BIASES WITH THE GENERALIZED EUCLIDEAN DISTANCE IN DISPARITY ANALYSES WHEN MISSING DATA IS ABUNDANT

Commented [RJB1]: An alternative would be "Biases with the generalized Euclidean distance in disparity analyses with high levels of missing data"

by OSCAR E. R. LEHMANN[1]*, MARTÍN D. EZCURRA[1,2]*, RICHARD J. BUTLER[2], *and* GRAEME T. LLOYD[3]

[1] Sección Paleontología de Vertebrados, CONICET−Museo Argentino de Ciencias Naturales "Bernardino Rivadavia", C1405DJR, Buenos Aires, Argentina; e-mails: lehmanncxii@gmail.com, martindezcurra@yahoo.com.ar

[2] School of Geography, Earth and Environmental Sciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK.

[3] School of Earth and Environment, University of Leeds, Leeds LS2 9JY, UK.

* Corresponding authors.

Abstract. The Generalized Euclidean Distance (GED) has been extensively used to conduct morphological disparity analyses based on palaeontological matrices of discrete characters. This is in part because some implementations allow the use of morphological matrices with high percentages of missing data without needing to prune taxa for a subsequent ordination of the data set. Previous studies have suggested that this way of using the GED may generate a bias in the resulting morphospace, but a detailed study of this possible effect was still lacking. Here, we test if the percentage of missing data for a taxon artificially influences its position in the morphospace, and if missing data affects pre- and post-ordination disparity measures. We find that this use of the GED creates a systematic bias, whereby taxa with higher percentages of missing data are placed closer to the centre of the morphospace than those with more complete scorings. This bias extends into pre- and post-ordination calculations of disparity measures and can lead to erroneous interpretations of disparity patterns, especially if specimens present in a particular time interval or clade have distinct proportions of missing information. We suggest that this implementation of the GED should be used with caution, especially in cases with high percentages of missing data. Results recovered using an alternative distance measure, Maximum Observed Rescaled Distance (MORD), are more robust to missing data. As a consequence, we suggest that MORD is a more appropriate distance measure than GED when analysing data sets with high amounts of missing data.

Keywords. Morphological disparity, distance measure, missing data, palaeontological matrices.

A large number of palaeontological studies dealing with morphological disparity have been published over the last decade, mainly driven by the widespread use of already available morphological matrices constructed for phylogenetic analyses. A critical step in most of these studies is the transformation of the morphological matrix into a distance matrix. This is done using a distance measure, such as Gower's coefficient (Gower 1971), the maximum observed rescaled distance (MORD; Lloyd 2016), or the Generalized Euclidean Distance (GED; Wills 1998), among others. The existence of missing data in the morphological matrix can produce distance matrices with missing entries when certain distance measures are used. This means that no distance could be calculated for certain pairs of taxa because of the absence of overlapping information between them, a rather common situation when dealing with palaeontological information. A complete distance matrix is necessary if an ordination is desired to reduce the dimensionality of the data set to produce a morphospace, calculate some common disparity measures, or both. A strength of the GED is that it operates by inferring missing distances and hence it is possible to always return a complete distance matrix, and in part because of this it has become one of the most extensively used distance measures (*e.g.* Wills 1998; Brusatte *et al.* 2011; Prentice *et al.* 2011; Thorne *et al.* 2011; Butler *et al.* 2012; Ruta *et al.* 2013a, b; Hetherington *et al.* 2015; Oyston *et al.* 2015; Marx and Fordyce 2016; Lamsdell and Sendel 2017; Ezcurra and Butler 2018). However, some concerns have been raised about the smoothing effect the GED may have on the distance matrix and the disparity estimates derived from it, especially when it is used to analyse matrices with a high percentage of missing data (Lloyd 2016, Ezcurra and Butler 2018).

In this study we explore the performance of the GED in disparity analyses based on discrete character matrices to address the concerns about its possible biases. This is especially important due to the widespread use of the GED. The behaviour of an alternative distance measure, MORD, is also analysed as a control against which to compare the GED. Three main objectives drove this study. First, to determine if a hypothesised bias resulting from the replacements made by the GED are indeed present; second, to explore if this bias produces a noticeable distortion in the morphospace; and finally, to determine if differences in the percentage of missing data between groups of taxa can explain which group is recovered as the most disparate.

MATERIALS

A data set of 158 published morphological matrices, each containing at least 20 taxa and 50 characters, was used for this study. These matrices comprise a wide spectrum of taxonomic groups, including plants, beetles, echinoids, actinopterygians, basal tetrapods, dinosaurs, birds, lepidosaurs, and mammals, among others, and were primarily conceived for cladistic phylogenetic analyses. They range from 20 to 223 taxa and from 50 to 486 characters. Only discrete characters are present. Most of the matrices are compiled at the personal web page of one of the authors (GTL; http://www.graemetlloyd.com/), were previously used by Wright et al. (2016), and all are available in Lehmann *et al.* (2019, SI 2), together with a complete list of references. The analyses were performed with different subsets of these matrices, as discussed below. Detailed information about their dimensions, distribution of missing data, and use is available in Lehmann *et al.* (2019, SI 3-4).

