

Hominin presence in Eurasia by at least 1.95 million years ago

Corresponding Author: Dr Sabrina Curran

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Attachments originally included by the reviewers as part of their assessment can be found at the end of this file.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Curran et al discuss evidence suggesting the presence of hominins in Romania approximately 2 million years ago, predating the widely accepted oldest Eurasian site of Dmanisi (Georgia). Their evidence includes cut-marked bones and new Uranium-Lead (U-Pb) dates of fossil dentine. The authors also use horse enamel isotope data for a coarse paleoclimate and paleoenvironmental reconstruction. Except for a few typos and awkward sentences, the manuscript is generally well-written. Figures need a little extra work (see below), but they are fairly well-crafted. Unfortunately, I must defer the evaluation of the hominin presence and U-Pb components of the paper to other reviewers, as they are outside my area of expertise. I can only say that their arguments and conclusions seem reasonable in light of the data to a non-expert.

My main concern with this manuscript is the oversimplification of the interpretation of the isotope data. More specifically, here are the two main issues:

1) Interpretation of enamel oxygen isotope data. The authors interpret lower-than-present precipitation $\delta^{18}\text{O}$ (calculated from horse enamel $\delta^{18}\text{O}$) as an increase in winter precipitation component to aquifer recharge relative to the present. This is one of the many possible explanations. Alternative possibilities include a lower mean annual temperature, and a change in air mass trajectories. These alternative possibilities must be discussed here in depth. Furthermore, there is no evidence that the horses were drinking groundwater. Finally, the uncertainty in the average $\delta^{18}\text{O}$ value of precipitation is largely underestimated. The authors need to perform a rigorous error analysis and propagate the errors from ALL sources.

2) Interpretation of enamel carbon isotope data. There are many issues with the calculation of MAP from enamel $\delta^{13}\text{C}$, including a) the possibility that the investigated taxon was selective in terms of its diet and habitat; b) the uncertainty in the enrichment factor between diet and enamel; and c) the uncertainty in the relationship between the $\delta^{13}\text{C}$ of modern C3 plants and MAP. Like with $\delta^{18}\text{O}$, the authors largely underestimate the error in the reconstructed MAP. I am inclined to say that because of all these issues these MAP calculations and paleoenvironmental reconstructions have such a large uncertainty that they are almost meaningless, especially if only one taxon is used like in this case. The authors should consider eliminating the calculated MAP from the manuscript unless they provide a robust estimate for the uncertainty.

These two issues must be fixed before the manuscript can be considered for publication in Nature Communications. Minor issues include hard to read figures and text. For instance, the text in Fig. 1 is hard to read (especially in the inset) and the “18” and “13” in Fig. 2 should be superscripts. Again, the manuscript also needs some minor editorial work to fix some typos and to rewrite some poorly structured sentences.

Reviewer #2

(Remarks to the Author)

The paper by Curran and colleagues titled “Hominin presence in Eurasia by at least 1.95 million years ago” presents a highly important find that hominins were indeed present and active outside of Africa at least 1.95 million years ago. I have thoroughly enjoyed reviewing this paper, and certainly think it is relevant and falls within the scope of this journal. As the authors themselves point out, however, extraordinary claims require extraordinary evidence to avoid and answer possible

scepticism. While I think the photographic evidence of the cut marks is very compelling, I will focus on assuring that the classifications and analyses are as rigorous as possible. For this reason, I will therefore dedicate my review solely to the taphonomic questions of the paper, as all other aspects are far from my area of expertise.

First of all, I think a terminological clarification is needed, as I find the use of the word score throughout the text to be quite problematic. The word “score” typically refers to carnivoran activity, as in “tooth score”, while I have very rarely seen people refer to cut marks as scores. It is fundamental that we are as clear as possible in the words we use to describe certain processes, so as to avoid confusion, and because of this I encourage the authors to use a different term that is less likely to be mistaken with other taphonomic alterations (we already have enough issues with taphonomic equifinality, we don’t want to add any “terminological equifinality” to the mix!). In the taphonomic atlas published by Fernández-Jalvo and Andrews, for example, the authors talk about a “linear” category, where both tooth scores and cut marks can be considered “linear marks” – they have a length that is more than twice their width. I think it is much more appropriate to use a term like this, than use a word that is already shared in the name of a different type of taphonomic alteration.

