Which camera-trap images to review? Estimating occupancy from a small part of a survey

Michael Tabak[1,2]

[1]Quantitative Science Consulting, LLC, Negaunee, Michigan, USA
[2]Biology Department, Northern Michigan University, Marquette, Michigan, USA

Correspondence: tabak@qsc.earth; ORCID 0000-0002-2986-7885

Abstract

Context. Camera-trap surveys produce more images than people can review. Computer-vision classifiers such as SpeciesNet label animals but miss some species, so people still review images.

Aims. To find how many and which images should be reviewed to reliably estimate species occupancy, and how a project can tell when it has reviewed enough.

Methods. On four fully labeled camera-trap surveys (in the USA, South Africa, and Malawi), I simulated partial reviews by choosing capture events in five ways: at random, with animals, spread across cameras, least certain first (as in active learning), and by confirming the suggestions an occupancy model counts. I then tested a two-step procedure for human-in-the-loop review: review two capture events per camera; then, in rounds, confirm the most confident suggestion of each unconfirmed species-site-week (a species at a site in a week), retraining a model of the project's species after each round.

Key results. SpeciesNet's suggestions alone gave occupancy within a mean absolute error (MAE) of 0.06 to 0.12 of a full review's where it knew the fauna, but missed Canada lynx and nyala; in Malawi, the MAE was 0.24. Reviewing 25% of images at random gave an MAE of 0.16 to 0.24 with only reviewed images counted as detections. Confirming suggestions came closest; reviewing the least certain first did little better than random. In the procedure, each capture event reviewed confirmed 0.6 to 1.0 new species-site-weeks until 92 to 98% were confirmed, then fewer than 0.2. Stopping there, after 15 to 27% of the images (1.1 to 1.8 capture events per species-site-week), occupancy, counting only reviewed images, had an MAE of 0.04 or less, lynx included.

Conclusions. Confirming suggestions approaches a full review's estimates because occupancy models need species-site-weeks, not images; reviewing the least certain images first spends effort where estimates gain least.

Implications. Projects can nearly match a full review's occupancy estimates by reviewing two capture events per camera, then confirming one suggestion per species-site-week until a round confirms fewer than one new species-site-week per five capture events. Researchers can expect to review about 1.5 capture events per species-site-week that the first step shows.

Keywords: active learning; camera traps; capture events; computer vision; human-in-the-loop review; occupancy; partial review; SpeciesNet

1 Introduction

Occupancy modeling, which estimates the share of sites a species occupies while accounting for imperfect detection, is a common analysis of camera-trap surveys (MacKenzie *et al.*, 2002; Burton *et al.*, 2015). A survey of a few dozen cameras over a few months can produce tens of thousands of images; these images must be labeled before they can become detection histories. Camera-trap users rank this among their greatest needs (Glover-Kapfer *et al.*, 2019).

Machine learning can do much of the work. Object detectors such as MegaDetector set aside the empty images that often make up most of a survey (Beery *et al.*, 2019; Fennell *et al.*, 2022; Hernandez *et al.*, 2024), and species classifiers (Norouzzadeh *et al.*, 2018; Tabak *et al.*, 2019; Willi *et al.*, 2019), such as SpeciesNet, suggest a species for each animal (Gadot *et al.*, 2024). Where a classifier knows a survey's fauna, occupancy models fitted to its labels with no human review can come close to those fitted to experts' labels (Whytock *et al.*, 2021; Thornton *et al.*, 2026). But classifiers lose accuracy at new locations and on unfamiliar species (Beery *et al.*, 2018, 2019; Schneider *et al.*, 2020; Tabak *et al.*, 2020), and occupancy from their labels differs most for rarely detected species (Thornton *et al.*, 2026). A project cannot tell how far to trust its classifier without reviewing some of its images, so human review remains standard practice (Thornton *et al.*, 2026). The question is how much to review.

Reviewing part of a survey is not a simple fix. If the images not yet reviewed count as non-detections, detection probability falls as less of the survey is reviewed. Occupancy models allow for imperfect detection, but with few detections per site their estimates are imprecise, often at their limits (MacKenzie & Royle, 2005; Guillera-Arroita *et al.*, 2010; Specht *et al.*, 2017; Kays *et al.*, 2020). Counting unreviewed machine labels as detections instead adds false positives, which bias occupancy high, mostly for rare and look-alike species (Royle & Link, 2006; Miller *et al.*, 2011; Wright *et al.*, 2020). Occupancy models that allow for false positives can combine confirmed and unconfirmed detections (Miller *et al.*, 2011; Chambert *et al.*, 2015; Clare *et al.*, 2019). For automated acoustic detections, confirming as little as 1% of them gave accurate estimates (Chambert *et al.*, 2018). Yet these models need enough confirmed detections to distinguish a species that is present but seldom confirmed from a false positive (Clement, 2016; Guillera-Arroita *et al.*, 2017).

Two things about occupancy suggest a better way to spend a review. First, an occupancy model needs only whether each species was detected at each site in each sampling occasion, not every image that shows it: a survey of 60 cameras over eight weeks may have tens of thousands of images but only a few hundred to a few thousand such detections. Second, a classifier's suggestions can point to those detections. A review that confirms one suggestion for each species at each site in each occasion could therefore give nearly the detection history of a full review at a small part of its cost. A classifier trained

on the reviewer's own labels, which learns the project's species and names (Norouzzadeh *et al.*, 2021; Tabak, 2026), can point better as the review goes on.

Systems that learn while people label often choose the images they are least sure of (Norouzzadeh *et al.*, 2021; Kellenberger *et al.*, 2020), which helps a classifier learn but need not find what an occupancy model needs. Haucke *et al.* (2026) recently showed that choosing reviews for the occupancy model is valuable. Their method, ACORN, treats each classifier score as evidence in a continuous-score occupancy model (Rhinehart *et al.*, 2022) and asks an expert to verify, one species at a time, the camera-days whose labels are expected to tell the most about that species' occupancy model. On iWildCam images with SpeciesNet's scores, it came close to a fully labeled survey with fewer reviews than random, highest-score, or uncertainty-based review. However, ACORN refits a Bayesian model of each species after each batch of reviews and asks about one species on one camera-day at a time, whereas people review images, naming every species in them at once.

Here I compare ways of choosing which images to review for occupancy models, on four camera-trap surveys in which people labeled every image. I propose and test a simple, two-step review procedure with a rule for when to stop. It reviews images for every species at once and needs no occupancy model while reviewing. Because every image already has a label, I could simulate any review: reviewing an image means saving its label from the published dataset, and the review of every image in the dataset is the standard to compare with. The surveys differ in how well SpeciesNet knows them. It was probably trained on one; of the other three, it knows the fauna of two, in North America and southern Africa, and often names the species of the third, in Malawi, as related species from elsewhere.

I ask three questions. First, how much does occupancy from a partial review differ from that of a full review, with and without suggestions for the images not reviewed, when the images to review are drawn at random? Second, does choosing the images differently (images with animals, spread across cameras, the least certain suggestions first, or the suggestions that the occupancy model counts) bring the estimates closer for the same effort? Third, how much review does the procedure need, and how can a project tell, from its own review, when it has reviewed enough?

2 Materials and methods

2.1 Software and models

I processed the images in WildObserve (version 0.7.6; Tabak, 2026), a desktop application for reviewing camera-trap images and fitting occupancy models. Briefly, MegaDetector v6 (compact variant; Hernandez *et al.*, 2024) finds animals, people, and vehicles in each image (by default, boxes scoring ≥ 0.5). For each animal box, the SpeciesNet v4.0.3a crop classifier (Gadot *et al.*, 2024) computes a fingerprint, the 1,280 numbers from its

final pooling layer, and SpeciesNet's own final layer turns it into scores for its 2,498 labels. Images taken no more than 60 s apart at one camera form a capture event, which the review shows together.

As a reviewer labels images in the software, WildObserve trains a *project model*, a multinomial logistic regression on the fingerprints of the reviewed boxes, which names only the project's own species. It retrains in about a second as labels are added and scores every box in the survey again without re-reading any image. On images that SpeciesNet had not seen, ten labels per species made it more accurate than SpeciesNet (Tabak, 2026). I ran the experiments with open code on the data that WildObserve exported (Data availability statement), fitting the same model with scikit-learn (Pedregosa *et al.*, 2011). In the simulations, the project model was retrained on all the reviews so far after the procedure's first step and after each round (below), and once more before its suggestions were counted.

From the reviewed images, and optionally the suggestions on images not yet reviewed, WildObserve builds detection histories and fits single-season occupancy models (MacKenzie *et al.*, 2002) by maximum likelihood, as well as occupancy models that allow for false positives among the suggestions (Miller *et al.*, 2011); its fits agree with those of the R package unmarked (Fiske & Chandler, 2011; Kellner *et al.*, 2023) to about four significant figures (Tabak, 2026). The open code fits the same models with SciPy (Virtanen *et al.*, 2020), and on the four full reviews its estimates agreed with WildObserve's to within 0.0001. A site is a camera's location, and each survey is divided into seven-day occasions from its first day; a site's effort in an occasion is the number of days its camera ran. A species is detected in an occasion if a reviewed image shows it, or, when suggestions count, if the top suggestion for an animal box on an image not yet reviewed (the project model's, once there is one) names it with a score ≥ 0.9. Only species on the project's species list count as suggested. A species detected at a site in an occasion is called a *species-site-week*.