METHODS

The following analyses were carried out using both the GED and the MORD for the generation of the distance matrices. In some cases, the ordination may become impossible when the MORD is used, as it is one of the distance measures that does not guarantee the creation of a complete distance matrix. To solve these cases, the TrimMorphDistMatrix function of Claddis (Lloyd 2016) was used to generate a complete distance matrix. This function iteratively removes the taxa responsible for the generation of most empty cells until the matrix is complete. These taxa were also trimmed from the GED matrix in order to generate comparisons with the same taxon sampling. This approach also makes the analyses presented here more conservative than if the GED were used directly on the untrimmed matrix, as it is commonly used, because the trimmed taxa tend to be the most incomplete, and thus the most prone to be affected by any bias generated by the GED.

All the analyses performed in this study were conducted in the programming environment R (R Core Team 2018), using custom-made functions and functions implemented in the package Claddis v. 0.2 (Lloyd 2016). The scripts used for this study are available in Lehmann *et al.* (2019). Statistical significance was assessed at α = 0.05.

Calculation of GED and MORD

The GED (d_{ij}) between a taxon *i* and a taxon *j* is calculated with the following formula:

$$d_{ij} = \sqrt{\sum_{k=1}^{v} \left(S_{ijk} W_{ijk}\right)^2}$$

where *k* is the current character, *v* is the total number of characters, S_{ijk} is the dissimilarity between taxon *i* and taxon *j* for character *k*, and W_{ijk} is the weight of character *k*. It can be used to generate a complete distance matrix by replacing missing dissimilarities with a weighted mean value $\bar{S}$ obtained from those dissimilarities that could be calculated, as follows:

$$\bar{S} = \frac{\sum_{k=1}^{v}\left(S_{ijk}W_{ijk}\delta_{ijk}\right)}{\sum_{k=1}^{v} range\ (X_k)W_{ijk}}$$

where δ_{ijk} is 1 if both taxa have a scored, non-missing value for character *k*, and 0 if any of them is missing, and *range(X*$_k$*)* represents the maximum possible dissimilarity for character *k*. Note that in Wills' (1998) original paper this weighted mean fractional univariate distance was intended to represent the taxon pair being considered only, but (as pointed out by Hopkins and St John 2018) Lloyd (2016) misinterpreted this as representing the mean for the whole matrix, i.e., all taxon pairs. Thus as Wills (1998) originally conceived the GED it would suffer from the same incomplete distance matrix problem as other metrics. However, here we use the common Lloyd (2016) implementation available in Claddis versions 0.1-0.2 as this is both the only currently offered implementation that will always return complete pairwise distances and the most frequently applied GED implementation in the recent literature.

The MORD is just a slight modification of Gower's Coefficient, yielding identical results to it when all the characters are binary or unordered (Lloyd 2016), and is calculated as:

$$d_{ij} = \frac{\sum_{k=1}^{v}\left(S_{ijk}W_{ijk}\right)}{\sum_{k=1}^{v} range\ (X_k)W_{ijk}}$$

The distances obtained with this formula were subjected to the arcsine square-root transformation, as is default in the R package Claddis (Lloyd 2016). A detailed account of the two distance measures used here is available in Lehmann *et al.* (2019, SI 1).

Effect on the morphospace

The first effect evaluated was if the use of the GED influenced the position of taxa in the morphospace. This was done for an empirical set of morphological matrices and for a simulated set of matrices created by including sequentially higher percentages of missing data into empirical matrices with originally very low percentages (<5%) of missing data.

Empirical matrices with missing data. This analysis had the objective of testing if there is an inverse correlation between the percentage of missing data for a taxon and its distance to the centroid of the morphospace when the GED is used to construct the distance matrix. If this is the case, then taxa with a higher percentage of missing data will be closer to the centroid than those with fewer missing entries, arguably because the GED replaces a higher percentage of their pairwise dissimilarities with a weighted mean value. We further considered whether correlations may be more pronounced when smaller amounts of variance are used in the calculation of distance to the centroid and that the bias might disappear after a certain threshold level of explained variance.

A total of 126 matrices with more than 5% of missing data were used for this analysis. The first step was the creation of the distance matrices, which was followed by their ordination with a principal coordinates analysis (PCoA; Gower 1966; Legendre and Legendre 1998: ch. 9) to generate a morphospace of reduced dimensionality (Wills 2001). The Lingoes correction (Lingoes 1971) was applied if negative eigenvalues were created. This correction is necessary if all eigenvalues are required for the interpretation of the data (or if the absolute value of the largest eigenvalue is larger than the smallest positive eigenvalue of the principal coordinate axes [PCos] of interest [Legendre and Legendre 1998: p. 437]), but it frequently causes a distinct reduction in the variance explained by the first few PCos (Hopkins 2016). As the PCoAs are frequently corrected in published disparity analyses (*e.g.* Prentice *et al.* 2011; Butler *et al.* 2012; Oyston *et al*. 2015), we decided to include the correction as a step in the disparity pipeline, but we note that other authors prefer to avoid the correction if it is not completely necessary (*e.g.* Hopkins and Smith 2015; Hopkins 2016). Thus, to test if the use of a correction could affect the results of our study, this particular analysis was also performed without any correction for negative eigenvalues. PCos with negative eigenvalues were excluded from the analysis.