I also think more evidence about how the collections were obtained and how they are currently being curated is also very important, which ties in to my next comment about the moulding process. The authors state that at least two moulds were taken per cut mark, however a recent study by Valtierra et al. (2023, DOI: 10.1111/arcm.12938) have shown that successive extraction of moulds can have an effect on the morphology of the traces being studied. While they showed that statistically these changes are not too important, this work was carried out on the morphological study of a single profile along the BSM, at a very specific point. It is also important to point out that with each mould taken, the natural variability of metric variables decrease, which could have a potential impact on certain analyses. While some of the variables used in the quantitative analysis of the present study are similar and should not, therefore, change too much, I am pretty sure that there will be a change in the morphology of the trace after each mould is taken. This leads me to ask, which moulds were used for the quantitative analysis? The first mould taken? Or the last mould taken? How were the fossil specimens cleaned as well? When performing such high resolution analyses, and considering variables such as volume, these different processes could have an impact. While I don’t think this would have too much of an impact, it is important to at least document and be aware of it.

Beyond overlying trephic modifications, I also think it is very important to consider the impact dendritic etchings and root marks may have had on the samples, especially when considering the methodology employed for quantitative BSM analysis and identification. For example, in VGr.0519, VGr.1557, VGr.1767, VGr.2275, VGr.2687 and VGr.2649 you can see overlapping biochemical alterations that would surely affect the calculation of BSM volume and other parameters, wouldn’t these parameters now be unreliable for the identification of the traces? To what extent should the analysis therefore include these variables? What happens when you perform calculations on some of the samples including or excluding some of these parameters? Do the classifications change? It is very true that almost any methodology dedicated to the high-resolution analysis of BSMs would be conditioned by these alterations, however, calculations such as volume are more likely to be effected by such changes to the cortical surface.

With regards to data acquisition, much more information is needed regarding the microscopic analysis, including the precise FOV and numerical aperture of the lens used and the windows scanned; variables that have been proven in the past by authors such as Calandra et al. (2019, DOI: 10.1038/s41598-019-42713-w) to effect the replicability of analyses. Also, if I recall correctly, the S Neox has the option of obtaining 3D models using either Focus Variation or Confocal techniques, that would also condition the quality of the 3D model produced, so the modality must be stated. Why was only 5x magnification used? What colour light was used? My own experience with a similar microscope found 5x and 20x to produce different results for parameters such as roughness and volume when doing use-wear analysis. For this reason I’d also argue that comparing data obtained from the S Neox and the Nanovea ST400 could definitely be considered problematic for some of the quantitative variables included in the study (again, see the work of Ivan Calandra). The authors state that the results should not be impacted because of “coarser scales of measurement” (which would probably answer my question of why was only 5x used), however this is yet to be tested, and when you calculate variables such as volume and roughness, then multiple researchers, including myself, would in fact disagree with this statement. Information about how many non-measured points during the scans are also required information, so as to properly assess the quality of the scans used for the analysis.

Regarding the post-processing of the 3D models obtained using the microscope, the author states that they removed form from the data, yet no indication about what polynomial was used to do this. I have to say that I think form removal for cut mark analysis is potentially very problematic, however I understand why it has been done considering some of the variables that have been measured. Form is removed in microwear studies because this directly has an impact on roughness parameters, and will also influence how easy it is to compute each traces volume, however when we are talking about the properties of the profile, form removal will have an effect on mark morphology.

With regards to the statistical analysis of the data, and classification tasks performed, I also have some concerns that need addressing. First of all, Quadratic Discriminant Analysis (QDA) should not be used for ordination, as has been done for supplementary figure 22. While use of QDA algorithms is perfectly acceptable for classification, it should not be used for ordination for a number of mathematical and epistemological reasons. QDA is a classification algorithm, so of course there will be separation between samples in the constructed and visualised feature spaces, because that’s what it is designed to do – force separation to exist. There is also a huge issue with these types of constructed feature spaces due to issues with sample size and imbalance. While the individual to variable ratio is acceptable for the dataset provided, the sample imbalance is large enough to be problematic, in particular between cut marks and percussion marks (side note – why are these two compared in general, are there people who consider equifinality to exist between percussion and cut marks?).

