ORIGINAL RESEARCH

Economical crowdsourcing for camera trap image classification

Pen-Yuan Hsing[1] (iD), Steven Bradley[2], Vivien T. Kent[3,4], Russell A. Hill[4], Graham C. Smith[5], Mark J. Whittingham[6], Jim Cokill[3], Derek Crawley[7], MammalWeb volunteers & Philip A. Stephens[1]

[1]Conservation Ecology Group, Department of Biosciences, Durham University, South Road, Durham DH1 3LE, United Kingdom
[2]Department of Computer Science, South Road, Durham DH1 3LE, United Kingdom
[3]Durham Wildlife Trust, Chilton Moor, Houghton-le-Spring, Tyne and Wear DH4 6PU, United Kingdom
[4]Department of Anthropology, Durham University, South Road, Durham DH1 3LE, United Kingdom
[5]National Wildlife Management Centre, Animal and Plant Health Agency, Sand Hutton Campus, Sand Hutton, York YO41 1LZ, United Kingdom
[6]School of Natural and Environmental Sciences, Newcastle University, Newcastle-Upon-Tyne NE1 7RU, United Kingdom
[7]The Mammal Society, 18 St John's Church Road, London E9 6EJ, United Kingdom

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 about nine concordant crowdsourced classifications per sequence. However, there is high variation among species. For highly recognizable 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.

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 (Dickinson et al. 2010; Bonney et al. 2014), 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. 2010, 2012; 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 utilizing 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 3 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 4000 'gold standard' images, classified by experts, to show that consensus classifications typically had an accuracy exceeding 97% (Swanson et al. 2015, 2016).

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.

In 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 (Geider et al. 2001; Gaston and Fuller 2008). The local relevance and lower charisma of these studies might make it harder to mobilize a large international userbase. As a result, it may be necessary to determine image contents with 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 (~550:1) and monitors mammals that are less diverse and may be considered 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 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: misidentification 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 classifications.

Materials and 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 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 organization focused on environmental conservation, education and engagement). Recruiting and training citizen scientists from local community groups such as the Durham Wildlife Trust is comparable 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

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 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 characterize the distribution and skewness of classification intensity by individual Spotters, we calculated the proportions of those who classified fewer than 100 images and greater than 1000 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 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 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 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 – 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 10 483 sequences (35 417 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 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 generalized 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 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': $A_{s,j} = O_{s,j} + B_j$). 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

$$\left(\sum_{j} P_{s,j} \bigg/ \sum_{j} T_j\right)$$

This parameter, which we designate as $\Pr(s)$ (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 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 for the full dataset (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 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 dataset would include non-independent data points (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 1000 random bootstrap samples of the dataset, 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 (bootstrapped) full dataset 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 recognizes the fact that species differ in both their detectability and their recognizability; 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 $V_{s,j} \sim P_{s,j} + O_{s,j} + B_j$ 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 $\Pr(s)$ 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 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:

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

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

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

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 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 3.5.1 (R Core Team, 2017) with the packages dplyr (Wickham et al. 2017), ggplot2 (Wickham 2016), lubridate (Grolemund and Wickham 2011), lme4 (Bates et al. 2015) and EnvStats (Millard 2013).

Results

As of 7 March 2018, MammalWeb citizen scientists had cumulatively deployed camera traps at 261 unique sites in North East England for 15238 camera trap days. This yielded 173315 images uploaded to our website. Since project inception, 265 Spotters (including those who deployed camera traps, i.e. Trappers) had contributed, via the MammalWeb website, 249425 classifications of the content of 115944 images (40709 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 (58.9%) of MammalWeb users ($n = 156$) classified less than 100 photos, whereas 11.3% of the users ($n = 30$) each classified more than 1000 photos (Fig. S3). The top 10% of Spotters ($n = 27$, 15 of whom were Trappers) contributed 84.9% of all classifications.

