen science (Bonney et al., 2014; Dickinson et al., 2010), there are successful projects across a variety of disciplines from meteorology (Hennon et al., 2014) to astronomy (Willett et al., 2013). Typically, these initiatives crowdsource data capture (i.e. volunteers as “sensors” in Goodchild, 2007), data classification (interpreting collected data) or, occasionally, a combination of both (Kosmala et al., 2016). Some may even include citizen scientists in data analyses (Haklay, 2013).

In the field of ecology, technological developments (Newman et al., 2012) and increasing recognition of the need for monitoring over large spatial and temporal scales (Conrad and Hilchey, 2011; Stephens et al., 2015) have led to a proliferation of ecological citizen science projects (Kosmala et al., 2016). Concurrent with this is growing concern over “volunteer” skill and the resultant quality of data (Cohn, 2008; Dickinson et al., 2012, 2010; Lukyanenko et al., 2016). Data capture can be improved through iterative protocol refinement or intensive training (Kosmala et al., 2016). In one case of community-managed resource monitoring, regular follow-up training for volunteers enabled them to produce data of quality comparable to that collected by professional scientists (Danielsen et al., 2014).

For data classification, quality can be improved by aggregating inputs from multiple users, especially when processing large datasets. For example, Snapshot Serengeti is an ecological research project utilising crowdsourced classifications to identify the contents of images taken by motion sensing camera traps deployed in Serengeti National Park. Researchers attracted over 28,000 online volunteers who, within three days, cast one million “votes” for what they thought was in the camera trap photos, equivalent to processing an 18-month backlog of images (Swanson et al., 2015). For each photo, a consensus classification was determined from votes cast by an average of 27 volunteers. They were then validated against almost 4,000 “gold standard” images, classified by experts, to show that consensus classifications typically had an accuracy exceeding 97% (Swanson et al., 2016, 2015).

The considerable success of Snapshot Serengeti might be due, in part, to project-specific factors. These include: (1) the presence in images of highly charismatic and diverse African megafauna which are novel to largely European and American audiences; (2) the low image to volunteer ratio (approximately 1.2 million images for 28,000 volunteers, or ~43:1); and (3) the long-established platform (<https://www.zooniverse.org/>) on which the project was hosted, with a large and dedicated international userbase.

75 By contrast, many citizen science projects focus on less charismatic faunas in areas of lower species diversity. Despite their lower diversity, focal communities may include species of conservation concern, as well as species that are locally common and, therefore, important contributors to ecosystem function (Gaston and Fuller, 2008; Geider et al., 2001). The local relevance and lower charisma of these studies might make it harder to mobilise a large international userbase. As a result, it may be necessary to determine image contents with  
80 fewer user classifications by crowdsourcing more economically.

An example of this is MammalWeb, a project in North East England that pilots the approach of involving local citizen scientists in monitoring mammals with camera traps. Participants engage in both data capture and data classification (camera trapping and classification of images) as defined by Kosmala et al. (2016). MammalWeb has a high image to classifier ratio (~580:1) and monitors mammals that are less diverse and may be considered  
85 less charismatic (Lorimer, 2007) than their African counterparts. Preliminary indications from the pilot period are that the deployment of camera traps by MammalWeb's citizen scientists can yield useful data. Examples include the identification of a raccoon (*Procyon lotor*), an invasive non-native species, subsequently trapped and re-homed by the United Kingdom's (UK) Department for the Environment, Food and Rural Affairs (DEFRA), and the contribution of thousands of new mammal records to the Environmental Records  
90 Information Centre (ERIC) for the North East of England.

Using data collected in the MammalWeb study, we investigated economical approaches to aggregating user input into consensus classifications. This included analysing species-level variations in the number of classifications (including different combinations of correct and incorrect classifications) needed to achieve consensus at various confidence levels, and differentiating between two types of incorrect classifications: mis-  
95 identification of a species or missing the presence of an animal altogether.

Relative to applying a generic consensus algorithm to all images, we showed that images of certain species could be retired more rapidly because (1) consensus was achieved with fewer classifications or (2) referral to expert classification may be preferable. Since MammalWeb combines data collection and classification in one citizen science project, we also examined whether this increased engagement affected the accuracy of  
100 classifications.

## Methods

### ***Project background and citizen scientist recruitment***

MammalWeb focuses on North East England, addressing a general dearth of mammal monitoring in an area (Croft et al., 2017) with a relatively limited fauna (14 wild mammal species *cf.* 40 in the Snapshot Serengeti  
105 data base; Swanson et al., 2015). Between March 2015 and March 2018, we recruited 79 citizen scientists across the region (centred around County Durham) to deploy camera traps for the MammalWeb project. They consisted mainly of Durham University staff and members of the Durham Wildlife Trust (a local non-governmental organisation focused on environmental conservation, education and engagement). Recruiting and training citizen scientists from local community groups such as the Durham Wildlife Trust is comparable  
110 to projects such as eMammal (Forrester et al., 2017). Many participants were retirees, and most reported curiosity about local wildlife as their motivation for joining. A small number of contributors were local primary and secondary school teachers using camera traps in their teaching.

### **Camera trap data capture and classification**

115 After training the citizen scientists to use a standard protocol, they were lent camera traps (primarily Browning  
Strikeforce, Reconyx Hyperfire, and Bushnell cameras) and self-selected sites on which to deploy them. During  
deployment, all cameras were set to burst mode and would typically take three images in quick succession per  
trigger. By default, most cameras included a 30 second pause before the next trigger. Volunteers uploaded their  
camera trap images to the MammalWeb website (<http://www.MammalWeb.org/>), and also submitted metadata  
120 such as the deployment time period, location, make and model of camera trap, and height of camera above  
ground.

Anyone with an Internet connection can register on MammalWeb to classify images (i.e., to be a “Spotter”),  
including those who deployed camera traps and uploaded photos (i.e., “Trappers”). Spotters were recruited  
through the same channels as Trappers, plus at public events and schools. Spotter classification effort varied  
from tens to thousands of images. Consequently, to characterise the distribution and skewness of classification  
125 intensity by individual Spotters, we calculated the proportions of those who classified fewer than 100 images  
and greater than 1,000 images. We also determined the relative contribution from the top 10% of Spotters in  
terms of classifications.

Uploaded camera trap photos taken less than 10 seconds apart were grouped into sequences, which typically  
(c. 84% of sequences) consisted of the three images taken in one burst (indeed, 94% of sequences are of length  
130 2 or 3). The contextual information provided by adjacent images in a sequence should aid classifications that  
would otherwise be problematic (Fig. S1). Therefore, MammalWeb’s classification interface is such that the  
"next photo" button takes a Spotter to the next photo *in the sequence* rather than to another randomly selected  
one in the global pool of images (Fig. S2). By going backwards and forwards through a sequence, Spotters  
may show greater accuracy in classifying the animals depicted since there is a greater chance of at least one  
135 clear image within the sequence. Users were encouraged to proceed only after they have classified all images  
in a sequence. Upon clicking “next sequence”, they were shown a randomly selected sequence from the global  
pool (or, optionally, the user’s own pool of uploaded photo sequences).

The classifications for each image in a sequence were aggregated into the classification for that sequence. For  
example, a three-image sequence where the images are sequentially classified as “blank”, “rabbit,” and “grey  
140 squirrel” will have “rabbit and grey squirrel” as its classification. We treated each sequence as the base unit of  
animal detection, and all analyses for classification accuracy and consensus classifications were conducted at  
the sequence level.

### **Determining classification accuracy**

We determined the accuracy of MammalWeb citizen scientists and assessed how the nature of a classification  
145 – correct and incorrect – may influence the calculation of a consensus. This was done by comparison with a  
"gold standard" set of classifications created by us, consisting of 10,478 sequences (34,841 images).

We calculated the probabilities of a user classification being correct for each species. For incorrect  
classifications, we examined, for each species, the proportions of classifications that were for another species  
or for the absence of any animal. With this information we also constructed a confusion matrix breaking down

150 cases of mistaken identifications by species, and calculating false negative (missing the presence of a species) and false positive (stating a species is present when it is not) rates.

We also compared classification accuracies of citizen scientists who deployed camera traps and uploaded images (“Trappers”) and those who did not. Within the Trapper group, we also investigated whether they were more accurate when classifying their own images versus those uploaded by others. Both comparisons used  
155 generalised linear mixed effects models, with a binary response (correct or incorrect), spotter type (spotter or trapper, or uploader or other trapper) as a fixed effect, and spotter identity as a random effect.