2.2 Surveys

I used four camera-trap surveys in which people had labeled every image (Table 1). Each camera's images formed one site. All four datasets are distributed by LILA BC (`https://lila.science/datasets`).

Caltech Camera Traps (CCT; Beery *et al.*, 2018): every image from the 60 camera locations in the southwestern United States that ran in the eight weeks from 1 August 2014, the eight weeks in which the most cameras ran. SpeciesNet was trained partly on the LILA BC datasets released up to Orinoquía Camera Traps in January 2022 (D. Morris, Google, pers. comm., 2026), which probably include CCT (released in 2018), so this survey is a best case. CCT's categories are broad (any deer is "deer"), so confusing two species within a category costs nothing.

WSU Lynx (Thornton *et al.*, 2026): 637 camera locations in north-central and north-eastern Washington, USA, in grid cells of two to nine cameras at least 1 km apart, run in the snow-off months of 2016 and 2017 and labeled by sequence in 26 categories; images of people and vehicles were removed before release. I drew 60 cameras at random from the 103 that ran throughout the eight weeks from 10 July 2016, the eight weeks of 2016 in which the most cameras ran, and took every image they recorded in that time. SpeciesNet was not trained on WSU Lynx (D. Morris, pers. comm.), but its fauna is North American, which SpeciesNet knows well, and it is the Washington survey on which Thornton *et al.* (2026) found occupancy from SpeciesNet's labels similar to the experts'. Sequences that its labelers could not identify were reviewed as animals without a species.

Associated Private Nature Reserves (APNR): Snapshot Safari's cameras in the private reserves bordering Kruger National Park, South Africa, from the Snapshot Safari 2024 Expansion (Pardo *et al.*, 2021), labeled by sequence by Zooniverse volunteers. SpeciesNet was not trained on APNR's images (D. Morris, pers. comm.), but it knows the region's animals from Snapshot Serengeti and from the first seasons of six other Expansion projects, among them Snapshot Kruger, which LILA BC had also released separately. The metadata has no capture times, so each image's capture time came from its EXIF data. From the fourth season, I took every image from the 55 cameras that ran throughout the eight weeks from 30 November 2019, when the most cameras ran; 52 of them took images in that time. I named the categories as a reviewer there would (e.g., "baboon" as chacma baboon, the only baboon there), and merged rhinoceroses into one category, since none was labeled white or black.

Nkhotakota Camera Traps (Appel *et al.*, 2025): 324,635 images from 69 camera locations in Nkhotakota Wildlife Reserve, Malawi, labeled by image. SpeciesNet was not trained on it (D. Morris, pers. comm.). I took every image from the eight weeks from 22 September 2020, when the most cameras ran: 63 cameras, 60 of which ran throughout.

Images from all four surveys were reduced to 1,024 pixels wide, with their capture times written into their metadata, and processed in WildObserve with its defaults. For effort, I took each camera as running from its first image to its last (at Nkhotakota, from the first to the last image on its memory cards).

2.3 Full and partial reviews

No one reviewed images for this study; every review was simulated. Throughout, *reviewing* an image means saving its label from the published dataset into a copy of the processed survey, so the simulated reviewer makes no mistakes beyond those already in the datasets' labels. I saved labels as follows. On an image of one species, every animal box that the detector found scoring ≥ 0.5 was named with that species, and the image's count of animals was spread over them. An image without animals was saved as empty. An image

Table 1: The four surveys. Each covered eight weeks; a site is a camera. *Species-site-weeks*: the detections that the full review gives an occupancy model (each species detected at a site in a seven-day occasion). *Species*: categories detected in the full review, and those scored (detected at three or more sites, with a 95% interval for occupancy no wider than 0.5).

	CCT	WSU Lynx	APNR	Nkhotakota
Region	SW USA	Washington	South Africa	Malawi
Weeks	Aug–Sep 2014	Jul–Sep 2016	Dec 2019–Jan 2020	Sep–Nov 2020
Cameras (sites)	60	60	52	63
Images	14,937	50,848	13,276	41,009
Capture events	9,579	12,703	5,528	19,863
Images with animals (%)	30	33	90	32
Species-site-weeks	385	1,144	1,269	1,556
Species detected (scored)	16 (10)	20 (13)	32 (18)	39 (20)
SpeciesNet trained on it	probably	no	no	no

showing several species, or a species on which the detector put no box, received a box over the whole frame for each species and was marked uncertain, which keeps it out of the project model's training but not out of the occupancy tables. Saving every image this way gave the *full review*, the standard with which every other estimate was compared. Every partial review took whole capture events, as people review bursts of images.

2.4 Ways of choosing the images

First, I reviewed 2, 5, 10, and 25% of each survey's images (and, at random, also 50%), five times for each share, choosing capture events by one of five strategies until the share was reached:

1. *Random*: capture events drawn at random from all of them.
2. *Animals*: capture events drawn at random from those in which the detector found an animal (a box scoring ≥ 0.5).
3. *By camera*: from those, one capture event from each camera in turn, in random order.
4. *Close calls*: in five equal rounds, the first drawn at random from the capture events with animals; after each round the project model was retrained, and the next took the capture events whose suggestions were least certain first, as WildObserve's *Close calls* list orders them (the smallest lead of any animal box's top suggestion over the runner-up, or zero when SpeciesNet and the project model disagreed). This is the uncertainty sampling of active learning.
5. *Confirm*: in five equal rounds, the first as in *Close calls*, with the project model retrained after each. Each later round took, in random order, the capture event

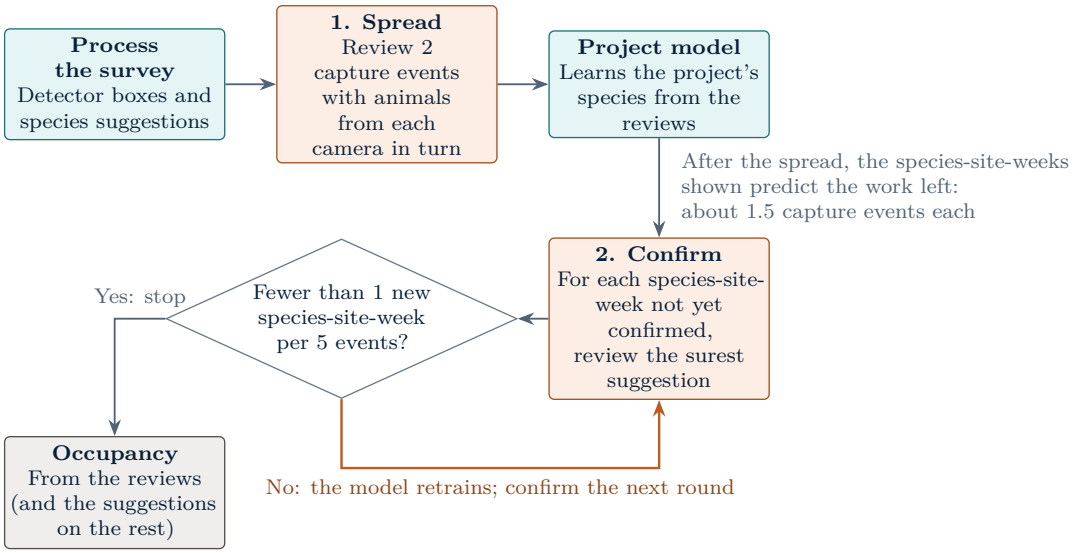


Figure 1: The review procedure for occupancy. Orange: the reviewer's steps; teal: the models; grey: the estimate.

holding the project model's most confident suggestion of each species-site-week that the suggestions showed but no reviewed image yet confirmed; if these did not fill the round, the rest of it took the least certain capture events, as in *Close calls*.

2.5 The review procedure

The procedure combines the strategies that did best (see Results) and adds a rule for when to stop. It has two steps (Fig. 1).

1. *Spread*: review up to k capture events with animals from each camera, one from each camera in turn, so that the project model learns the common species from every site. I tried $k = 2$, 5, and 10.

2. *Confirm*: train the project model; then, in rounds, review the capture event holding the most confident suggestion of each species-site-week that the suggestions show but no review has confirmed, retraining the model after each round. A round takes whole capture events until it reaches 2% of the survey's images; a reviewed event that does not show the suggested species leaves its species-site-week unconfirmed, so a later round may try the next most confident suggestion there. (WildObserve's *For occupancy* list, which runs this procedure, has no rounds: it judges the yield over the last 50 capture events reviewed from it, a number within the range of round sizes at which the procedure stopped here, 1 to 251 capture events.)

