Estimation of missing ordinal data: A comment on Wissler et al. (2022)

Sébastien Villotte, Frédéric Santos

Estimation of missing ordinal data: A comment on Wissler et al. (2022)

Running title: A comment on Wissler et al. (2022)

Authors: Frédéric Santos (1), Sébastien Villotte (2, 3, 4)

Affiliations:

1. Univ. Bordeaux – CNRS – MCC, UMR 5199 PACEA, Pessac, France
2. UMR7206 Éco-Anthropologie, CNRS, MNHN, Université Paris Cité. Musée de l'Homme, 17 place du Trocadéro, Paris, 75116 France
3. Quaternary environments and Humans, OD Earth and History of life, Royal Belgian Institute of Natural Sciences, rue Vautierstraat 29, Brussels, B-1000 Belgium
4. Unité de Recherches Art, Archéologie Patrimoine, Université de Liège, allée du six août 10, Liège, B-4000 Belgium

Correspondence to: Frédéric Santos, Univ. Bordeaux – CNRS – MCC, UMR 5199 PACEA, Bâtiment B8, Allée Geoffroy Saint-Hilaire, CS 50023, 33615 Pessac Cedex, France. Email: frederic.santos@u-bordeaux.fr

Keywords: entheseal changes, imputation, missing data

Main text

In their recent article published in the *AJBA*, Wissler et al. (2022) presented an in-depth review of several methods of multiple imputation for missing ordinal or continuous data, and provided detailed and useful guidelines to handle missing data efficiently, a crucial issue in our disciplinary field. Their study showed 1) that imputation of continuous data leads to a substantially lower error than with ordinal data, and 2) that imputation errors (especially when assessed using the normalized root mean square error, NRMSE) are strongly related the amount of missing values.

Their article also invited the community to evaluate these methods in other contexts. In particular, they acknowledged that the sample sizes might be considerably smaller in practical applications; and also advocated for further investigation of the performance of these methods on ordinal paleopathology data (p. 358).

We propose here a two-fold comment on this study. From a statistical wiewpoint, we discuss the accuracy metric used in the original paper, for the NRMSE formula used in it might have led to misleading interpretations. From a biological viewpoint, we assess the performance of multiple imputation by predictive mean matching (PMM) on a previously published ordinal paleopathology dataset (Villotte & Santos, 2023), consisting in entheseal changes recorded on a trichotomic scale. The ordinal dataset used in Wissler et al. (2022) was composed of ordinal values for different conditions (porotic hyperostosis, cribra orbitalia, periodontal disease, linear enamel hypoplasia, and periosteal lesions of the tibia), and it appeared possible to us that their results—for this dataset—may be partly related to the fact that these conditions have different aetiologies. We expected that applying their approach for a similar dataset (i.e., ordinal and paleopathogical), but for which there is a substantial correlation between the markers under study, would lead to a better accuracy in the

estimated values, i.e., closer to those obtained by Wissler *et al.* for continuous data. We thus decided to test this hypothesis using ordinal scores for 9 fibrocartilaginous entheseal changes described in a recent study (Villotte & Santos, 2023). These changes are usually correlated (Villotte, 2009), likely due to the fact that they are all correlated with the individual age-at-death (Villotte & Santos, 2023).

All analyses presented below were performed using R 4.3.1 (R Core Team, 2023), and are fully reproducible using R notebooks available on GitLab (https://gitlab.com/f-santos/imputation-of-ordinal-missing-values).

Assessing success of imputation methods

In Wissler et al. (2022), from an initial and complete dataset *T* of true ordinal values, various percentages of missing data (ranging from 5% to 40%) were added at random in *T*, resulting in incomplete datasets *S*. Then, the values missing in *S* were predicted using various algorithms, resulting in turn in imputed datasets *I*. Wissler *et al.* evaluated the success of these imputation methods using "*normalized root mean square error (NRMSE), which measures the difference between predicted and observed values*" (p. 354). However, several definitions of this metric do exist, and there is no proper standard in the literature. Wissler *et al.* used the R package `{hydroGOF}` (Zambrano-Bigiarini, 2020) to compute the NRMSE. As per the package's documentation, it defines this metric as follows:

$$NRMSE = 100 \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(T_i - I_i)^2}}{sd(T_i)},$$

where T_i and I_i are *i*-th values among the true and imputed datasets respectively, among the whole set of *N* values under study.