### **Evaluating consensus classifications**

For consensus classifications, we determined the following for each sequence,  $j$ :  $T_j$  (“total classifications”), the total number of unique classifications for the sequence;  $P_{s,j}$  (“present”), the number of unique classifications  
160 indicating species  $s$  is present in one or more photos within the sequence;  $O_{s,j}$  (“other”), the number of unique classifications indicating that species not including  $s$  are present in the sequence;  $B_j$  (“blank”), the number of unique classifications indicating that the sequence is devoid of animals. The total number of classifications for a sequence is thus:  $T_j = P_{s,j} + O_{s,j} + B_j$ . These numbers allowed us to determine the number of classifications indicating a species’ presence in a sequence ( $P_{s,j}$ ) and the number indicating its absence (“absence”:  $A_{s,j} =$   
165  $O_{s,j} + B_j$ ). We then used this information for four separate analyses.

First, using all sequences in our gold standard set that were identified as containing species  $s$ , we asked what proportion of classifiers (“Spotters”) agreed with this designation ( $\sum_j P_{s,j} / \sum_j T_j$ ). This parameter, which we designate as  $\text{Pr}(s)$  (the probability that species  $s$  is correctly identified in a sequence), serves as a crude indicator of which species are typically most (or least) readily identified within our focal fauna. For each gold standard  
170 species  $s$ , we also examined classifiers’ incorrect classifications to determine the relative proportions of those that were misclassifications (given by  $O_{s,j}$ ) versus failed detections (given by  $B_j$ ). This comparison serves to indicate how the potential for classifiers to overlook or misclassify varies among species.

Second, we used binary logistic regression to assess how the presence of a species in an image sequence is related to the number of classifications indicating its presence and absence. We conducted this analysis both  
175 for the full data set (across all species) and then separately for different species. Specifically, we determined whether the number of classifications indicating presence ( $P_{s,j}$ ) and absence ( $A_{s,j} = O_{s,j} + B_j$ ) of a given species (or no species at all) in a sequence was related to its true presence in, or absence from, the sequence. This model can be represented as  $V_{s,j} \sim P_{s,j} + A_{s,j}$ , where  $V_{s,j}$  is a binomial indicator that species  $s$  is truly present in ( $V_{s,j} = 1$ ) or absent from ( $V_{s,j} = 0$ ) sequence  $j$  (and the error has a binomial distribution). Where  
180 multiple species have been identified to occur in sequence  $j$ , there may of course be multiple species in the image. This would not be a problem, as both users and gold-standard classifiers can classify multiple species in any image (and so, for two species  $a$  and  $b$  that occur in sequence  $j$ ,  $0 \leq P_{a,j} + P_{b,j} \leq 2T_j$ ). Far more commonly, however, where multiple species have been identified to occur in sequence  $j$ , one or more of those species has been designated in error. Here, using the entire data set would include non-independent data points  
185 (because, where species  $a$  and  $b$  are both identified as being in sequence  $j$ , even though only one of them is actually in the sequence, model  $V_{a,j} \sim P_{a,j} + A_{a,j}$  is necessarily the converse of model  $V_{b,j} \sim P_{b,j} + A_{b,j}$ ). To

avoid this issue, we created 1,000 random bootstrap samples of the data set, stratified by sequence, in which all sequences were represented only once. We analysed each bootstrap sample as described above, and report mean and standard deviations of their Akaike information criteria (AICs, Akaike, 1974). Analyses of the  
 190 (bootstrapped) full data set suggested strong support (based on AIC scores; see Results) for an influence of the pictured species  $s$  on the relationship between confidence in classifications and  $P_s$  and  $A_s$ . To determine the effect of this variation among species, we analysed data on the more commonly occurring species using only the subset of sequences for which at least one user has indicated the presence of the focal species.

Third, we investigated whether, for a given species  $s$  in sequence  $j$ , the impact on confidence of classifications for other species (“false positives”,  $O_{s,j}$ ) differs from that of blanks (“false negatives”,  $B_j$ ). This analysis recognises the fact that species differ in both their detectability and their recognisability; thus, classifications representing confusion over a species’ identity might reduce confidence in the species’ presence to a different extent to classifications suggesting that no animal species occurred in the sequence. This analysis used binary logistic regression, as described above; this time, the focus was on comparing the performance of the model  
 195  $V_{s,j} \sim P_{s,j} + O_{s,j} + B_j$  with that of the simpler model  $V_{s,j} \sim P_{s,j} + A_{s,j}$ .

Fourth, we determined the rate at which we can retire sequences of species from the pool of sequences to be classified, given a target confidence threshold. This was based on two sources of information. Specifically, we used  $\text{Pr}(s)$  from our first analysis as an estimate of the probability that any new classification would be for the pictured species. We also used fitted models of the form  $V_{s,j} \sim P_{s,j} + A_{s,j}$  to estimate the number of  
 200 classifications needed ( $R$ ) to achieve a given level of confidence  $C$ . For a given number of classifications indicating absence of a species in a sequence ( $A_{s,j} = \{0,1,2,3\}$ ), it is possible to identify the number of classifications for the species’ presence ( $P_{s,j}$ ) which would be required to give the desired confidence that the species is present:

$$R_{C,s,j} \sim P_{s,j} + A_{s,j}$$

210 The probability that this combination of classifications will be obtained is then:

$$\text{Pr} \left( A_{s,j}, P_{s,j} \mid \text{Pr}(s) \right) = \binom{A_{s,j} + P_{s,j}}{P_{s,j}} \text{Pr}(s)^{P_{s,j}} (1 - \text{Pr}(s))^{A_{s,j}}$$

The average number of classifications needed before a sequence containing a given species can be retired from the pool for classification is then given by the average sum of  $A_{s,j} + P_{s,j}$  for  $A_{s,j} = \{0,1,2,3\}$ , weighted by the probability with which each is obtained, plus the probability that none of these criteria are satisfied, multiplied  
 215 by the number of classifications we would accept before removing the sequence from the classification pool. We can then compare the implications of different approaches and target confidence thresholds for the speed at which sequences can be considered classified.

All data processing, analyses, and modelling was conducted in R 3.4.1 (R Core Team, 2017) with the packages dplyr (Wickham et al., 2017), ggplot2 (Wickham, 2016), lubridate (Grolemund and Wickham, 2011), lme4  
 220 (Bates et al., 2015) and EnvStats (Millard, 2013).

## Results

As of 7 March 2018, MammalWeb citizen scientists had cumulatively deployed camera traps at 234 unique sites in North East England for 15,368 camera trap days. This yielded 172,428 images uploaded to our website. Since project inception, 273 Spotters (including those who deployed camera traps, i.e. Trappers) had contributed, via the MammalWeb website, 286,844 classifications of the content of 121,537 images (42,077 sequences). For the images with at least one classification, the median number of classifications was 2 (IQR: 1 – 3, maximum: 33). The majority of classifications were submitted by a small number of Spotters (Fig. S3). More than half (57.7%) of MammalWeb users ( $n = 157$ ) classified less than 100 photos, while 12.5% of the users ( $n = 34$ ) each classified more than 1,000 photos (Fig. S3). The top 10% of Spotters ( $n = 27$ , 15 of whom were Trappers) contributed 84.8% of all classifications.

At the sequence level, 21 species have been classified in our data set. For most of the species in sequences with a gold standard, >90% of user-provided classifications were correct (Fig. 1a). Badgers (*Meles meles*) were recognised by almost 95% of classifiers and only five species were correctly classified by <80% of users. Species vary markedly in whether incorrect classifications are due to missing the presence of an animal ( $B_j$ ) or mistaking it for another species ( $O_{s,j}$ ) (Fig. 1b). For instance, most of the erroneous classifications of sequences containing brown hares (*Lepus europaeus*) were due to mistaken identification (58 out of 65 incorrect classifications; Fig. 1). By contrast, 96% of misclassifications of small rodents (a shared designation in MammalWeb for species of <500 g in body mass, principally rats, *Rattus norvegicus*; mice *Apodemus sylvaticus* and *Mus musculus*; and voles, *Microtus agrestis*) were due to them being missed altogether (470 out of 490 incorrect classifications where small rodents were present according to the gold standard; Table 1).

Among Spotters, those who also deployed camera traps and uploaded photos (“Trappers”) were slightly more accurate in their classifications (Fig. 2a). In addition, Trappers were more accurate when classifying images they had obtained than those uploaded by other Trappers (Fig. 2b).

Analyses of the data across species showed that both the number of classifications indicating presence and the number indicating absence of a species provide important information about the probability with which that species is actually in a sequence (Fig. 3). On the global level, when a single classification has been submitted indicating a species’ presence, it is about 95% likely that the species in question does appear in the sequence. Predictably, more classifications for the species being present increase the likelihood that it is there, while more classifications for its absence have the opposite effect (Fig. 3).

