Cazalis et al. investigate the effectiveness of protected areas on birds in North America using a large-scale citizen science monitoring program. I read this manuscript with interest. The authors address a relevant question related to the effectiveness of protected areas in conserving North American breeding bird assemblages, and for this purpose they use a relevant approach that differs from what has been used mostly in previous assessment studies.

Although I am confident about the relevance of this manuscript, I think there are some issues to sort out, and clarifications are needed throughout the manuscript. I also have several methodological concerns regarding this study, and I think this may be addressed before being further considered for publication.

> it seems to me that the proportion of sites under protection is relatively low compared to the proportion of sites under no protection status, and even lower when considering sites according to main habitat (main vegetation structure). I would expect this type of bias in sampled sites to obviously lead to a lack of power, and potentially of significancy, in the analysis of PA effectiveness. Also, I am wondering why the authors did not consider the major habitat type based on the proportion of land cover classes within buffer surrounding the entire routes? it would increase the sample size of PA sites. If the authors were worried about the representativeness of a single major habitat across the entire BBS routes, they could have considered using the shapefile provided with BBS data '*bbsrte_2012_alb*' (which corresponds to polylines) to split entire BBS routes into segments (or sections, to use the same terminology as in the manuscript) and then assigned major vegetation structure and PA proportion to each of these segments, instead of using only small sections of BBS routes.
See my detailed comments below.

> Although I get the purpose of using a phylogenetic regression - especially in the context of this study - I think this should be clearly justified (e.g. to account for potential phylogenetic relatedness among species). Also, I am wondering why the authors implemented two different types of phylogenetic models (Brownian and Lambda). I did not find any information in the ms. that could justify the use of multiple phylogenetic models… Justification is needed. Finally, as recommended by Ives & Garland (2010), PLR model coefficients were estimated using bootstrap. see my detailed comments below.

DETAILED COMMENTS

L78: compare to what? inside vs. outside PAs I guess… Please specify

L132: "all three meta-analyses": which one? you change paragraph, so it would help to cite the studies again, or specify something like "discussed above", to help the reader.

L159: "positive effect of PAs" - what kind of effects? provide examples. Given the studies cited, it might be relevant to specify those are mostly designation effects…

L184: "effectiveness to select the most interesting sites" - it does seem to me that the designation effect that this sentence refers to, is mentioned for the first time here or at least not clearly stated/ defined in the Introduction. This should be better defined in the Intro. Indeed, the expected differences between protected and non-protected sites may result from either the 'designation effect' (the initial state of biodiversity) or from the 'protection effect' (protection efficiency per se). On the first case, it will be assess through a comparison of the bird assemblages (or biodiversity) inside and outside the protected sites, and on the latter case, it will require the comparison of the state of the response variable before and after the designation of PAs… Given the purpose of this study, I think this would be worth clarifying this in the Introduction.

L187: I didn't find any mention about the data from first-year observers. How were BBS route-year data from first-year observer treated? Same for data collected during poor weather.
According to Kendall et al. (1996), and following other studies using NA BBS data (e.g. Huang et al. 2014; Wood et al. 2015), records surveyed by first year observers should be removed to minimize observer bias.

L191: “25-mile routes”. I suggest add or replace by the correspondence in kilometers.

L195: Given the landscape characteristics across North America (often one dominant land cover type over broad extents) I am wondering whether this represent a high proportion of routes, i.e. those with multiple land cover types.

L213: where does this threshold of 5 years come from? did you perform sensitivity analyses to test this? or was it an arbitrary value?

L241-242: what is the sample size? how many sites classified as within PAs and how many as outside PAs? Looking at the map presented in Appendix S1, it seems there are very few site located in protected areas overall… what’s the proportion of these within each habitat type? this disequilibrium in the number of BBS routes considered as “protected” vs. “unprotected” may affect the analyses.
Also I am wondering what would be the proportion of BBS routes with PAs cover > 50% if the routes were considered in their entirety…
The authors may want to have a look at the paper from Wood et al. (2014), as the methodology to define protected sites differ and as such, the ratio protected/unprotected sites does not seem to match (at least the number of protected routes seems way larger in their study).

L256: I have some concern regarding the temporal mismatch between the bird data and the land cover data. How this delay of ca. 10-15 years may not affect the results? I would suggest using the MODIS-based GLCC instead, for a better match in the temporal extents of the data sets…

L270-271: argument and reference towards the biases related to altitude and productivity were already presented in the intro, no need for this here.

L291: “fully” vs. “unprotected route” - what does it mean exactly? 100% vs. 0%? Please clarify.

L305: Were all the BBS routes considered to calculate this human-affinity index, or only the subset of the routes’ sections further considered for the analyses? in the latter case, this may bias the estimation of how sensitive species might be to human disturbance if all survey routes are not gathered to provide such estimate (?), and as such this may not be totally representative of the habitat preferences or sensitivity of each species to human disturbances… This should be at least discussed.

