Response to the reviews : Content

Responses to the review by the anonymous reviewer# 1

General notes. The manuscript presents a new R package, and associated shiny application, called "trident". The aim of this package is to analyse dental microwears using various measurements to finally detect which measurements are the best to discriminate the dataset into categories based on the diet, taxonomy etc. To illustrate the different applications of this package, the authors analysed three datasets. They detailed the analytical protocol and the results for each case (A, B and C) as well as how the results highlight the power of discrimination of trident.

What makes the trident R package of high interest is that it gathers various dental microwear texture analyses (DMTA) in one place. It considerably eases the investigations and the shiny application helps the users that are not used with R language, providing .SUR files as inputs. Even if I have a very positive opinion on this manuscript, I still have multiple questions and comments that are detailed below and I wish the authors answer to.

Abstract. It synthesises well the aim of the study in overall, but there are few information that are too detailed and few other information that are missing in my opinion. I would advise to remove the sentence about the remove of polynomial surfaces (l. 25-26). I don't think specifying the number of variables and parameters (l. 26-27) is useful but it would be good to detail the "five different methods" (l. 28; isn't it four instead, as said at l. 207?). Finally, I would advise the authors to add a first sentence that reminds the reader about the interest of dental microwear in ecological investigations.

Responses: We added a sentence to refresh the reader regarding dental microwear textures.

We also removed the degree of the polynomial surfaces. We rephrase the sentence listing the number of parameters/variables. However, regarding the number of methods to be used for classifying the variables, we indeed have used 5 of them. This is explained in Material and Method/DMTA with trident/ Classification of variables/

Introduction. This section is well written in my opinion. I don't think Figure 1 is necessary in the main text. Could it be replaced by a reference? Or put in supplementary information? More details are required to better understand this figure. Indeed, there is no information related to the studied taxa and their ecology. The units of each variable (epLsar and asfc) should also be explained.

Responses: We agree with the reviewer. Figure 1 was complex because we included too many species so the main purpose of the Figure is blurred, indeed. We here proposed a figure with only 3 species and we rephrased the caption. We think we go now straight to the point.

trident methodology. I think the flowchart in Figure 2 would be more useful if it depicted the whole methodology of trident package, especially when the variables are ranked depending their ability to discriminate categories. Each case A, B and C could then be applied on this general flowchart, using coloured arrows for example.

Response: We understood that we did not specify that this paper is not only an analytic study but rather a focus on the presentation of new software with a case study for which the framework differs from each other, and so does the analytic process.

We have tried to make the figure according to the reviewer's comments. However, it is indeed too complex to read.

We also think the new supplementary file as requested by the first reviewer that illustrates screenshots of the 3 case studies is indeed a bonus for readers. We think it somehow responds to the reviewer's comment.

I am starting to get lost from lines 280 to 292. Why proposing all these protocols that have only subtle differences?

Response: Here we propose 5 ways to rank variables. Why? The comments of the reviewer make it clear we need to explain why there is 5 ways.

The 1st and the 2nd ways of ranking are equivalent as the first mode is for variables whose distribution respects conditions required for parametric tests. The purpose of these two modes is just to separate discriminant from non-discriminant variables. To make things easier for users, variables are ranked by the p values of the ANOVA (or Kruskal Wallis analysis). One could be interested to have a first glance at the discriminant variables and most discriminating ones whatever the number of groups involved in the study.

3rd choice: To go further in classifying discriminant variables, we can classify them by the number of significant differences. For instance, a variable could be well classified with the p values of the ANOVA but not so well when the number of discriminant groups is considered. We want here to target

the variables that discriminate the highest number of groups. Then, among variables discriminating the same number of groups, we ordered them by the increasing values of the mean of the significant p-values of the post hoc (HSD Tukey). For instance, with 4 discriminant groups among 7, the mean of the 6 p values of the 6 pairs of significantly different samples is computed for each discriminant variable. They are then ordered. One could then run a PCA on trident or export the dataset with selected variables to any other external software to run any analysis types of his/her choice.