The above analysis is based on a model of the form  $V_{s,j} \sim P_{s,j} + A_{s,j}$ . However, models that included, also, the pictured species ( $s^*$ ) as a fixed factor, outperformed the simpler model ( $\Delta\text{AIC} = 203.9$ ,  $\text{SD} = 28.51$ ). Consequently, we also analysed the relationship between image contents and numbers of classifications for individual species. Twelve species (including “nothing”, or blank ( $B$ ), i.e., where no image in the sequence contained an animal) appeared in more than 200 gold standard sequences and so were analysed at the species level. For the different species, there was marked variation in the meaning of different combinations of classifications indicating presence and absence (Fig. 4). In particular, some designations (e.g., small rodents) require larger numbers of classifications for their presence to confer confidence in their appearance in the sequence (e.g.  $P = 3$  for 97.5% confidence), but classifications for their absence ( $A$ ) make relatively little

260 difference (Fig. 4). Other species, such as red foxes, need few classifications for their presence to instil  
confidence that they are truly present but small numbers of differing classifications substantially undermine  
that confidence (Fig. 4). Notably, increases in the number of classifications indicating that the sequence  
contains “nothing” do not materially increase the likelihood of consensus being correct (Fig. 4). Even with 10  
265 classifications indicating that the sequence contains “nothing”, the level of confidence does not rise above  
97.5%. Any dissenting classifications, indicating that there is “something” in the sequence, have a very high  
impact on confidence that the sequence is indeed devoid of animals.

Models for individual species differed when separating classifications for absence (*A*) into those for other  
species (*O*) and those for no animals (*B*) (Fig. S4). For six species, doing so produced a better-supported model  
(Table S1). For instance, the model for hedgehogs ( $\Delta\text{AIC} = 3.51$ ) suggests that classifications for the lack of  
any animals (*B*) reduce confidence to a greater extent than mistaken identifications (Fig. 5 and Fig. S5). Models  
270 for species such as grey squirrels ( $\Delta\text{AIC} = 31.25$ ) or small rodents ( $\Delta\text{AIC} = 34.87$ ) suggest that mistaken  
identification reduces confidence more than missing the presence of the animal altogether (Fig. 5 and Fig. S5).

Globally (without regard to specific species), 45.2% of sequences can be retired with 97.5% confidence after  
five classifications and a further 20.6% of sequences could be retired after nine (Table S2). At the 99%  
confidence level, 28.1% of sequences can be retired after eight classifications (Table S2). The implication of  
275 these analyses is that, on average, 9.2 classifications would be needed per sequence to retire them with 97.5%  
confidence, but an average of 14.1 classifications are required for 99% confidence. If algorithms for sequence  
retirement are sensitive to the species most likely to be pictured, 87.4% or more of sequences containing highly  
recognisable species, such as badgers, could be retired after just two classifications (with 97.5% confidence)  
(Table S3). However, less recognisable species would need many more classifications to instil confidence  
280 (Table S3). For example, only about 40% of sequences classified as hedgehogs can be retired at 97.5%  
confidence even after 18 classifications (Table S3).

## Discussion

There is a trend for citizen science projects to crowdsource data classification. The question of how  
proliferating projects can obtain confident classifications from a finite group of contributors suggests that more  
285 economic ways of utilising user input would be beneficial. Data from the MammalWeb project suggest that  
individual classifiers are typically highly accurate and that a reliable consensus could be reached with  
approximately nine classifications per sequence. Moreover, we show that greater economy could be obtained  
by treating different species separately, and by discriminating between classifications that conflict over the  
identity of the pictured species, and classifications suggesting no species is present. Here, we discuss our results  
290 and their implications for crowdsourced image classification, increasing the classification rate, and large-scale  
mammal monitoring.

### ***Implications for crowdsourced image classification***

The majority of MammalWeb’s camera trap image classifications originated from relatively few contributors  
(Fig. S3), a pattern common among scientific crowdsourcing efforts (Sauermaun and Franzoni, 2015). That  
295 the top 10% of MammalWeb classifiers (“Spotters”) contributed 84.8% of all classifications is comparable to



the average of 79% from a survey of seven projects on the Zooniverse citizen science platform (Sauermann and Franzoni, 2015).

Notably, Spotters who also helped to deploy camera traps (“Trappers”) were slightly more accurate in their classifications (Fig. 2a). This might be assumed to occur because citizen scientists involved in both the data capture and classification stages of the project are engaged to a higher level (Haklay, 2013) than those involved only in classification. Alternatively, it could reflect the fact that many Trappers are nature enthusiasts since they were recruited from a local nature-based organisation (similar to Forrester et al., 2017). However, the data show that this difference arises principally because Trappers were more accurate in classifying images captured by themselves (Fig. 2b). This is possibly due to direct access to those images on their own computers, where they can be scrutinised to a greater extent than on our website. It is also possible that these Trappers are simply more familiar with the fauna at sites where they deployed camera traps, although the vertebrate biota across North East England shows limited spatial variation.

We showed that the accuracy of volunteer-contributed classifications is generally high (Fig. 1). With only one classification indicating the presence of a species, the likelihood is about 95% that the species is indeed present (Fig. 3). For a given sequence where the species present is known, true positive rates are generally high, which also suggests high accuracy (Table 1). In spite of this accuracy, to confer higher confidence in consensus classifications, multiple classifications are required per sequence. Specifically, without an algorithm that distinguishes between species, sequences in our data set can be retired from the classification pool after an average of 9.2 classifications (for an accuracy of  $\geq 97.5\%$ ) or 14.1 classifications (for  $\geq 99\%$  accuracy) (Table S2). Given that there is some evidence that different types of votes against the presence of a species may carry different weight (and, in particular, that votes for the absence of any species of interest are generally less damaging to confidence than votes for a different species; Fig. 5), more elaborate approaches accounting for the nature of dissent might substantially improve these figures.

For some species, the number of classifications can be substantially reduced (e.g. 97.5% confidence with just two classifications indicating the presence of a badger, Fig. 4); for other species, however, larger numbers would be required and an early transfer to expert classification might be preferable (Table S2). Species-level differences were also evident when differentiating the impacts from mis-identification (i.e., the false positive identification of a species) or mistakenly stating that no animal was present (i.e. false negative) (Fig. 5, S5 and Table 1). A good example of the complications around false positives is given by brown hares. We found that brown hares are relatively poorly-recognised in our data set. In fact, they are commonly confused with rabbits (*Oryctolagus cuniculus*), the more frequently-occurring lagomorph in the region. Although our analyses suggest that the majority of sequences containing rabbits could be removed after only three or six classifications (depending on the desired confidence level), this overlooks the possibility that brown hares might be of more interest, would need many more classifications to compel confidence, and could be overlooked if apparent rabbit sequences are retired rapidly. More data would be required to assess this problem, especially in relation to the specific probability with which hares are classified as rabbits (and the resultant probability that a sequence could achieve consensus on a rabbit being pictured, even if a hare is the actual subject).

With these analyses, we illustrated the importance of considering (1) the entire *combination* of classifications for the presence and absence of a species when calculating consensus classifications, and (2) the potential

335 usefulness of a species-specific approach to doing so rather than applying a single algorithm to the entire dataset. An additional benefit is that even though an animal may be more or less evident in different images, achieving consensus for a sequence would let us retire all of its constituent images without needing consensus on each one.

340 One finding that might be very general to crowdsourced classifications is that far more classifications are required to classify with confidence a sequence having no subjects of interest, than to classify with confidence a sequence that does contain animals. Indeed, even 10 uncontested classifications suggesting that a sequence is devoid of animals is not sufficient to impart 97.5% confidence in that designation (Fig. 4). That contrasts with the other species considered in Fig. 4, which require between two and four uncontested classifications to give high confidence that they are actually present. As we noted above, more efficient algorithms for  
345 crowdsourcing reliable classifications should probably discriminate between the weight attributed to disagreements over whether a species is present and disagreements over the identity of a pictured species.

### ***Increasing the classification rate***

Our analyses suggest that a higher ratio of classifiers to images will be necessary before MammalWeb can be expanded and expected to contribute to timely and informative ecological analyses. In particular, our analyses  
350 suggest that, without distinguishing species, at least five or an average of 9.2 classifications will be required per sequence for 97.5% confidence in consensus. In the first 120 weeks of the project, we accumulated new sequences at a rate of approximately 370 per week, and new sequence classifications at a rate of approximately 1,324 per week; this yields a ratio of approximately 3.6 classifications per sequence. This suggests that one option to ensure that classifications keep pace with accumulating image data is to increase our classifier pool  
355 by a factor of approximately 2.5, relative to the number of camera trappers. At present, we have approximately 3.5 classifiers to every trapper, so this would need to increase to approximately 9:1. Such an increase should inform any efforts to extend the reach of the MammalWeb project and can be built on existing work that seeks to understand citizen scientist motivations and to promote their continued involvement (Eveleigh et al., 2014; Everett and Geoghegan, 2016; Jennett et al., 2016; Wald et al., 2016).