Subsequently, the Euclidean distance was calculated between each taxon and the centroid of the morphospace, which is its origin of coordinates and also the mean of the PCos. This was done employing the PCos that correspond to accumulated variance from 5% to 100% in 5% increments, and the particular case of the variance accumulated in the first three PCos, which are the most commonly used to graphically represent the morphospace. Finally, the Spearman's correlation coefficient was calculated between the percentage of missing data for each taxon and the corresponding distance to the centroid of the morphospace. This was calculated for each distance measure and each level of explained variance.

Empirical matrices with simulated missing data. To determine if the results obtained in the empirical analyses could be simulated, some of the scored data entries in 33 matrices with less than 5% of missing data were randomly replaced with missing values. These matrices have no possible meaningful correlation between the percentage of missing data of their taxa and the distances of taxa to the centroid. This is for either of two reasons: (1) most of the taxa had no missing data, which caused the majority of the distances to the centroid to be evaluated based on a percentage of missing data equal to zero; or (2) there were no missing entries in the whole matrix and, therefore, no standard deviation (and thus no correlation coefficient) could be calculated.

To ensure a rather realistic distribution of missing values across the taxa, a beta distribution with parameters $\alpha = 6$ and $\beta = 10.67$ was used to draw the proportion of missing data that each taxon would have in the morphological matrix. The beta distribution has a domain of [0, 1], and the selected parameters ensured an expected value of 0.35, which is approximately the median proportion of missing data observed in the empirical matrices used in the previous analysis. To avoid unrealistic scenarios, such as a taxon or character with one or none scorings, the minimum number of non-missing characters scored per taxon was limited to three and the minimum number of non-missing scorings for a character to two. The disparity analysis protocol was the same as with the empirical data set, with the distance to the centroid calculated using all the PCos. The entire protocol was repeated 200 times on each of the 33 matrices.

Effect on disparity measures

We explored the effect that the GED may have on four commonly-used disparity measures: the pre-ordination measures mean pairwise distance (MPD) and weighted mean pairwise distance (WMPD), and the post-ordination measures sum of variances and sum of ranges. Details about the calculation of these metrics and some of their interpretations can be found in Lehmann *et al.* (2019, SI 1). First, we evaluated how the disparity values of a matrix varied as increasing percentages of entries were randomly replaced with missing values. Secondly, we studied how the values of the disparity measures for groups of taxa from the same matrix varied as the groups were assigned different percentages of missing entries.

Whole matrix analysis. This analysis was designed to test if the use of the GED induces a bias associated with the percentage of missing data of the morphological matrix into the chosen disparity indices.

A total of 33 matrices with less than 5% of missing data were used for this analysis. The first step was the calculation of the four disparity measures for the unmodified matrices in order to have a baseline value to which compare the subsequent results. The next step involved the random replacement of between 10% and 60% of the entries of the matrix with missing values, with increments of 10%, and the calculation of the disparity measures for each case. The replacement with missing entries started from the original, unmodified morphological matrix every time (a matrix with 60% of missing data was assigned all its missing values at once, not six times increasing by 10%). This protocol was replicated 200 times for each level of percentage of missing data.

The MPD and the WMPD were calculated without trimming the distance matrix, but the calculation of the sum of variances and the sum of ranges sometimes required the trimming of taxa, as the PCoA was a necessary step for their calculation. As a result, the disparity estimates for these post-ordination measures are comparable with each other but not with those of the pre-ordination measures because they were calculated on potentially different taxon samplings. The sum of variances and the sum of ranges were calculated using all the PCos.

Groups of taxa from the same matrix. This analysis was designed to evaluate if the disparity measures calculated for groups of taxa within a matrix would be higher for groups with smaller percentages of missing data, and lower for groups with higher percentages of missing data. If this is the case, the effect may be attributed to the fact that the weighted mean value calculated by the GED would generate a stronger bias in the groups with higher percentages of missing data, thus making them more homogeneous when compared with groups with smaller influence from the replacement value.

To conduct this analysis, matrices that could be broken into five equally-sized, reasonably large (10 or more taxa) groups were chosen. These matrices were also required to have less than 5% of missing data. Seven of the previously used matrices with less than 5% of missing data fit these criteria, and two additional matrices with an originally higher percentage of missing data but with incomplete taxa removed until the threshold was met were also considered, summing up a total of nine matrices. Five equally-sized groups of taxa were randomly chosen for each matrix. If the original number of taxa in a matrix was not divisible by five, the maximum number of taxa that allowed generating equally-sized groups was randomly chosen.