4th mode: this is very similar to the 3rd one. However, it is made to emphasize specific pairs. Indeed, we do not want to select all of the discriminant variables but we specifically target pairs of species. For instance, it can be used for a study including 5 species among which we have hard time discriminating two of them with traditional dental microwear textures while we suspect differences in feeding preferences as they occupy different habitats. In such a case, we can run trident using the ranking of variables by targeting the ones discriminating these two species with significant overlaps.

5th mode. TOP3 gives the 3 most discriminant variables for each pair of samples, so it cannot be based on p-value of ANOVA (or Kruskal Wallis), but on the post hoc (here Tukey HSD) test for each pair. In most cases when comparing a few groups, three parameters appear to be enough to choose two low-correlated parameters for a biplot. So let's say the easiest and fastest way to find out which and how variables discriminate the groups.

l. 280-283: I don’t understand the two sentences and so the protocol. What type of arrangement is it at line 280; what is this calculation of mean p-value at lines 281 and 282; what is the post-hoc tests at line 282; what p values are mentioned at line 283?

l. 284-285: what is “a given pair”? What is a “post-hoc p value”?

l.286-292: what data are used to perform Tukey’s HSD?

Response: I think we need to clarify what are post hoc test: HSD Tukey or LSD Fisher tests. In most of the time, we prefer to use the HSD Tukey because it is more conservative than LSD. Taking into account the high number of discriminant variables, we can be conservative by using the HSD Tukey instead the LSD Fisher test?

Case-specific analyses. To be honest, I have not fully understood the applied protocol for each case. A justification of the methodological choices is also missing: why choosing different protocols for these cases?

Response: This paper shows the application of a new tool. So we have chosen different case studies with specific frame to illustrate the possibilities offered by the software.

The first case has been chosen because there is a double analysis of crushing and shearing facets. We extract the most discriminant variables using the TOP 3 option on each of the two datasets, then they are fused using trident, and then a single PCA is produced, still using trident. This allows us to see either if one of the two facets bears more discriminant variables or if they are complementary for discriminating groups.

Second case: after ranking the variables, one could see that several variables can be issued from the same texture parameter: for instance, it could be the min.25, the mean, the max 25, the SD from the parameter Sal. Here, we have then used only the most discriminant variables per parameter, meaning the most discriminant statistics per parameter. For instance, "min.25_Sal".

Third case: this is not so different from the second case although the sample size and numbers of groups are lower. It however differs in our aim to integrate fossils in a PCA built only with modern species. Paleobiologists very often use such a workflow. The two-folded analysis (model + insertion of the fossil) obliges us to not use box-cox but log-transformation on raw data. Indeed, box-cox transformation being sample-dependent, we could not have box-cox transformed the whole set of data (from both extant and extinct specimens) and then run a PCA with a portion of them (extant specimens).

The authors keep the variables that are best to discriminate the categories in the different datasets. It is interesting to know which variables differentiate the most the categories, but what is finally the biological meaning of the variables? For me it is still a "jungle of parameters" if we cannot connect the variables to specific type of dental microwears and so to a specific diet.

Response:

We understand. However, we now propose a table describing parameters.

It would have been interested to compare the retained variables between the three cases to see the similarities/differences that could be imputed to the diet but also to the way the animals process the food. Perhaps the correlation circles in Figures 3-5 could be simplified in replacing the name of the variables by the type of microwear they refer to.

Response: We understand the purpose here. It makes sense, but we do think this paper should be seen as some kind of software solution "tutorial" with few independent case studies. There are too much parameters here (experimental vs. wild data, wild data with detailed individual life history traits vs. wild data from large database).

Case A.

l. 147 "… was fed 30% of barley seeds": does it mean the diet was 70% base + 30% barley, or 100% base + 30% barley? I have the same question for cases B and C.