At the sequence level, 21 species have been classified in our dataset. For most of the species in sequences with a gold standard, >90% of user-provided classifications were correct (Fig. 1A). Badgers (*Meles meles*) were recognized by more than 95% of classifiers and only four 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 (59 out of 66 incorrect classifications; Fig. 1). In 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 (473 out of 494 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
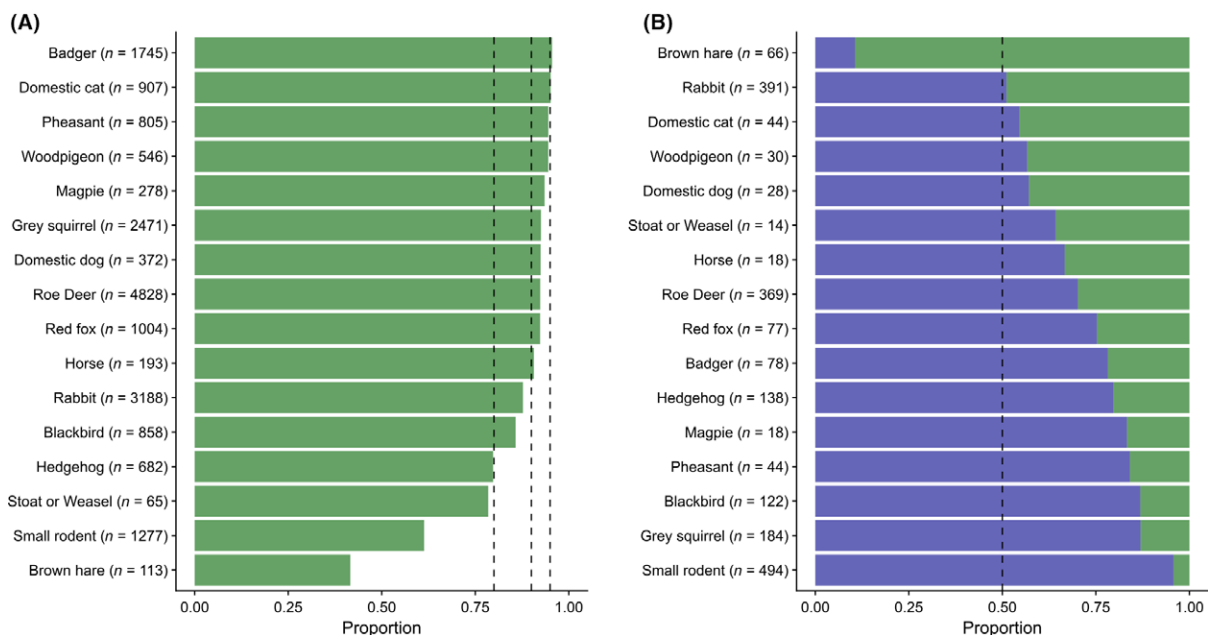

Figure 1. (A) Proportional accuracy of submitted classifications across the whole pool of sequences with gold standard classifications. Sample sizes (n) represent the number of classifications provided for sequences in which the gold standard indicates that the named species is present. Vertical lines 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.

Table 1. Confusion matrix for accuracies of commonly classified species.

Gold standard	Badger (1745)	Blackbird (858)	Domestic cat (907)	Grey squirrel (2471)	Hedgehog (682)	Pheasant (805)	Rabbit (3188)	Red fox (1004)	Roe Deer (4784)	Small rodent (1277)	Nothing (6353)	False positive rate
User classifications												
Badger (1680)	.955				.003		.001	.001	.001		.000	.008
Blackbird (773)		.858	.001	.000		.001	.001		.000	.001	.003	.048
Domestic cat (886)	.001		.951				.001	.007			.001	.026
Grey squirrel (2379)	.001	.005	.001	.926	.003	.004	.012			.002	.004	.039
Hedgehog (578)	.001		.006	.001	.798		.002			.008	.001	.059
Pheasant (773)						.945	.002				.001	.016
Rabbit (2905)	.002	.003		.003	.019	.002	.877	.003	.002	.001	.002	.037
Red fox (968)	.002	.001	.008	.000			.001	.923	.003	.002	.001	.042
Roe Deer (4513)	.002	.002		.000			.003	.004	.932		.003	.012
Small rodent (836)			.001	.001	.016		.003	.001		.613	.004	.063
Nothing (7770)	.035	.124	.026	.065	.161	.046	.063	.058	.054	.370	.975	.203
False negative rate	.045	.142	.049	.074	.202	.055	.123	.077	.068	.387	.025	

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. E.g. For badgers, there are 1680 user classifications indicating their presence of which 0.8% 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).
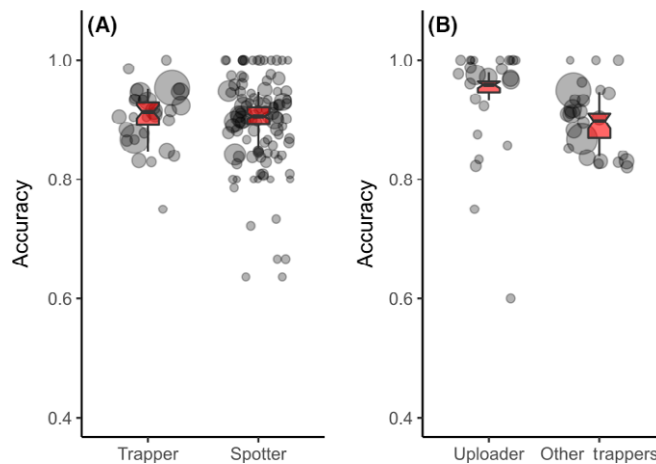