360 One alternative to increasing the relative size of the classifier pool is to encourage higher classification effort from existing users. Species-specific algorithms for sequence retirement could be problematic in this regard. For example, some of the more recognisable species in our data set are also some of the more charismatic. If these sequences are removed more rapidly than others, the data set could rapidly become biased towards less charismatic species, more indistinct photos, and images devoid of animals. Preliminary evidence from  
365 Snapshot Serengeti suggests that moderate numbers of images devoid of wildlife can actually increase classifier-engagement, by ensuring the relative rarity and novelty of wildlife images (Bower et al., 2015). By contrast, MammalWeb volunteers routinely cite animal-free images (about 41% of all sequences, based on gold standard classifications) as a deterrent to classification. It would be useful to investigate the source of this difference in the reported impacts of blank images on motivation. This may be related to the charisma of the  
370 animals being monitored, whether a project involves citizen scientists in both data capture and classification, user interface design, or inaccuracies in self-reporting.

The importance of sequences devoid of animals is clear (Fig. 4). Given the high proportion (41% according to the gold standard) of blank sequences in our data set (and many other camera trap data sets), it is clear that the

375 relatively low confidence with which blank sequences can be classified will have a major impact on the overall  
speed at which sequences can be retired without a species-specific classification algorithm. Options for  
reducing the proportion of blanks in the data set include asking Trappers – who are more accurate at classifying  
their own images (Fig. 2) – to pre-screen their data and remove blanks before upload, or using an automated  
algorithm to do so (see further below).

380 One further possibility for overcoming limitations to classification effort is to use the data set to identify  
classifiers who have very high accuracy, giving a higher weighting to their votes, or preferentially tasking them  
with classifying more difficult images. User skill level was accounted for in one of the Bayesian consensus  
models by Siddharthan et al. (2016), requiring 3.2 classifications per image to achieve 91% confidence. Some  
crowdsourcing platforms (e.g. van der Wal et al., 2016) include automated checking and training functionality  
385 with computer-generated structured feedback for volunteers, which could help to increase individual accuracy  
and reduce required numbers of classifications.

### ***Implications for large-scale mammal monitoring***

In contrast to some other taxa, mammals have not been routinely monitored at a community level in the UK  
(Battersby and Greenwood, 2004; Croft et al., 2017). Over the past two decades, mammals have been recorded  
by many of the volunteers who conduct the British Trust for Ornithology's (BTO) Breeding Bird Survey (BBS)  
390 (Harris et al., 2016). However, given the nocturnal habits and generally low detectability of many mammals,  
the relatively short period during which the daytime-only BBS is carried out means that many species will be  
missed where they occur, and site-specific changes could be highly subject to stochasticity. Camera trapping  
would deliver a substantially richer picture of mammal occurrence in space and time and, ultimately, an  
approach like MammalWeb could be used to monitor mammals at a national level. In spite of this,  
395 MammalWeb was deliberately implemented at a local level to determine the feasibility of the approach. Our  
analyses suggest that the approach taken by MammalWeb should be feasible with modest efforts to increase  
the engagement or accuracy of existing classifiers, or the ratio of classifiers to images. The system could,  
consequently, be extended – but, at least given the current approach, it would be important to increase  
recruitment of classifiers to a greater extent than recruitment of camera trappers.

400 More generally, mammal monitoring using camera traps continues to grow globally (Rowcliffe and Carbone,  
2008), and there are increasing calls for more systematic and widespread approaches to the challenge  
(Steenweg et al., 2017). Crowdsourcing image classification is one solution to this challenge, and MammalWeb  
is one of several platforms that engages citizens for wildlife image classification. Others include Instant Wild  
(<http://www.edgeofexistence.org/instantwild/>, reviewed in Verma et al., 2016), Zooniverse (Simpson et al.,  
405 2014), eMammal (McShea et al., 2015), iSpot (Silvertown et al., 2015) and BeeWatch (van der Wal et al.,  
2016). While our findings regarding accuracy for specific species might not generalise to other platforms, the  
approach to crowdsourcing classifications should.

There are several reasons why our approach might compare favourably to previous algorithms, especially on a  
species-by-species basis. As previously discussed, our classifiers are largely local to North East England and  
410 so are likely to be highly familiar with the small number of species commonly occurring on camera traps in  
the area. This can be seen in the high accuracy of their classifications (Fig. 2), especially from those who do  
the camera trapping (Fig. 1). Moreover, classifiers on MammalWeb are shown entire sequences of images,

potentially benefiting from contextual information across the sequence. Whether this provides a measurable benefit and, if so, to what extent, would be straightforward to determine with a platform that can easily be adjusted to show photos either individually or in sequence. Overall, our requirement for as few as five classifications per sequence for 97.5% confidence (if an animal is present) shows greater achievable efficiency than consensus algorithms employed where efficiency is not a strong requirement (Swanson et al., 2016).

Researchers frequently point to image classification as a major barrier to making best use of their camera trapping data. As camera trapping increases in scope, the demand for citizen scientists to assist with image classification is also likely to increase. Whether supply can keep pace with demand is unclear but it is likely that more and larger projects will compete for a finite pool of classifiers, with projects focused on less charismatic or conservation-relevant faunas struggling to meet demand. More refined approaches to training volunteers and making use of their data (e.g. van der Wal et al., 2016) should help. In addition, automated techniques to assist with image recognition may become necessary to alleviate the classification challenge. This need will be even more pronounced as those running camera trapping studies embrace more complex forms of analysis, such as those requiring animal speed and distance detection (Howe et al., 2017; Rowcliffe et al., 2016). Automated solutions are starting to emerge but, so far, have been proprietary (Kays *pers. comm.*), require manual image pre-processing (Yu et al., 2013), or yield very high false positive rates (Price Tack et al., 2016). Whilst there is likely to be low transferability of species-detection algorithms among studies, experience at MammalWeb provides a strong motivation for change detection algorithms (Radke et al., 2005) simply to highlight (and remove) photos unlikely to contain wildlife; as discussed above, this process could substantially reduce the average number of classifications required to retire sequences. Knowing the presence and identity of wildlife within sequences could provide a data set useful for training machine learning algorithms that are under development (Thom, 2017).

In summary, we believe MammalWeb has demonstrated the viability of a local citizen science camera trapping project that can sustainably monitor wildlife. Importantly, we have shown the benefits of considering species level differences when calculating consensus classifications including the relative impacts from false positive and false negative classifications. Our findings regarding the importance to retirement rates of reducing the proportion of 'blank' sequences in the data set are highly likely to generalise across projects. Other differences from past citizen science projects, including involving citizen scientists in data capture and classification, the methods we used for crowdsourcing data classifications, and our insights into the use of sequence-level classifications to improve retirement rates of photos, are also of value to future monitoring initiatives.

## Acknowledgements

We gratefully acknowledge C. Branston, M. Dawson, L. Gardner, and C. Neal for their assistance with the MammalWeb project. We also thank two anonymous reviewers for their valuable criticisms during the preparation of this paper. This work is supported by the United Kingdom Heritage Lottery Fund, the British Ecological Society, and a Durham University Doctoral Scholarship for P.-Y. Hsing.