Note that there are two types of pairs (T_i, I_i): (1) if S_i was a missing value, then I_i is a predicted value which should slightly differ from T_i; (2) if S_i was not a missing value, then T_i, S_i and I_i are exactly the same value. An issue in the previous definition is that both types of pairs are used to compute the NRMSE, so that this indicator will strongly depend on the percentage of missing data in the dataset *S*, but will not necessarily be a good estimate of the accuracy of the predicted values.

Indeed, if *S* contains 5% of missing values, the NRMSE will be computed on pairs of data points that are strictly equal 95% of the time, which will necessarily result in a very low error, even if the imputation algorithm gave quite bad predictions on the 5% of missing data. Conversely, if *S* contains 40% of missing values, the NRMSE will probably be much higher, just because it was computed on less pairs that are equal by construction. More generally, if we denote *p* the percentage of missing data in *S*, and if we suppose that the error made by the imputation algorithm on the missing values remains globally constant regardless of *p*, the NRMSE will all the same linearly increase with *p*. This explains that, in Figures 1 and 5 from Wissler et al. (2022), we can note an almost linear dependency between the percentage of missing data and the NRMSE. However, this does not mean that the "true" prediction error on missing values linearly increases as a function of *p*, and this does not even mean that the prediction error increases at all depending on *p*.

Thus, other R packages implement a more realistic and useful NRMSE definition: for instance, the package `{MissForest}` (Stekhoven & Bühlmann, 2012) uses the following one, as per its documentation:

$$NRMSE = \sqrt{\frac{mean((T_i - I_i)^2)}{var(T_i)}},$$

where mean and var "*are used as short notation for the empirical mean and variance computed over the missing values only*". Actually, this formula is almost the same as in Wissler *et al.* if the NRMSE is computed only on the pairs of values that are missing in *S*.

This has a substantial impact on the conclusions one can draw from the simulations performed in the original article, as shown by a re-analysis of the ordinal data made publicly available in Wissler et al. (2022). Following their initial design, we simulated five incomplete datasets for each given percentage of missing values (from 5% to 40%), and then imputed these missing values artificially introduced. To avoid unnecessary redundancies, we considered hereafter only one of the five imputation methods studied in the original article, namely the predictive mean matching (PMM) method.

As shown in Figure 1, the NRMSE actually depends very weakly on the proportion of missing values introduced in the original dataset, as long as the remaining values represent a sufficient sample to train the algorithm. Even if the R package they used implement a potentially misleading success indicator, Wissler *et al.* did provide insightful discussion and references to support this fact in their article (p. 359).

Although we only evaluated one method of multiple imputation here (PMM), similar results can be obtained for the the other methods studied in the original article (results not shown).

Evaluation of PMM on entheseal changes

In a second step, we applied the same approach on a set of 9 fibrocartilaginous entheseal changes, recorded on a trichotomic scale, all of them being located on the right side of the individuals. As Figure 2 shows, the values predicted by the PMM algorithm were closer to the true values than for Wissler et al. (2022) data, for all percentages of missing values, and for both measures of accuracy (NRMSE and proportion of false classification). In particular, the stages of entheseal changes were correctly estimated in about 70% of the cases; this proportion is substantially lower (about 40%) for the markers used in Wissler et al. (2022).