Figure 2. Of the citizen scientists who classified at least 10 sequences, (A) those who deployed camera traps (30 'Trappers', 13446 classifications) were marginally more accurate at image classification than those who did not (102 'Spotters', 12100 classifications) but this effect was not supported (ΔAIC = −1.49, model weight = 0.32, relative to a model that did not account for the Spotter type). (B) There was strong support for the finding that 26 Trappers who classified images they uploaded ('Uploaders', 2578 classifications) were more accurate than Trappers who classified images uploaded by other Trappers ('Other Trappers', 10136 classifications) (ΔAIC = 66.28, 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 summarize 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).

the likelihood that it is there, whereas 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 (ΔAIC = 196.74, SD = 19.99). Consequently,

we also analysed the relationship between image contents and numbers of classifications for individual species. Twelve species (including 'nothing', or blank (B), that is 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,
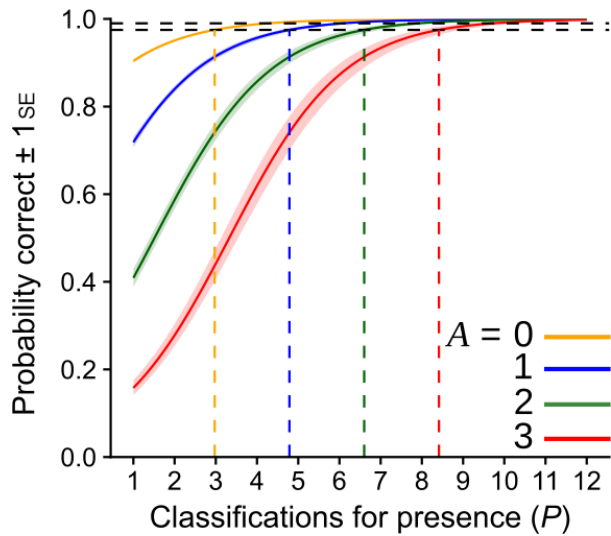
Figure 3. Global-level relationship between the number of classifications for the presence (*P*) and absence (*A*) of a given species in a sequence and the probability that it is indeed in the sequence. Solid lines show the mean relationship (over 1000 bootstrapped samples) between the probability (predicted 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 ± 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%.

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 difference (Fig. 4). Other species, such as badgers, 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 5 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 eight species, doing so produced a better-supported model

(Table S1). Coefficient values suggest the relative reduction in confidence resulting from classifications for no animals (B) and those for other species (O) (Fig. S5). Classifications for other species (O) have a particularly strong effect on confidence for badgers, red foxes, and domestic cats (Fig. 5 and Fig. S5).

Globally (without regard to specific species), 42.9% of sequences can be retired with 97.5% confidence after four classifications and a further 21.4% of sequences could be retired after seven (Table S2). At the 99% confidence level, 34.7% of sequences can be retired after five classifications (Table S2). The implication of these analyses is that, on average, 7.2 classifications would be needed per sequence to retire them with 97.5% confidence, while an average of 9.1 classifications are required for 99% confidence. If algorithms for sequence retirement are sensitive to the species most likely to be pictured, 88.1% or more of sequences containing highly recognizable species, such as badgers, could be retired after just two classifications (with 97.5% confidence) (Table S3). However, less recognizable species would need many more classifications to instil confidence (Table S3). For example, only about 85% of sequences classified as small rodents can be retired at 97.5% confidence even after six classifications (Table S3).