Response: The diet was 70% base + 30% barley. Cases B and C are wild animals so we have no clues about the exact food proportion in their diet the few days or weeks before their death.

l. 307 "Variables that passed the multi check were classified": what are the other variables and are they removed from the further classification? Wouldn't it be better to favour a Kruskal-Wallis test to keep them in the analyses? Same question for cases B (l. 317) and C (l. 330-331).

Response: The multi check includes the analysis of variances following parametric and non-parametric ways. So, we don't have to decide. The distributions of variables are the only criteria to select parametric and non-parametric ways.

l. 356 "the other groups are distinctly separated": this can be tested using a MANOVA but it depends what the authors want to test, either a significant difference on the first PCA axes (even if it gathers only 60% of the variance), or a significant different based on the input variables that were used to build the PCA.

Response: We agree. It can be tested with MANOVA on either PCs or on variables themselves. However, it has no meaning because the very variables used to set up the PCA (Figure 3) were indeed selected because they show the largest differences between groups arranged by pair for the groups of dental facets (see Tables 3 and 4). Those variables have been targeted by the trident software using the TOP3 option (see Methods section).

I cannot read Tables 3 and 4. What do the columns "p value ANOVA" and "Post-hoc p values" refer to? I don't understand the ranking of the variables since they should be arranged based on their p-value: for example in Table 3, why Sk2 (p-value=0.04) is above Ssk (p-value=0.02) for the pair Co-CK?

Response: We understand it can be confusing. So we make the simpler tables and refer to the Method section in the table captions.

l. 459: the separation between the groups has to be tested (see above).

Response: it does not have meaning because the PCA was built with the most discriminant variables pre-identified through the multicheck. Tables 3 and 4 show that for most of the variables, differences are supported by the most conservative post hoc tests.

Case B.

l. 323-324: why removing correlated parameters in this case (and in case C, l. 335-336) and not in case A? And why not integrating this step to the R script/trident protocol since it is used in 2/3 of the cases?

Response: In Cases B and C, we selected a single list of discriminant variables for which it was quite easy to extract correlated variables. In case A, there are N lists of discriminant variables (from which TOP3s are extracted) for N/2 pair of groups and each of two dental facet types. Thus, the number of variables to extract would have been high and counterproductive as extracting one variable for a given pair would have handicapped the possibility of discriminating different groups in other pairs. However, this is possible as trident proposes tables of correlation.

l. 385-394: since the diet of the taxa is one of the main focus of the results, I think it would be useful to add this information in Figure 4.

Response: We think that the results section is sufficiently explicit. See below

"When comparing the means between species (Fig. 4C), the most folivorous species (*Trachypithecus auratus*, *Colobus guereza* and *Piliocolobus badius*) have the lowest PC1 values. They are followed by terrestrial graminivorous papionines *Papio hamadryas* and *Theropithecus gelada*, then *Nasalis larvatus*, *Semnopithecus entellus* and *Trachypithecus cristatus*. The three latters species are also folivorous but present higher Asfc2 values in our sample, indicating the opportunistic consumption of seeds (Thiery et al., 2021). This is supported by the surprisingly large breadth of PC1 value dispersion for these three species, especially T. cristatus. Then, opportunistic terrestrial cercopithecines and papionines show higher PC1 values, with the highest values found in the hard seed predator *Lophocebus albigena* (Lambert et al., 2004) and *Macaca*

sylvanus, one of the most granivorous macaque (Kato et al., 2014).”

l. 483-484: if there is a continuum, then there is a significant difference between the two extremes that can be statistically tested. This continuum is observed only on PC1 which accounts for ~43% of the overall variance: the variables that contribute the most to this PC should be used as input to test this between-group difference.

Response: we understand. However, we should here keep in mind that the PCA has been built with the most discriminant species. For instance, in table 5 the first variable is the most discriminant as the p value of the ANOVA (<0.01) and the number of significant differences (=37) prove it.

l. 503: I don’t see any biomechanics in the discussion.

Response: I agree. But once again, the paper is first some kind of tutorial. The exploration of factors driving differences in dental microwear textures is another topic here

Case C. The choice of the different taxa is not justified in the main text: do they share phylogenetic relationships? Ecological similarities?

