Peer Community In
Ecology

Accounting for false positives and negatives in monitoring data from sensor networks and eDNA

Damaris Zurell *based on peer reviews by* ***Jonathan Rose***, ***Saoirse Kelleher*** *and 2 anonymous reviewers*

Célian Monchy, Marie-Pierre Etienne, Olivier Gimenez (2024) Using informative priors to account for identifiability issues in occupancy models with identification errors. bioRxiv, ver. 3, peer-reviewed and recommended by Peer Community in Ecology. https://doi.org/10.1101/2024.05.07.592917

Submitted: 11 May 2024, Recommended: 06 January 2025

Cite this recommendation as:
Zurell, D. (2025) Accounting for false positives and negatives in monitoring data from sensor networks and eDNA. *Peer Community in Ecology*, 100699. 10.24072/pci.ecology.100699

Published: 06 January 2025

Biodiversity monitoring increasingly relies on modern technologies such as sensor networks and environmental DNA. These high-throughput methods allow biodiversity assessments with unprecedented detail and are especially useful to detect rare and secretive species that are otherwise difficult to observe with traditional survey-based methods. False negatives through imperfect detection are a typical problem in survey data and depend on intrinsic characteristics of the species, site characteristics of the survey site as well as survey characteristics (Guillera 2017). While imperfect detection might be reduced in modern sensor data and eDNA data, also these types of data are by no means error-free and may bare other challenges. In particular, the bioinformatics and image classification approaches used for species identification from these data can induce a higher rate of false positives than would be expected in expert-based survey data (Hartig et al. 2024).

Occupancy models (or occupancy-detection models) have been widely used to map species distributions by fitting a hierarchical model that estimates the paramaters of both the species-environment relationship and an observation submodel. They account for false negatives by inferring detectability from the detection history of a survey location, for example from replicate visits or multiple observers (Guillera 2017). These basic occupancy-detection models assume no false positive errors in the data. Other authors have proposed extensions for false positives that typically rely on unambiguous (known truth) information for some sites or observations (Chambert et al. 2015).

In their preprint, Monchy et al. (2024) propose an extension of classic occupancy models that considers a two-step observation process modelling the detection probability at occupied sites and the associated identification probability, separated into the true positive identification rate and the true negative identification rate. Using a simulation approach, the authors compare the effectiveness of a frequentist (maximum likelihood-based) and Bayesian approach for parameter estimation and identifiability, and additionally test the effectiveness of different priors (from non-informative to highly informative). Results of the maximum-likelihood approach indicated biased parameter estimates and identifiability problems. In the Bayesian approach, inclusion of prior information greatly reduces biases in parameter estimates, especially in detection and positive identification rate.

Importantly, informative priors for the identification process are a by-product of the classifiers that are developed for processing the eDNA data or sensor data. For example, species identification from acoustic sensors is based on image classifiers trained on labelled bird song spectrograms (Kahl et al. 2021) and as part of the evaluation of the classifier, the true positive rate (sensitivity) is routinely being estimated and could thus be readily used in occupancy models accounting for false positives. Thus, the approach proposed by Monchy et al. (2024) is not only highly relevant for biodiversity assessments based on novel sensor and eDNA data but also provides very practical solutions that do not require additional unambiguous data but recycle data that are already available in the processing pipeline. Applying their framework to real-world data will help reducing biases in biodiversity assessments and through improved understanding of the detection process it could also help optimising the design of sensor networks.

