Decision for round #2 : *Revision needed*

Dear authors,

Your revised manuscript has now been evaluated by the two previous reviewers and they are overall very happy with your revision. Once you have adressed their minor suggestions, I will be happy to recommend your preprint.

In addition please address my following minor comments:

1) When running the code "EAS_bird-v1.1/m-rieger-EAS_bird-f1e6cc6/EAS_bird.R" I get following error message by the end: "There is no final model for 'Black Redstart'"

Thanks for pointing that out! There were some discrepancies whether the final model should be labelled as TRUE or 'yes', we corrected that and uploaded a new version.

2) Line 687: change "re_formual = NA" with "re_formula = NA"

We changed that.

3) Lines 800 and 1548: isn't it four, not five, species?

That is correct, this was a leftover from our earlier manuscript version. We corrected that.

I look forward to receiving your revision,

Matthieu

Reviewer 1

Most of the concerns with the procedures of the previous versions have been addressed and improved. In general, this is a thoroughly conducted study where a lot of effort have been put into trying to deal with the challenges of the data. While it is not so clear that the components of the analysis are particularly novel (except perhaps for the specifics of the treatment of observer effects), it is an overall nicely worked case study for a fairly complex data set.

Thank you for your feedback and the time you put into the review! Your thoughts and suggestions were helpful and we hope that we were able to incorporate them satisfactorily.

L125-129. This suggests that the biogeographic regions are not used which could lead to the question of why you mention them here. You state later how you use them, but would be good to give a hint here.

We changed that paragraph accordingly.

L146. 2021 should be 2011?

Correct, thanks for spotting that.

L200. What basis dimension did you use for the smoothers?

We used the default basis dimensions in brms/mgcv which are k = 10. We added this information.

L210. Why do you think that landscape type is the main driver of difference in trends? I.e. why does landscape type “account for spatial bias”?

We expect different trends between natural regions, since these regions are differently affected by changes in habitat quality due to climate change and human intervention. We account for the unbalanced sampling in the post-modelling process (not in the model per se). We changed that paragraph accordingly.

L235. With only 2-4 observations per site, it would be next to impossible to provide a meaningful estimate of autocorrelation.

Thank you, we added this argument to the paragraph.

L244. Should explain that tree depth refers to the Hamiltonian Monte Carlo algorithm in Stan.

We added this information.

L257. What was the computational cost of estimating the models in general, and particularly running the cross validation? Could be useful to know for readers considering an approach like this.

The computational cost is indeed quite high. One model per species (without cross-validation) took approximately 1-5 minutes (depending on the model structure and complexity of the species abundance distribution), whereas the cross-validation highly increased the runtime, lasting between 10 to 40 minutes per model and species. We used a computer with 16 threads, running either 4 chains with 4 threads in parallel (model), or 4 models each with four chains and one thread per chain (cross-validation). We added a remark.

L295. This choice means that the uncertainty of the index is not an uncertainty of change. In the panels in Fig 6 the uncertainties (blue and yellow) are therefore not comparable, they mean different things.

That is true, we added remarks in the paragraph and in the figure caption of Fig. 6 to point that out.

Fig 5. It seems like more surveys are being flagged as negative compared to positive. Could imbalance of the flagging affect the corrections?

In fact, there are even slightly more positively flagged surveys than negative (79 vs. 72, compared to 497 flagged as ‘none’). Given these rather balanced values we don’t expect unwanted effects on the (direction of) correction.

L420-423. It is not just the choice of baseline, the smooth trend does not capture difference in abundance between the first few years and later years.

We corrected the paragraph accordingly.

L471. How do you know it is effective? I.e. that there are not other spatial causes of bias that your model is missing.

That is correct. We changed the sentence accordingly.

L487-492. This begs the question of what the reason for the change in the observer effects may be. You addressed this in the response, but it seems like important information to add here.

Thanks for pointing this out. We added this information at the end of the paragraph.

Reviewer 2

General comments

I really appreciated the comprehensive and complete answer of the authors, I would like to thank them for the time they dedicated to answer my comments and the serious of their work. I think the modification done by the authors have improved a lot the quality of the manuscript. I don't have major comments left that I did not raise. However, I would like to bring last thoughts about the observer effect that author used.

Thanks to the authors for their explanations in their answer that have clarified my mis-understanding of their observer effects construction. I still think that this index might capture other effects than the observer effects. Some years can be exceptionally bad for bird reproduction in general, often due to climatic factors. The observer effects proposed here could remove these "bad years" from the population trend and absorb them, while their increasing/decreasing frequency would be very informative and should be included in the population trend. Could authors justify with some supplementary figures the threshold of 25%? How variable is the summed abundance across years for a given site? Is there any sites followed during several years by the same participant? If yes, these sites could be used to analyse the expected variability across year.