Thirteen different simulations of missing data distribution (Table 1) were conceived to provide a diverse scenario of distributions of missing data among the groups of taxa. The simplest simulations involved the random replacement of a given percentage of cells with missing data, equivalent to the procedure explained in the previous section. The other simulations assigned different proportions of missing data to each group of taxa. For example, Simulation 6 assigned 30% of missing data to Group 1, randomly distributed among all its taxa. Progressively smaller amounts of missing data, with steps of 5%, were assigned to the other groups. This was designed to introduce a total of 20% of missing data for the whole matrix, making Simulations 4–6 comparable because they all have the same overall quantity of missing data and only differ in their distribution.

To assign the numeration to the groups, a procedure that ensured the independence between the numeration and the percentage of missing entries introduced was implemented. The four disparity measures were calculated for each group, and then the groups were ordered and numbered in a way that made the Spearman's rank correlation coefficient between the disparity values for each measure and the sequence 1 to 5 equal to 0. Then, 200 replicates were made for every combination of matrix and simulation, and the disparity measures were calculated for each group in every case. The post-ordination measures were calculated employing all the PCos.

The Spearman's rank correlation coefficient was calculated to compare the values of each disparity measure and the group number. Because of the design of the simulations, the baseline measures (*i.e.* with no simulation of missing data) and the group number have a correlation coefficient of 0. In the simulations in which the distribution of missing data is heterogeneous between the groups, the percentage of assigned missing data was the highest for Group 1 and sequentially lower, with the lowest percentage in Group 5. If the disparity measures calculated for groups of taxa within a matrix is indeed higher for groups with smaller percentages of missing data, and lower for groups with higher percentages of missing data, the Spearman's correlation should be closer to 1, and equal to 1 in the cases in which the ranking of the disparity measure is completely driven by the distribution of missing data between the groups.

RESULTS

Effect on the morphospace

Empirical matrices with missing data. A total of 87.3% of the studied matrices showed a significant and negative correlation between the distance to the centroid and the percentage of missing data in the taxa when all the PCos were considered and the GED was used as the distance measure (Figures

1 and 2; see Lehmann *et al.* 2019, SI 5 for the complete results). This contrasts with the results produced by the MORD, in which 79.4% of the matrices presented a non-significant correlation between the variables. These percentages are similar to those obtained using only the first three PCos (Figure 2). When using increasing amounts of variance (up to using all the PCos), the same result is retained, *i.e.* mostly negative and significant correlations with the GED and mostly non-significant correlations with the MORD (Figure 3; see Lehmann *et al.* 2019, SI 6 for the complete results). The results of this analysis do not appear to be affected by the use of the Lingoes correction for negative eigenvalues, as the percentages are similar when it is used and when no correction is applied (Lehmann *et al.* 2019, SI 5).

Empirical matrices with simulated missing data. The simulations showed a trend similar to that observed in the analysis with the matrices with more than 5% of missing data, in which the GED produces a high proportion of significant correlations with negative coefficients, and the MORD mostly non-significant correlations (Figure 4; see Lehmann *et al.* 2019, SI 7 for the complete results).

Effect on disparity measures

Whole matrix analysis. Both distances produced consistent results for the pre-ordination measures of disparity (Figure 5). The GED generated the highest value for the complete matrix, with progressively lower average disparities as the percentage of missing data increased. The MORD showed very similar average disparity values across the range of percentage of missing data studied, with some matrices having a small dip at around 60% of missing data (Lehmann *et al.* 2019, SI 8). This was expected because the MORD explicitly takes into account the amount of missing data present.

The post-ordination measures presented a more complex pattern when the GED was used. The sum of variances generally decreased as the percentage of missing data increased, starting from the highest value if no additional missing data was introduced. However, this was not always the case. In some of the matrices, the values of the sum of variances for the unmodified matrix were lower than those with added missing data, while in others the maximum average value was reached with the addition of 20–30% of missing data (Lehmann *et al.* 2019, SI 8). Nevertheless, the average sum of variances tended to diminish after adding 30–40% of missing data. The sum of ranges presented patterns similar to those of the sum of variances for each matrix.

When the MORD was used, the average sum of variances increased with higher percentages of missing data in all the studied matrices. The sum of ranges presented a similar behaviour, but in some cases the average disparity decreased during the addition of 10–30% of missing data and after the introduction of higher percentages of missing data it increased.

Groups of taxa from the same matrix. When the GED was used, the nine studied matrices showed correlations progressively closer to 1 as the difference in missing data between the groups increased (Figure 6; Lehmann *et al.* 2019, SI 9 for the complete results). This effect was more pronounced for the pre-ordination disparity measures, which also presented similar rankings of the disparity of the groups when the missing data was homogeneously-distributed with respect to the ranking without addition of missing data. The post-ordination measures showed a similar behaviour, but not as pronounced. The average correlations with homogeneously-distributed missing data were negative, though non-significant.