What the authors should use instead, if they wish to visualise the multivariate differences between samples, is perform and visualise a PCA with a biplot, or simply present boxplots for each variable separately. If they want to use the QDA figure, then it must be used as a complementary representation of the data. It's also important to ensure that the feature space is constructed using cross-validation, otherwise separation is much more likely to be spurious in nature. For an example, see <https://doi.org/10.1007/s11692-020-09494-x> - although this paper was done for the purpose of discussing use of similar techniques in geometric morphometrics, the algebraic properties of the ordination technique are very similar, while the input to the figures in this paper are not landmark data but instead standard multivariate distributions and are thus completely comparable to the present case study. Finally, an ordination technique like these ones should not be presented without including in the axes labels the percentage of informative that is captured by each dimension, otherwise they are very hard to interpret.

For the evaluation of how well a classification algorithm works before being applied to archaeological data (presented in the first paragraph of page 22), the "accuracy" metric is often misleading, especially in cases where the imbalance is notable. Instead the authors must present other metrics such as precision, recall, sensitivity or specificity, that are more informative of other parts of the confusion matrix

Some minor comments:

In the results section of the main article the authors state that a reduced number of bones presenting polishing is proof of limited transport. Polish does not necessarily imply movement, it's the action of water-born abrasive particles altering the overall surface of the bone, this could occur with a low flowing course of water over a prolonged period of time without moving the bones at all. In the supplementary materials the authors mention rounding as well. I think it's more informative to simply present all of the water-based alterations to cortical surfaces in the main text as well.

In the last paragraph of the taphonomy section in the results the authors mention that some specimens have "more degraded surfaces", it would be nice if the nature of this degradation be clarified – is it due to fluvial processes, for example?

Have percussion marks or evidence of anthropogenic fracture patterns been identified in the site?

On line 227 the authors include some citations of papers that support the notion that sites exist with identified cut marks but no associated lithic assemblages. Among these cited papers, they include BK from Olduvai, however BK does contain a huge amount of lithics published as far back as the Leakeys, and is even mentioned in the abstract of the very paper they cite. I can assure you BK has lithics, while Mary Leakey even noted the presence of handaxes from this site. You also cite Dikika, which is very problematic, and many would even argue that not only are there no lithics, but no cut marks either

In summary, I once again wish to thank the authors for their work, I truly have no problem at all with the validity of their finds, I just hope to ensure that the presentation of results are as rigorous, robust, and transparent, as possible. I have no problem disclosing my name as a reviewer - Lloyd A. Courtenay

Reviewer #3

(Remarks to the Author)

The paper presented by Curran and colleagues, entitled "Hominin presence in Eurasia by at least 1.95 million years ago," presents clear evidence of ancient anthropogenic activity in the form of butchery marks from several sites in the Olteț River Valley in Romania. The remains come mainly from a single Grăunceanu assemblage, although they also provide data from some other penecontemporaneous sites. A list of U-Pb dates establishes an age of 1.95 Ma. This work offers highly relevant information for understanding human dispersal prior to 2 Ma and provides palaeoenvironmental data, through isotopic analysis, suggesting conditions conducive to the development of such human movements.

The authors present bone remains with unquestionable cut marks, which have been analyzed using both quantitative and qualitative methods. Nevertheless, it is surprising that the authors state that the fossils were examined under strong, low-angled light for evidence of bone surface modifications. It would have been more appropriate to use microscopic observational methods (stereomicroscope, ESEM, digital microscopes, etc.). It seems plausible to suggest that this is the reason why the analysis method of Rodrigo et al., 2009, has been modified, with the aim of avoiding the microscopic variables of the analysis. It is likely that had a microscopic approach been employed, a greater number of marks would have been identified. The marks attributed to cut marks are clearly visible to the naked eye, and therefore it is reasonable to hypothesize that a microscopic analysis would have identified a greater number of evidence, as well as helped to attribute some of the probable marks to high-confidence marks.