Discussion

There is a trend for citizen science projects to crowd-source data classification. The question of how proliferating projects can obtain confident classifications from a finite group of contributors suggests that more economic ways of utilizing 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 and their implications for crowd-sourced 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 crowd-sourcing efforts (Sauermann and Franzoni 2015). That the top 10% of MammalWeb classifiers ('Spotters')
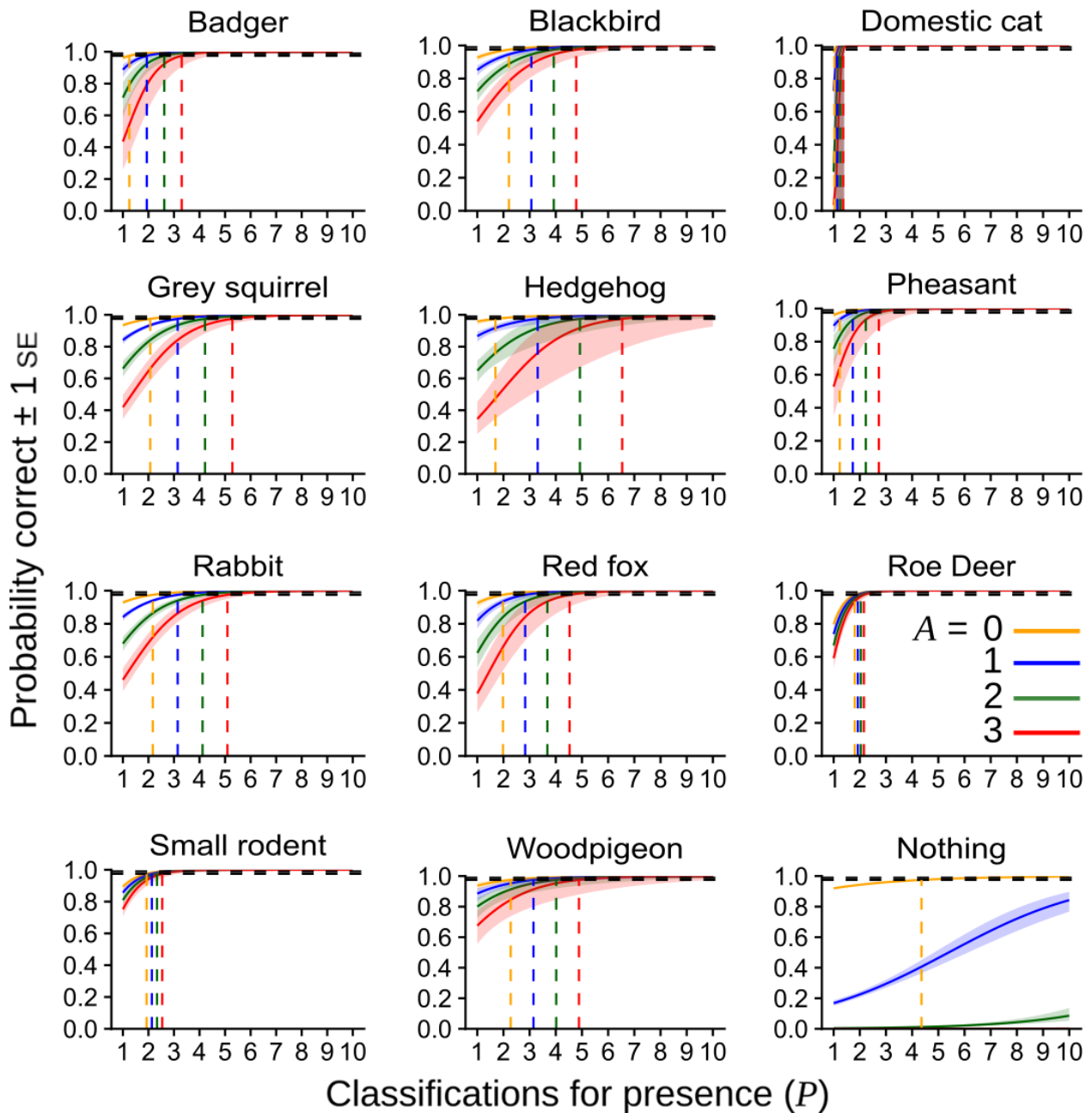
Figure 4. Species-level relationship between the number of classifications indicating the presence (*P*) and absence (*A*) of a given species, and the probability that it appears in a sequence. Solid lines show the mean relationship between the probability (predicted by the fitted model) that a species 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%.