Discussion and conclusion

Comparing the results obtained using the two datasets represented in Figure 2 is not trivial, since the markers under study have a different number of stages. Unlike the entheseal changes from Villotte & Santos (2023), most of the biological markers used in Wissler et al. (2022) were recorded on scales containing at least four values. This may be a partial explanation for the higher proportion of false classification obtained for this dataset (Fig. 2). Further studies may be necessary to assess more precisely the impact of the number of stages retained to record a given trait on the accuracy of the imputation methods, or to compare the accuracy of ordinal missing data imputation in various biological contexts, but where all traits have the same number of levels.

Overall, the very concept of imputation assumes that missing values can be inferred using the information provided by the other traits of the same individual, and the correlations among traits observed on the other individuals. Thus, the stronger the intercorrelation among the variables, the more we can expect an efficient imputation of missing values. However, especially in large datasets, some traits can be only weakly correlated to all other traits, which makes difficult the imputation of missing values on these traits. A concept of imputability measure has been developed to that end in other disciplinary fields (Liao et al., 2014) to identify those missing values that should be more prone to imputation errors. However, this method still lacks a clean and easy-to-use implementation in R or Python, and has not been tested yet–to the best of our knowledge–on biological traits.

References

Liao, S. G., Lin, Y., Kang, D. D., Chandra, D., Bon, J., Kaminski, N., Sciurba, F. C., & Tseng, G. C. (2014). Missing value imputation in high-dimensional phenomic data: Imputable or not, and how? *Bmc Bioinformatics*, *15*(1), 346. https://doi.org/10.1186/s12859-014-0346-6

R Core Team. (2023). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.*

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest: Non-parametric missing value imputation for mixed-type data. *Bioinformatics (Oxford, England)*, *28*(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

Villotte, S. (2009). *Enthésopathies et activités des hommes préhistoriques: Recherche méthodologique et application aux fossiles européens du Paléolithique supérieur et du Mésolithique*. Archaeopress.

Villotte, S., & Santos, F. (2023). The effect of age on entheseal changes: A study of modifications at appendicular attachment sites in a large sample of identified human skeletons. *International Journal of Osteoarchaeology*, *n/a*(n/a). https://doi.org/10.1002/oa.3197

Wissler, A., Blevins, K. E., & Buikstra, J. E. (2022). Missing data in bioarchaeology II: A test of ordinal and continuous data imputation. *American Journal of Biological Anthropology*, *179*(3), 349–364. https://doi.org/10.1002/ajpa.24614

Zambrano-Bigiarini, M. (2020). *hydroGOF: Goodness-of-fit functions for comparison of simulated and observed hydrological time series*. Zenodo. https://doi.org/10.5281/ZENODO.839854