The authors propose that some of the striae are attributable to high-confidence cut marks, while others may be associated with probable or unknown marks. Although they employ a quantitative methodology, they utilize a similarity-by-association approach to ascertain the origin of the marks as either anthropogenic or non-anthropogenic. It would be preferable for the authors to not use association to attribute the marks, but rather to only apply the proposed analytical principles. It is unclear why the teeth were excluded from the taphonomical analysis. If one of the objectives was to identify butchering marks, the teeth may show evidence of cuts related to the cheek defleshing and/or tongue removal. In addition, some of them could have been dated, obtaining direct dates from remains modified by lithic tools.

In the text (third paragraph of the Taphonomy section), the number of remains with high-confidence cut marks is initially indicated as seven. However, later in the text, it is stated that there are eight remains, and these are also the number of

fossils that appear in Table S2. Therefore, it is necessary to correct the text.

In some instances, the observed marks appear to be clearly anthropic in origin. However, the analysis suggests otherwise. This discrepancy may be a consequence of the macroscopic analysis developed by the authors and the omission of certain microscopic variables that could have biased the results. A good example is the metopodial VGr.0429. Upon examination of the image (Fig. S12), it appears to be a clear anthropic disarticulation mark. However, the analysis suggests that it is, in fact, a tooth mark. It is recommended that these probable cutmarks be reanalyzed under microscopic tools (a stereomicroscope, at a minimum).

VGr.1490 is described as a cutmark with a high probability and a high confidence, yet it appears both in the text (page 6 of SI) and in table S2 as a probable cut mark. Could this be an error? In the SI description, the authors even assign the activity carried out by hominins, tendon extraction.

Why do the authors attribute the mark observed on specimen VGr.2052 (Fig. S18) to trampling if the analysis indicates that it is a cut mark with a high probability? This suggests that the authors sometimes adhere to the result of the quantitative analysis, but at other times, the observation of the researchers takes precedence.

Given the paucity of information on the stratigraphic context of these ancient collections, it is understandable that the authors refrain from offering interpretations regarding the type of occupation or the mode of access to the resources. However, it is crucial to clarify the authorship of the marks and their age.

Tables 1 and S7 are invaluable for the compilation of all ancient sites with evidence of hominids or anthropogenic activity (lithics and butchery marks) and the corresponding bibliography. It is recommended to include additional bibliographical references concerning published extended taphonomic studies of the cited sites:

- For the Ain Boucherit site (Algeria):

Cáceres, I., Chelli Cheheb, R., van der Made, J., Harichane, Z., Boulaghraief, K., Sahnouni, M., 2023. Assessing the subsistence strategies of the earliest North African inhabitants: evidence from the Early Pleistocene site of Ain Boucherit (Algeria). *Archaeological and Anthropological Sciences* 15, 87. <https://doi.org/10.1007/s12520-023-01783-8>

For the Sima del Elefante site (Spain):

Blasco, R., Blain, H.A., Rosell, J., Díez, J.C., Huguet, R., Rodríguez, J., Arsuaga, J.L., Bermúdez De Castro, J.M., Carbonell, E., 2011. Earliest evidence for human consumption of tortoises in the European Early Pleistocene from Sima del Elefante, Sierra de Atapuerca, Spain. *Journal of Human Evolution* 61, 503–509. <https://doi.org/10.1016/j.jhevol.2011.06.002>

Huguet, R., Saladié, P., Cáceres, I., Díez, C., Rosell, J., Bennàsar, M., Blasco, R., Esteban-Nadal, M., Gabucio, M.J., Rodríguez-Hidalgo, A., Carbonell, E., 2013. Successful subsistence strategies of the first humans in south-western Europe. *Quaternary International* 295, 168–182. <https://doi.org/10.1016/j.quaint.2012.11.015>.

The evidence presented is of significant interest and should be published. The utilization of microscopy to analyze the attributes of the cut marks is recommended in order to identify additional marks, correctly assign probable ones, and enhance the images presented. It is also recommended to add some high-resolution images of examples of trampling or carnivore toothmarks.

Reviewer #4

(Remarks to the Author)

Hominin presence in Eurasia by at least 1.95 million years ago, by Curran et al.