The simulations conducted with the MORD showed that the differences in the percentage of missing data among the groups have little effect on the pre-ordination measures (Figure 6). For these disparity measures, almost all of the mean correlation coefficients were not significantly different in each level of percentage of missing data and no clear bias was present. In five of the nine studied matrices the post-ordination measures progressively showed correlations closer to -1, which suggest a reversed order of disparity from the base ranking. This implies that the groups with higher percentages of missing data are recovered as more disparate than those with lower percentages, which is the inverse pattern recovered for the GED.

DISCUSSION

The effect of the GED in disparity analyses with missing data

The results presented here reveal that the GED generally produces a significant negative correlation between the percentage of missing data of a taxon and its distance to the centroid. In other words, the more information is lacking for a taxon, the closer it will be placed to the centre of the morphospace (Figure 7). This has been observed in the morphological matrices with a wide range of distributions of missing data (Lehmann *et al.* 2019, SI 4), and in the almost complete morphological matrices with the simulated inclusion of missing entries. Also, the presence of this bias seems to be independent of the number of PCos used to calculate the distances to the centroid, so the use of only certain PCo axes cannot solve this issue. In the analysis with groups with different percentages of missing data, the GED showed a bias for all the studied matrices and for all the disparity measures explored. A result consistent with this study has been recently reported by Ezcurra and Butler (2018) for an empirical morphological matrix.

The bias caused by the use of the GED may have profound consequences for the study of morphological disparity in the fossil record, as palaeontological data matrices, particularly those focused on fossil vertebrates, typically have a high proportion of missing data (Lloyd 2016). If the missing data are randomly distributed in the matrix, the effect will be diluted among all the taxa, but if they are concentrated in some taxa (which is likely to be the case in many empirical data matrices), the interpretations of the disparity analysis may be flawed. For example, if a clade is represented by more incomplete fossil specimens in a particular time interval, the disparity measures calculated from the GED could bias the results towards an interpretation in which this clade has a lower disparity in this time interval than in others with more complete specimens.

In five of the nine studied matrices for groups with different percentages of missing data, the MORD showed a bias towards recovering the groups with more missing data as more disparate in the case of the post-ordination measures. This pattern is congruent with the observation of Ezcurra and

Butler (2018) that in an empirical data set some taxa with high percentages of missing data tended to be pushed to the edge of the morphospace and away from the centroid when the MORD was used. However, this bias can be mitigated by avoiding ordination altogether, as advocated by Lloyd (2016), and as shown here in the pre-ordination (*i.e.* ordination-free) disparity measures.

Contradictory results to previous studies

Ciampaglio *et al.* (2001) studied the behaviour of the MPD and the sum of variances with different amounts of missing data as part of the most thorough evaluation of the performance of several disparity measures published so far. In that analysis, both measures showed no variation in their average values as the percentage of randomly-distributed missing entries changed from 0 to 25%, by increments of 5%. However, the results recovered in our study show a remarkably different behaviour, with the mean values of the MPD decreasing as the percentage of missing data increases in the simulations and the sum of variances showing a similar but more complex behaviour (Figure 5).

An important difference is how Ciampaglio *et al.* (2001) and we simulated the missing data. Ciampaglio *et al.* (2001) simulated the missing entries by selecting iteratively a column of the matrix (*i.e.* the scores of a character for all taxa) and then an entry of that column (*i.e.* a character score for a single taxon). This selected entry was replaced by the mean of all the other scores in the column, and the entire procedure was repeated until the desired proportion of "missing data" was achieved (Ciampaglio *et al.* 2001: p. 698). However, this method is inadequate because in empirical matrices the missing entries are not added sequentially. The successive replacement of missing entries with the mean values of the column could never be applied to an empirical matrix, so the results yield no information for comparisons with possible protocols of missing data handling. Also, it must be noted that the matrices are effectively complete, as no scoring is actually missing.

In our study, the missing data is simulated by the direct elimination of information from the morphological matrix. No attempt is made to replace the missing entries in the morphological matrix; this is possible but outside the scope of this study. This leaves the handling of the missing information to the calculation of the distance measure. The GED replaces the missing dissimilarities, not the missing entries of the original data matrix; and the MORD does not replace any missing information, it just rescales the distances with the number of available comparisons between a pair of taxa (see Lehmann *et al.* 2019, SI 1). The GED then squares and sums the dissimilarities, and finally calculates the square root of this value. This series of transformations complicates a straightforward prediction of the behaviour of the calculated disparity measures when randomly-distributed missing entries are added.