## References

- 450 Akaike, H., 1974. A new look at the statistical model identification. *IEEE Trans. Autom. Control* 19, 716–723. <https://doi.org/10.1109/TAC.1974.1100705>
- Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting Linear Mixed-Effects Models Using **lme4**. *J. Stat. Softw.* 67. <https://doi.org/10.18637/jss.v067.i01>
- Battersby, J.E., Greenwood, J.J., 2004. Monitoring terrestrial mammals in the UK: Past, present and future, using lessons from the bird world. *Mammal Rev.* 34, 3–29.
- 455 Bonney, R., Shirk, J.L., Phillips, T.B., Wiggins, A., Ballard, H.L., Miller-Rushing, A.J., Parrish, J.K., 2014. Next steps for citizen science. *Science* 343, 1436–1437. <https://doi.org/10.1126/science.1251554>
- Bower, A., Maidel, V., Lintott, C., Swanson, A., Miller, G., 2015. This image intentionally left blank: Mundane images increase citizen science participation, in: 2015 Conference on Human Computation & Crowdsourcing. Presented at the Conference on Human Computation & Crowdsourcing, San
- 460 Diego, California, United States.
- Cohn, J.P., 2008. Citizen science: Can volunteers do real research? *BioScience* 58, 192–197. <https://doi.org/10.1641/B580303>
- Conrad, C.C., Hilchey, K.G., 2011. A review of citizen science and community-based environmental monitoring: issues and opportunities. *Environ. Monit. Assess.* 176, 273–291. <https://doi.org/10.1007/s10661-010-1582-5>
- 465 Croft, S., Chauvenet, A.L.M., Smith, G.C., 2017. A systematic approach to estimate the distribution and total abundance of British mammals. *PLOS ONE* 12, e0176339. <https://doi.org/10.1371/journal.pone.0176339>
- Danielsen, F., Jensen, P.M., Burgess, N.D., Altamirano, R., Alviola, P.A., Andrianandrasana, H., Brashares, J.S., Burton, A.C., Coronado, I., Corpuz, N., Enghoff, M., Fjeldså, J., Funder, M., Holt, S., Hübertz, H., Jensen, A.E., Lewis, R., Massao, J., Mendoza, M.M., Ngaga, Y., Pipper, C.B., Poulsen, M.K., Rueda, R.M., Sam, M.K., Skielboe, T., Sørensen, M., Young, R., 2014. A multicountry assessment of tropical resource monitoring by local communities. *BioScience* 64, 236–251. <https://doi.org/10.1093/biosci/biu001>
- 470 Dickinson, J.L., Shirk, J., Bonter, D., Bonney, R., Crain, R.L., Martin, J., Phillips, T., Purcell, K., 2012. The current state of citizen science as a tool for ecological research and public engagement. *Front. Ecol. Environ.* 10, 291–297. <https://doi.org/10.1890/110236>
- Dickinson, J.L., Zuckerberg, B., Bonter, D.N., 2010. Citizen science as an ecological research tool: Challenges and benefits. *Annu. Rev. Ecol. Evol. Syst.* 41, 149–172. <https://doi.org/10.1146/annurev-ecolsys-102209-144636>
- 480 Eveleigh, A., Jennett, C., Blandford, A., Brohan, P., Cox, A.L., 2014. Designing for dabblers and deterring drop-outs in citizen science, in: Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems, CHI '14. Presented at the SIGCHI Conference on Human Factors in Computing Systems, ACM, New York, New York, USA, pp. 2985–2994. <https://doi.org/10.1145/2556288.2557262>
- 485 Everett, G., Geoghegan, H., 2016. Initiating and continuing participation in citizen science for natural history. *BMC Ecol.* 16, 15–22. <https://doi.org/10.1186/s12898-016-0062-3>
- Forrester, T.D., Baker, M., Costello, R., Kays, R., Parsons, A.W., McShea, W.J., 2017. Creating advocates for mammal conservation through citizen science. *Biol. Conserv.*, The role of citizen science in biological conservation 208, 98–105. <https://doi.org/10.1016/j.biocon.2016.06.025>
- 490 Gaston, K.J., Fuller, R.A., 2008. Commonness, population depletion and conservation biology. *Trends Ecol. Evol.* 23, 14–19. <https://doi.org/10.1016/j.tree.2007.11.001>
- Geider, R.J., Delucia, E.H., Falkowski, P.G., Finzi, A.C., Grime, J.P., Grace, J., Kana, T.M., La Roche, J., Long, S.P., Osborne, B.A., others, 2001. Primary productivity of planet earth: Biological

- 495 determinants and physical constraints in terrestrial and aquatic habitats. *Glob. Change Biol.* 7, 849–882.
- Goodchild, M.F., 2007. Citizens as sensors: The world of volunteered geography. *GeoJournal* 69, 211–221. <https://doi.org/10.1007/s10708-007-9111-y>
- 500 Greenwood, J.J.D., 2007. Citizens, science and bird conservation. *J. Ornithol.* 148, 77–124. <https://doi.org/10.1007/s10336-007-0239-9>
- Grolemund, G., Wickham, H., 2011. Dates and times made easy with lubridate. *J. Stat. Softw.* 40, 1–25. <https://doi.org/10.18637/jss.v040.i03>
- Haklay, M., 2013. Citizen science and volunteered geographic information: Overview and typology of participation, in: Sui, D., Elwood, S., Goodchild, M. (Eds.), *Crowdsourcing Geographic Knowledge*. Springer Netherlands, pp. 105–122. [https://doi.org/10.1007/978-94-007-4587-2\\_7](https://doi.org/10.1007/978-94-007-4587-2_7)
- 505 Harris, S.J., Massimino, D., Newson, S.E., Eaton, M.A., Marchant, J.H., Balmer, D.E., Noble, D.G., Gillings, S., Procter, D., Pearce-Higgins, J.W., 2016. The breeding bird survey 2015 (No. 687). British Trust for Ornithology, Thetford, United Kingdom.
- Hennon, C.C., Knapp, K.R., Schreck, C.J., Stevens, S.E., Kossin, J.P., Thorne, P.W., Hennon, P.A., Kruk, M.C., Rennie, J., Gadéa, J.-M., Striegl, M., Carley, I., 2014. Cyclone center: Can citizen scientists improve tropical cyclone intensity records? *Bull. Am. Meteorol. Soc.* 96, 591–607. <https://doi.org/10.1175/BAMS-D-13-00152.1>
- 510 Howe, E.J., Buckland, S.T., Després-Einspenner, M.-L., Kühl, H.S., 2017. Distance sampling with camera traps. *Methods Ecol. Evol.* <https://doi.org/10.1111/2041-210X.12790>
- 515 Jennett, C., Kloetzer, L., Schneider, D., Iacovides, I., Cox, A., Gold, M., Brian, F., Eveleigh, A., Mathieu, K., Ajani, Z., Talsi, Y., 2016. Motivations, learning and creativity in online citizen science. *J. Sci. Commun.* 15.
- Kosmala, M., Wiggins, A., Swanson, A., Simmons, B., 2016. Assessing data quality in citizen science. *Front. Ecol. Environ.* 14, 551–560. <https://doi.org/10.1002/fee.1436>
- 520 Lorimer, J., 2007. Nonhuman charisma. *Environ. Plan. Soc. Space* 25, 911–932. <https://doi.org/10.1068/d71j>
- Lukyanenko, R., Parsons, J., Wiersma, Y.F., 2016. Emerging problems of data quality in citizen science. *Conserv. Biol.* 30, 447–449. <https://doi.org/10.1111/cobi.12706>
- McShea, W.J., Forrester, T., Costello, R., He, Z., Kays, R., 2015. Volunteer-run cameras as distributed sensors for macrosystem mammal research. *Landsc. Ecol.* 1–12. <https://doi.org/10.1007/s10980-015-0262-9>
- 525 Millard, S.P., 2013. *EnvStats - An R package for environmental statistics*, 2nd ed. Springer-Verlag, New York, New York, United States.
- Newman, G., Wiggins, A., Crall, A., Graham, E., Newman, S., Crowston, K., 2012. The future of citizen science: emerging technologies and shifting paradigms. *Front. Ecol. Environ.* 10, 298–304. <https://doi.org/10.1890/110294>
- 530 Price Tack, J.L., West, B.S., McGowan, C.P., Ditchkoff, S.S., Reeves, S.J., Keever, A.C., Grand, J.B., 2016. AnimalFinder: A semi-automated system for animal detection in time-lapse camera trap images. *Ecol. Inform.* 36, 145–151. <https://doi.org/10.1016/j.ecoinf.2016.11.003>
- R Core Team, 2017. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
- 535 Radke, R.J., Andra, S., Al-Kofahi, O., Roysam, B., 2005. Image change detection algorithms: A systematic survey. *Trans. Image Process.* 14, 294–307. <https://doi.org/10.1109/TIP.2004.838698>
- Ratcliff, J., 2008. *The transit of Venus enterprise in Victorian Britain*. Pickering & Chatto, London, United Kingdom.
- 540 Rowcliffe, J.M., Carbone, C., 2008. Surveys using camera traps: Are we looking to a brighter future? *Anim. Conserv.* 11, 185–186. <https://doi.org/10.1111/j.1469-1795.2008.00180.x>