This is a very interesting work with some exciting results suggesting early evidence of human presence in East Europe by > 2 Ma, i.e. much older than thought until very recently. Please note that the evaluation of the cut marks on bones and whether there are human made is beyond my area of expertise. I will mostly focus my review on the chronology and general context of the fossils.

As a general comment, I would say that the manuscript is overall well written and easy to follow. The results and data presented apparently support the interpretation proposed by the authors. Actually, the main issues are in my opinion not related to what is shown in the work, but rather to what is NOT shown. In that regard, the absence of a detailed contextualization of the fossils is my major concern. I get that the fossils are coming from old excavations and that the various sites are most likely no longer accessible, but we have virtually no idea about their stratigraphic position, and whether the fossils represent one single or various assemblages. In other words, the integrity of the fossil assemblages is simply unknown. This major question mark about the integrity of the fossil assemblage, and whether all fossils correspond to a single depositional event, is actually not addressed in the present work, since the fossils dated with U-Pb are not those that have cut marks. While one may assume that they have all the same age, we actually don't know, and this should be verified by running additional analyses on the cut-marked bones. The same comment applies to the stable isotope analyses performed on the equid teeth. Direct U-series dating should be performed in order to confirm its antiquity and make sure it can be closely associated to the other fossils. In that regard, I genuinely wonder whether Rare Earth Elements could also help demonstrate that all fossils have the same origin and come from the same stratigraphic unit.

The whole context is briefly summarized in Methods, but remains (too) vague. The authors mention that the location of various localities have been 'reidentified', therefore it should be possible to get further information about the context, and establish a proper stratigraphic column, description of the sediment, with some additional information about the formation of the fossil horizon and the depositional environment for the 3 localities. A geological or geomorphological map of the ORV with the stratigraphic correlations of the various sections would also be very useful.

The lack of independent age control is also another major concern. There is some circular reasoning in using the biochronology (which is based on independently dated assemblages) to support the numerical ages obtained here. The uncertainty on biochronological inferences is usually minored and not really taken into account, and especially on large mammals. I doubt very much that the precision on the biochronological age is sufficient to undoubtedly position the site around 2.2-1.9 Ma, within a small 0.3 Ma age window. This would need further explanation with a dedicated section in SM, and in particular about the robustness of the age control on the localities used as chronological tie points. Rodents may possibly help refining the chronology of the assemblage (e.g. Maul & Markova QI, 2007), but seems like from Table S1 that there are too few remains. As mentioned above, the absence of independent age control adds a non-negligible uncertainty to the dating results, especially when knowing that dental tissues behave as open system, and may show U leaching, which would artificially age the sample. I had a quick look at Andreescu et al. QI (2013), and I didn't find the existing age control on the Tetoiu Fm very convincing to securely confirm its chronological attribution to the Early Pleistocene.

On an unrelated matter, I see in Table S5 that several samples are in secular equilibrium, while other do show U-leaching (i.e., beyond equilibrium values). For example, this is the case for sample VGr.2256-2. My first reaction would be that this outcome may seriously question the reliability of the U-Pb age of 1.95 Ma obtained on that same tooth. I think this aspect would deserve some additional discussion in the manuscript. From the information given in SM, I understand that both solution and laser ablation analyses were carried out on the teeth. Some samples were microdrilled. I am not sure to understand why, and this could be clarified in the text perhaps. It is also unclear to me how LA analyses were performed, either by drilling vertically within the dentine, or through mapping an area of the occlusal surface? If so, can we see the maps with the spatial distribution of the various ratios? Please add some extra explanations in SM. The exact location of the samples analyzed for U-series (given in Table S5) within the teeth could also be given. Fossil teeth are known to be highly heterogeneous in terms of spatial distribution of radioelement, and especially old teeth (e.g., Duval et al. QG, 2011) and it is unclear to me how this uncertainty was evaluated. Some domains within the dentine or enamel may locally show very distinct U-series data, which would bias the age results. It would be strongly advisable to sample other parts of each tooth to evaluate this variability and help identify more recent uranium uptake or leaching overprints.