We replicated the procedure of missing data introduction of Ciampaglio *et al.* (2001) to quantitatively compare their results to ours (see Lehmann *et al.* 2019, SI 10). Our results with the protocol of Ciampaglio *et al.* (2001) are markedly different from those obtained by the original authors. The relatively constant values of these disparity measures with increasing quantities of missing data could not be recovered here, and we found that the absolute values of the disparity measures tended to decrease as more missing data is introduced, resembling the results of our simulations. For the sum of variances, the results were robust to either using or ignoring the Lingoes correction. The calculation protocol of these measures also differs between both studies (see Lehmann *et al.* 2019, SI 1), but the analysis we conducted showed that no matter which formula is used, the results are consistent (see Lehmann *et al.* 2019, SI 10).

Thus, neither the different protocol of missing data introduction, the differences in the calculation of the disparity measures, nor the use of a correction for negative eigenvalues (in the case of the sum of variance) seem to account for the differences found.

The MORD as a more robust distance measure for dealing with missing data

Many methods have been explored to deal with large amounts of missing data (see Smith *et al.* 2014 and discussion therein). It is certainly difficult to imagine an elegant and unbiased way to do this, especially because any imputation will most likely induce some kind of bias. In this study, we investigated what happens when the GED is used to deal with missing data and found that the solution it proposes induces an important bias in the pre- and post-ordination disparity measures and the morphospace.

The GED was not designed to deal with matrices with large proportions of missing data, as the use of a weighted mean dissimilarity was suggested for cases 'Where the proportion of missing data is relatively small [...]' (Wills 1998: p. 471). The overall percentage of information being added is higher than that of missing data found in the morphological matrix, because the GED replaces missing character dissimilarities between pairs of taxa, not missing entries. For example, all the matrices that had more than 31.5% of missing data in the original data matrix had more than 50% of missing dissimilarities in the distance matrix. This fact implies that more artificial values were used for the construction of their distance matrices than the information actually provided by the morphological matrix. This does not mean that the distance measure itself yields inadequate results. The main problem found in this work is that the weighted mean values added to permit the creation of a complete distance matrix skews the results as was previously discussed, particularly when the GED is applied to matrices with a large amount of missing data.

The use of alternative distance measures, such as the MORD, may alleviate the effect of missing data. The MORD did not generate such a remarkable trend of correlations nor induce any visible systematic bias in our analyses, at least in the majority of our results. If this is combined with the suggestion that it retains more of the original signal of the data than other distances (Lloyd 2016), the MORD seems to be a better alternative than the GED in the construction of the distance matrix for a disparity analysis, particularly if the matrix being analysed has a high proportion of missing data. An issue with the use of the MORD is that it does not guarantee a complete distance matrix, thus precluding the ordination of the complete data set and the creation of a morphospace if missing distances are present. In these cases, some taxa may have to be trimmed to continue with the analysis or, as Lloyd (2016) recommended, ordination-free approaches can be pursued instead. Another unexplored solution may be to combine the Wills (1998) and Lloyd (2016) interpretations of the GED into a new "hybrid" GED such that missing distances are inferred from either the rest of the information for the taxon pair (if available) or the global mean (all taxon pairs) if not. Thus a complete pairwise distance matrix could still be returned, but the potential biases shown here could be minimised.

The GED showed a systematic bias for the vast majority of the matrices studied, and the MORD showed some hints of bias when coupled with the post-ordination disparity measures. From these results, it is clear that, whichever distance is chosen, an adequate screening of the distance matrices and the ordination results is extremely important to detect any possible bias, and avoid artefactual conclusions, especially when post-ordination disparity measures are desired. The use of the MORD may not solve the issue of incomplete distance matrices in these scenarios, but the apparent solution provided by the GED seems to be unreliable not only for post-ordination disparity measures, but also for some common pre-ordination ones. We thus suggest that the GED should be used with caution, in particular when the morphological matrix has a high percentage of missing data.

Acknowledgements. We thank two anonymous reviewers for their useful comments about the manuscript, and Sally Thomas for editorial suggestions. RJB is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement 637483 (ERC Starting Grant TERRA).

DATA ARCHIVING STATEMENT

Data for this study, including data sets, scripts, and complete graphical results are available in the Dryad Digital Repository: https://doi.org/10.5061/dryad.4cv1421.

SUPPORTING INFORMATION

Additional Supporting Information can be found in the online version of this article:

Appendix S1. Formulas and discussion of the distances and disparity measures used in this study.

Appendix S2. References of the matrices employed in this study.

Appendix S3. Table with the basic information of the matrices employed in this study.

Appendix S4. Distribution of missing data in the matrices used in this study.

Appendix S5. Scatter plots for the 126 matrices with more than 5% of missing data showing the Euclidean distance to the centroid for each taxon against its percentage of missing entries, and histogram showing the results of the analysis without the Lingoes correction.

Appendix S6. Scatter plots for the 126 matrices with more than 5% of missing data showing the correlation coefficient between the Euclidean distance to the centroid for each taxon and its percentage of missing entries against the percentage of explained variance of the PCos used to calculate the distance to the centroid.

Appendix S7. Histograms for the 33 matrices with less than 5% of missing data showing the percentage of replications with significant and negative, non-significant, and positive and significant correlations between the Euclidean distance to the centroid and the percentage of missing entries of the taxa.