Response: We disagree here. The material section is indeed clear.

“The Bauges Natural Regional Park is a typical subalpine massif located in the French Alps (….) Mandibles were collected at the same locality, during a short period (for more details, see Merceron, Berlioz, et al., 2021), representing a hypothetical fossil assemblage composed of different species occupying different small-scale habitats (open alpine grassland, bushland, shrubland, deciduous, mixed, coniferous forests) in a common geographical range.”

l. 337-338: “The remaining variables were used for a PCA. At this point, the surfaces of *Gazellospira torticornis* were added as supplementary individuals to the PCA”: so the retained variables for the extinct species are the same than the retained variables for the extant species?

Response: Yes.

l. 420-427: the authors mention significant differences between the extinct species and some of the extant taxa. These differences can be statistically tested with a MANOVA, based on the input variables that build the PCA.

Response: Yes

Reproducibility. I was not able to use the shiny app as no .SUR file was available as supplementary material. I also tried to use trident package using the .TXT files given as supplementary material. The categories for each case is not easily available as I had to build data frames based on the information from Tables S1-S3. I found the function "trident.arrange" a bit cryptic since there is no detail about the available parameters for the argument "by".

Response: the new version corrects this. However, be careful as specific versions of R and specific packages have to be installed.

It would be great if the authors provided the script they used to analyse the data on R and/or the .SUR files so the reader could reproduce the results with the shiny app. The functions could be mentioned in the main text of the manuscript so the reader would know exactly what function to use for each methodological step.

Response: the R script is provided in GitHub and the raw data (Plu.plux, mnt, and sur files) are available in a CNRS/MNHN "InDores" repository.

Finally, as the retained variables are specific for each data set, how the analyses can be reproduced within a given taxonomic group? For example, if the data set of the domestic pigs is increased with new data (new specimens or even new diets), the retained variables might change. Then, it would be impossible to compare the new results with the old ones.

Response: yes, every study will generate different sets of significant variables. This is the weak and strong point here. The retained variables might change (but not necessarily), and it does not make the comparison impossible. Yes, combining data from different studies (unless the full dataset is provided) would be impossible, but not to compare them. Depending on the research question, it could be interesting to check for similarities/dissimilarities in the sets of significant variables between studies. If we include a new group of pigs fed with another type of herbs, do we detect the same (or similar) variables distinguishing pigs fed on corn silage from pigs fed on seeds; or do we detect some other variables, if yes which ones? This could make sense biologically.

Additional remarks.

l. 114: "… and measure 16 variables". I see only 15 variables in Table 2.

Response: Yes, table 15 include the 15 statistics for a given parameter, but the 16th is the single parameter value calculated for the whole surface.

l. 230 and 231: I don't understand this sentence. What are the 360 computed variables?

Response: as we explored variations of parameters over each surface by using sub-sampling, we have 1 global value + 15 statistics values (Mean, kurtosis, …; generated from N sub surfaces) for each of the texture parameters (Asfc2, Sa, Sal, …) so it make a high important of variables as a parameter such as Sa is then present with its mean, skewness, standard deviation among others. trident runs a routine to target the most discriminant ones to reduce the dataset.

l.232-238: I don't think Box.1 is useful since it is the only box in the whole manuscript. The text could be integrated in the main text.

Response: We agree and integrate this section in the main text.

l. 494-496: I don't think this information is useful here.

Response: We keep this information in the text because although the surface-view tool is not included in the graphic interface, one could use it through the trident R package.

Response to the review comments made by Mugino Kubo.

Thank you very much for the opportunity to review this very interesting paper. This research is closely related to my interests and the research that I am currently working on.

In this manuscript, a program package called "trident" developed by the authors for statistical comparison of dental microwear is introduced and its usefulness is demonstrated using three examples. As a researcher deeply involved in dental microwear research, I would like to express my sincere respect to the authors for developing such a wonderful program. This is because recent 3D dental microwear texture analysis (DMTA) involves the calculation of parameters that characterize microwear properties from the surface, but this computation is highly dependent on paid software (MountainsMap, sold by the French company DigitalSurf), which is a potential barrier to entry in this field (because of its high functionality, MountainsMap is very expensive).