contributed 84.9% 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
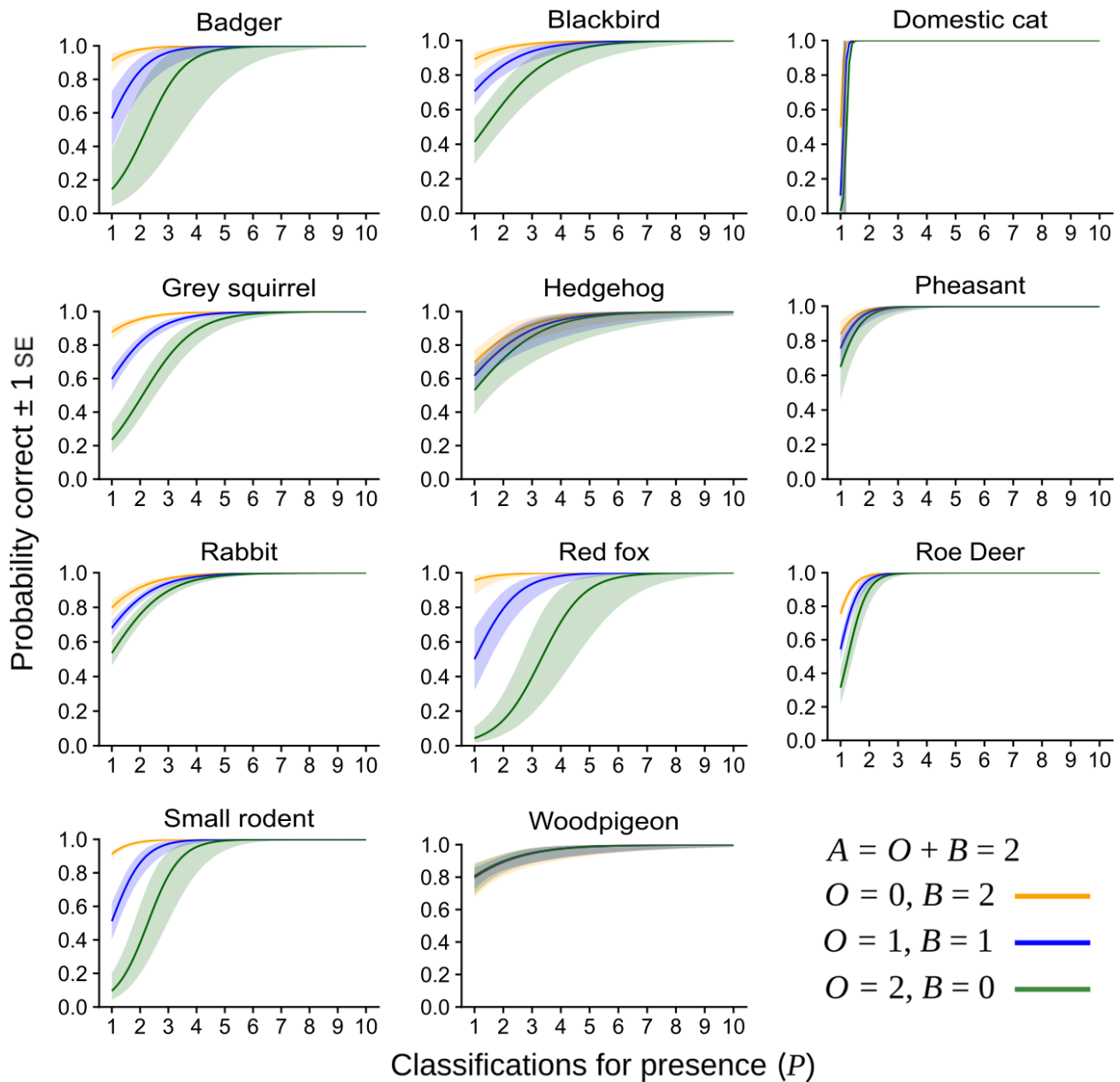
Figure 5. Implications of distinguishing between different types of classifications indicating that a species is absent (*A*). For some typically highly detectable species, such as the badger, classifications suggesting that no animal is present in the sequence ('false negatives', *B*) are more damaging to confidence than are classifications suggesting that the pictured species is some other species ('false positives', *O*). For visually 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.

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 organization (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 scrutinized 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 like-lihood 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. Specifi-cally, without an algorithm that distinguishes between species, sequences in our dataset can be retired from the classification pool after an average of 7.2 classifications (for an accuracy of ≥97.5%) or 9.1 classifications (for ≥99% accuracy) (Table S2). Given that there is some evi-dence that different types of classifications against the presence of a species may carry different weight (and, in particular, that classifications for the absence of any spe-cies of interest are generally less damaging to confidence than classifications 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 S3). Species-level differences were also evident when differentiating the impacts from misidentification (i.e. the false-positive identification of a species) or mistakenly stating that no animal was present (i.e. false negative) (Fig. 5, Fig. S5 and Table 1). A good example of the complications around false positives is given by brown hares. We found that brown hares are rel-atively poorly recognized in our dataset. 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 four 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 over-looked if apparent rabbit sequences are retired rapidly. More data would be required to assess this problem, espe-cially in relation to the specific probability with which hares are classified as rabbits (and the resultant probabil-ity 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 usefulness of a species-specific approach to doing so rather than applying a single algorithm to the entire dataset. An addi-tional 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.

