Marrot et al. 2023

2023-10-16

The manuscript focuses on spatially explicit modeling of the lifetime reproductive success (LRS) of a clownfish population at Kimbe Island, Papua New Guinea. Though I am not an expert on this topic, from the introduction I understand that such studies are mostly isolated to captive populations of organisms, whereas the current manuscript models LRS in a wild population. As a consequence, there are various components to both the ecological and sampling process that need to be accounted for with an appropriate statistical model, which here includes the presence of many zeros as well as a spatially structured sampling design. My background is a combination of formal education in statistics and ecology, so that I am not able to comment on many of the biological aspects of the study. Instead, I will focus on commenting on textual errors and outlining textual improvements, as well as provide my general thoughts on the modeling. Overall I think the authors have done a good job with the manuscript, and I commend them on their efforts, but there is room for improvement.

Besides some typographical errors and few sentences that I found difficult to understand, the study is reasonable well written. The methods section includes many important details for understanding the analysis, and though the results section is short it seems to include the necessary components. The discussion has a large focus on the spatially explicit modeling, and consequences of lacking to account for the spatial properties of the sampling process. This is briefly done by comparing the results of a spatially explicit model and a model that does not account for such properties. There is also a paragraph in the introduction with that angle, though the introduction is generally more biologically focused.

Overall, I do not agree with many of the statements made with respect to the analysis, nor do I find the analysis particularly novel (which, to be clear, is fine). Including spatial components in statistical models is mainstream in biology and many of its empirically focused subfields. However, the present study attempts to convince the reader that the modeling is groundbreakingly novel, complex or "state of the art". My primary recommendation is to shift that focus and to bring the study more in line with its biological narrative. The attempt to convince Biologists that not accounting for the consequences of spatially structure designs is also not new, and since it is based on a single study, and is not exhaustively discussed, distracts from the main narrative of clownfish lifetime reproductive success. In the discussion the authors state that the inference drawn from their study would have been negatively impacted if it had not accounted for its spatially structured design in the analysis, but insufficiently go into detail as to in what ways it would be different, for a study focused on the consequences of spatially structured sampling designs in evolutionary biology.

My second major comment relates to the difference between data and model. In various places the authors confuse properties of the data with (violations in) assumptions of their statistical model. In the results (for example) it is stated that the data follow a Poisson distribution where the sample mean of the data is used as estimator for the first moment of a Poisson distribution. However, in one of the following sentences the authors state that there are excess zeros relative to that same Poisson distribution. Generally, the authors use properties of the raw data to motivate their choice of statistical model, without verifying if those choices indeed are required due to assumption violations in the statistical modeling approaches that they apply. This is the case both for the zero-inflated component, as well as the spatially structured design. Hence, I would like to see more evidence of violated residual assumptions for a Poisson GLM (violation of non-independence of observations as well as deviation from Poissonanity), before the authors move to a spatially structured zero-inflated Poisson GLM.

Generally, the authors should be aware of the difference between calibrating a model for prediction, and calibrating a model for inference. A model that predicts well is rarely suited for performing inference.

While the authors seem to be interested in performing inference, the methods they apply are mostly suitable for performing predictions. AIC is a information criteria that finds models that predict well, and model-averaging is usually very suitable to find a model that predicts well but is rarely useful for inference (to support this statement I have also included a link to a recent preprint by Ben Bolker below, but I encourage the authors to further study this issue). My suggestion is to remove the model-averaging component from this study entirely; for inference the full model often has considerably better properties. This combines well with my suggestion in the following paragraph.

On a more technical note, the authors fit a GLM with eigenvectors from the neighbourhood matrix included as predictors. Consequently, the eigenvectors and associated slope parameters are treated as fixed-effects. Usually, unmeasured effects (such as spatial effects) are treated as random instead of fixed. A more appropriate approach might be 1) to treat the slopes of the eigenvectors as random effects coming from the same distribution, or 2) use a spatially distributed random effect instead (note: this is implemented for ZIP models in the glmmTMB R-package amongst others). Personally, I would advice the authors to go with the second approach as it allows for a more explicit relationship between distance of sampling units and counts (in terms of an autocorrelation function) and it circumvents the issue of sensitivity of the analysis to the number of included eigenvectors, which I consider an unresolved issue to the approach the authors have taken.

Finally, the authors mention that the dataset is the result of long-term sampling, but no other mention to a temporal component is made. It would be nice (but not completely necessary) if the temporal nature of the dataset could be elaborated on a little bit in reply to this review.

Detailed comments

l40: The first few sentences outline a knowledge gap, so I suggest to rewrite this sentence in a way that makes clear that this study addresses that.

l42: I suggest (in the entire paper) to avoid such vague references as “here” or “there” as it tends to be confusing to the reader. Explicitly name the object/place that you reference, please.

l44-46: This sentence is relatively ambiguous. Please improve along the lines of “However, habitat can in fact be understood as comprised of various components . . . , so that the exact driver of LRS is unclear”.

l48-49: “state of the art” feels like overselling what is in practice a spatially structured GLM.

l57: changes -> change.

l59: mean -> means, changes -> change.

l60: please avoid the use of “parameter” is a non-statistical context here.

l67: this statement implies in this article adaptive potential is “directly estimated”, but how exactly is not clear to me. If it is directly estimated, I would like to see that explained more clearly in the manuscript. If it is not, this sentence should be removed.

l75-82: good sentence.

l87: affect -> affects.