While the paper is clearly written and has no major problems, I believe it is important to further clarify the reliability and usefulness of this program in order for it to be widely used, and some suggestions for manuscript revision are provided below. Some of these require additional calculations, graphing, figures, etc., and I encourage the authors to include them in the revised manuscript. Other minor comments are noted in the PDF.

I hope to see the revision published in PCIPaleo.

1) Prerequisite for trident

Authors explained that trident can read the SUR file, but on p. 9 L. 183-184 the authors write "All surfaces were preprocessed following Mercecron et al, (2016)". It is necessary to clearly state whether software other than trident is required to prepare the SUR files for analysis and what further processing was done, rather than just showing a citation. This is important to indicate to the reader whether the entire analysis can be completed using only trident, or whether the data can only be used after it has been acquired with profilometers and then pre-processed with software other than trident.

Response: In the present case, we have used pretreated surfaces that were already published (Louail et al. 2021; Merceron et al. 2021a, 2021b; Hermier et al. 2020; Thiery et al. 2021). In these earlier studies, surfaces were treated with LeicaMap, MountainsMap-derived software provided with the DCM8 surface profilometer (which is a Sensofar based microscope) However, one could have pretreated surfaces with alternative open-access software such Gwyddion.

We have complemented the appropriate section in Material and Methods/Surface acquisition.

2) Comparison of calculated values with the standard analysis software MountainsMap

As mentioned earlier, MountainsMap is widely used as the standard analysis software for DMTA. Therefore, if the values calculated with trident are shown to be consistent with those shown in previous studies, users who have already used MountainsMap can use trident with

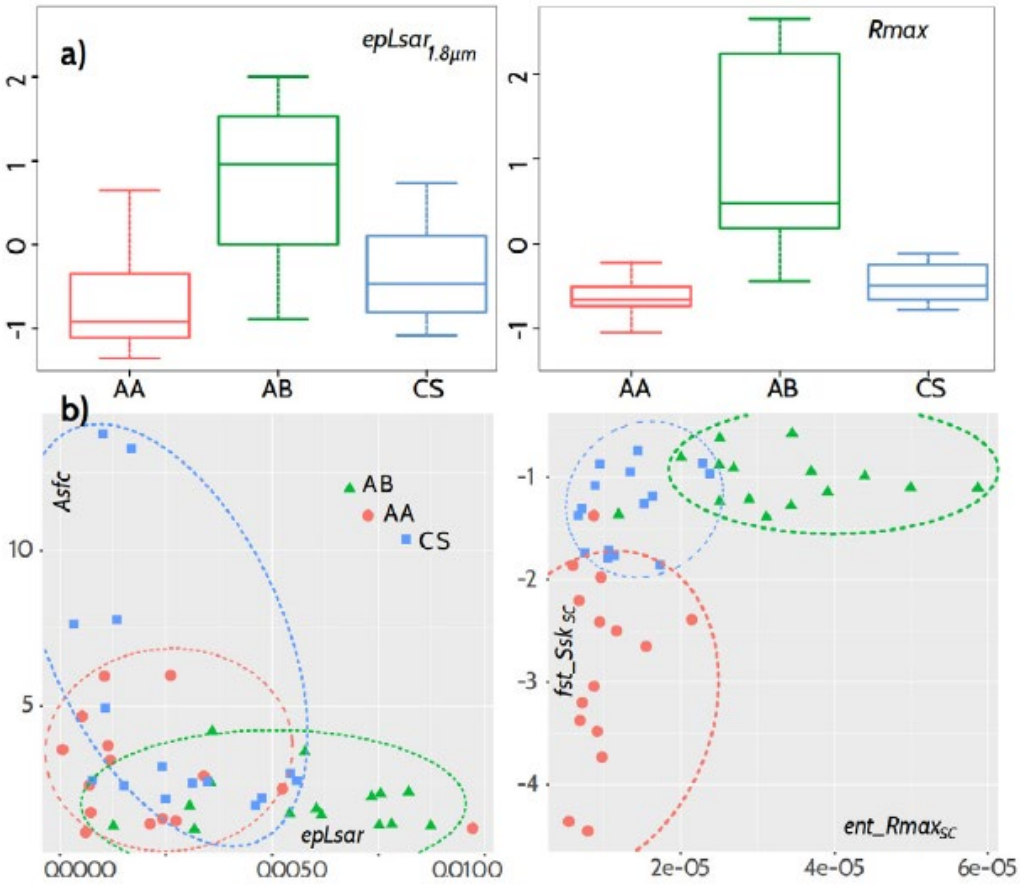
confidence. Specifically, for basic parameters such as Sq, it would be good to show that the values calculated by MountainsMap and those calculated by trident match on some (e.g. N=10) surface data (without subsampling of the surface).