- Rowcliffe, J.M., Jansen, P.A., Kays, R., Kranstauber, B., Carbone, C., 2016. Wildlife speed cameras: Measuring animal travel speed and day range using camera traps. *Remote Sens. Ecol. Conserv.* 2, 84–94. <https://doi.org/10.1002/rse2.17>
- 545 Sauermann, H., Franzoni, C., 2015. Crowd science user contribution patterns and their implications. *Proc. Natl. Acad. Sci.* 201408907. <https://doi.org/10.1073/pnas.1408907112>
- Siddharthan, A., Lambin, C., Robinson, A.-M., Sharma, N., Comont, R., O'mahony, E., Mellish, C., Wal, R.V.D., 2016. Crowdsourcing without a crowd: Reliable online species identification using Bayesian models to minimize crowd size. *ACM Trans. Intell. Syst. Technol.* 7, 1–20.
- 550 <https://doi.org/10.1145/2776896>
- Silvertown, J., Harvey, M., Greenwood, R., Dodd, M., Rosewell, J., Rebelo, T., Ansine, J., McConway, K., 2015. Crowdsourcing the identification of organisms: A case-study of iSpot. *ZooKeys* 480, 125–146. <https://doi.org/10.3897/zookeys.480.8803>
- Simpson, R., Page, K.R., De Roure, D., 2014. Zooniverse: Observing the world's largest citizen science platform, in: *Proceedings of the 23rd International Conference on World Wide Web, WWW '14 Companion*. ACM, New York, New York, United States, pp. 1049–1054. <https://doi.org/10.1145/2567948.2579215>
- 555 Steenweg, R., Hebblewhite, M., Kays, R., Ahumada, J., Fisher, J.T., Burton, C., Townsend, S.E., Carbone, C., Rowcliffe, J.M., Whittington, J., Brodie, J., Royle, J.A., Switalski, A., Clevenger, A.P., Heim, N., Rich, L.N., 2017. Scaling-up camera traps: Monitoring the planet's biodiversity with networks of remote sensors. *Front. Ecol. Environ.* 15, 26–34. <https://doi.org/10.1002/fee.1448>
- 560 Stephens, P.A., Pettorelli, N., Barlow, J., Whittingham, M.J., Cadotte, M.W., 2015. Management by proxy? The use of indices in applied ecology. *J. Appl. Ecol.* 52, 1–6. <https://doi.org/10.1111/1365-2664.12383>
- 565 Swanson, A., Kosmala, M., Lintott, C., Packer, C., 2016. A generalized approach for producing, quantifying, and validating citizen science data from wildlife images. *Conserv. Biol.* 30, 520–531. <https://doi.org/10.1111/cobi.12695>
- Swanson, A., Kosmala, M., Lintott, C., Simpson, R., Smith, A., Packer, C., 2015. Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. *Sci. Data* 2. <https://doi.org/10.1038/sdata.2015.26>
- 570 Thom, H., 2017. Unified detection system for automatic, real-time, accurate animal detection in camera trap images from the arctic tundra (Master's thesis). The Arctic University of Norway.
- van der Wal, R., Sharma, N., Mellish, C., Robinson, A., Siddharthan, A., 2016. The role of automated feedback in training and retaining biological recorders for citizen science. *Conserv. Biol.* 30, 550–561. <https://doi.org/10.1111/cobi.12705>
- 575 Verma, A., van der Wal, R., Fischer, A., 2016. Imagining wildlife: New technologies and animal censuses, maps and museums. *Geoforum* 75, 75–86. <https://doi.org/10.1016/j.geoforum.2016.07.002>
- Wald, D.M., Longo, J., Dobell, A.R., 2016. Design principles for engaging and retaining virtual citizen scientists. *Conserv. Biol.* 30, 562–570. <https://doi.org/10.1111/cobi.12627>
- 580 Wickham, H., 2016. *ggplot2 - elegant graphics for data analysis*, 2nd ed, Use R! Springer International Publishing.
- Wickham, H., Francois, R., Henry, L., Müller, K., 2017. *dplyr: A Grammar of Data Manipulation*.
- Willett, K.W., Lintott, C.J., Bamford, S.P., Masters, K.L., Simmons, B.D., Casteels, K.R.V., Edmondson, E.M., Fortson, L.F., Kaviraj, S., Keel, W.C., Melvin, T., Nichol, R.C., Raddick, M.J., Schawinski, K., Simpson, R.J., Skibba, R.A., Smith, A.M., Thomas, D., 2013. Galaxy Zoo 2: Detailed morphological classifications for 304122 galaxies from the Sloan Digital Sky Survey. *Mon. Not. R. Astron. Soc.* 435, 2835–2860. <https://doi.org/10.1093/mnras/stt1458>
- 585 Yu, X., Wang, J., Kays, R., Jansen, P.A., Wang, T., Huang, T., 2013. Automated identification of animal species in camera trap images. *EURASIP J. Image Video Process.* 2013, 52. <https://doi.org/10.1186/1687-5281-2013-52>
- 590





## Figure captions

**Figure 1.** (a) Proportional accuracy of submitted classifications across the whole pool of sequences with gold standard classifications. Sample sizes ( $n$ ) represent the number of classifications provided for sequences in which the gold standard indicates that the named species is present. Vertical lines show (from left to right) 80, 90 and 95% accuracy across all classifications of these sequences. (b) Proportions of incorrect classifications (classifications indicating absence of the true species in a sequence) that were for another species (green) or the absence of any animal (blue). Vertical line is 50%. Sample sizes ( $n$ ) are the number of incorrect classifications.

**Figure 2.** Of the citizen scientists who classified at least 10 sequences, (a) those who deployed camera traps (34 “Trappers”, 10,385 classifications) were marginally more accurate at image classification than those who did not (87 “Spotters”, 9,425 classifications) but this effect was only weakly supported ( $\Delta\text{AIC} = 1.7$ , model weight = 0.70, relative to a model that did not account for the Spotter type). (b) There was strong support for the finding that 24 Trappers who classified images they uploaded (“Uploaders”, 1,921 classifications) were more accurate than Trappers who classified images uploaded by other Trappers (“Other Trappers”, 7,469 classifications) ( $\Delta\text{AIC} = 138.3$ , model weight = 1.00, relative to a model that did not account for the Spotter type). In both panels, each data point represents a different individual; point size reflects relative numbers of classifications. Boxes and whiskers summarise predicted accuracy levels across individuals (line across each box indicates the median and the box boundaries indicate the interquartile range, IQR; whiskers identify extreme data points that are not more than 1.5 times the IQR on both sides; dots are more extreme outliers).

**Figure 3.** Global-level relationship between the number of classifications for the presence ( $P$ ) and absence ( $A$ ) of a given species in a sequence and the probability that it is indeed in the sequence. Solid lines show the mean relationship (over 1,000 bootstrapped samples) between the probability (predicted by the fitted model) that a species is present in the sequence and the number of classifications for that species ( $P$ ), for 0 (orange line), 1 (blue line), 2 (green line) and 3 (red line) classifications indicating the species is absent ( $A$ ). Polygons around the lines show  $\pm$  mean SE across the bootstrapped samples. Dashed horizontal lines show probabilities of 0.975 and 0.99. Corresponding dashed vertical lines show the number of classifications for the species required to give a confidence of 97.5%.

**Figure 4.** Species-level relationship between the number of classifications indicating the presence ( $P$ ) and absence ( $A$ ) of a given species, and the probability that it appears in a sequence. Solid lines show the mean relationship between the probability (predicted by the fitted model) that a species is present in the sequence and the number of classifications for that species, for 0 (orange line), 1 (blue line), 2 (green line) and 3 (red line) classifications indicating the species is absent. Polygons around the lines show  $\pm$  mean SE. Dashed horizontal lines show probabilities of 0.975 and 0.99. Corresponding dashed vertical lines show the number of classifications for the species that are required to give a confidence of 97.5%.

**Figure 5.** Implications of distinguishing between different types of classifications indicating that a species is absent ( $A$ ). For some typically highly-detectable species, such as the badger, classifications suggesting that no animal is present in the sequence (“false negatives”,  $B$ ) are more damaging to confidence than are classifications suggesting that the pictured species is some other species (“false positives”,  $O$ ). For visually-

530 distinctive species, such as the grey squirrel, the converse is true. For species that are seldom overlooked or misclassified, classifications indicating their absence count equally, regardless of whether they are for other species or no animals at all.