Appendix S8. Results from the 33 morphological matrices for the studied disparity measures against the proportion of randomly-distributed missing entries added to morphological matrices.

Appendix S9. Results of simulations with groups of taxa and different distributions of missing entries.

Appendix S10. Procedure to compare the protocol of missing data generation of Ciampaglio *et al.* (2001) and of this study, and related figures.

REFERENCES

Baron, M. G., Norman, D. B. and Barrett, P. M. 2017. A new hypothesis of dinosaur relationships and early dinosaur evolution. *Nature*, **543**, 501-506.

Brusatte S. L., Montanari, S., Yi, H. and Norell, M. A. 2011. Phylogenetic corrections for morphological disparity analysis: new methodology and case studies. *Paleobiology*, **37**, 1–22.

Butler, R. J., Brusatte, S. L., Andres, B. and Benson, R. B. J. 2012. How do geological sampling biases affect studies of morphological evolution in deep time? A case study of pterosaur (Reptilia: Archosauria) disparity. *Evolution*, **66**, 147–162.

Ciampaglio, C. N., Kemp, M. and McShea, D. W. 2001. Detecting changes in morphospace occupation patterns in the fossil record: characterization and analysis of measures of disparity. *Paleobiology*, **27**, 695–715.

Ezcurra, M. D. and Butler, R. J. 2018. The rise of the ruling reptiles and ecosystem recovery from the Permian-Triassic mass extinction. *Proceeding of the Royal Society of London Series B, Biological Sciences*, published online 13 June 2018. doi:10.1098/rspb.2018.03.61.

Gower, J. C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53**, 325–338.

Gower, J. C. 1971. A general coefficient of similarity and some of its properties. *Biometrics*, **27**, 857–874.

Hetherington, A. J., Sherratt, E., Ruta, M., Wilkinson, M., Deline, B. and Donoghue, P. C. J. 2015. Do cladistic and morphometric data capture common patterns of morphological disparity? *Palaeontology*, **58**, 393–399.

Hopkins, M. J. and Smith, A. B. 2015. Dynamic evolutionary change in post-Paleozoic echinoids and the importance of scale when interpreting changes in rates of evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 3758–3763, published online 23 February 2015. doi:10.1073/pnas.1418153112

Hopkins, M. J. 2016. Magnitude versus direction of change and the contribution of macroevolutionary trends to morphological disparity. *Biological Journal of the Linnean Society*, **118**, 116–130.

Hopkins, M. J. and St John, K. 2018. A new family of dissimilarity metrics for discrete character matrices that include inapplicable characters and its importance for, disparity studies. *Proceedings of the Royal Society B*, **285**, 20181784.

Lamsdell, J. C. 2015. Horseshoe crab phylogeny and independent colonizations of fresh water: ecological invasion as a driver for morphological innovation. *Palaeontology*, **59**, 181-194.

Lamsdell, J. C. and Selden, P. A. 2017. From success to persistence: identifying an evolutionary regime shift in the diverse Paleozoic aquatic arthropod group Eurypterida, driven by the Devonian biotic crisis. *Evolution*, **71**, 95–110.

Legendre, P. and Legendre, L. 1998. *Numerical Ecology*. 2nd edition, Elsevier, Amsterdam.

Lehmann, O. E. R., Ezcurra, M. D., Butler, R. J. and Lloyd, G. T. 2019. Biases with the generalized Euclidean distance in disparity analyses when missing data is abundant. Dryad Digital Repository. doi: https://doi.org/10.5061/dryad.4cv1421

Lingoes, J. C. 1971. Some boundary conditions for a monotone analysis of symmetric matrices. *Psychometrika*, **36**, 195–203.

Lloyd, G. T. 2016. Estimating morphological diversity and tempo with discrete character-taxon matrices: implementation, challenges, and future directions. *Biological Journal of the Linnean Society*, **118**, 131–151.

Marx, F. G. and Fordyce R. E. 2015. Baleen boom and bust: a synthesis of mysticete phylogeny, diversity and disparity. *Royal Society Open Science*, **2**, 140434. doi :10.1098/rsos.140434.

Oyston, J. W., Hughes, M., Wagner P. J., Gerber S. and Wills, M. A. 2015. What limits the morphological disparity of clades? *Interface focus*, **5**, 20150042. doi:10.1098/rsfs.2015.0042.

Pérez, D. E. 2018. Phylogenetic relationships of the family Carditidae (Bivalvia: Archiheterodonta). *Journal of Systematic Palaeontology.* In press. doi: 10.1080/14772019.2018.1532463.

Prentice, K. C., Ruta, M. and Benton, M. J. 2011. Evolution of morphological disparity in pterosaurs. *Journal of Systematic Palaeontology*, **9**, 337–353.

R Core Team, 2018. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. URL: https://www.R-project.org/.