L308: What is the purpose of testing these two types of phylogenetic models? Would the Brownian motion model not be enough to test for an effect of the phylogenetic relatedness among species on the response to PA?… Please explain

L310: what kind of trees? (e.g. primary backbone tree of Hackett) Please specify.

L311: why choosing to use this specific method (maximum clade credibility) to summarise phylogenetic trees, instead of another common method such as using a consensus tree ? both methods can be criticized, and depending on the case, they may end on very similar results. However, depending on the number of trees selected, and then depending on the shape of the posterior density, you may not sample the highest probability topology… A consensus tree could have a higher credibility than any of the trees in your selection. Do you have any thought, additional information about this?

L314: if neither a table nor a plot are presented as results, I think estimate coefficients would be appreciated…

L350-L353: this corresponds to result interpretation and should be more relevant I think in the Discussion… To counteract this lack of significance, I think it would be worth considering the entire BBS routes, as I mentioned above already. This lack of significance (and power regarding the analyses) could in somehow raise questions about the robustness - and the relevance - of the results presented here.

L423-424: and why is this so? what's the difference between these two phylogenetic models that could explain results differ… this echoed what I am saying above, and ask for more details about these two models and why they are both considered in this study.

L427-431: I don't think this is needed here. It is all "theoretical" (would not require the exact same model), and not even feasible/done through this study. Answering this type of question would rather require to consider analysing a metric such as the Phylogenetic Diversity (e.g. Zupan et al 2014; Thuiller et al. 2015). I would suggest removing this section, but I leave it up to the authors…

L432-433: in other words, the species difference in PAE_{For} relies on phylogenetic relatedness among species

L502-505: maybe considering eBird data for example could be useful, even as a joint analyses of both data sets…

WRITING

I would suggest maybe asking for an external English proofread, or at least a very careful reading of the manuscript. According to me, some pieces of the text could definitely be made simpler and easier to read.

Need to be consistent in the use of present or past tense throughout the manuscript, and even when referring to the study.

L16: I would suggest rephrasing as "using neighbouring sites to protected areas as counterfactuals"

L27-28: remove "area the one avoiding human activities"

L80: "populations inside these areas" instead of "in them"

L102: "act as"

L104: please rephrase: "to the former if it had not been protected"

L120-123: need to rephrase. I suggest "Indeed, ignoring this could lead to comparing sites that are not expected to have similar biodiversity regardless of their protection (e. g. protected grasslands or unprotected forests), without neglecting the effect of protected areas by preventing habitat changes (e. g. deforestation or urbanization)."

L182: "we use the term"

L185: "in creating" or "to create"

L469: "compared" - past tense

L482-484: this sentence is not clear to me… "in relation to pairwise comparisons"??

L486: I suggest rephrasing: "our results emphasize that it is impossible to clearly measure the effectiveness of PAs in conserving species diversity without defining precisely what is expected of them"

L500: "in North America"

Appendix S1: "herbaceous" in the legend

Suggested References:

Huang Q, Swatantran A, Dubayah R, Goetz SJ (2014) The Influence of Vegetation Height Heterogeneity on Forest and Woodland Bird Species Richness across the United States. PLoS ONE 9(8): e103236. doi:10.1371/journal.pone.0103236

Kendall WL, Peterjohn BG, Sauer JR (1996) First-Time Observer Effects in the North American Breeding Bird Survey. The Auk 113: 823–829.

Thuiller W, Maiorano L, Mazel F, Guilhaumon F, Ficetola GF, Lavergne S, Renaud J, Roquet C, Mouillot D. 2015 Conserving the functional and phylogenetic trees of life of European tetrapods. Phil. Trans. R. Soc. B 370: 20140005.

Wood, E. M., Pidgeon, A. M., Radeloff, V. C., Helmers, D. , Culbert, P. D., Keuler, N. S. and Flather, C. H. (2014), Housing development erodes avian community structure in U.S. protected areas. Ecological Applications, 24: 1445-1462. doi:10.1890/12-1992.1

Wood, Eric M., Anna M. Pidgeon, Volker C. Radeloff, David P. Helmers, Patrick D. Culbert, Nicholas S. Keuler, and Curtis H. Flather. “Long-Term Avian Community Response to Housing Development at the Boundary of US Protected Areas: Effect Size Increases with Time.” J Appl Ecol, July 2015, n/a–n/a. https://doi.org/10.1111/1365-2664.12492.

Zupan L, Cabeza M, Maiorano L, et al. Spatial mismatch of phylogenetic diversity across three vertebrate groups and protected areas in Europe. Divers Distrib. 2014;20(6):674-685.