References:

Thierry Chambert, David A. W. Miller, James D. Nichols (2015), Modeling false positive detections in species occurrence data under different study designs. Ecology, 96: 332-339. https://doi.org/10.1890/14-1507.1

Gurutzeta Guillera-Arroita (2017) Modelling of species distributions, range dynamics and communities under imperfect detection: advances, challenges and opportunities. Ecography, 40: 281-295. https://doi.org/10.1111/ecog.02445

Florian Hartig, Nerea Abrego, Alex Bush, Jonathan M. Chase, Gurutzeta Guillera-Arroita, Mathew A. Leibold, Otso Ovaskainen, Loïc Pellissier, Maximilian Pichler, Giovanni Poggiato, Laura Pollock, Sara Si-Moussi, Wilfried Thuiller, Duarte S. Viana, David I. Warton, Damaris Zurell D, Douglas W. Yu (2024) Novel community data in ecology - properties and prospects. Trends in Ecology & Evolution, 39: 280-293. https://doi.org/10.1016/j.tree.2023.09.017

Stefan Kahl, Connor M. Wood, Maximilian Eibl, Holger Klinck (2021) BirdNET: A deep learning solution for avian diversity monitoring. Ecological Informatics, 61: 101236. https://doi.org/10.1016/j.ecoinf.2021.101236

Célian Monchy, Marie-Pierre Etienne, Olivier Gimenez (2024) Using informative priors to account for identifiability issues in occupancy models with identification errors. bioRxiv, ver.3 peer-reviewed and recommended by PCI Ecology https://doi.org/10.1101/2024.05.07.592917

Reviews

Evaluation round #2

DOI or URL of the preprint: https://doi.org/10.1101/2024.05.07.592917
Version of the preprint: 2

Authors' reply, 08 December 2024

We have deeply appreciated to read your cheerful comments about the last version of the manuscript, and we sincerely thank you for that. We are also grateful for your precise last comments to finally improve the reading fluidity. We have accounted for all of your comments, and we provide a response to two of them.

> *Comments from recommender Damaris Zurrell*
>
> - L241-243: this sentence seems incomplete, especially with the inset "whose default prior", I suggest double-checking

We simply wanted to precise that we accounted for the default non-informative prior among the 4 prior we counted. We rearranged the sentence to clarify it.

> Review by Jonathan Rose
>
> Lines 270–272 – The statement here that the two informative priors perform comparably does not match what I see in Figure 4C and 4D. The estimates of p and wA appear to have much smaller bias (closer to zero) in Figure 4D with the highly informative prior than in 4C with the weakly informative prior.

We thank you to have noticed this ambiguity. We were thinking about the estimation of the occupancy probability Ψ when we talked about the comparable performance between the 2 priors, and not relatively to the estimates of p and wA. We edited this sentence to remove this ambiguity.

Download tracked changes file

Decision by Damaris Zurell, posted 27 November 2024, validated 29 November 2024

Thank you very much for this revised preprint and please accept my apologies for the delay in communicating the reviews and decision.

The two previous reviewers and I have carefully read through the revision and are all very satisfied with the changes undertaken and only suggest some minor edits. Overall, this is a highly relevant and interesting manuscript that provides important insights and clear recommendations to guide occupancy models based on novel community data and considering false positives. I am looking forward to the final version.

Minor comments/edits:

- Abstract/end of 2nd paragraph "Several model extensions aim to address these potential errors": consider editing to "Several model extensions have been proposed to address these potential errors"
- L37: edit "once data are processed"
- L72: I suggest a full stop after "performance metrics)." and start the new sentences e.g. with "In this case, eliciting informative prior"
- L110: consider adding a half-sentence that additional to poor quality there might be other reasons for misidentification of species, e.g. image classification errors for very similar species
- Fig. 2 caption: consider changing the caption to e.g. "Identifiability issues in Site Occupancy Model accounting for false-positive and false-negative errors in the identification layer". This way the main message of the figure would be clearer to the reader.
- L241-243: this sentence seems incomplete, especially with the inset "whose default prior", I suggest double-checking
- L279: edit "sources of errors"
- L342: edit "a parameter may cause bias"

Reviewed by Saoirse Kelleher, 04 November 2024

This manuscript has been much improved in this round of revisions, many thanks to the authors for their detailed responses to the comments from the previous round. The adjustments to the article's structure make it a much more cohesive read, and changes to the introduction and discussion clarify where this paper stands with respect to the occupancy modelling literature. In particular, the characterisation of the unique problems associated with novel sampling techniques like eDNA and autonomous recorders is much improved. Importantly, key points in the methods have also been clarified, particuarly around the Bayesian methods. At this point all of my substantiative concerns have been resolved, and I have only a few minor notes for clarity.

