Biogeosciences Discuss.,
https://doi.org/10.5194/bg-2017-503-RC1, 2018

Biogeosciences Discussions

Interactive comment on "Gap-filling a spatially-explicit plant trait database: comparing imputation methods and different levels of ecological information" *by* Rafael Poyatos et al.

Anonymous Referee #1

Received and published: 10 January 2018

This paper compares the performance of different imputation methods for a trait dataset. Although this topic has already been studied before, the authors bring three novelties: (i) they use a spatially-explicit dataset (ii) which includes intra-specific information, and (iii) they use of a set of measures to assess imputation performance in terms of multivariate trait structure. This type of analysis is interesting, in a time where gap-filling methods are used more and more often to impute trait datasets. However, the paper would need some clarifications and better justifications, in particular for the choice of the methods. Some proper synthesis of the results is missing (shown by the total number of figures in main text and appendix) and no tests are performed, giving the general impression that none of the methods performs better than others. There is

also a number of problems in the results and figure presentations that should be fixed. Please refer to the comments below for more detail, in particular my points 7 to 10 are quite major.

Another major point is the use of "ecological information" in the imputation process. The authors show that adding this information to the imputation improves its quality, as one would expect. However, in most ecological papers, authors usually look at the relationships between traits and this "ecological information". This introduces a serious problem of circularity on any analysis using trait data imputed with "ecological information". The authors never mention this potential bias, do not suggest in which type of analyses such dataset could be used and more importantly, they do not test for this potential bias. A simple test would be to look at the relationship between traits and "ecological variables" in the complete dataset, in the datasets with missing data and in the imputed ones. This would give an idea about the importance of the bias and warn (or not, depending on the results) the users about it.

Abstract

1- L.16: "functional biogeography" is not mentioned elsewhere in the paper

2- L.16: "they offer specific challenges in terms of data imputation": these are not mentioned in the paper neither, it would actually be interesting to discuss these specific challenges

Introduction

Overall, the introduction is well written and clear, it just lacks some details on some aspects (see below).

3- Please provide a reference and explanation for the following statement L.56: "they all alter, to different degrees, the univariate trait distributions and the covariance structure of the dataset". Please also explain why multiple imputations do a better job in conserving trait distributions and covariances.

4- L60: Please cite which multiple imputation (MI) techniques are better to preserve structure and distribution of traits. MICE includes a high number of different algorithms, some accounting for variables distribution, others for interactions between variables, etc.. And actually some MICE techniques perform even worse than single imputations (e.g. see Stekhoven & Buhlmann 2012). MICE alone does not say much about the method and it would be important to specify quite early in the paper that it is MICE-PMM which is tested.

5- L 62. I would not define kNN as a sophisticated method, it's in fact quite a simple one

6- L90: sampling date is not really "ecological information", also most of the predictors mentioned were not really introduced before.

Methods

In general, the methods are clearly explained but some things are understandable only when reading the supplementary information. The methods are quite dense, so it is OK to have some descriptions in the appendix. However, they should be self-understandable. Figure 1 is very helpful and important. I also think that the evaluation metrics are very interesting. However, I did not understand some of the choices, which I think should be better justified (see below).

7- A quite major point is that the authors state that in the real dataset missing data is biased towards leaf traits deliberately. So data are not MCAR in the dataset (l.120). We also know that in trait datasets, values are often MAR (Nakagawa & Freckleton 2008). But the authors then remove data completely at random (L121). We also know that imputation methods are not designed for MCAR data (van Buuren & Groothuis-Oudshoorn, 2011). So why the authors chose to remove data MCAR? I would suggest to introduce data at random with the same structure as in the original dataset. This would be a fairer test than just removing data completely at random. It is important to see how the methods behave when data is MAR.

8- Another point is that there is no comparison with the dataset including missing data. The question: "should I impute or not" is an important one so it would be good to know how well the imputation methods perform (in terms of KGE and structure) when compared to just using the dataset with gaps.

9- Related to the previous, at the end, users would like to know which method is the best, considering together NRME, distributions, correlations, structure and regardless of the trait considered or the % of missing data. This could be analysed using a mixed model with the identity of the metric (NRMSE, KGE..etc) and the missing dataset identity as random factors.

10- Given that the dataset has intraspecific variation and this is presented as a novelty both in the introduction and the "implication" section, I would have expected an analysis of the error at the species level. Also, it should be noted that adding species identity assumes that interspecific variation is higher than intraspecific variation, which is OK, but should maybe be stated somewhere.

11- L. 126: why m=5? MICE authors recommend to use at least 10 iterations (van Buuren& Groothuis-Oudshoorn, 2011). I saw later that this was explained in the appendix, I think that it should be at least mentioned in the main text.

12- L.138: similar to the previous point, it would be good to state that k=7 comes from an additional analysis (described in the appendix), otherwise the "7" is quite obscure. All these omissions make the reading/methods understanding quite challenging.

13- The MICE paragraph (l.151) should give more details on what is described in appendix S4.