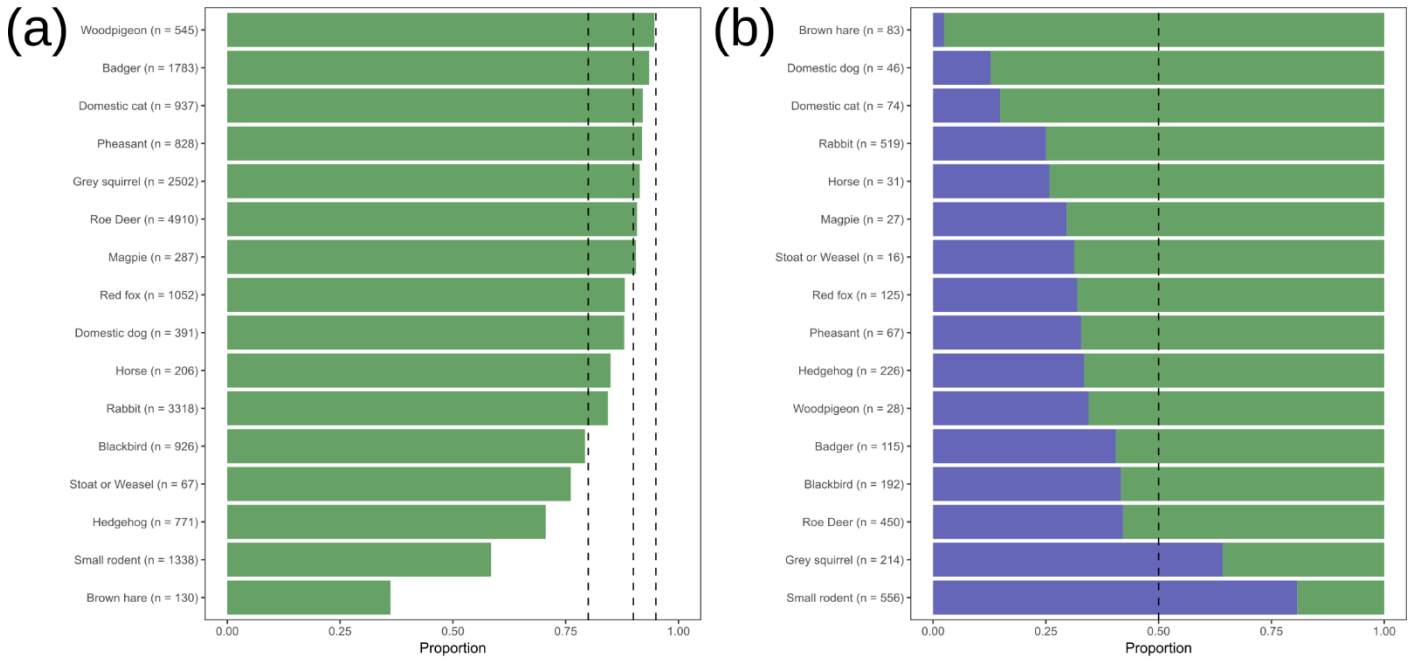
# Tables

535 **Table 1.** Confusion matrix for accuracies of commonly classified species. Shaded cells are true positive rates representing the probability of a user classification being correct given an image of a certain species. False negative rates are the inverse (including stating there is nothing when an animal is present), and false positive rates are how often a species is identified when it is not there. Numbers of classifications are in parentheses. For example, for badgers, there are 1694 user classifications indicating their presence of which 1.6% are incorrect (false positives). There are 1,745 classifications where badgers are truly present, of which 95.5% were correct identified (true positives), and 4.5% where they were not identified (false negatives).

540

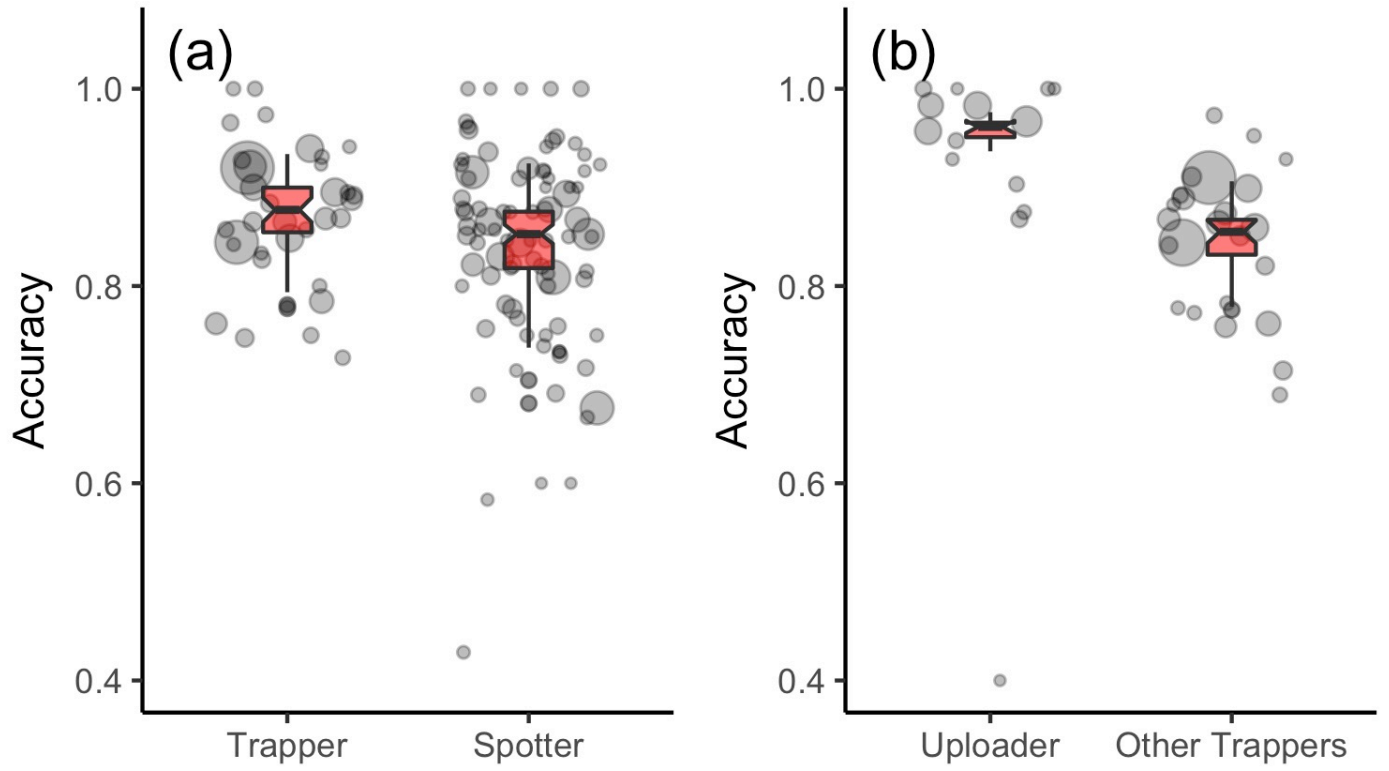
User classifications	Gold standard	Badger (1745)	Blackbird (864)	Domestic cat (907)	Grey squirrel (2485)	Hedgehog (689)	Pheasant (819)	Rabbit (3213)	Red fox (1004)	Roe Deer (4787)	Small rodent (1272)	Nothing (6344)	False positive rate
Badger (1694)		.955				.023		.001	.001	.001		.000	.016
Blackbird (790)			.850	.001	.001		.002	.002		.001	.001	.002	.071
Domestic cat (886)		.001		.951				.001	.007			.001	.026
Grey squirrel (2402)		.001	.007	.001	.920	.003	.005	.013		.001	.002	.004	.048
Hedgehog (580)		.001		.006	.001	.790		.002			.008	.001	.062
Pheasant (793)					.001		.929	.003		.001		.001	.040
Rabbit (2928)		.002	.009		.004	.026	.009	.871	.003	.002	.001	.002	.044
Red fox (971)		.002	.001	.008	.000	.004		.001	.923	.003	.002	.001	.045
Roe Deer (4526)		.002	.002		.001		.009	.004	.004	.931		.003	.015
Small rodent (836)				.001	.001	.017		.003	.001		.615	.004	.065
Nothing (7806)		.035	.119	.026	.063	.136	.040	.062	.058	.053	.369	.977	.206
False negative rate		.045	.150	.049	.080	.210	.071	.129	.077	.069	.385	.023	