Specific comments

1. Main text, 'The current best': What does 'best' mean here? Please be more specific.
2. What is the definition of 'ephemeral traces' here? There several occurrences of this term throughout the manuscript and I guess this should be explained somewhere.
3. U-Pb dating, ' 2.01 ± 0.20 Ma to 1.87 ± 0.16 '. Please indicate at least once in the main text whether errors given are 1 or 2σ . 'Ma' missing after ' 1.87 ± 0.16 '
4. U-Pb dating, 'which is ubiquitous in fossil teeth and bone samples': Do you mean at this specific site? Please clarify the sentence. There are actually many works showing the timing of the U uptake may significantly vary within a given dental tissue or bone (e.g., Grün and Taylor Ancient TL 1996; Grün, RM 2009; Duval et al., QG, 2011), between the various tissues of a given tooth, and between fossils from the same site/horizon.
5. U-Pb dating, 'Grăunceanu took place >1.95 Ma and potentially $>2.01 \pm 0.20$ Ma': actually these two ages mentioned here are virtually the same given the associated error. There is no need in my opinion to mention them both, one is sufficient.
6. U-Pb dating, 'These results make these sites the most securely dated localities from Early Pleistocene Europe': the concept of 'securely dated' is very relative. The absence of independent age control and the presence of minimum ages on fossils do not make the sites better constrained than others in my view.
7. Discussion, 'Our paleoenvironmental data for Grăunceanu indicate a more open, arid environment with some nearby woodlands and water sources': do you mean more open and arid than today? Please specify. Additionally, the estimated MAP for Grăunceanu is higher than current conditions, which seems to be in contradiction with the statement above on a more arid environment.
8. Discussion, 'While it is clear that the hominin presence in Eurasia at this time was not extensive': Why is it clear? And what do you mean by extensive? There is actually early evidence of human presence in Europe from Ukraine to Spain.
9. Fig. 3 caption: please specify whether errors are displayed at 1 or 2σ .
10. Fig. S24: I am not sure to understand the interest of showing elemental concentrations in dental tissues? What kind of information does it bring?
11. Fig. S24: Metal concentration or elemental concentration?
12. Please would it be possible to show instead the maps of U concentrations, U234/U-238 and apparent age for each of the dated samples?
13. Fig. S25: How was the present-day U234/U-238 measured?
14. Table S1: Does the composition of the fossil assemblage come from previous studies (if so, please cite the relevant references), or from the direct study of the specimens by the authors? If the data given for this table are coming from a new taxonomic revision of the fossil assemblage, then a dedicated section including a proper systematic paleontology study should be included.
15. Table S4: which U-234/U-238 values were used for the disequilibrium corrected age? This may be clarified, as Table S5 shows that there are several values available for each sample.
16. Table S5:

- Please clarify whether MXF and UMF samples have all been measured through solution analyses, or UMF samples have been analyzed using Laser Ablation?
- Where (=which lab) was Th-230/Th-232 ratio measured?
- Not sure I understand well the U mobility column. Do 'Closed' mean 'secular equilibrium', and 'U gain' mean that the ratio are still below secular equilibrium?
- The apparent U-Th age can be added to this table. Out of curiosity, how can we explain that the same tooth may provide an apparent U-Th age of <500 ka and a U-Pb age of > 1.9 Ma. This would probably deserve some additional comments in SM.

17. I have my doubts on the reliability of the age of many localities mentioned in that table. Can the authors at least add one column mentioning the dating methods employed to constrain the age of the site (e.g., U-Pb, palaeomagnetism, biochronology, ESR, TCN, etc.)?

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

I have read the revised manuscript by Curran et al., titled "Hominin presence in Eurasia by at least 1.95 million years ago." As I mentioned in my previous review, the manuscript is generally well-written and of broad interest. However, I am unable to comment on the robustness of the evidence for hominin presence or the U-Pb ages, as these techniques fall outside my expertise.

My main concern with the original version related to the interpretation of the stable isotope data. The authors have provided a thorough and detailed response to my comments, and I am overall satisfied with the revisions, which include:

- Calculations of the oxygen isotope composition of meteoric precipitation 1.95 Ma using three different transfer functions, along with an uncertainty estimate.
- A brief discussion justifying the conclusion of a higher winter precipitation component to aquifer recharge relative to today.
- Estimates for precipitation amounts in the driest and wettest months, rather than an average for the whole year.