I ran the procedure five times for each k on each survey. Each run continued until no unconfirmed suggestion was left, a round confirmed fewer than 0.02 new species-site-weeks per capture event (below any rule worth trying), or 40% of the images were reviewed

(no run reached this). After the first step and after every round, I recorded the images and capture events reviewed, the species-site-weeks confirmed, and the estimates. A real project does not have a full review to compare with, so it needs a rule that it can apply to its own review. I evaluated, on the recorded runs, rules that stop after the first round that confirmed fewer than y new species-site-weeks per capture event it reviewed, for $y = 0.1$, 0.2, 0.3, and 0.5. To see whether a project can predict its work, I also counted, after the first step, the species-site-weeks that the reviews and the project model's suggestions showed between them.

2.6 Estimates and comparison

From each partial review I estimated occupancy three ways: from the reviewed images alone, counting images not yet reviewed as non-detections, with the standard model (MacKenzie *et al.*, 2002); with the project model's suggestions counted on the other images, with the same model; and with those suggestions in a model that allows for false positives (Miller *et al.*, 2011), in which detections in reviewed images are certain and suggestions are not. Detection probability depended on the days a camera ran in the occasion (Appendix S1). For the fixed shares, I also estimated occupancy with SpeciesNet's suggestions, without a project model, and from SpeciesNet's suggestions with no review at all. SpeciesNet's labels counted toward each survey's categories by taxonomy: a label counted toward the narrowest category containing its taxon, a broader label only when a single category lay within it, and any bird label toward a survey's bird category; another species of a category's genus counted as a different species. In the procedure's runs, I also estimated occupancy with the project model's scores as evidence instead of a cut-off (Rhinehart *et al.*, 2022), as WildObserve's score-based option does: two Normal distributions fitted to occasions' highest scores, as in ACORN (Haucke *et al.*, 2026), gave each occasion that no review settled a log Bayes factor (Appendix S1).

For each species, I compared occupancy from each partial review with occupancy from the full review. I scored only the species that the full review estimated well: those it detected at three or more sites, whose model fit converged without reaching a limit (an occupancy of 0 or 1), and whose 95% interval for occupancy was no wider than 0.5. SpeciesNet's estimates left out the categories that none of its labels counted toward: "small mammal" on WSU Lynx, and "other animal" and "small antelope" on Nkhotakota. For each estimate, I report the *mean absolute error* (MAE): the mean, over the scored species, of the absolute difference between a species' occupancy and its occupancy from the full review, averaged over the draws (occupancy runs from 0 to 1, so an MAE of 0.1 is 10 percentage points). In places I also report the median absolute error, and how often a partial review's 95% interval held the full review's estimate. A species that a run never detected was given an occupancy of 0. I also counted the species-site-weeks that a review

confirmed, and the share of the full review's that it found.

2.7 Declaration of the use of AI

I used Claude (Anthropic; Claude Opus models through Claude Code, September–November 2026) as an assistant for coding, review, and formatting text. Under my direction, it reviewed the code for the simulations, both in WildObserve's benchmark tools and in the open code. It also assisted with writing unit tests. Claude reviewed the text in this manuscript, suggested edits, and helped with formatting for this journal. I designed the study and its experiments and chose the surveys. I checked the code and its results, including that the open code reproduced WildObserve's own runs (Software and models), and I take full responsibility for the content. No AI tool created or altered the surveys' images or labels.

3 Results

3.1 SpeciesNet's suggestions with no review

Where SpeciesNet knew the fauna, its suggestions alone gave occupancy close to the full review's on average: the MAE was 0.06 on CCT, 0.09 on WSU Lynx, and 0.12 at APNR (median absolute error 0.03 on each; Fig. 2, dotted lines). On Nkhotakota, the MAE was 0.24 (median 0.11), largely because SpeciesNet named several of the reserve's species as related species from elsewhere: it called the yellow baboons, seen at 60 of the 63 sites, "Kinda baboon" on about 4,000 boxes scoring ≥ 0.9, bushpigs "wild boar", and many duikers and other antelopes "white-tailed deer". The averages hid failures for single species on the other surveys. SpeciesNet's suggestions detected no Canada lynx on WSU Lynx, where the full review put lynx occupancy at 0.13, and seem to have called them bobcats (bobcat occupancy rose from 0.52 to 0.63). At APNR, SpeciesNet called the chacma baboons "olive baboon" or "Kinda baboon", and the nyala other ungulates. Only 8 baboon boxes scored ≥ 0.9 as a baboon, too few detections to estimate from, so baboon occupancy went to its limit of 1 (full review, 0.39). Nyala occupancy fell to 0 (0.49).

3.2 Reviewing at random

Reviewed images alone, drawn at random, gave occupancy far from the full review's until half of a survey was reviewed (Fig. 2): the MAE was 0.36 to 0.41 with 5% reviewed and 0.10 to 0.12 with 50%. With 5% reviewed, the 95% intervals held the full review's estimate for only 40 to 77% of the species that had an interval. Random review spent most of its effort on empty images: about a third of the reviewed images showed an animal, except at APNR, where 90% of all images did. Counting suggestions for the rest helped. With 5%

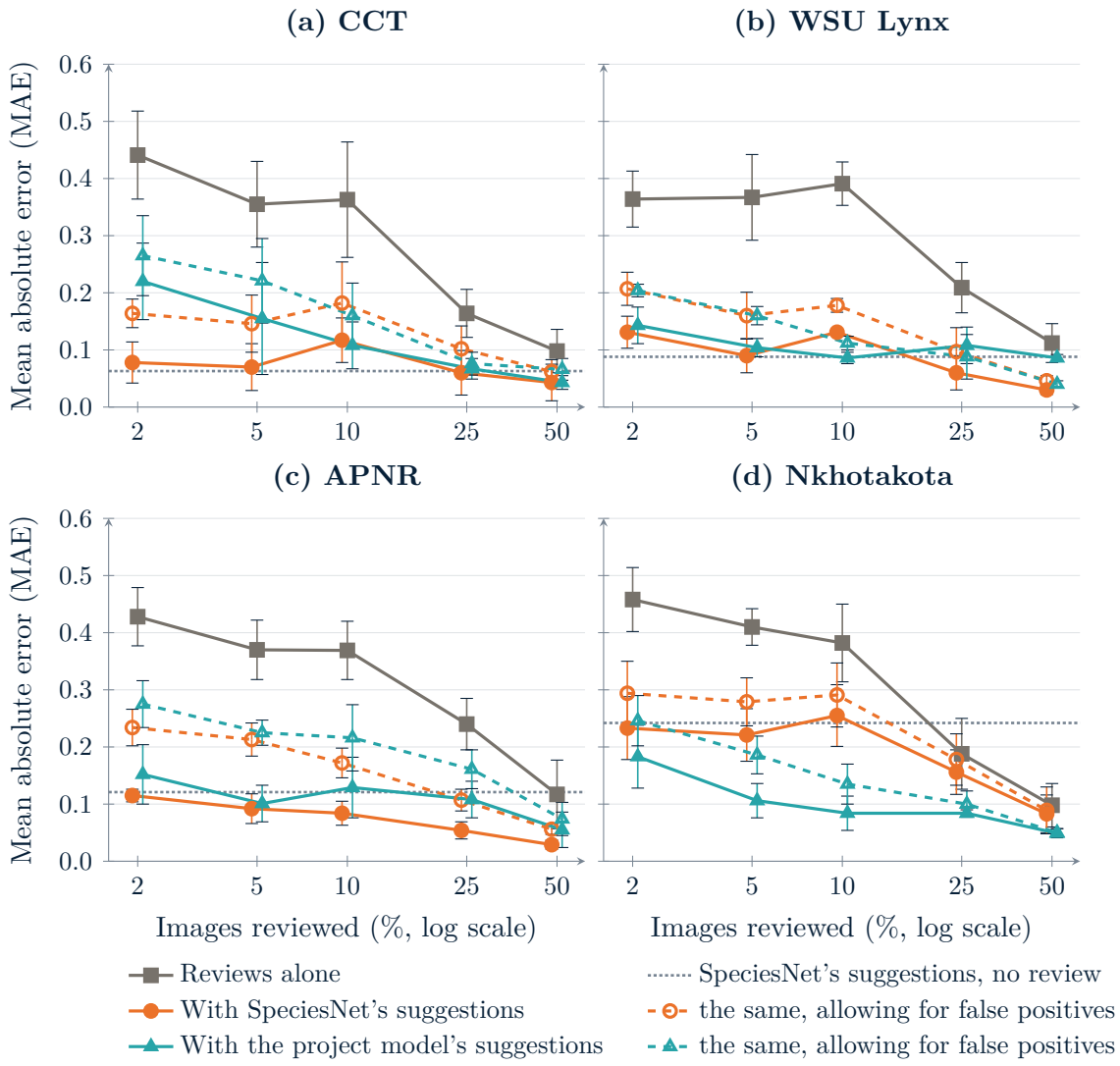


Figure 2: The mean absolute error (MAE) of occupancy estimated from part of a survey's review, against a full review, by the share of images reviewed, when capture events are reviewed at random, on (a) CCT, which SpeciesNet was probably trained on; (b) WSU Lynx and (c) APNR, whose fauna it knows; and (d) Nkhotakota, whose species it often misnames. Points: means over five draws; bars: ±1 standard deviation. Grey: reviewed images alone. Orange: with SpeciesNet's suggestions on the images not reviewed; teal: with the project model's, trained on the reviewed images. Dashed lines: allowing for false positives among the suggestions. Dotted line: SpeciesNet's suggestions with no review at all.

reviewed, on the three surveys whose fauna SpeciesNet knew, the MAE was 0.07 to 0.09 with SpeciesNet's suggestions and 0.10 to 0.15 with the project model's. On Nkhotakota, the project model's suggestions halved the MAE, to 0.11 from 0.22 with SpeciesNet's.