One finding that might be very general to crowd-sourced 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, five or more uncontested classifications suggesting that a sequence is devoid of animals is needed 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 three uncontested classifications to give high confi-dence that they are actually present. As we noted above, more efficient algorithms for crowdsourcing reliable clas-sifications should probably discriminate between the weight attributed to disagreements over whether a species is present and disagreements over the identity of a pic-tured 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 infor-mative ecological analyses. In particular, our analyses sug-gest that, without distinguishing species, at least four or an average of 7.2 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 1324 per week; this yields a ratio of approximately 3.6 classifi-cations per sequence. This suggests that one option to ensure that classifications keep pace with accumulating image data is to increase our classifier pool by a factor of approximately 2.5, relative to the number of camera trap-pers. At present, we have approximately 3.5 classifiers to every trapper, so this would need to increase to approxi-mately 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).

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 recognizable species in our dataset are also some of the more charismatic. If

these sequences are removed more rapidly than others, the dataset could rapidly become biased towards less charismatic species, more indistinct photos and images devoid of animals. Preliminary evidence from 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). In contrast, MammalWeb participants 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 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 (31.4% according to the gold standard) of blank sequences in our dataset (and many other camera trap datasets), it is clear that the 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 dataset 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).

One further possibility for overcoming limitations to classification effort is to use the dataset 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 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) (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, 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.

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. 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 generalize 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 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. 1), especially from those who do the camera trapping (Fig. 2). 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 four 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 (Rowcliffe et al. 2016; Howe et al. 2017). 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 dataset useful for training machine learning algorithms that are under development (Thom 2017; Norouzzadeh et al. 2018).

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 dataset are highly likely to generalize 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.

Acknowledgments

Data Accessibility

The data (crowdsourced classifications and gold standard photos) used in this article are shared under the Creative Commons Attribution-ShareAlike 4.0 license in this repository (DOI: 10.17605/OSF.IO/ZNM6K): https://osf.io/znm6k/

The code used for analysing consensus classifications and producing the relevant figures and tables is shared under the GNU GPLv3+ license in this git repository: https://gitlab.com/penyuan/consensus_classifications_MammalWeb/

References

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **19**, 716–723.

Bates, D., M. Mächler, B. Bolker, and S. Walker. 2015. Fitting linear mixed-effects models using **lme4**. *J. Stat. Softw.* **67**, 1–48.

Battersby, J. E., and J. J. Greenwood. 2004. Monitoring terrestrial mammals in the UK: past, present and future, using lessons from the bird world. *Mammal Rev.* **34**, 3–29.

Bonney, R., J. L. Shirk, T. B. Phillips, A. Wiggins, H. L. Ballard, A. J. Miller-Rushing, et al. 2014. Next steps for citizen science. *Science* **343**, 1436–1437.

Bower, A., V. Maidel, C. Lintott, A. Swanson, and G. Miller. 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 Diego, California, United States.

Cohn, J. P. 2008. Citizen science: can volunteers do real research? *Bioscience* **58**, 192–197.

Conrad, C. C., and K. G. Hilchey. 2011. A review of citizen science and community-based environmental monitoring: issues and opportunities. *Environ. Monit. Assess.* **176**, 273–291.

Croft, S., A. L. M. Chauvenet, and G. C. Smith. 2017. A systematic approach to estimate the distribution and total abundance of British mammals. *PLoS ONE* **12**, e0176339.

Danielsen, F., P. M. Jensen, N.D. Burgess, R. Altamirano, P. A. Alviola, H. Andrianandrasana, et al. 2014. A multicountry assessment of tropical resource monitoring by local communities. *Bioscience* **64**, 236–251.

Dickinson, J. L., B. Zuckerberg, and D. N. Bonter. 2010. Citizen science as an ecological research tool: challenges and benefits. *Annu. Rev. Ecol. Evol. Syst.* **41**, 149–172.

Dickinson, J. L., J. Shirk, D. Bonter, R. Bonney, R. L. Crain, J. Martin, et al. 2012. The current state of citizen science as a

tool for ecological research and public engagement. *Front. Ecol. Environ.* **10**, 291–297.