Figures

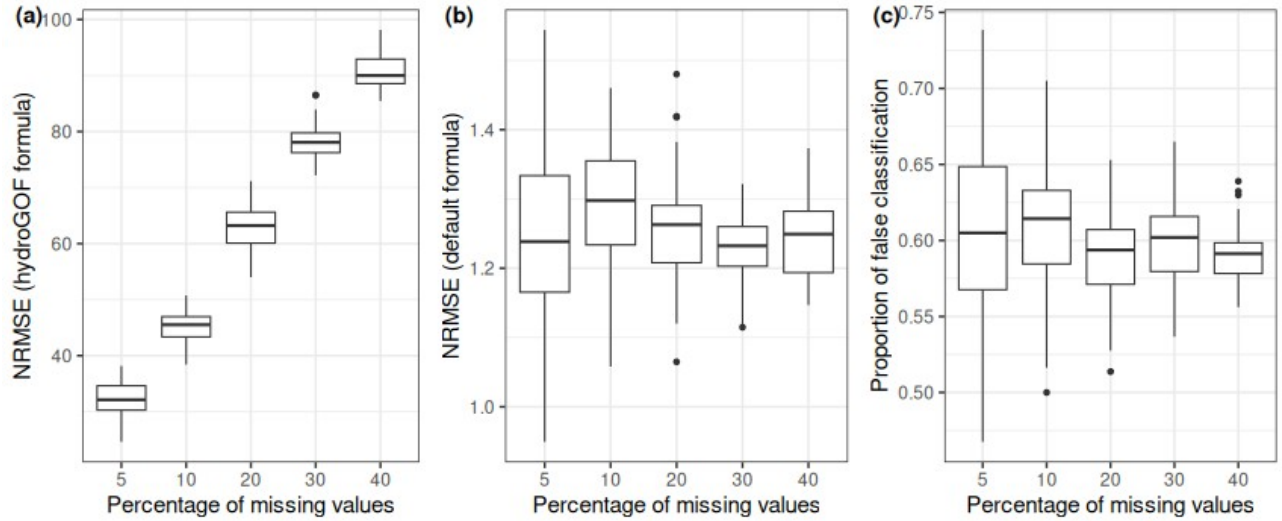

Figure 1: Re-analysis of the ordinal data published in Wissler et al. (2022). Evaluation of the accuracy of imputation by predctive mean matching with (a) the NRMSE formula used in the original article, (b) an alternative NRMSE formula proposed in the present article, (c) the proportion of false classification.

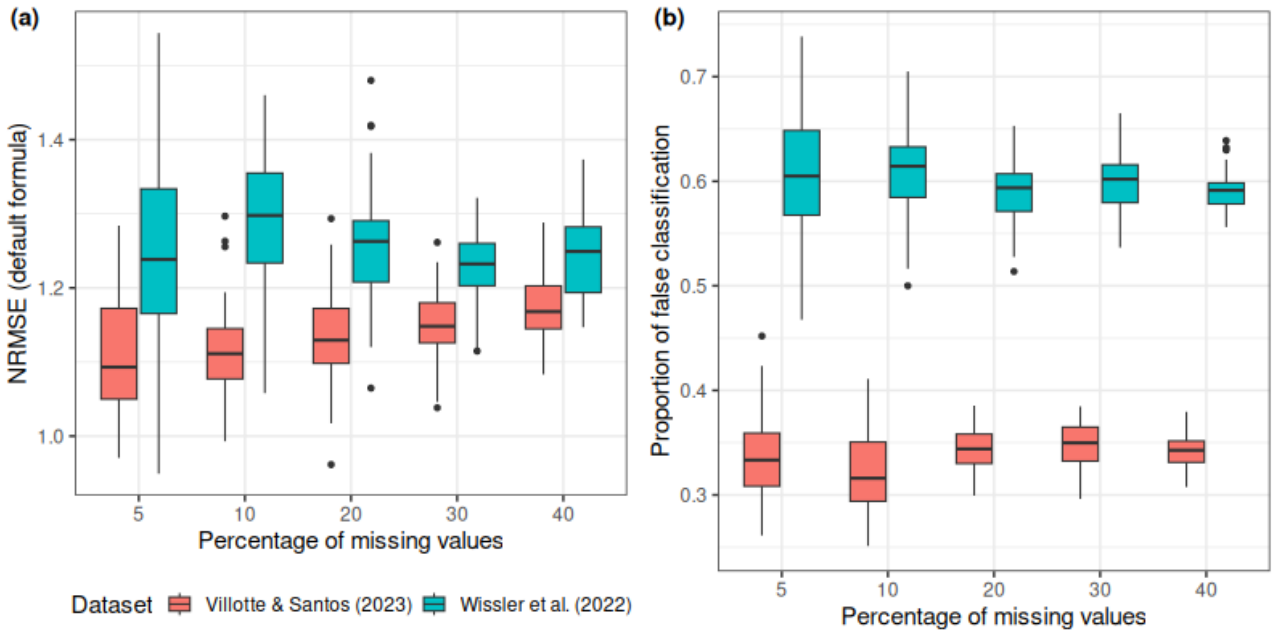

Figure 2: Comparison of the accuracies for estimating missing data by predictive mean matching in Wissler et al. (2022) data, and in Villotte and Santos (2023) data (fibrocartilaginous entheseal changes) respectively.