3.3 Ways of choosing the images

How the images were chosen mattered most for occupancy from the reviewed images alone (Fig. 3; Table 2). Confirming suggestions brought it closest: with 10% of the images reviewed, an MAE of 0.07 to 0.22, against 0.36 to 0.39 at random, and with 25%, 0.09

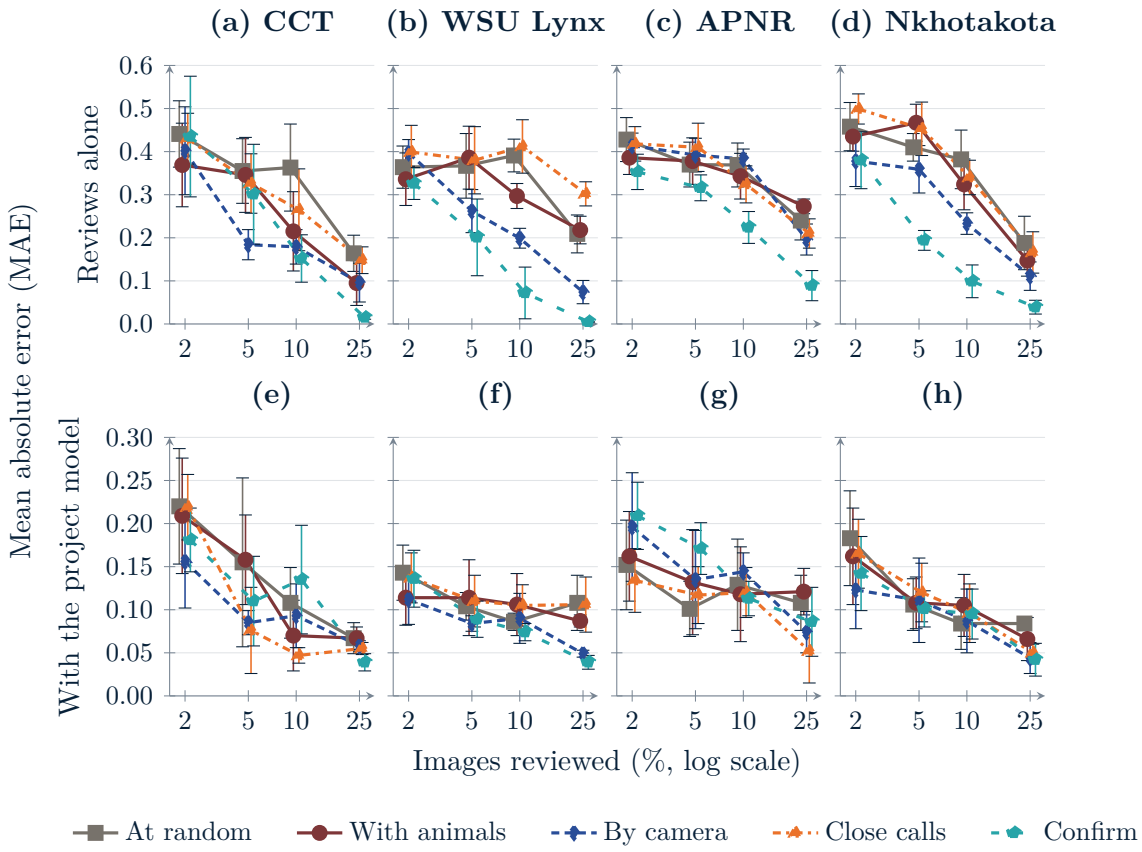


Figure 3: The MAE of occupancy against a full review for each way of choosing the images to review, from the reviewed images alone (top row) and with the project model's suggestions on the rest (bottom row), on the four surveys (columns, as in Fig. 2). Grey: capture events at random; maroon: at random from those in which the detector found an animal; slate blue: from those, each camera's in turn; orange: the closest calls first, in five rounds; teal: the most confident suggestion of each species-site-week that no review yet confirmed, in five rounds. Points and bars as in Fig. 2. The two rows have different scales.

or less, against 0.16 to 0.24. It was the only strategy that brought the MAE of the reviewed images alone below 0.1 on all four surveys with 25% or less reviewed (Table 2). Spreading the review across cameras came next. Reviewing capture events with animals at random helped little, because the detector also boxed fixed objects such as stumps on empty frames, so that only 35 to 52% of the images reviewed showed an animal (except at APNR). Reviewing the closest calls first was no better than that: it favored boxes on which SpeciesNet and the project model disagreed, which were often boxes on empty frames. In one draw on CCT, only 2 of the 435 images it chose showed an animal.

With the project model's suggestions counted, the strategies differed little with 10% or less of the images reviewed (MAE averaged over the surveys, 0.15 to 0.17 with 2% reviewed, 0.10 to 0.13 with 5%, and 0.09 to 0.10 with 10%; Fig. 3, bottom row), and no strategy was best on every survey. With 25% reviewed, confirming suggestions and spreading the review across cameras came closest (MAE 0.04 to 0.09 and 0.04 to 0.07, against 0.07 to 0.11 at random).

Table 2: The review that each way of choosing images needed: the smallest share of a survey's images (%) that, once reviewed, brought the MAE of occupancy below 0.1 and kept it there at every larger share run; > means it never did, up to the largest share run (50% for random review, 25% for the others).

		CCT	WSU Lynx	APNR	Nkhotakota
Reviews alone	At random	50	>50	>50	50
	With animals	25	>25	>25	>25
	By camera	25	25	>25	>25
	Close calls	>25	>25	>25	>25
	Confirm	25	10	25	10
With SpeciesNet	At random	25	25	5	50
	With animals	2	25	10	>25
	By camera	5	5	10	25
	Close calls	10	25	5	>25
	Confirm	2	5	5	10
With the project model	At random	25	50	50	10
	With animals	10	25	>25	25
	By camera	5	5	25	10
	Close calls	5	>25	25	10
	Confirm	25	5	25	10

3.4 The review procedure

After the first step, two capture events with animals from each camera (0.8 to 3.8% of the images), the reviewed images alone were still far from the full review (MAE 0.34 to 0.39), but the project model's suggestions brought the MAE to 0.14 to 0.19. Each round then confirmed about 0.6 to 1.0 new species-site-weeks for every capture event it reviewed (an event can show more than one species), until most of the full review's species-site-weeks had been confirmed. The yield then fell abruptly, to below 0.2 (Fig. 4). Stopping at that point, after 15 to 27% of the images had been reviewed, the procedure had confirmed 92 to 98% of the full review's species-site-weeks. The reviewed images alone gave an MAE of 0.04 or less (under 0.01 on WSU Lynx), and with the project model's suggestions, 0.03 to 0.05 (Table 3). These are means over five draws; in single draws the MAE reached 0.12 at APNR and 0.07 on Nkhotakota, where sparse detections put aardvark and leopard (APNR) or southern reedbuck (Nkhotakota) at an occupancy of 1. The procedure recovered the species that SpeciesNet had named wrongly: lynx occupancy was 0.13 (full review, 0.13), chacma baboon occupancy 0.39 (0.39), and nyala occupancy 0.43 (0.49). At about the same effort and from the reviewed images alone, the procedure came closer to the full review than random review did. After its first round past 10% of the images (about 11%), its reviewed images alone had an MAE of 0.07 to 0.25, against 0.36 to 0.39 with 10% reviewed at random; with the project model's suggestions, the two were similar (0.07 to 0.13 against 0.08 to 0.13).

Table 3: Where the review procedure stopped (the first round that confirmed fewer than 0.2 new species-site-weeks per capture event); means over five draws. *First step*: capture events with animals reviewed from each camera before confirming. *Images*: share of the survey's images reviewed. *Events per species-site-week*: capture events reviewed per species-site-week of the full review. *Found*: share of those species-site-weeks confirmed. *MAE*: from the reviewed images alone, with the project model's suggestions, and with its scores as evidence.