Response: thank you for this suggestion. We have inserted as supplementary data a file (supplementaryMaterial_Software comparisons.txt) containing the height parameters calculated with both trident and LeicaMap v.8.2. The reader can see that it fits exactly although the values calculated with trident are always provided with the official unit (Meter) while the MountainMap derived solution provide adapted values; here in micrometer rather than in meters.

3) Advantages of trident

The superiority of trident's new method (subsampling of the surface and then obtaining the dispersion indices of the parameters, which are then used as new parameters) over the traditional univariate comparison of DMTA parameters, or PCA with multiple parameters without ranking the variables should be shown.

Response. The routine (including ANOVAs and post hoc tests) targeting the most discriminant variables upstream of the PCA is evidence of the efficiency of trident. We make mention of the earlier studies done with this routine (Francisco et al., 2021a, 2021b; Louail et al. 2021; Merceron et al. 2021a). In all of these cases, the most discriminant variables are often the ones describing the dispersion of the parameters. Thus, the groups are much better separated than with the traditional method using a single value per surface. To exemplify this, we will refer to the figure below (from Francisco et al. 2018. Surf. Topogr.: Metrol. Propr. 6: 015002). You can see Rmax (higher for anisotropic surfaces) is much more efficient to discriminate *Alcelaphus buselaphus* (AB) from the other species and the value of the first 5% quantile of the height skewness (fst_Ssk_{SC}) is more efficient to discriminate *Alces alces* (AA) from the other species; *Alces alces* having significantly more sub-areas with flat surface with few deep microwear scars.



For example, Fig. 4C shows a box plot of PC1 using the trident method, but next to it is a box plot using univariate parameter (e.g. Sq), the variable with the largest significant difference between groups using the conventional method, or the conventional PCs, showing that trident's PC1 better captures the trend of the feeding groups. Similarly in Fig. 5, a comparison with the conventional method would be more convincing.

Response. Here we agree with the reviewer. This is a good idea. We integrate such comparison in Figure 3 (with a biplot Asfc vs epLsar), Figure 4 (with a boxplot on Asfc) and Figure 5 (with two boxplots in Sal and Sk parameters).

4) Screenshot of trident in use

The Supplement includes a manual for trident, which describes in detail the interface and usage of trident, but it would be easier for readers to understand what the program is like if the main manuscript also includes screenshots of trident in use.

Response. We fully agree with the reviewer. Now we integrate a piece of new supplementary information that illustrates with screenshots the 3 case studies. We do believe this will be a bonus for readers. Thanks for this idea

None of the above comments require a great deal of effort on the part of the authors. Again, it is hoped that this software will lead to more DMTA research cases. The search for parameters that better illustrate differences in food properties (including derived parameters obtained by subsampling surfaces) will also provide a better understanding of how differences in food properties produce microwear features, i.e., the etiology of microwear. These future prospects should be discussed as a perspective at the end of the Discussion

Response. We understand the comments here. But as mentioned the benefit of the process has been discussed in earlier publications and we prefer to focus the scope on the software presentation.