Referee: 1

Comments to the Author(s)

Dear authors,

I applaud your intention of showcasing the use of data imputation methods and Bayesian phylogenetic comparative analysis. I agree that these methods have the potential to improve investigations into trait evolution and that researchers should be made more aware of them and consider adding them to their tool kits. However, I think that the methodological pipeline that you propose is missing some key elements that mean that it does not currently provide the "solid basis for an improved approach to comparative phylogenetic studies" [L392] that you envisage. To inspire others to adopt these methods (and convince them of the validity of your results) you need to set an example of best practices and provide more of an orientation to the methods for people who are not familiar with them.

van Buuren (2012) provides reporting guidelines for data imputation methods. These are described in Nakagawa, S., 2015. Missing data: mechanisms, methods, and messages, in: Fox, G.A., Negrete-Yankelevich, S., Sosa, V.J. (Eds.), Ecological Statistics: Contemporary Theory and Application. Oxford University Press, Oxford, which is a book chapter that is well worth reading if you can access it. It provides a more extended discussion of the subject of data imputation than the Nakagawa & Freckleton (2008) paper that you cite. You have satisfied many of van Buuren's reporting guidelines, but certain key ones are missing.

We thank the reviewer for the important issues raised and we have addressed all of them comprehensively withing an extended supplementary material. We have modified the manuscript to accommodate these recommendations and we are convinced that the paper is of a better quality due to all the recommendations provided by the referee.

Firstly, the imputation of missing data should be preceded by a detailed assessment of missingness in the data set, to understand exactly where the missing data are, whether they are clustered in any non-random way, what the mechanisms driving the missingness may be, and what biases would actually be introduced if complete case data were used. This is important so that researchers can understand exactly what effect data imputation will have on their data set, and whether it is appropriate. It may not be if, for instance, large portions of particular taxonomic groups would end up being represented only by imputed data. I do not agree with your suggestion that data imputation "is an approach that unequivocally can be useful in any comparative study" [L373] - researchers need to take steps to understand exactly what effect data imputation will have and whether it is appropriate to use on their particular data set. I would have expected to see such an assessment of missingness in your paper, along with figures showing the detailed distribution of missingness in the data set and across the phylogeny used (to see if missingness is clustered in particular taxonomic groups), and perhaps also correlation plots showing the relationship between missingness and the variables in the data set, to see how they may be interrelated. The histogram showing overall missingness in each variable is insufficient. Perhaps the heatmap of the "Pattern" of missingness would provide more insight, but you do not provide a caption with the figure to explain what it is showing. You may find the {naniar} R package useful for conducting a detailed assessment of missingness in your data set, as well as the miss.phylo.d function from the {sensiPhy}package to evaluate whether missingness in the data is distributed non-randomly across the phylogeny.

We have added detailed assessment of missingness to the supplement with multiple graphs. We reported the missing data per geographic origin and more clear presentation of the % of missing cases. We also added a graph presenting collinearity between missing data. We have incorporated the suggestions from the reviewer along with some of the suggestions from other papers. [1]

As to why we use multiple imputations and not full cases only:

“from a theoretical point of view, multiple imputation is *always* to be preferred over listwise and pairwise deletion, and that reasons of researchers to prefer listwise deletion are based on misunderstandings about multiple imputation.” [2].

Another paper confirming that imputations are always to be preferred to complete cases only - [3]

Also missing was a clear description of the process and model used for the data imputation itself. The description of the process [L172-178] was opaque to me, and the model is only vaguely described when you say that you used as predictor variables the 13 variables in the data set with less than 35% missingness [L195-196]. Incidentally, by saying that you used as predictor variables the 13 variables with less than 35% missingness, does this mean that brain size was included as a predictor variable in the data imputation process? I believe that it is not good practice to use what will be the dependent variable in a subsequent analysis as part of the imputation procedure, as it could introduce circularity into analyses.

We also included a more detailed description of the imputation process in the main text and a bit more of the rationale behind the usage of multiple imputations. Also using the dependent variable for imputations seems to follow the good practice for multiple imputations and also, disregarding the % of missingness, multiple imputations seem to work well. [4]

“What researchers holding this misconception do not realize about multiple imputation is that the model used for multiple imputation is not meant as a conceptually meaningful model. Multiple imputation is only used to accurately describe the relations and structures found in the data, and impute data with similar properties. As long as a variable correlates with another variable with missing data, it is a potential candidate for a predictor in the imputation model. It does not matter that we are not interested in the prediction of one variable from the other in the subsequent analysis, or that this prediction is even nonsensical. All an imputation model does is (a) determine that in general a high age coincides with a high income, (b) when age is missing for someone with a high income, infer that this person’s age must probably be high as well, and (c) consequently impute a high value for age.” [2]

A crucial next step after imputation is to evaluate the validity of the resulting data. You say that the results are "usually realistic" [L184], but there should be a formal validation procedure. For instance, a random sample of the imputed data could be compared to qualitative descriptions of species, to check that the estimations are plausible. A sensitivity analysis should also be conducted to ensure that the imputed data were not biased by particular characteristics of the data set, phylogeny, or imputation method used. An analysis of the complete case data would also usually be presented alongside the analysis of the imputed data, so the effect of the imputed data on the statistical models and overall findings could be evaluated. Again, however, all these things are absent.