	First step	Images	Events per	Found	MAE		
Survey	per camera	(%)	species-site-week	(%)	alone	with model	with scores
CCT	2	20	1.7	96	0.02	0.03	0.01
	5	22	2.0	97	0.02	0.03	0.02
	10	24	2.2	97	0.01	0.03	0.02
WSU Lynx	2	16	1.8	98	0.00	0.03	0.00
	5	16	1.8	97	0.00	0.03	0.01
	10	17	1.9	97	0.01	0.03	0.01
APNR	2	27	1.1	97	0.03	0.03	0.03
	5	28	1.2	96	0.03	0.03	0.04
	10	30	1.2	97	0.01	0.01	0.01
Nkhotakota	2	15	1.3	92	0.04	0.05	0.05
	5	15	1.3	92	0.04	0.04	0.05
	10	16	1.4	92	0.03	0.04	0.04

The share of images that the procedure needed differed between the surveys, from 15% on Nkhotakota to 27% at APNR, where nearly every image showed an animal. The capture events it needed per species-site-week of the full review differed less, from 1.1 to 1.8 (Table 3), and measured this way the surveys' progress nearly coincided (Fig. 4b, d). A project can estimate its total review before confirming. After the first step, the reviews and the project model's suggestions together showed 0.76 to 1.31 times as many species-site-weeks as the full review had, and the procedure stopped after 1.2 to 1.8 capture events per species-site-week shown then (1.5 on average).

A larger first step took more review in all and did little better. With 10 capture events per camera, the procedure stopped after 16 to 30% of the images, with an MAE from the reviewed images alone of 0.01 to 0.03 (Table 3).

Rules of 0.1 and 0.2 new species-site-weeks per capture event stopped at nearly the same round. A rule of 0.3 sometimes stopped a round early (on CCT with 5 events per camera, an MAE of 0.04 rather than 0.02), and 0.5 stopped early on CCT and WSU Lynx (on CCT with 2 events per camera, 0.09 rather than 0.02).

3.5 False positives and scores as evidence

Allowing for false positives among the suggestions lowered the MAE (by more than 0.005) in only 14 of 168 combinations of survey, strategy, share reviewed, and kind of suggestion; otherwise it raised the MAE or barely changed it (Fig. 2, dashed lines). All 14 had at least

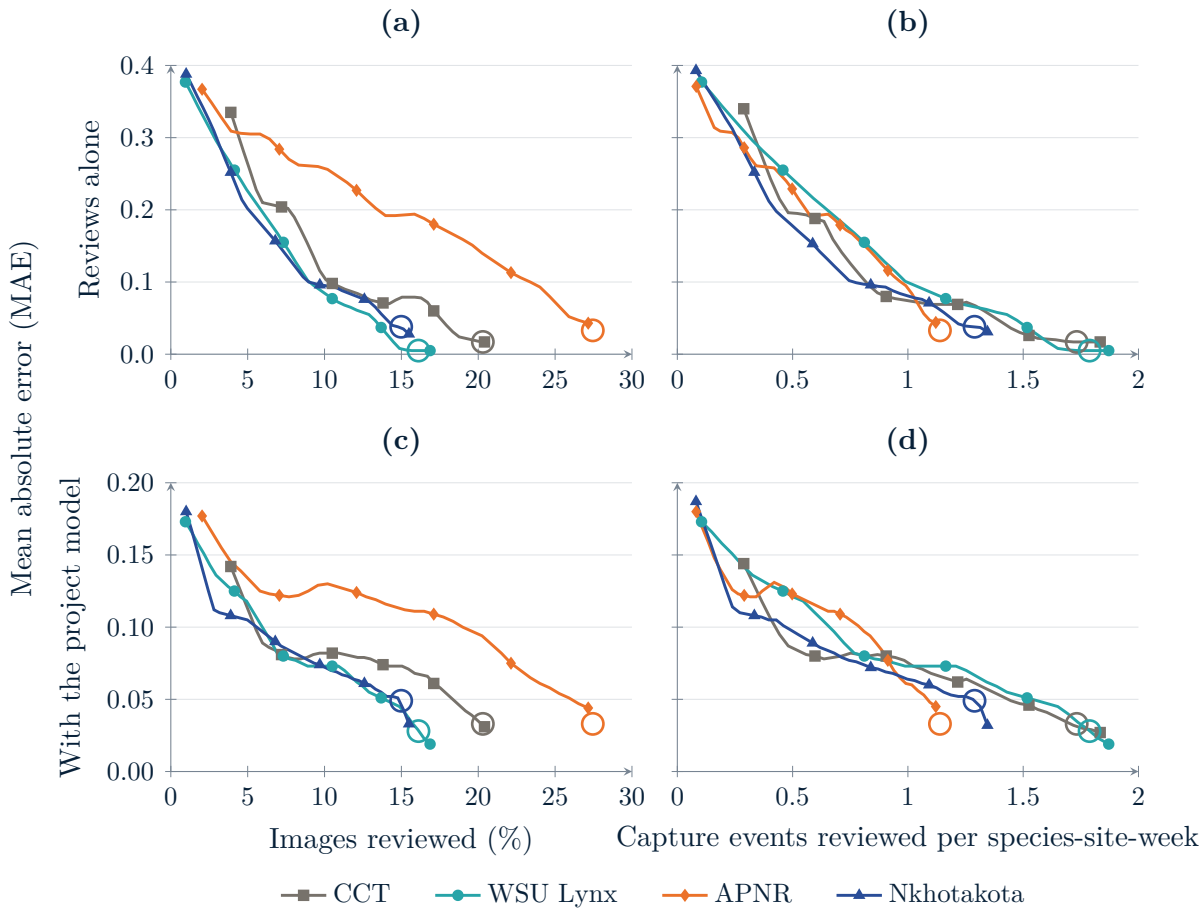


Figure 4: The review procedure's progress, with two capture events per camera in the first step: the MAE of occupancy against the full review, from the reviewed images alone (a, b) and with the project model's suggestions on the rest (c, d), against the share of the survey's images reviewed (a, c) and the capture events reviewed per species-site-week of the full review (b, d). Each line is the mean over five draws; open circles mark where the procedure stopped, on average, when a round confirmed fewer than 0.2 new species-site-weeks per capture event. Grey: CCT; teal: WSU Lynx; orange: APNR; slate blue: Nkhotakota. The two rows have different scales.

10% reviewed and all but one used the project model's suggestions; most were reviews that confirmed suggestions or spread across cameras, and most were on WSU Lynx. In the procedure, it helped on WSU Lynx midway (MAE 0.02 against 0.07 after about 11% of the images), where the confirmed detections it needs had accumulated, but not at APNR.

Weighing the project model's scores as evidence, instead of counting its suggestions over the cut-off, rarely helped either. After the first step, its MAE was 0.16 to 0.33 (counting suggestions, 0.14 to 0.19); after about 11% of the images, 0.03 to 0.14 (0.07 to 0.13), lower only on WSU Lynx (0.03 against 0.07); and at the stop, 0.05 or less, near the reviewed images alone (Table 3).

4 Discussion

Reviewing only the camera-trap images that an occupancy model needs gave estimates close to a full review's for a small part of its effort. The procedure, a small review spread across cameras followed by rounds that confirmed the most confident suggestion of each species-site-week, stopped by its own rule after 15 to 27% of each survey's images, with a mean MAE ≤ 0.04 from the reviewed images alone. Reviewing a similar share at random left the MAE several times higher, and reviewing the least certain images first, as systems that learn while people label often do, did little better than random. The choice of images mattered most when occupancy came from the reviewed images alone; with the project model's suggestions counted for the rest, the strategies differed little with $\leq 10\%$ reviewed.

4.1 Why confirming suggestions works

An occupancy model needs only each species' detection history: whether it was detected at each site in each occasion. Once a review has confirmed every species-site-week that the full review has, the reviewed images alone give the full review's detection history, and further review cannot change the estimates. A random review spends most of its effort on empty images and on repeated detections of common species, and still misses many species-site-weeks, so the model must infer how much it missed from too few detections. Confirming one suggestion per species-site-week goes straight to what the model needs, and the classifier only has to point to the right images. Each confirmation also teaches the project model the project's own species, so it points better as the review progresses, including to species that a general classifier names incorrectly: the procedure recovered the Canada lynx that SpeciesNet missed, and the baboons that it named as species from elsewhere.

The same reasoning explains when to stop. While unconfirmed species-site-weeks remain, most capture events offered confirm one; once nearly all are confirmed, the events offered hold mostly false suggestions, and the yield falls abruptly (Fig. 4). A project can watch that yield in its own review, without a full review to compare with; on every survey here the yield fell below 0.2 once 92 to 98% of the full review's species-site-weeks had been confirmed (Table 3). The number of species-site-weeks, more than the number of images, predicts how much review is needed: measured per species-site-week, the four surveys' progress nearly coincided (Fig. 4b, d).

4.2 Recommendations

These results suggest that projects can follow the review procedure (Fig. 1). Before reviewing, choose the sites and the occasion length (here seven days), since they define the species-site-weeks. Then review in two steps:

1. *Spread*: review two capture events with animals from each camera in turn, and train a model of the project's species on them.

2. *Confirm*: in rounds (here about 2% of the images each), review the capture event with the most confident suggestion of each species-site-week not yet confirmed; retrain after each round.

Stop when a round confirms fewer than one new species-site-week for every five capture events reviewed, and estimate occupancy from the reviews. A larger first step helped little: ten capture events per camera took more review in all and gave estimates at most slightly closer at the end (Table 3). A project that stops early can count the project model's suggestions on the rest: after about 11% of the images, these gave an MAE of 0.07 to 0.13, about as close as random review with the same suggestions (0.08 to 0.13). Unlike random review, though, the procedure can continue to estimates that need no suggestions.