- It may be more clear if Figure 2 is moved below the 'identifiability issues' paragraph. In reading it, I personally (and inaccurately) connected this figure with the 'Classical estimation with a frequentist approach' paragraph. That's possibly just be my error, but perhaps worth considering if it may cause confusion for others.
- Just a couple of minor typographic issues I noted in the text:
  - L270: Duplication in *"... **mean mean** bias in the median ..."*
  - L336: Word missing in *"using an informative **[prior]*** is *necessary"*
  - L349: Double period at the end of *"noise generated during processing is essential**..**"*

Beyond those minor issues, I think this is a promising and important addition to the occupancy modelling literature and would be happy to see it published.

Reviewed by Jonathan Rose, 17 October 2024

Title and abstract
Does the title clearly reflect the content of the article? [X] Yes, [ ] No (please explain), [ ] I don't know
Does the abstract present the main findings of the study? [X] Yes, [ ] No (please explain), [ ] I don't know
Introduction
Are the research questions/hypotheses/predictions clearly presented? [X] Yes, [ ] No (please explain), [ ] I don't know
Does the introduction build on relevant research in the field? [X] Yes, [ ] No (please explain), [ ] I don't know
Materials and methods
Are the methods and analyses sufficiently detailed to allow replication by other researchers? [X] Yes, [ ] No (please explain), [ ] I don't know
Are the methods and statistical analyses appropriate and well described? [X] Yes, [ ] No (please explain), [ ] I don't know
Results
In the case of negative results, is there a statistical power analysis (or an adequate Bayesian analysis or equivalence testing)? [X] Yes, [ ] No (please explain), [ ] I don't know
Are the results described and interpreted correctly? [X] Yes, [ ] No (please explain), [ ] I don't know
Discussion
Have the authors appropriately emphasized the strengths and limitations of their study/theory/methods/argument? [X] Yes, [ ] No (please explain), [ ] I don't know
Are the conclusions adequately supported by the results (without overstating the implications of the findings)? [X] Yes, [ ] No (please explain), [ ] I don't know

I enjoyed reading the revised manuscript. The authors adequately addressed my comments on the original submission and I think the manuscript has greatly improved thanks to their revisions. Please see a few minor comments below.

Additional comments

Abstract, line 4 – Delete “with” so the text reads “suit specific sampling designs…”

Line 25 – Change “needs” to “need” to match plural data.

Line 236 – Change “diluted informative prior” to “weakly informative prior” for consistency.

Lines 270–272 – The statement here that the two informative priors perform comparably does not match what I see in Figure 4C and 4D. The estimates of p and wA appear to have much smaller bias (closer to zero) in Figure 4D with the highly informative prior than in 4C with the weakly informative prior.

Line 336 – Add the word prior after “informative”

Evaluation round #1

DOI or URL of the preprint: https://doi.org/10.1101/2024.05.07.592917

Version of the preprint: 1

Authors’ reply, 02 October 2024

Download author’s reply

Decision by Damaris Zurell, posted 17 June 2024, validated 17 June 2024

Revise

This is an interesting preprint discussing and testing methods for accounting for false positives in monitoring data. With the advent of automated monitoring methods (imaging, acoustics, eDNA), the problem of correcting inferred parameters and occupancy rates for false positives becomes ever more relevant. Thus, the study is very timely and can provide an important contribution to the current scientific literature on the topic. This is also acknowledged by two independent reviewers who have read the study with great interest. However, the reviewers also identified several issues that need improvement. They provide critical and very constructive feedback that I am convinced will greatly improve the manuscript and I highly recommend the authors to take the suggestions on board to the best of their ability.

Additionally, I would like to add a few points myself that mostly concern the overall structure and presentation. I am looking forward to seeing a revised version of this manuscript.