We performed a careful visual inspection and added the following to the supplement:

“Visual inspection and evaluation of ‘realistic’ imputation values of numerous species in all 25 imputed datasets was performed at random, and the imputed data seems to be realistic. Due to the nature of the missing data, many of the traits are difficult to compare with ‘realistic values’ (i.e. metabolic rate, play behavior). Due to the nature of multiple imputations, many values (including categorical variables) were different in different imputed datasets i.e. the same species was imputed as, for example, arboreal in one dataset, but as terrestrial in another.”

Although we added a complete case PGLS analysis to the supplement, there are a few papers that unequivocally show that multiple imputations solve most of the problems with missingness (disregarding whether it is missing at random, missing completely at random or missing not at random) and are always to be preferred to deletion and usage of complete cases only. [2, 3]

More generally, I think that you need to provide a more extended discussion of data imputation to familiarise people with the subject. In relation to this, I was surprised that you did not discuss Penone et al. (2014). Imputation of missing data in life-history trait datasets: which approach performs the best? Methods in Ecology and Evolution 5, 961–970, as this would seem to be important background to your argument for the adoption of data imputation methods. Such a discussion should also make people more aware of the things they should be considering when deciding whether to use data imputation in their research because, as I said above, it should not be applied unthinkingly.

We extended the paragraph on multiple imputations in the discussion. As the imputations are not the sole focus of the paper, we cannot devote more than a single paragraph on this technique. We are using an approach called phylogenetic multiple imputations which is quite different to the ones discussed in Penone (2014). We have added an explanation how phylogentic non-independence is considered during the multiple imputation process, but unfortunately the paper describing the method is not published so we could not provide a reference to the comprehensive description of the method.

As with the data imputation section, I also feel that elements are missing from your description of the MCMCglmm analyses. As with the imputation model, you do not explicitly describe the specification of the models that were analysed, and only do so informally in lines 251-264. It is not even clearly stated how brain and body size were included in the analyses, e.g., whether body size was included as a covariate or whether you used residuals from a brain-body regression. One must look in the code provided as supplementary material to find clear information about the models, but it should be front and centre in the paper.

This is an important remark and we have included full model specifications to the supplementary material and mentioned the general model structure in the main text.

The extra space that would be needed to add in the missing elements could be gained by removing the analysis of the mode of evolution in different radiations. Currently, this does not seem relevant to the main subject of the paper - investigating the drivers of mammalian brain evolution. It feels like it should be its own paper. If this is not appropriate, then I think it needs to be weaved together with the other analyses to a greater extent. At the moment, the only link made between the two that I can see is in lines 318-319.

We have the paragraph linking the two analysis in the main text.

My final comment is not about any missing methodological elements, but about some confusion I have about the implicit hypothesis of your study and your conclusions.

As I understand it, you say that the principal benefit of studying the causes of brain evolution in marsupials is that marsupials exhibit little variation in reproductive traits, meaning that they provide a sort of natural control that can help determine whether the correlations between socioecological variables and brain size found in placental mammals are simply due to these variables being correlated with reproductive variables, which are then in turn correlated with brain size. The hypothesis is that, if this is indeed the case, then in marsupials, which exhibit little variation in reproductive traits, there will be no correlation between reproductive traits and brain size, and this will "break" the correlation between socioecological variables and brain size. However, you find a correlation between a reproductive trait and brain size. This suggests to me that there actually is significant variation in some reproductive traits across marsupials, enough to vary strongly with brain size. The fact that you do not find correlations between socioecological variables and brain size cannot, therefore, be because the connection via reproductive traits has been "broken", because it has not. It must be for some other reason. Yet you conclude that your results show the importance of reproductive traits in mediating the effects of other variables on brain size. I do not see how this conclusion follows from your results, since your original hypothesis was not supported. Perhaps I have simply misunderstood something somewhere, but I think this apparent dissonance between your initial hypothesis and your results and conclusion needs to be clarified somehow.

The reviewer raises an important point here and we have addressed this in the main text in order to clear any misunderstandings regarding the confounding effect of reproductive modes on brain size variation.

In conclusion, I think that your paper has great potential to be a showcase for the use of data imputation methods and Bayesian phylogenetic comparative analysis, but I think it needs to be further fleshed out before it can become that.

More ancillary comments are listed below:

- In line 119 you say that “Data on brain volumes were derived from measurements of endocranial volumes (ECV)”, and in lines 135-139 you talk about how ECV data might need correction. This suggests that you may have adjusted the ECV data in some way, but you do not explain how.