WildObserve's *For occupancy* list runs both steps: it offers capture events in the procedure's order, shows how many species-site-weeks are left to confirm and about how many capture events that will take, and says when the reviewer can stop, once fewer than 10 of the last 50 capture events reviewed from it confirmed a new species-site-week. The open code implements every step outside WildObserve (Data availability statement).

Models that allow for false positives (Miller *et al.*, 2011; Chambert *et al.*, 2015) lowered the MAE only once the review had confirmed many detections, which such models need to tell true detections from false ones (Clement, 2016), and even then only on some surveys. Chambert *et al.* (2018) needed to confirm only a small share of automated detections, in part because their model also used the number of detections in each occasion; the model here used only whether an occasion had one.

4.3 No review and active learning

Where SpeciesNet knew a survey's animals, its suggestions alone came close on average (MAE 0.06 to 0.12; Fig. 2, dotted lines), as automated labels have before (Whytock *et al.*, 2021; Thornton *et al.*, 2026). The averages hid the failures that matter most to a survey, however: the procedure was needed to recover the lynx for which the WSU survey was designed, and the baboons and nyala at APNR. Thornton *et al.* (2026) likewise found the largest differences for rarely detected species, and a project cannot know in advance which of its species a general classifier will miss. The procedure's first step can reveal them: a species that the reviewer keeps correcting is one that the general classifier gets wrong.

Systems that learn while people label often choose the images they are least sure of, to help a classifier learn (Norouzzadeh *et al.*, 2021; Kellenberger *et al.*, 2020). For occupancy from the reviewed images alone, that choice performed poorly in this experiment (Fig. 3; Table 2). The least certain images were often boxes on empty frames, where the two

models disagreed, and even uncertain images of animals are seldom the detections that the occupancy model lacks. What to review should follow from the estimate that the review is for.

Haucke *et al.* (2026) reached the same conclusion by another route. ACORN chooses the reviews expected to tell the most about each species' occupancy model, and it beat both uncertainty sampling and random review. Reviewing the highest-scoring camera-days, the baseline nearest to confirming suggestions, was surprisingly strong, and ACORN beat it mainly with small budgets, by not spending reviews on detections that the model could already predict. Confirming only the species-site-weeks that no review has confirmed avoids the same waste without a model of each species. The two approaches differ in what they ask of a project. ACORN refits a Bayesian model of each species and its covariates as the review goes on, and its use of the classifier's scores as continuous evidence brought its estimates closer to a full labeling before any review. The procedure here needs only a count that a reviewer can keep, reviews every species in an image at once, and lets the classifier learn the project's own species, which recovered those that SpeciesNet missed. Weighing the project model's scores as evidence helped little here (Table 3), perhaps because ACORN measured the scores' gain against reviewed labels alone, whereas here they were compared with a project model's confident suggestions, which already carry much of that information.

4.4 Limitations

The reviews were simulated from the datasets' own labels, so the simulated reviewer made no mistakes of its own, and the full review with which every estimate was compared inherits the labels' errors. APNR's labels are volunteers' consensus, and WSU Lynx and APNR were labeled by sequence. Additionally, I measured effort in images and capture events, not in reviewers' time. Empty images are quicker to review, so the procedure, which reviews mostly animals, costs more time per image than a random review. Furthermore, I fitted single-season, single-species models without covariates on occupancy, whereas Thornton *et al.* (2026) and Haucke *et al.* (2026) also compared species–environment relationships. The stopping rule and the planning figure (1.5 capture events per species-site-week) were chosen on the same four surveys on which they were tested, all with seven-day occasions. Finally, a species-site-week that no model suggests with confidence cannot be confirmed, and neither can a species that the first step never meets. Here, the procedure left 2 to 8% of the full review's species-site-weeks unconfirmed. Projects that care about rare species should also look for species that the review has not met, for example with WildObserve's *New species?* list.

4.5 Implications

A camera-trap project that needs occupancy estimates need not review every image, or trust a general classifier's labels without review: it can confirm the detections that the occupancy model needs, in the order that a model of its own species suggests, and stop when the review stops confirming new ones. The work is predictable from the review's first step, about 1.5 capture events for each species-site-week it shows, and the project can tell from its own review when it is done. The procedure presented here can be valuable for teams that cannot review every camera-trap image: they could estimate occupancy for the species they detect soon after each survey, rather than waiting for a full review.

Acknowledgments

I thank Dan Morris for confirming which datasets SpeciesNet was trained on; the California Institute of Technology, the WSU Lynx team, Snapshot Safari and its volunteers, the Nkhotakota Camera Traps team, African Parks, Malawi's Department of National Parks and Wildlife, and LILA BC for making the surveys available.

Author contributions

MT: conceptualization, methodology, software, formal analysis, data curation, visualization, writing (original draft), and writing (review and editing).

Conflict of interest

MT owns Quantitative Science Consulting, LLC, which develops WildObserve, the software used in this study.

Declaration of funding

This research received no specific funding.

Ethics statement

This study reanalyzed images from four published camera-trap surveys that are publicly available (Data availability statement). No fieldwork was done and no animals were captured, handled, or disturbed, so animal ethics approval was not required.

Data availability statement

The four camera-trap surveys are public on LILA BC (`https://lila.science`) under the Community Data License Agreement (Permissive): Caltech Camera Traps (Beery *et al.*, 2018), WSU Lynx (Thornton *et al.*, 2026), the Snapshot Safari 2024 Expansion (Pardo *et al.*, 2021), and Nkhotakota Camera Traps (Appel *et al.*, 2025). The code that runs every experiment from the data exported from WildObserve, with the export script, is at `https://github.com/QScienceC/occupancy-camtrap` (MIT License). The exported data and the results are archived on Zenodo (`https://doi.org/10.5281/zenodo.23734806`). WildObserve is available from the developer (`https://www.qsc.earth/WildObserve`); its Community edition (version 0.9.0 or later), free for non-commercial use, includes the *For occupancy* list and the project model. Preliminary results for earlier versions of this analysis are presented in the preprint describing WildObserve (Tabak, 2026).

References

Appel, C.L., Subramanian, A., Koning, J.S., Ngosi, M., Sullivan, C.M., Levi, T. & Lesmeister, D.B. (2025) Developing custom computer vision models with Njobvu-AI: A collaborative, user-friendly platform for ecological research. *Ecological Applications* **35**, e70096.

Beery, S., Morris, D. & Yang, S. (2019) Efficient pipeline for camera trap image review. *arXiv preprint* **arXiv:1907.06772**.

Beery, S., Van Horn, G. & Perona, P. (2018) Recognition in Terra Incognita. *Computer Vision – ECCV 2018*, pp. 472–489, Springer International Publishing.

Burton, A.C., Neilson, E., Moreira, D., Ladle, A., Steenweg, R., Fisher, J.T., Bayne, E. & Boutin, S. (2015) Wildlife camera trapping: a review and recommendations for linking surveys to ecological processes. *Journal of Applied Ecology* **52**, 675–685.

Chambert, T., Miller, D.A.W. & Nichols, J.D. (2015) Modeling false positive detections in species occurrence data under different study designs. *Ecology* **96**, 332–339.

Chambert, T., Waddle, J.H., Miller, D.A.W., Walls, S.C. & Nichols, J.D. (2018) A new framework for analysing automated acoustic species detection data: occupancy estimation and optimization of recordings post-processing. *Methods in Ecology and Evolution* **9**, 560–570.

Clare, J.D.J., Townsend, P.A., Anhalt-Depies, C., Locke, C., Stenglein, J.L., Frett, S., Martin, K.J., Singh, A., Van Deelen, T.R. & Zuckerberg, B. (2019) Making inference

with messy (citizen science) data: when are data accurate enough and how can they be improved? *Ecological Applications* **29**, e01849.

Clement, M.J. (2016) Designing occupancy studies when false-positive detections occur. *Methods in Ecology and Evolution* **7**, 1538–1547.

Fennell, M., Beirne, C. & Burton, A.C. (2022) Use of object detection in camera trap image identification: Assessing a method to rapidly and accurately classify human and animal detections for research and application in recreation ecology. *Global Ecology and Conservation* **35**, e02104.

Fiske, I. & Chandler, R. (2011) unmarked: An R package for fitting hierarchical models of wildlife occurrence and abundance. *Journal of Statistical Software* **43**, 1–23.

Gadot, T., Istrate, Ș., Kim, H., Morris, D., Beery, S., Birch, T. & Ahumada, J. (2024) To crop or not to crop: Comparing whole-image and cropped classification on a large dataset of camera trap images. *IET Computer Vision* **18**, 1193–1208.

Glover-Kapfer, P., Soto-Navarro, C.A. & Wearn, O.R. (2019) Camera-trapping version 3.0: current constraints and future priorities for development. *Remote Sensing in Ecology and Conservation* **5**, 209–223.

Guillera-Arroita, G., Lahoz-Monfort, J.J., van Rooyen, A.R., Weeks, A.R. & Tingley, R. (2017) Dealing with false-positive and false-negative errors about species occurrence at multiple levels. *Methods in Ecology and Evolution* **8**, 1081–1091.