l89: I am not sure what maintenance means in this context.

l90: unclear what “most” refers to; most clownfish researchers?

l93: an odd place to mention “statistical framework” here, it feels again like overselling the modeling in this study.

l102: I am not sure what the authors refer to with “in this case”.

l103: here -> in this study.

l108: “a previous study”.

l113-116: this is an ambiguous sentence, please rewrite. For example, what is “relative detailed contribution”?

l125: “450ha” not information that is of interest here; the introduction can be relatively high level and broad.

l127-128: I consider pseudo replication and spatial autocorrelation as two separate issues. Indeed, both result in non-independence of observations, but that is about all that these issues have in common. The “Haining” reference is a whole book, please cite more specifically: is there a specific page or chapter that the authors want to refer to perhaps?

l129: “a ..24 studies”, I do not see the relevance of this information for the introduction.

l120-134: I find this a lot of text to say “we fitted a GLM”.

l213-214: this statement is problematic. There is a difference between distribution of data and the calculation of its sample mean, and the assumed distribution of a model and its estimator for the first moment of the assumed distribution. Please revise.

l214: is this “maximum” of 20 the possible upper bound for the number of self-recruits? I.e., is the Poisson assumption of an upper bound at infinity unrealistic for this application?

l215-217: the authors first state that the data follow a Poisson distribution, but then state that it actually does not (excess zeros relative to a Poisson process). Please correct this.

l217-221: the authors seem to confuse 1) excess zeros in the data, and 2) the need to account of those excess zeros in the model because the Poisson assumption is violated as shown by (e.g.,) residual diagnostics. I would advice the authors read Warton (2005): Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data, and rethink this paragraph. I would also request that they include residual diagnostics (with randomized quantile residuals, the DHARMa R-package will return this for ZIP models fitted with glmmTMB).

l221-223: I do believe that is true, but AIC finds the model that predicts best, not the model that is best for inference or provides the most valid inference. Hence, I do not find this statement very convincing and would advice the authors to take a difference approach to motivate their choice of a zero-inflated model, which is more aligned to their goal of ecological inference instead of prediction.

l239-240: the authors at length discuss the repercussions of fitting a non-spatial model to data that suffer from spatial structuring, but include non-spatial models in their model-averaging. I suggest to rethink this.

l252: I am a fan of such a “simple” approach of just visualizing the distribution of counts to convey the issue of spatial structuring, but from fig 1A it is not apparent to me that this dataset indeed suffers from spatial structuring. What I would expect to see is that points more closely together have higher counts than points further away. Generally, I do see a divide between the left and right sides of the figure; the counts on the right side seem higher, but it would be good if this is further elaborated on by the authors. Generally, Figure 1A misses scales; it is difficult to read (surrounding geographical context is missing).

l252-255: OK, the figure is clear but it is not clear what model the RHS is retrieved from, and thus does not clarify if the spatially explicit terms in the model are necessary.

l263: I am not a fan of abbreviating terms unnecessarily, or in ways that are not frequent in the various branches of ecology or statistics, as here. It does not improve the readability of the text, in fact it (usually) makes for a text that is harder to read through, espeically when used in combination with other abbreviations that the reader might be unfamiliar with (for me here: LRS). A 11 letter word is abbreviated to a 3 letter word, and one that is only used in a limited section of the text (mostly the methods). I suggest to use “eigenvectors” as usual.

l278: covariate -> covariates.

l289-290: the authors do not state how they calculated Moran’s I. I suppose this was on the residuals of some model (not clear which model from the whole subset of models, or if it is from the final model-averaged solution), but which type of residual is not stated.

l293: as I think I understand from another part in the text (I admit, I may misunderstand), the “spatial model-averaging procedure” included models without eigenvectors from the neighborhood distance matrix. This seems odd, as the authors (at length, in the introduction) describe the fact that biases might be introduced to estimates when not accounting for spatial autocorrelation in a model when that is required. In essence, the authors use a method that they admit in the manuscript to be biased.

l303-306: I have no idea what these sentences mean or are trying to convey. I would suggest an alternative, but struggle with that too. So, I broadly suggest for the authors to rewrite these sentences.

l318-319: This interpretation is incorrect; the Poisson process can also generate zeros, so this statistic that is retrieved from the zero-inflated component of the model is correctly interpreted as “the probability that fish will not produce a self-recruit not due to the Poisson process” or similar.

l327-328: this is a sentence for in the methods.

l335: remove “however”; the meaning of the sentence remains the same.

l344-345: It is not clear to me how the inference in this manuscript would be different if the authors would not have included the spatial terms in their model. This statement is probably true, but the authors could try to better connect it to their own results.

l363-364: This is not a convincing statement. There are few ecologists/evolutionary biologists or statisticians that would be surprised that including 29 additional variables in a model explains more variation.

Appendix

Figure 1: these are not statistical distributions (densities) these are histograms of data.

Zenodo repository

- Available in review portal under “data for results” but does not seem to actually include data
- Please include scripts as R files, not as text files
- The data availability statement notes that Rdata files are available in the zenodo repo, yet all I can see is scripts and the manuscript
- Please provide scripts that use the RData files instead of the raw data, if the raw data will not be deposited anywhere
- Please trim the package lists; lme4 is loaded (and lmerTest) but not used in the script as far as I can tell
- Note that although some model objects are called “glmm” these are in facts glms without random effects
- try to run: does it work? file paths anywhere? a-b runnable?
- self-contained?