We have not adjusted the data, as all volumes are derived from ECV, so if there is an error, it is supposed to be consistent throughout the whole sample.

- Acronyms (e.g., BM, OU, EB) need to be introduced with the first usage of the full term.

We have introduced them at the first usage.

- Variable names need to be used consistently to avoid confusion. Sometimes you use "Status" and sometimes "Vulnerability"; sometimes you use "Hibernation" and sometimes "torpor"; sometimes you use "activity period" and sometimes "diurnality", etc.

Thanks for the observation! We have chosen to use Torpor (yes/no) and activity period. We usually use Vulnerability, and we use the 3 different statuses to describe the 3 levels of vulnerability (See the table explaining the variables).

- In line 339 you say that your play behaviour data contain more than 80% imputed values, but elsewhere you say that that variable only has 68% missingness.

This has been corrected to the correct percentage – 68!

Referee: 2

Comments to the Author(s)

This paper uses a novel approach to find correlates of brain size evolution in marsupial mammals, by imputing missing data of predictor variables in a phylogenetic model. I cannot judge whether this method is justified or whether the imputation may lead to either exaggerating or “blurring” the patterns in available data. Perhaps an expert in statistics may help with this point. However, the materials and methods are very well described and thus the study is certainly valid. To me, it seems that there is no information added from imputation, and thus we cannot expect additional insight. But imputation may help to combine a larger number of variables in a single model. The results of such an approach should nevertheless be regarded as less reliable than those from original data.

Thanks for the comment! We have addressed the issue of why imputations are always better than using only complete cases with listwise deletion in some of the comments to reviewer 1.

In the current paper, I would therefore like to see the also the results of each original predictor variable (non-imputed data) in the models

brain size ∼ predictor \* body mass, in an appendix.

Thanks for the suggestion. Even though this is not necessary (see some of the comments to reviewer 1 and the attached papers on imputations, and the following up paragraph) we had attached PGLS analysis of the complete cases only to the supplement.

But see: Why we should use the results of MI instead of deletion:

“As already pointed out, results from listwise deletion suffer from a loss of power, but more important, they may be biased under MAR and NMAR. When carried out correctly, multiple imputation results in more power than listwise deletion, it completely corrects for bias under MAR, and partly corrects for bias under NMAR. This increase in power and correction for bias could explain possible differences in results between multiple imputation and listwise deletion.” [2]

We have also added full model specifications to the supplement material.

Using a baysian approach (MCMCglmm) for phylogenetic analyses is not less prone to the problem of robusticity than classic PGLS. In both methodologies, large contrasts in a variable between closely related species may have a disproportionate influence on the results, although this is less visible the more complex the analyses are designed. Thus, data quality remains of utmost importance, and the authors did a good job to compile a large sample of high quality. They even address the issue of a potential discepancy between ECV and brain mass, which has been found in koalas, but not in any other species so far. It would be certainly interesting to study this in more species, as seasonal variation in brain size has been found e.g. in some small placental mammals.

Thanks for the remark and this is most certainly a good suggestion!

Drawing conclusions from the analyses is a bit tricky. Actually, the negative results for any of the different realm models do not tell us much. They may stem from low power (although I don’t know how to assess power in such a complicated statistical approach), or from not including covariates that are likely to be correlated with brain size, even if their effect is not reaching a significant level. But these points are mentioned in the discussion, which is well written and considers all relevant literature.

Overall, in my view this is a careful, well written study which certainly merits publication and will be of broad interest, even though it does not proclaim any flashy new findings. Its merit is the thoughtful, new methodological approach on a newly compiled comprehensive dataset. It convincingly shows that there are many unsolved questions about brain size evolution, for which insights from marsupial mammals must not be neglected.

Thanks for the kind words and we hope we had addressed all of reviewer 2 concerns!

1. Bondarenko I., Raghunathan T. 2016 Graphical and numerical diagnostic tools to assess suitability of multiple imputations and imputation models. *Stat Med* **35**(17), 3007-3020. (doi:10.1002/sim.6926).

2. van Ginkel J.R., Linting M., Rippe R.C.A., van der Voort A. 2020 Rebutting Existing Misconceptions About Multiple Imputation as a Method for Handling Missing Data. *Journal of Personality Assessment* **102**(3), 297-308. (doi:10.1080/00223891.2018.1530680).

3. Janssen K.J., Donders A.R., Harrell F.E., Jr., Vergouwe Y., Chen Q., Grobbee D.E., Moons K.G. 2010 Missing covariate data in medical research: to impute is better than to ignore. *J Clin Epidemiol* **63**(7), 721-727. (doi:10.1016/j.jclinepi.2009.12.008).

4. Kontopantelis E., White I.R., Sperrin M., Buchan I. 2017 Outcome-sensitive multiple imputation: a simulation study. *BMC Medical Research Methodology* **17**(1), 2. (doi:10.1186/s12874-016-0281-5).