Abstract needs revising (also see reviewer comments): although the middle part provides a nice summary of the problem at hand, it takes too much space and the abstract is slightly imbalanced as the simulation study is only mentioned on the fly and no results and lessons learned (recommendations) are provided.

Introduction:

- the streamlining in the introduction could be improved to make very clear that occupancy-detection models have been around for some time but the problem of false positive (identification) problem is rarely addressed although becoming more and more relevant with the novel types of monitoring data. All this information is provided in the introduction but currently L7-25 read a bit like a history of occupancy models instead of clearly stating that this is the status quo but new data provide new challenges.

- L41-51 feel a bit out of place and I would recommend some restructuring with the previous paragraph to provide a clear story arc ranging from novel data peculiarities (image classification, acoustics, eDNA, etc) to mis-identification, which then can lead smoothly to models accounting for false positives (L54 onwards).

- in L56-72 the authors discuss more conventional solutions for addressing false positive like ”reference sites”. However, this part does not reflect on peculiarities of novel monitoring/sensor data that often rely on

AI or bioinformatics approaches for classification. As the motivation for this study hinges so much on novel sensor data, it would be helpful to also mention more explicitly here why these data in particular may have increased levels of false positives (see e.g. https://doi.org/10.1016/j.tree.2023.09.017).

- L79-83: as the study also uses a simulation approach to test the different approaches, this objective should be explicitly stated here.

Structure/Presentation:

As voiced by #1, the structure could be improved and clearer signposts could help directing readers. I want to echo this and encourage the authors to more clearly separate different types of sections, e.g. model formulation (or concept) from simulation study. The simulation study could also follow a more classical structure of first explaining the methods and then the results. However, with clearer signposts, the more narrative structure could also work.

- L135-225: these sections use very short paragraphs containing only 1-2 sentences. As result, the text may appear incoherent and like a list of unconnected throughts. Here, the text should be revised and coherence be improved, also better balancing paragraph size.

Figures:

Not all Figure captions are stand-alone, meaning they do not provide enough information to understand the content of the figure without consulting the manuscript text. For example, Fig. 3 caption does not explain the notations (e.g. w_A and p_hat), and does not explain ”with/without constraint” (similar issues with Figs. 1 and 4, and appendix figures). By contrast, Fig. 2 caption is much more complete and stand-alone. Please carefully check and revise all figure captions.

Reviewed by anonymous reviewer 2, 11 June 2024

Summary

The manuscript ‘Using informative priors to account for identifiability issues in occupancy models with identification errors’ introduces a new parameterisation of the single-season occupancy model with parameters for occupancy, detection probability, true-identification probability, and false-identification probability. The model is applied to simulated data in the maximum likelihood and Bayesian frameworks, with informative priors for true-identification rate used in the latter. The approach introduced here differs from other occupancy models extended to account for false positive detections in that it does not require additional datasets or a source of true positive detections in the data. As the authors note, this is particularly applicable to studies using camera traps, acoustic monitors, or eDNA sampling – all increasingly popular data sources for occupancy models.

General comments

Writing

Given the need for clarity in such a methods-focused paper, the writing in this article could be made more direct and concise – there are areas where meaning is difficult to interpret due to the phrasing. More specifically, many sentences are overly long with many commas and could easily be separated into independent clauses.

Abstract

I think that this abstract would be challenging to interpret for people who are not very familiar with the occupancy model and particularly its false-positive extensions. These models and their limitations in their standard form need to be more generally introduced. It may also be slightly too long, with some content more appropriate for the discussion.

Article structure

The layout of the manuscript, including headings, is unconventional and make it challenging to interpret. It is difficult to tell where the introduction and model formulation end and where the simulation studies begin, and within the sections 'Classical estimation from the identification layer (L135)' and 'Using an informative prior to address identifiability issues (L176)' the methods, results, and discussion components are heavily intertwined. These should be more explicitly separated; one example of where this occurs is L140-152, where elements of methods, results, and discussion occur in a confusing order.