Ruta M., Angielczyk K. D., Fröbisch J. and Benton M. J. 2013a. Decoupling of morphological disparity and taxic diversity during the adaptive radiation of anomodont therapsids. *Proceedings of the Royal Society of London Series B, Biological Sciences*, **280**, 20131071. doi:10.1098/rspb.2013.1071.

Ruta M., Botha-Brink J., Mitchell S. A. and Benton M. J. 2013b. The radiation of cynodonts and the ground plan of mammalian morphological diversity. *Proceedings of the Royal Society of London Series B, Biological Sciences*, **280**, 20131865. doi:10.1098/rspb.2013.1865.

Smith, A. J., Rosario, M. V., Eiting, T. P. and Dumont, E. R. 2014. Joined at the hip: Linked characters and the problem of missing data in studies of disparity. *Evolution*, **68**, 2386–2400.

Thorne, P. M., Ruta, M. and Benton M. J. 2011. Resetting the evolution of marine reptiles at the Triassic-Jurassic boundary. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 8339–8344.

Toljagic, O. and Butler, R. J. 2013. Triassic-Jurassic mass extinction as trigger for the Mesozoic radiation of crocodylomorphs. *Biology Letters*, **9**, 20130095.

Wills, M. A. 1998. Crustacean disparity through the Phanerozoic: comparing morphological and stratigraphic data. *Biological Journal of the Linnean Society*, **65**, 455–500.

Wills, M. A. 2001. Morphological disparity: a primer. *In* Adrain, J. M., Edgecombe G. D. and Lieberman, B. S. (Eds.) Fossils, phylogeny, and form. Springer, Boston, MA, USA.

Wright, A. M., Lloyd, G. T. and Hillis, D. M. 2016. Modeling character change heterogeneity in phylogenetic analyses of morphology through the use of priors. *Systematic Biology*, **65**, 602-611.

FIGURE CAPTIONS

Figure 1. Scatter plots of distance to centroid against percentage of missing data in taxa of three selected matrices. The distances are calculated as the Euclidean distance in the morphospace with the corresponding number PCos, and are scaled to unit to facilitate comparisons. Regression lines are shown for each data set. “S” denotes a significant Spearman’s correlation between variables and “NS” a non-significant correlation, with α = 0.05. The distances are calculated based on the first 3 PCos and for all the PCos.

Figure 2. Histograms showing the percentage of matrices with significant and negative (black), non-significant (light grey), and significant and positive (dark grey) Spearman’s correlations for the GED and the MORD, with α = 0.05. The distances are calculated based on the first 3 PCos and all the PCos.

Figure 3. Variation in the Spearman’s correlation coefficient between the percentage of missing data in taxa and their Euclidean distance to the centroid through increasing percentages of explained variance for three selected data matrices. Filled symbols denote significant correlations, while open symbols indicate non-significant correlations, with α = 0.05.

Figure 4. Histograms showing the percentage of replications with significant and negative (black), non-significant (light grey), and significant and positive (dark grey) correlations between the Euclidean distance to the centroid and the percentage of missing data in taxa for three selected matrices, calculated for the GED and the MORD, with α = 0.05. Approximately 35% of the entries in the matrices were replaced with missing data, and this procedure was repeated 200 times for each matrix and distance measure. The distance to the centroid was calculated using all the PCos.

Figure 5. Disparity values for the studied disparity measures against the proportion of randomly-distributed missing entries added to morphological matrix of Pérez (2018), calculated from distance

matrices generated from the GED and the MORD. Each dot represents the mean of 200 replications. The values are scaled to unit for each distance and measure to allow an easier visual comparison. Vertical bars indicate the 95% confidence interval for the mean. Values of the percentage of missing data are treated as different categories, with the GED shifted slightly to the left and the MORD slightly to the right in order to avoid overlappings.

Figure 6. Results of simulations with groups of taxa and different distributions of missing data for the matrix of Pérez (2018). Each dot represents the mean of 200 values, with bars indicating the 95% bootstrap for the mean. The dashed vertical line represents the reference correlation of 0 (see text for a detailed explanation). Values closer to 1 and -1 indicate a stronger influence of the differences in the missing data per group in the ranking of the disparity measures. The column '%' indicates the percentage of missing data for the matrices in the corresponding simulations and the column 'S' indicates the simulation number according to Table 1.

Figure 7. Effect of the missing data in a disparity analysis. A: bidimensional morphospace generated with the original matrix of Pérez (2018), with only 3.7% of missing or inapplicable entries. B: bidimensional morphospace generated after the addition of 35% of missing entries to the taxa marked with black dots, producing their displacement (dotted lines) towards the centroid. C: bidimensional morphospace generated after the addition of 35% of missing entries to the entire matrix (following a beta distribution; same procedure as for the simulations, see text), showing that the spatial structure of the groups is diluted. The numbers in parentheses indicate the percentage of variance explained by each PCo. The morphospaces of B and C were rotated to match A via Procrustes to accommodate for differences in the rotation and scale of the points cloud.

Table 1. Thirteen simulations applied to five randomly-generated groups in selected matrices.