The figure 5 shows that there is an important temporal shift in observer effects from negative to positive, how can authors explain this? Did participants change behaviour? Did the kind of participants change over time? This is likely to affect the resulting population trends.

The editor and authors will decide if it is relevant to discuss again this, but since authors aim to propose a general method to analyse bird counts data, I think the solution they propose should be adapted to most of the possible datasets, and I am not sure this observer effect is.

In addition to this comment, the manuscript would need some minor modifications to be ready for publications: figure caption could be more detailed, authors used a lot of acronyms, homogenisation of the terms used to describe variables ("regions", "landscape", etc.). Also, I did not miss anything I think there is an error in the formula of the population trend (see my comment below about line 695).

Thank you for your valuable feedback and thoughts on our revised version! We hope, we satisfactorily addressed all your concerns. You will find answers concerning your general comments in the 'specific comments' section.

Specific comments

Lines 131-132: "Moreover, TRIM is restricted to categorical covariates, requiring climate or landscape composition covariates to be transformed into categories (Bogaart et al. 2020)." If I got it well it is now what authors do also (landscape is a categorical variable), so it is a common limit between TRIM and their model. For this reason i would raise that point in discussion rather than in introduction.

This might be a misunderstanding due to our use of the terms 'landscape', 'habitat' and 'region', which we now reformulated to avoid confusion (see also our answer for line 218). We here refer to landscape composition data which we integrated in the PCA, as explained in section 2.2. The resulting PCs are continuous covariates, which can be integrated in our model approach.

Line 181: results are given for both regions together (25 to 36) but the maximum number of sites was given per region (43 and 47), which is a bit confusing.

Thanks for pointing this out. The maximum values referred to the two topmost maximum values which were rather high and different from other years (for both regions together). We see that this might be confusing and now only kept the topmost maximum value 47.

Lines 190-194: I think I got what authors did here but it took me a bit of time. The verb "merge" is not really explicit. I think the lack of clarity comes from the superposition of old and new methods version. The biogeographic regions occupy a big place in Figure 1 and beginning of the methods but they are not used in the analyses. This is confusing. I would rewrite the methods and presentation of the dataset focusing on variables that will be used in the analyses. For example, here I would say something like: "Site were distributed into six natural regions (Fig. 1) and located or not in a Metropolitan area. We combined these two layers of information to produce a landscape categorical variable with 7 levels, all sites within the metropolitan area being classified as "urban" while other were classified according to the natural region they belong."

We changed the paragraph and Fig. 1 accordingly, so the main focus is now on the natural regions including the metropolitan area.

Lines 218: Here landscape is re-used again for other variable than the one defined line 192. I would avoid that and use another word for one of them. Either stick to natural region in the first place, or use "habitat" in the second place.

Thanks for pointing that out. We now renamed the combined natural region and metropolitan information as 'natural regions'. The term 'landscape' is reserved for landscape composition data (used in PCA) and 'biogeographical region' for the two biogeographical regions atlantic and continental. We adjusted the whole manuscript accordingly and tried to avoid speaking of 'regions' in general, since that might cause confusion which region we are referring to.

Figure 2: Caption says biogeographical region while actual what is plotted is the “landscape categorical variable” defined by authors in lines 192-194. Also, acronyms should be defined in the figure caption. The abbreviation for Continental is still in German (kon).

The term ‘biogeographical region’ is here correctly used since it refers to the right panel, where natural region shares are plotted per biogeographical region as well as NRW. We added acronym definitions and changed ‘kon’ to ‘con’ (the same accounts to Fig. 4, which also had the German abbreviation).

Lines 280-301: Thanks to the authors for their explanations in their answer and the details about that observer index. I still think that this index might capture other effects than the observer effects. Some years can be exceptionally bad for bird reproduction in general, often due to climatic factors. A deviation from 25% of the mean could thus be expected? If yes, the observer effects proposed here would remove these “bad years” from the population trend and absorb them, while their increasing/decreasing frequency would be very informative and should be included in the population trend. Could authors justify with some supplementary figures the threshold of 25%? How variable is the summed abundance across years for a given site? Is there any sites followed during several years by the same participant? If yes, these sites could be used to analyse the expected variability across year.

It might be that the datasets authors analyse here do not present these variations, but since they aim to propose a general method to analyse bird counts data, I think the solution they propose should be adapted to most of the possible data.