Figure layout could also be improved (e.g. Figure 3 should be moved to the prior section), and references to figures in the supplement (as in L169) should more explicitly note when figures are supplementary. The caption for Figure 1 should also be expanded to include definitions of the included parameters, as it currently cannot be independently interpreted.

Occupancy models and passive sensors

At times the paper seems to overstate (likely inadvertently) the connection between occupancy models and passive sensor data. This is most prominent in the abstract and introduction (L1-25), but occurs elsewhere too (e.g. L97). More clearly stating a) what occupancy models are, b) how passive sensor data differs from other presence/absence data types, and c) what issues those differences produce could help to clarify this.

Model formulation

The parameterisation of the occupancy model is described nicely, but it should be more clearly delineated in the text where the model diverges from the standard occupancy model (c. L94). Describing the detection and identification processes in more general terms rather than with respect to passive sensors may also be appropriate, as there is little reason why this model could not also be applied to standard field surveys performed by humans.

Bayesian models

The methods for the Bayesian simulation study require more detail in the main text. Most importantly the priors used for occupancy probability, detection probability, and false-identification probability must be included. One of the biggest questions many readers will ask is how sensitive this model is to bias induced when the informative priors used are not appropriate for the data – this is mentioned in L76-77 but is not explored in the simulations nor further commented on in the discussion. This should be expanded upon, at least with respect to limitations in the discussion.

Conclusion

The conclusion does not feel specific to the model defined in this article. It could be more explicitly stated how this manuscript contributes to the management of the identifiability issues commonplace with passive sensor.

Detailed comments

Abstract: *"the naïve occupancy model does not account for false detection"*

Specify that this model does not account for false positive detection to increase clarity, as false negatives are accounted for.

Abstract: *"Overall, what is at stake is enhancing statistical methods together with sampling noninvasive technologies, in a way to provide ecological outcomes suitable for conservation decision-making."*

This sentence is a bit strangely phrased; 'what is at stake' could be replaced with "the objective of this article is to ..."

Caption for Equation 1: *"... formulation of the occupancy model (Royle and Kéry, 2007) ..."*

It should be noted that the citation for Royle and Kéry 2007 is for the multiseason implementation of the occupancy model, although this part of the formulation is the same for the single and multiseason versions.

L149 *"It has been showed ..."*

"it has been shown"

L193-196: *“We study 3 types of priors for parameter wA … We introduce 2 different prior distributions for the probability to correctly identify the species …”*
These two sentences are unclear and seem repetitive. The first says three types of priors are used, the second that two prior distributions are used. This could be simplified to reduce ambiguity.

L239 *“ … and above all they are not specific, …”*
This phrasing is somewhat unclear; maybe “are not species-specific”?

L253: *“… by reducing data processing time, potential identification errors are introduced …”*
Further elaboration required on how this may occur

L260: *“… we need feedback on the performance of the identification process …”*
‘Feedback’ doesn’t quite fit in the context – ‘information’ or something else may be more appropriate.

L266: *“involving inputs on the detection process in the form of a prior”*
There is somewhat more consistency needed on separating the ‘detection’ and ‘identification’ processes – in this article, only priors on the identification parameter appears to be included, not detection.

PCI Questions:

- Does the title clearly reflect the content of the article? **Yes**
- Does the abstract present the main findings of the study? **Yes**
- Are the research questions/hypotheses/predictions clearly presented? **Yes**
- Does the introduction build on relevant research in the field? **Yes**
- Are the methods and analyses sufficiently detailed to allow replication by other researchers? **No - the supplementary R script provided is sufficient, but the main text requires further details on all priors used for the Bayesian implementation.**
- Are the methods and statistical analyses appropriate and well described? **Yes**
- Are the results described and interpreted correctly? **Yes**
- Have the authors appropriately emphasized the strengths and limitations of their study/theory/methods/argument? **No - further discussion is needed on the potential implications of poorly-defined informative priors.**
- Are the conclusions adequately supported by the results (without overstating the implications of the findings)? **Yes**

Download the review

Reviewed by anonymous reviewer 1, 05 June 2024

Download the review