Eveleigh, A., C. Jennett, A. Blandford, P. Brohan, and A. L. Cox. 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

Everett, G., and H. Geoghegan. 2016. Initiating and continuing participation in citizen science for natural history. *BMC Ecol.* **16**, 15–22.

Forrester, T. D., M. Baker, R. Costello, R. Kays, A. W. Parsons, and W. J. McShea. 2017. Creating advocates for mammal conservation through citizen science. *Biol. Conserv.* **208**, 98–105.

Gaston, K. J., and R. A. Fuller. 2008. Commonness, population depletion and conservation biology. *Trends Ecol. Evol.* **23**, 14–19.

Geider, R.J., E. H. Delucia, P.G. Falkowski, A. C. Finzi, J. P. Grime, J. Grace, et al. 2001. Primary productivity of planet earth: biological 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.

Greenwood, J. J. D. 2007. Citizens, science and bird conservation. *J. Ornithol.* **148**, 77–124.

Grolemund, G., and H. Wickham. 2011. Dates and times made easy with lubridate. *J. Stat. Softw.* **40**, 1–25.

Haklay, M. 2013. Citizen science and volunteered geographic information: overview and typology of participation. Pp. 105–122 in D. Sui, S. Elwood, M. Goodchild, eds. *Crowdsourcing Geographic Knowledge*. Springer, Netherlands.

Harris, S. J., D. Massimino, S. E. Newson, M. A. Eaton, J. H. Marchant, D. E. Balmer, et al. 2016. *The breeding bird survey 2015 (No. 687)*. British Trust for Ornithology, Thetford, United Kingdom.

Hennon, C. C., K. R. Knapp, C. J. Schreck, S. E. Stevens, J. P. Kossin, P. W. Thorne, et al. 2014. Cyclone center: can citizen scientists improve tropical cyclone intensity records? *Bull. Am. Meteorol. Soc.* **96**, 591–607.

Howe, E. J., S. T. Buckland, M.-L. Després-Einspenner, and H. S. Kühl. 2017. Distance sampling with camera traps. *Methods Ecol. Evol.* **8**, 1558–1565.

Jennett, C., L. Kloetzer, D. Schneider, I. Iacovides, A. Cox, M. Gold, et al. 2016. Motivations, learning and creativity in online citizen science. *J. Sci. Commun.* **15**, 1–23.

Kosmala, M., A. Wiggins, A. Swanson, and B. Simmons. 2016. Assessing data quality in citizen science. *Front. Ecol. Environ.* **14**, 551–560.

Lorimer, J. 2007. Nonhuman Charisma. *Environ. Plan. Soc. Space* **25**, 911–932.

Lukyanenko, R., J. Parsons, and Y. F. Wiersma. 2016. Emerging problems of data quality in citizen science. *Conserv. Biol.* **30**, 447–449.

McShea, W. J., T. Forrester, R. Costello, Z. He, and R. Kays. 2015. Volunteer-run cameras as distributed sensors for macrosystem mammal research. *Landsc. Ecol.* **31**, 55–66.

Millard, S. P. 2013. *EnvStats - An R package for environmental statistics, 2nd ed.* Springer-Verlag, New York, New York.

Newman, G., A. Wiggins, A. Crall, E. Graham, S. Newman, and K. Crowston. 2012. The future of citizen science: emerging technologies and shifting paradigms. *Front. Ecol. Environ.* **10**, 298–304.

Norouzzadeh, M. S., A. Nguyen, M. Kosmala, A. Swanson, M. S. Palmer, C. Packer, and J. Clune. 2018. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences* 201719367.

Price Tack, J. L., B. S. West, C. P. McGowan, S. S. Ditchkoff, S. J. Reeves, A. C. Keever, et al. 2016. AnimalFinder: a semi-automated system for animal detection in time-lapse camera trap images. *Ecol. Inform.* **36**, 145–151.

R Core Team, 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Radke, R. J., S. Andra, O. Al-Kofahi, and B. Roysam. 2005. Image change detection algorithms: a systematic survey. *Trans. Image Process.* **14**, 294–307.

Ratcliff, J. 2008. *The transit of Venus enterprise in Victorian Britain*. Pickering & Chatto, London, United Kingdom.

Rowcliffe, J. M., and C. Carbone. 2008. Surveys using camera traps: are we looking to a brighter future? *Anim. Conserv.* **11**, 185–186.

Rowcliffe, J. M., P. A. Jansen, R. Kays, B. Kranstauber, and C. Carbone. 2016. Wildlife speed cameras: measuring animal travel speed and day range using camera traps. *Remote Sens. Ecol. Conserv.* **2**, 84–94.