Thank you, we appreciate your concerns and thoughts!

Generally, we consider a decline (or increase) of a whole bird community by 25% within 6 years a possible, but extremely rare scenario (e.g., after major habitat deterioration). We are therefore confident that the vast majority of such cases in our dataset represent “true” between-observer variation rather than hidden true abundance shifts. Also bear in mind that our method does allow quite substantial variation in total abundances between years without penalization because also the extremely low (or high) abundance values are part of the mean total abundance calculation of the given site that is the basis for our observer effect correction. Since the current number of repetitions per site span just between 2 and 4, the low (or high) annual abundance has a high influence on the total mean. Thus, a single year has to deviate by even more than 25% from the mean total abundance of the other years to be flagged as “extreme”. These aspects are already addressed and mentioned in section 4.3 in the Discussion.

Note further that our approach is not removing possibly “bad” years from the calculations, nor are they artificially set e.g. to some mean value. Yet, our approach buffers to some degree against excessive positive or negative effects on the overall trend calculation. Finally, the detection of solitary exceptionally good or bad years – as possible in monitoring programs with yearly surveys, so that random intercepts per year can be calculate – is not possible with our smoothing approach in the first place.

To make the observer effect calculations more transparent, we now added another supplementary figure (Fig. 4.3), showing how observer effects vary between years, the total number of previous surveys an observer already did in EAS irrespective of the given site (as a measure of experience with the programme) and how often a site was surveyed by the same observer (as a measure of familiarity with the site). Note that we have an extremely low share of sites that have been visited repeatedly by the same observer.

The figure 5 shows that there is an important temporal shift in observer effects from negative to positive, how can authors explain this? Did participants change behaviour? Did the kind of participants change over time? This is likely to affect the resulting population trends.

Thanks to you and referee 1 for pointing this out. We added our answer from the last revision to the first paragraph of the discussion about observer effects.

Line 693: "Based on pairwise differences of simulated abundances…" Do authors mean predicted abundances?

Yes, that is correct. We changed the sentence.

Line 695: I f I did not miss anything this formula means that the trend is a difference of density? So, it is not a decline/increase per unit of time? Most often to calculate a trend we divide by the duration of the temporal window $(Nij - Ni(j - \Delta t))/\Delta t$. Authors could also calculate growth rates, that is a bit more meaningful for comparison among species that do not have the same abundance:

$100 \times ((Nij/Ni(j - \Delta t))^{(1/\Delta t)} - 1)$.

The formula calculates the absolute change in abundance per km² across 12 years, so it is a decrease/increase per time (namely, across the 12 years and not per year). We already added a note to the respective figures to avoid this pitfall, but we agree that this might lead to confusion. We now divided the difference in abundance per km² by the temporal window, following your first suggestion, and updated the respective figures (Fig. 4 and Fig. 5), so that trend estimates are now given as mean abundance change per km² per year.

Line 754: is that acronym (DDA) used after that? The authors defined a lot of acronyms along the methods and could try to limit this number. If an acronym is not used afterwards, there is no need to define it. Otherwise, the reader might try to keep all of them in memory thinking that it will be useful for latter and waste focusing abilities.

Thanks for pointing this out. We used the acronym 'DDA' only once afterwards, therefore we replaced it with its full name and deleted the acronym.

Figure 4: it would be nice to have the unit of the population trends on the y-axis. Especially because the formula used in the methods is not clear about how authors calculate the trends.

See comment above, the figure now displays mean annual changes in abundance per km².

Figure 6: I think figure caption could be a bit more explicit for the reader, and could be used as reminder of the meaning of the acronyms, for example, instead of "correlation of annual indices…" authors could say "Correlation between our annual index of abundance (EAS) and the one form the German Common Bird Monitoring scheme (Mhb)…"

We added remarks in the figure caption (first sentence) to point out that EAS is based on ‘our data’ and MhB used for comparison.

Figure 2.1 of the supplementary materials: Authors present the coefficient associated with the effect of PC1, PC2 and PC3. However, interpreting a polynomial function from its coefficient is extremely hard. Even if the polynomial is only of order 2, I would instead or in addition of presenting the coefficient, present the predicted polynomial in itself (abundance as a function of PC1, for average level of other variables), it would be more informative. It would also be informative to remind the reader which variable compose essentially each of the PCA axes.

We agree that the interpretation of linear and quadratic terms might be difficult. We added further descriptions to the figure caption and a new figure showing predicted abundances per PC score, as suggested (Fig. 2.2 in supplement).