# Figures



545

Figure 1



550 **Figure 2**

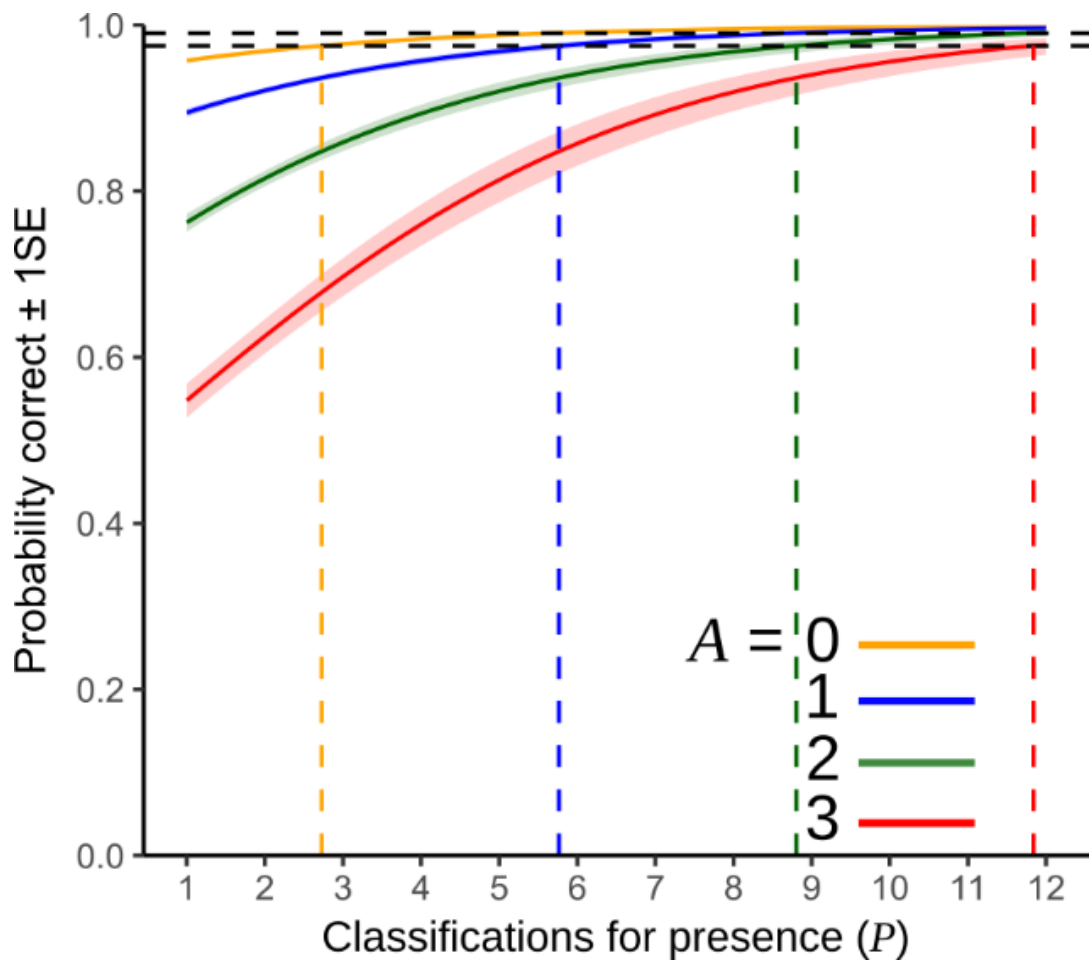
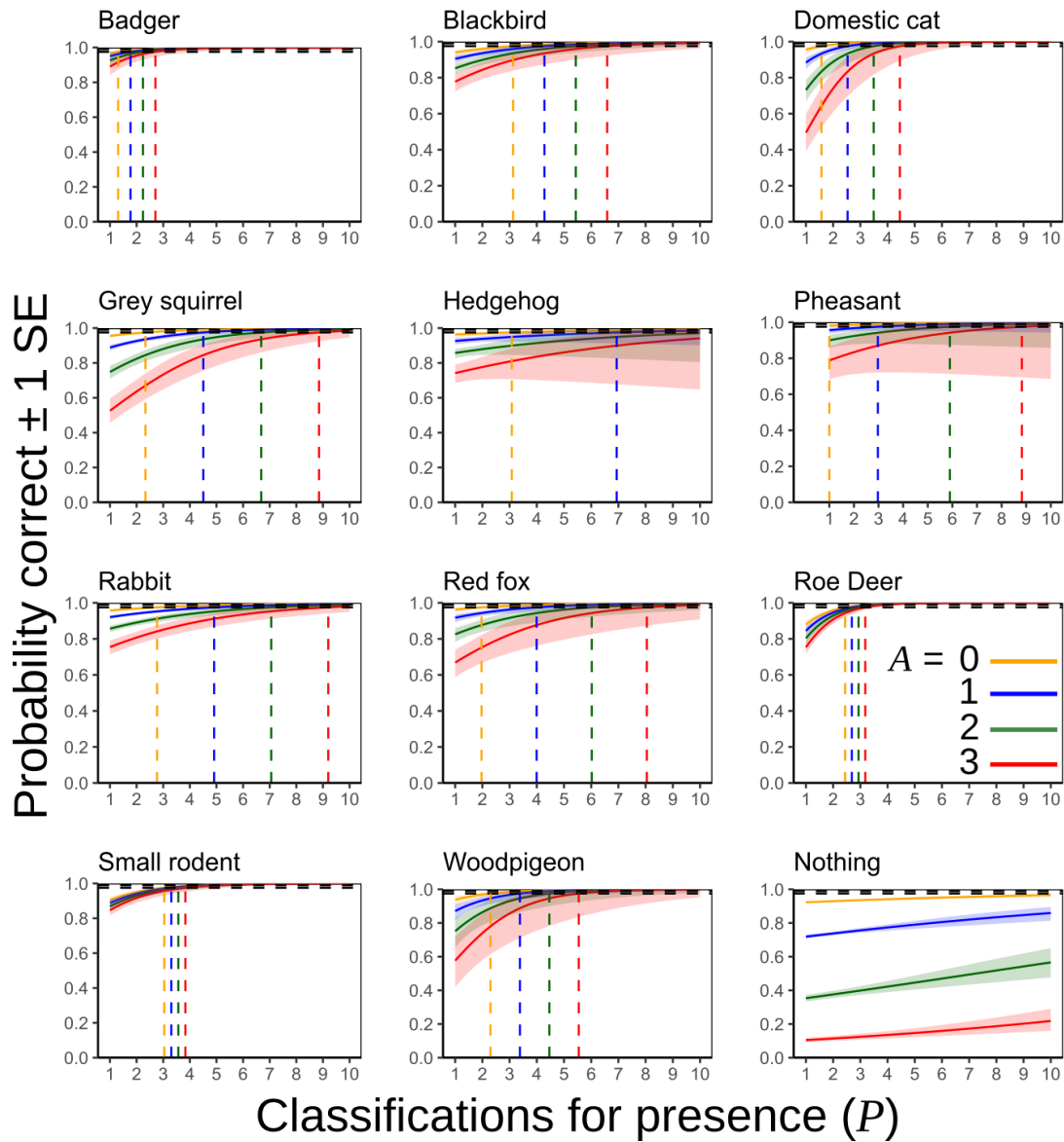


Figure 3



555

Figure 4

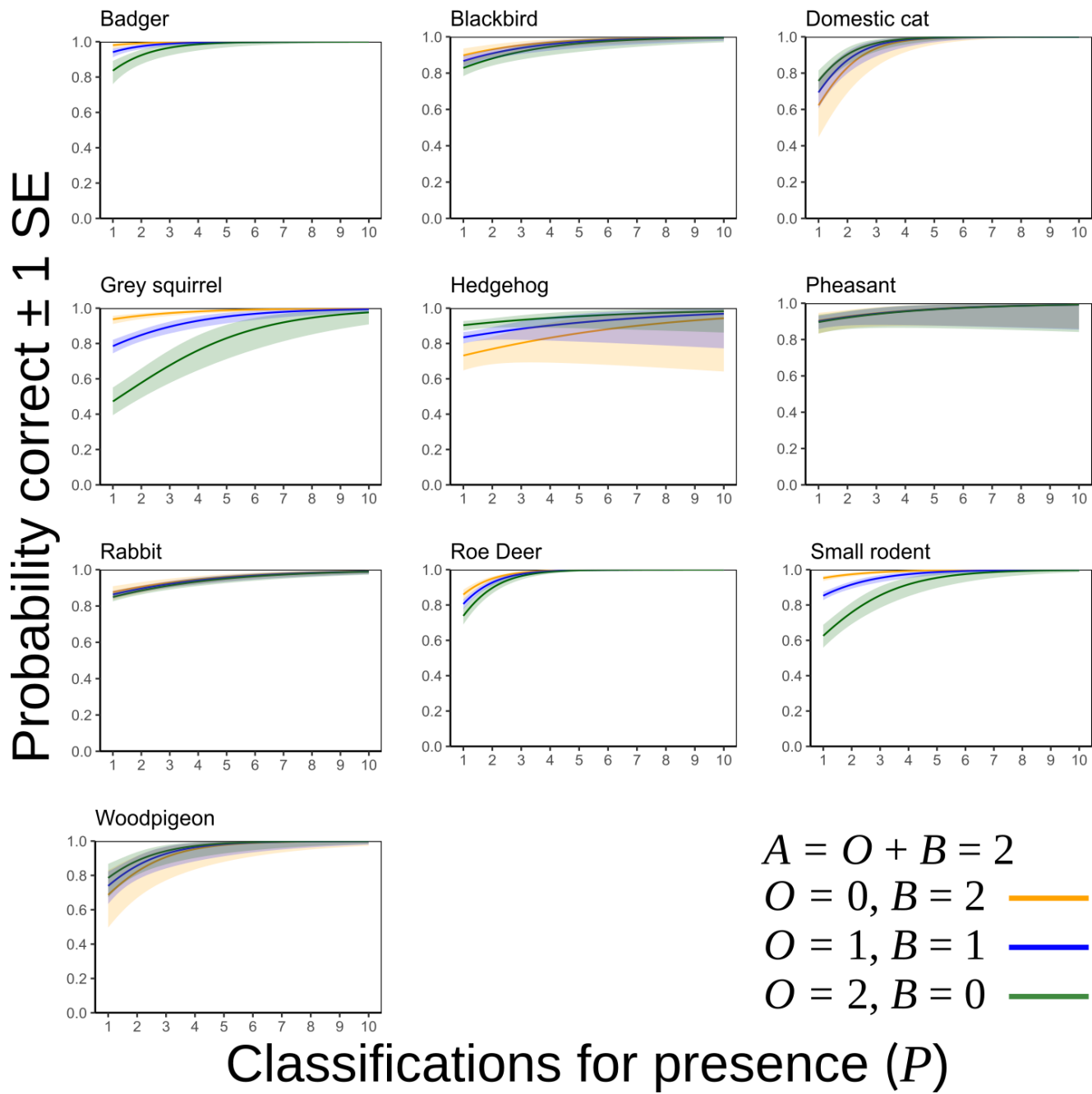


Figure 5