Guillera-Arroita, G., Ridout, M.S. & Morgan, B.J.T. (2010) Design of occupancy studies with imperfect detection. *Methods in Ecology and Evolution* **1**, 131–139.

Haucke, T., Harrell, L., Kay, J., Clapp, M. & Beery, S. (2026) Targeted review for AI-assisted biodiversity surveys: Active continuous-score occupancy modeling. arXiv preprint arXiv:2609.25657.

Hernandez, A., Miao, Z., Vargas, L., Beery, S., Dodhia, R., Arbelaez, P. & Ferres, J.M.L. (2024) Pytorch-Wildlife: A collaborative deep learning framework for conservation. *arXiv preprint* **arXiv:2405.12930**.

Kays, R., Arbogast, B.S., Baker-Whatton, M., Beirne, C., Boone, H.M., Bowler, M., Burneo, S.F., Cove, M.V., Ding, P., Espinosa, S., Gonçalves, A.L.S., Hansen, C.P., Jansen, P.A., Kolowski, J.M., Knowles, T.W., Lima, M.G.M., Millspaugh, J., McShea, W.J., Pacifici, K., Parsons, A.W., Pease, B.S., Rovero, F., Santos, F., Schuttler, S.G., Sheil, D., Si, X., Snider, M. & Spironello, W.R. (2020) An empirical evaluation of camera trap study design: How many, how long and when? *Methods in Ecology and Evolution* **11**, 700–713.

Kellenberger, B., Tuia, D. & Morris, D. (2020) AIDE: Accelerating image-based ecological surveys with interactive machine learning. *Methods in Ecology and Evolution* **11**, 1716–1727.

Kellner, K.F., Smith, A.D., Royle, J.A., Kéry, M., Belant, J.L. & Chandler, R.B. (2023) The unmarked R package: Twelve years of advances in occurrence and abundance modelling in ecology. *Methods in Ecology and Evolution* **14**, 1408–1415.

MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Royle, J.A. & Langtimm, C.A. (2002) Estimating site occupancy rates when detection probabilities are less than one. *Ecology* **83**, 2248–2255.

MacKenzie, D.I. & Royle, J.A. (2005) Designing occupancy studies: general advice and allocating survey effort. *Journal of Applied Ecology* **42**, 1105–1114.

Miller, D.A., Nichols, J.D., McClintock, B.T., Grant, E.H.C., Bailey, L.L. & Weir, L.A. (2011) Improving occupancy estimation when two types of observational error occur: non-detection and species misidentification. *Ecology* **92**, 1422–1428.

Norouzzadeh, M.S., Morris, D., Beery, S., Joshi, N., Jojic, N. & Clune, J. (2021) A deep active learning system for species identification and counting in camera trap images. *Methods in Ecology and Evolution* **12**, 150–161.

Norouzzadeh, M.S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M.S., Packer, C. & Clune, J. (2018) Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences* **115**, E5716–E5725.

Pardo, L.E., Bombaci, S.P., Huebner, S., Somers, M.J., Fritz, H., Downs, C., Guthmann, A., Hetem, R.S., Keith, M., Le Roux, A., Mgqatsa, N., Packer, C., Palmer, M.S., Parker, D.M., Peel, M., Slotow, R., Strauss, W.M., Swanepoel, L., Tambling, C., Tsie, N., Vermeulen, M., Willi, M., Jachowski, D.S. & Venter, J.A. (2021) Snapshot Safari: A large-scale collaborative to monitor Africa's remarkable biodiversity. *South African Journal of Science* **117**, 8134.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, É. (2011) Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830.

Rhinehart, T.A., Turek, D. & Kitzes, J. (2022) A continuous-score occupancy model that incorporates uncertain machine learning output from autonomous biodiversity surveys. *Methods in Ecology and Evolution* **13**, 1778–1789.

Royle, J.A. & Link, W.A. (2006) Generalized site occupancy models allowing for false positive and false negative errors. *Ecology* **87**, 835–841.

Schneider, S., Greenberg, S., Taylor, G.W. & Kremer, S.C. (2020) Three critical factors affecting automated image species recognition performance for camera traps. *Ecology and Evolution* **10**, 3503–3517.

Specht, H.M., Reich, H.T., Iannarilli, F., Edwards, M.R., Stapleton, S.P., Weegman, M.D., Johnson, M.K., Yohannes, B.J. & Arnold, T.W. (2017) Occupancy surveys with conditional replicates: an alternative sampling design for rare species. *Methods in Ecology and Evolution* **8**, 1725–1734.

Tabak, M. (2026) WildObserve: Camera-trap software that learns species labels during review and estimates occupancy and abundance. bioRxiv preprint.

Tabak, M.A., Norouzzadeh, M.S., Wolfson, D.W., Newton, E.J., Boughton, R.K., Ivan, J.S., Odell, E.A., Newkirk, E.S., Conrey, R.Y., Stenglein, J., Iannarilli, F., Erb, J., Brook, R.K., Davis, A.J., Lewis, J., Walsh, D.P., Beasley, J.C., VerCauteren, K.C., Clune, J. & Miller, R.S. (2020) Improving the accessibility and transferability of machine learning algorithms for identification of animals in camera trap images: MLWIC2. *Ecology and Evolution* **10**, 10374–10383.

Tabak, M.A., Norouzzadeh, M.S., Wolfson, D.W., Sweeney, S.J., Vercauteren, K.C., Snow, N.P., Halseth, J.M., Di Salvo, P.A., Lewis, J.S., White, M.D., Teton, B., Beasley, J.C., Schlichting, P.E., Boughton, R.K., Wight, B., Newkirk, E.S., Ivan, J.S., Odell, E.A., Brook, R.K., Lukacs, P.M., Moeller, A.K., Mandeville, E.G., Clune, J. & Miller, R.S. (2019) Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods in Ecology and Evolution* **10**, 585–590.

Thornton, D., Morris, D., King, T., Perera-Romero, L., Anderson, A., Garcia-Anleu, R., Fitkin, S. & Vynne, C. (2026) Identification of camera trap images by artificial intelligence and human experts produces similar multi-species occupancy models. *Journal of Applied Ecology* **63**, e70370.

Virtanen, P., Gommers, R., Oliphant, T.E. *et al.* (2020) SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods* **17**, 261–272.

Whytock, R.C., Świeżewski, J., Zwerts, J.A., Bara-Słupski, T., Koumba Pambo, A.F., Rogala, M., Bahaa-el din, L., Boekee, K., Brittain, S., Cardoso, A.W., Henschel, P., Lehmann, D., Momboua, B., Kiebou Opepa, C., Orbell, C., Pitman, R.T., Robinson, H.S. & Abernethy, K.A. (2021) Robust ecological analysis of camera trap data labelled by a machine learning model. *Methods in Ecology and Evolution* **12**, 1080–1092.

Willi, M., Pitman, R.T., Cardoso, A.W., Locke, C., Swanson, A., Boyer, A., Veldthuis, M. & Fortson, L. (2019) Identifying animal species in camera trap images using deep learning and citizen science. *Methods in Ecology and Evolution* **10**, 80–91.

Wright, W.J., Irvine, K.M., Almberg, E.S. & Litt, A.R. (2020) Modelling misclassification in multi-species acoustic data when estimating occupancy and relative activity. *Methods in Ecology and Evolution* **11**, 71–81.

Supplementary material for

Which camera-trap images to review? Estimating occupancy from a small part of a survey

Appendix S1. The occupancy models

Every estimate in the paper came from one of three single-season, single-species models, with no covariates on occupancy, fitted by maximum likelihood: the standard model (S1.2), a model that allows for false positives among the suggestions (S1.3), and a score-based model (S1.4). WildObserve fits all three (Tabak, 2026); the estimates reported come from the open code, which fits the same models and agreed with WildObserve to within 0.0001 on the four full reviews. Table S1 lists which model each estimate used.

S1.1 Detection histories

A site $i = 1, \ldots, N$ is a camera's location. Its survey is divided into seven-day occasions $j = 1, \ldots, J$ from its first running day, and d_{ij} is the number of days its camera ran in occasion j. An occasion with $d_{ij} = 0$ is missing: it does not enter the likelihood, and the products over j below run over the occasions with $d_{ij} > 0$ only. Images taken outside the occasions do not count.

For a species, $y_{ij} = 1$ if it was detected at site i in occasion j: a reviewed image taken there shows it (a *confirmed* detection), or, when suggestions count, a box on an image there not yet reviewed has the species as its top suggestion with a score of 0.9 or more. Only species on the survey's list count as suggested. Otherwise $y_{ij} = 0$, so an image not yet reviewed, and without such a suggestion, counts as a non-detection.