Sauermann, H., and C. Franzoni. 2015. Crowd science user contribution patterns and their implications. *Proc. Natl Acad. Sci.* **201408907**, https://doi.org/10.1073/pnas.1408907112.

Siddharthan, A., C. Lambin, A.-M. Robinson, N. Sharma, R. Comont, E. O'mahony, et al. 2016. Crowdsourcing without a crowd: reliable online species identification using Bayesian models to minimize crowd size. *ACM Trans. Intell. Syst. Technol.* **7**, 1–20.

Silvertown, J., M. Harvey, R. Greenwood, M. Dodd, J. Rosewell, T. Rebelo, et al. 2015. Crowdsourcing the identification of organisms: a case-study of iSpot. *ZooKeys* **480**, 125–146.

Simpson, R., K. R. Page, and D. De Roure. 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

Steenweg, R., M. Hebblewhite, R. Kays, J. Ahumada, J. T. Fisher, C. Burton, et al. 2017. Scaling-up camera traps: monitoring the planet's biodiversity with networks of remote sensors. *Front. Ecol. Environ.* **15**, 26–34.

Stephens, P. A., N. Pettorelli, J. Barlow, M. J. Whittingham, and M. W. Cadotte. 2015. Management by proxy? The use of indices in applied ecology. *J. Appl. Ecol.* **52**, 1–6.

Swanson, A., M. Kosmala, C. Lintott, R. Simpson, A. Smith, and C. Packer. 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.

Swanson, A., M. Kosmala, C. Lintott, and C. Packer. 2016. A generalized approach for producing, quantifying, and validating citizen science data from wildlife images. *Conserv. Biol.* **30**, 520–531.

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 Artic University of Norway.

Verma, A., R. van der Wal, and A. Fischer. 2016. Imagining wildlife: new technologies and animal censuses, maps and museums. *Geoforum* **75**, 75–86.

van der Wal, R., N. Sharma, C. Mellish, A. Robinson, and A. Siddharthan. 2016. The role of automated feedback in training and retaining biological recorders for citizen science. *Conserv. Biol.* **30**, 550–561.

Wald, D. M., J. Longo, and A. R. Dobell. 2016. Design principles for engaging and retaining virtual citizen scientists. *Conserv. Biol.* **30**, 562–570.

Wickham, H. 2016. *ggplot2 - elegant graphics for data analysis, 2nd ed, Use R!* Springer International Publishing, Switzerland.

Wickham, H., R. Francois, L. Henry, and K. Müller. 2017. dplyr: A Grammar of Data Manipulation.

Willett, K. W., C. J. Lintott, S. P. Bamford, K. L. Masters, B. D. Simmons, K. R. V. Casteels, et al. 2013. Galaxy Zoo 2: detailed morphological classifications for 304122 galaxies from the Sloan Digital Sky Survey. *Mon. Not. R. Astron. Soc.* **435**, 2835–2860.

Yu, X., J. Wang, R. Kays, P. A. Jansen, T. Wang, and T. Huang. 2013. Automated identification of animal species in camera trap images. *EURASIP J. Image Video Process.* **2013**, 52.

Supporting Information

Additional supporting information may be found online in the supporting information tab for this article.

Figure S1. A sequence of camera trap images taken in burst mode of a red fox (*Vulpes vulpes*). When shown in isolation, the left-hand and middle images in this sequence might achieve high levels of consensus regarding their content. By contrast, the right-hand image would be hard to classify and might be subject to considerable uncertainty regarding its focal subject.

Figure S2. MammalWeb camera trap image classification ('Spotter') interface.

Figure S3. The majority of classification effort was contributed by relatively few users.

Figure S4. Relationship between classification confidence and the number of classifications for the presence (P) and absence (A) of certain species, with the classifications for absence split into those for other species (O) and blank (i.e. containing no vertebrates) (B).

Figure S5. Coefficient values ($\pm$ mean SE) for models that distinguish between the effects on classification confidence of those for 'other species' (O) and 'blank' (B).

Table S1. Impact of separating classifications for absence (A) model term into those for other species (O) and blank (B). Positive ΔAICs (bold font) indicate that increasing the number of parameters by having separate O and B terms is justified by the improved model fit.

Table S2. Calculations for numbers of sequence-level classifications needed (CN) to achieve target confidence level across the global pool of images.

Table S3. Calculations numbers of sequence-level classifications needed (CN) to achieve target confidence level for commonly pictured species.