14- Why some variables are added sequentially whether others are included in a factorial design? (L175-180). Please clarify

15- Why lithology is not included into the RegKrig? And why is topography included in ordKrig? Isn't ordKrig only based on spatial coordinates? Please clarify.

16- KGE is a very interesting metric. However, since high values represent better performance (contrary to NRMSE and deltaCORMAT) I would either use –KGE for the figures or remind in figure captions that high values=better performance (e.g. fig4).

17- l233: for the imputation of the whole dataset the authors use m=50 imputations because it is recommend to choose a value closed to the missingness rate. Why this same rule was not applied for the simulations as well?

18- It would also be important to test the correlations between environmental variables, to see at which point it is interesting or not to use highly correlated variables in the imputation.

Supplementary information (SI) I would recommend to reduce this part, 32 pages of material is a lot. I would suggest to try to reduce it to the most important results.

19- Sentence L.106 of the appendix is misleading, its seems that the authors use MICE-RF in the paper.

20- Maybe it would be clearer to merge s3 and s4

21- Please also place all references at the end of the appendix and not just after each paragraph, this makes it difficult to read the (already long) document.

22- S4: I do not see how PMM performs better than the other methods. Fig s4 shows that there is almost no difference between methods and some are better than others for some traits. Fig s5 and s6 only seem to show that mice_PAS is worse than the others. And nothing is supported by statistical tests. I would suggest to revise these analyses or provide a more complete explanation of why PMM was used. Please also add in the caption what the traits are or provide the complete trait names in the figure itself.

Results and discussion

This section is not very clear and sometimes the results are simply a description of

the figures without any effort to synthesise what happens. The choice for the figures is also not consistent (see point 23). It is titled “results and discussion” but the discussion is almost absent. For instance, I missed some information on why some methods perform better than others or why some traits show better results. Please see some more detailed comments below.

23- Figures in the main text are a bit confusing. Fig2 is NRMSE and 4 methods, fig3 is deltaCoeff and 4 methods, fig4 is KGE, but this time with 11 methods, fig5 is deltaCoeff with 11 methods, fig6 is KGE with 7 methods and fig7 is deltaCoeff with 7 methods. This is quite inconsistent and we miss some comparisons (e.g. mean and smean). I would suggest to put together in a different form (e.g. as in figs10) and show all the metrics (NRMSE, KGE, deltaCoeff) for each comparison. Also, the results about distributions (Kolmogorov-Smirnov tests) are not performed for all methods and only shown in the appendix.

24- In the whole first paragraph (l237) I am not sure that the differences highlighted by the authors are actually significant differences. For instance at 10% I only see Mean being slightly worse than the other methods for LMA and WD, and OrdKrig for WD. OrdKrig also does not seems to perform better than others for most traits. Same remark for Fig3. Maybe a test looking at the effects of the method corrected by the trait identity would provide more general results.

25- L250: from what I see in fig s8 mice is not the method with lower NRMSE or deltaCormat: kNN is the best performing method except at around 50% of NAs, where OrdKrig seems to perform better. This seems to be the case also for KGE where kNN performs as good as the other methods (fig s7). However it performs bad when looking at distributions. I also do not see mice being better than kNN or OrdKrig in Fig.3 (as stated L.251). I would suggest revising this paragraph and carefully checking all results for similar problems.

26- L 266: the best performance of mice_ctsp is not really visible in fig s11. - Fig s10

is also not easily readable, I would suggest to jitter the points (geom_jitter in ggplot2). Fig s12 is better (but jittering would help as well).

27- Mice-ctsp is discussed and presented several times (l.266, 291, 297, 306). I would suggest to merge together information. Also the 3.3 sections seems redundant with the 3.2. Maybe showing together fig6 and 7 would save some redundancy.

28- paragraph L 306 does not belong to section 3.3, which is about comparing mean, mice and knn. A separate section would be more meaningful.

29- L328: “MICE informed by relevant ecological variables outperforms”: this was not properly tested as no analyses for this are provided

Minor comments

- L165: information about kNN is missing

- L234: “per missing value” should be “per missing dataset”

- L190: “statistical evaluation”, no stats are actually performed for the evaluation

- To facilitate the reading, please cite the exact figures of the supplementary and not only the section (e.g. do not refer to just to s5 or s6 but to fig.s7 or s10).

- All figures: please spell out the trait names or write it in the caption

References

Nakagawa, S. and Freckleton, R. P.2008. Missing inaction: the dangers of ignoring missing data, Trends in Ecology & Evolution, 23(11), 592–596

Stekhoven, D. J. and Bühlmann, P. 2012. MissForestâĂŤnon-parametric missing value imputation for mixed-type data. Bioinformatics 28: 112–118

van Buuren, S. and Groothuis-Oudshoorn, K., 2011, mice: Multivariate Imputation by Chained Equations in R, Journal of Statistical Software, 45(3)

Interactive comment on Biogeosciences Discuss., https://doi.org/10.5194/bg-2017-503, 2017.