S1.2 The standard model

The standard model (MacKenzie *et al.*, 2002) has a site occupied ($z_i = 1$) with probability ψ, and, at an occupied site, the species detected in occasion j with probability p_{ij}:

$$\operatorname{logit} \psi = \beta_0, \qquad \operatorname{logit} p_{ij} = \alpha_0 + \alpha_1 d_{ij}. \tag{1}$$

The term in d_{ij} is included only when the camera-days vary among the survey's occasions; otherwise $\operatorname{logit} p_{ij} = \alpha_0$. A site's likelihood is

$$L_i = \psi \prod_j p_{ij}^{y_{ij}} (1 - p_{ij})^{1 - y_{ij}} + (1 - \psi)\, \mathbf{1}\Big[\textstyle\sum_j y_{ij} = 0\Big], \tag{2}$$

and the likelihood is the product of the L_i. This is the model that the R package unmarked fits as `occu(~days ~1)` (Fiske & Chandler, 2011; Kellner *et al.*, 2023).

S1.3 Allowing for false positives

When suggestions count, a detection is either confirmed by a reviewed image or rests on suggestions alone. The model of Miller *et al.* (2011) with multiple detection states (unmarked's `occuFP`, with type 3 data) records an occasion's outcome as $w_{ij} = 2$ (confirmed), $w_{ij} = 1$ (detected by suggestions alone), or $w_{ij} = 0$ (not detected). At an occupied site, the species is detected with probability p_{ij}, as in S1.2, and a detection is confirmed with probability b_{ij}; at an unoccupied site, suggestions alone detect it with probability f, and nothing is confirmed:

$$\Pr(w_{ij} \mid z_i = 1) = \begin{cases} 1 - p_{ij} & w_{ij} = 0 \\ p_{ij}(1 - b_{ij}) & w_{ij} = 1 \\ p_{ij}\, b_{ij} & w_{ij} = 2 \end{cases} \qquad \Pr(w_{ij} \mid z_i = 0) = \begin{cases} 1 - f & w_{ij} = 0 \\ f & w_{ij} = 1 \\ 0 & w_{ij} = 2 \end{cases} \tag{3}$$

$$\operatorname{logit} b_{ij} = \gamma_0 + \gamma_1 r_{ij}, \qquad \operatorname{logit} f = \kappa,$$

where r_{ij} is the share of the occasion's images with animals that had been reviewed (an image with animals being, if reviewed, one the review gave an animal, and if not, one on which the detector found an animal): the more of an occasion's animal images were reviewed, the likelier a true detection there was confirmed. The term in r_{ij} is included only when r_{ij} varies among the occasions in which the species was detected. A site's likelihood is

$$L_i = \psi \prod_j \Pr(w_{ij} \mid z_i = 1) + (1 - \psi) \prod_j \Pr(w_{ij} \mid z_i = 0), \tag{4}$$

whose second term is 0 at a site with a confirmed detection. The model was fitted only for a species that some reviewed image showed; for any other, the standard model's estimate stood in.

S1.4 The score-based model

In the procedure's runs, the score-based model weighed the project model's scores as evidence instead of counting suggestions over the cut-off (Rhinehart *et al.*, 2022), with the scores' calibration fitted apart from the occupancy model, as in ACORN's decoupled model (Haucke *et al.*, 2026). For a species, an occasion is *settled* if a reviewed image there shows the species ($y_{ij} = 1$) or if no image there not yet reviewed has a box with a suggestion ($y_{ij} = 0$; WildObserve gives suggestions to the five most confident boxes of

each image). Every other occasion is *weighed*: its score s_{ij} is the highest log-odds that the project model gives the species among those boxes.

If g_1 and g_0 are the densities of an occasion's highest score with and without the species, a weighed occasion contributes $p_{ij}\, g_1(s_{ij}) + (1 - p_{ij})\, g_0(s_{ij})$ to an occupied site's likelihood and $g_0(s_{ij})$ to an unoccupied one's. Dividing both by $g_0(s_{ij})$, which does not depend on the model's parameters, gives the evidence as a log Bayes factor, $\ell_{ij} = \log g_1(s_{ij}) - \log g_0(s_{ij})$, and a site's likelihood

$$L_i = \psi \prod_{j \text{ settled}} p_{ij}^{y_{ij}} (1-p_{ij})^{1-y_{ij}} \prod_{j \text{ weighed}} \left[(1-p_{ij}) + p_{ij}\, e^{\ell_{ij}}\right] + (1-\psi)\, \mathbf{1}\left[\textstyle\sum_{j \text{ settled}} y_{ij} = 0\right], \quad (5)$$

with ψ and p_{ij} as in S1.2.

The densities g_1 and g_0 are Normal, fitted by expectation–maximization to three groups of scores: the highest score in each occasion with a confirmed detection (with the species), in each settled occasion without one (without it), and in each weighed occasion (a mixture of the two). Their parameters are then held fixed while occupancy is fitted. The project model learned from the reviewed boxes, so its scores on them run higher than on boxes it has not seen. The two Normal distributions are therefore first fitted to the weighed occasions' scores alone, and that fit is kept while it implies three or more detections among the weighed occasions and its mean with the species lies at least halfway from its mean without the species to the mean of the confirmed detections' scores; otherwise the settled occasions' scores join the fit. A species with fewer than three confirmed occasions, or whose fitted mean with the species does not lie above the other, has no score-based estimate, and the standard model's estimate stands in. The standard deviations are kept at 0.05 or more, and beyond either mean the evidence is held at least as strong as at that mean ($\ell_{ij} \geq \ell(\mu_1)$ for $s_{ij} > \mu_1$, and $\ell_{ij} \leq \ell(\mu_0)$ for $s_{ij} < \mu_0$), since a ratio of two Normal densities with different spreads would otherwise turn back.

S1.5 Fitting

Each model's negative log-likelihood was minimized with its analytic gradient by L-BFGS-B, in SciPy (Virtanen *et al.*, 2020), stopping when no component of the gradient exceeded 10^{-4}, as WildObserve's optimizer does (rather than on a small relative change in the likelihood, SciPy's default, which on large surveys stops before the gradient is small). A fit counted as converged if no component of the gradient exceeded 10^{-3} at its end, and occupancy as at its limit if $|\operatorname{logit} \hat{\psi}| > 8$ ($\hat{\psi}$ below 0.0003 or above 0.9997). The standard error of $\operatorname{logit} \hat{\psi}$ came from the inverse of the Hessian, taken by central differences of the gradient, and the 95% interval was $\operatorname{logit} \hat{\psi} \pm 1.96$ standard errors, transformed back, as unmarked's `predict()` gives it. A fit at its limit, with a Hessian that was not positive definite, or with a standard error above 10 on the logit scale, had no interval.

The fits started from ψ equal to the share of sites with a detection (in S1.3, a confirmed one; in S1.4, a detection or positive evidence), and p equal to the share of occasions with a detection at the sites with any detection, each kept between 0.05 and 0.95, with $\alpha_1 = \gamma_0 = \gamma_1 = 0$ and $\kappa = -3$. Where a species' detections let detection probability approach a step in camera-days, a fit carried further reaches a slightly higher likelihood with parameters running off toward infinity, and for a species detected at only one or two sites the likelihood can have a second, interior optimum that this start does not find. Both affect few species and draws.

Table S1: The estimates in the paper, the detections each counted, and the models fitted.

Estimate	Detections	Models
Reviewed images alone	reviewed images; the rest as non-detections	standard
With the project model's suggestions	reviewed images, and suggestions on the rest	standard; false positives
With SpeciesNet's suggestions	reviewed images, and suggestions on the rest	standard; false positives
SpeciesNet alone	SpeciesNet's suggestions, with no review	standard
Score-based (the procedure's runs)	reviewed images, and scores on the rest	score-based

References

Fiske, I. & Chandler, R. (2011) unmarked: An R package for fitting hierarchical models of wildlife occurrence and abundance. *Journal of Statistical Software* **43**, 1–23.

Haucke, T., Harrell, L., Kay, J., Clapp, M. & Beery, S. (2026) Targeted review for AI-assisted biodiversity surveys: Active continuous-score occupancy modeling. arXiv preprint arXiv:2609.25657.

Kellner, K.F., Smith, A.D., Royle, J.A., Kéry, M., Belant, J.L. & Chandler, R.B. (2023) The unmarked R package: Twelve years of advances in occurrence and abundance modelling in ecology. *Methods in Ecology and Evolution* **14**, 1408–1415.

MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Royle, J.A. & Langtimm, C.A. (2002) Estimating site occupancy rates when detection probabilities are less than one. *Ecology* **83**, 2248–2255.

Miller, D.A., Nichols, J.D., McClintock, B.T., Grant, E.H.C., Bailey, L.L. & Weir, L.A. (2011) Improving occupancy estimation when two types of observational error occur: non-detection and species misidentification. *Ecology* **92**, 1422–1428.

Rhinehart, T.A., Turek, D. & Kitzes, J. (2022) A continuous-score occupancy model that incorporates uncertain machine learning output from autonomous biodiversity surveys. *Methods in Ecology and Evolution* **13**, 1778–1789.

Tabak, M. (2026) WildObserve: Camera-trap software that learns species labels during review and estimates occupancy and abundance. bioRxiv preprint.

Virtanen, P., Gommers, R., Oliphant, T.E. *et al.* (2020) SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods* **17**, 261–272.