I would only recommend that the authors clarify in the text that the Kohn et al. equation was developed for mean annual precipitation (MAP), and its validity for reconstructing seasonal precipitation is untested. Additionally, while Kohn (2010) did not provide uncertainties for his equation, these are discussed in Kohn and McKay (2012, Paleo 3).

Overall, I am satisfied with this revised version and recommend the manuscript for publication in Nature Communications.

Reviewer #2

(Remarks to the Author)

I am grateful to once again have the chance to review the paper by Curran and colleagues, and am happy to see that they have replied in a constructive way to each of my comments and concerns, clarifying or rectifying the majority of the points I raised. There are some points, however, that have not fully been touched upon, while others that the authors have well justified in their letter to me, the reviewer, but should still be clarified in the text. I would just like to point out that if these questions have occurred to me, they can easily occur to somebody else, and our objective should therefore be to have a text that is as transparent and informative as possible, so as to avoid any further scepticism or questions. I reply here to some points that I still think are worth debating, but are not entirely an issue if they are not confronted in the text, while I insist on some other points that I feel are fundamental and must still be clarified.

Author's response to reviewer: The molds measured by Pante were the second mold taken from each specimen.

Reply: Please include this in the paper, there is no reason why this very simple sentence cannot be included, thus responding to the reviewer's (my) concerns.

Author's response to reviewer: Valiterra et al. 2023 used different methods of data collection and analysis than those that are applied here so their results are not comparable to this study. Their methods replicate those of Maté-González et al., 2015, which used a low resolution structured light scanner and only data extracted from profiles.

I debate: The authors don't think there is a correlation between depth, width, and volume. This is a potentially mathematically flawed statement; please take a triangle, change its width and height, and see how any minute change will also have an impact on the area. Regardless of whether the measurements taken are not precisely the same, if other authors note that there is a metric change to cut marks due to the moulding process, then one cannot completely disregard the study because slightly different measurements are being taken by another study. Each of these metrics are inherently dependent on the global properties of a mark, even if measured on only a single profile, or the entire mark. Also, if other authors were able to observe changes in the metric properties of cut marks at such a 'low resolution', then wouldn't these changes be more pronounced at a higher resolution? Half the time, lower resolution techniques are not able to differentiate between slight variations in variables, but high resolution techniques are, so if a low resolution technique is able to detect something, then it would be reasonable to assume that the high resolution technique is more likely to detect such a pattern. This can be simply acknowledged in the revised version of the text.

Author's response to reviewer: They also tested the impact of multiple moldings on modern and not fossilized bones that have surfaces that are much more resistant to change due to their lithified states. We agree with the reviewer that our

molding process likely did not impact our results, but we appreciate their research directed at determining the impact of molds on mark morphology.

Reviewer's reply: I understand the reply, however the authors have not clarified anything in the text to support this. I have studied materials at 3 Ma that present no sign of mineralization at all, and are equally as fragile as a modern bone, while on the contrary I also currently study materials that are only 20 ka that are completely mineralized. (1) Just because bones have fossilized, doesn't make their cortical surface more or less fragile, while (2) not all sites present the same degree of mineralization. The authors haven't mentioned at any point the mineralization of the bones in their paper. At this moment in time, we have to take the authors word that the moulding process has not damaged the bone. Transparency is needed and the authors should at least clarify some of these points. Also, if they are certain that the two moulds are the same and no changes have occurred, then the first and the second mould of the same mark should produce the same metric results, could this not be tested, just in case?

Overall reviewer's reply – The authors reply with some valid points to me, the reviewer, but these points are still left unattended in the text, leaving all other readers of this body of work in the dark. They could easily clarify these points, in most cases with a single sentence. I repeat – if these concerns and questions could have occurred to me, they can probably occur to somebody else. We need to be as transparent as possible in our analyses, especially in the case of such extraordinary findings. Including the sentence "The molds measured were the second mold taken from each specimen", is a very simple statement to include, with an additional acknowledgement about the potential risks, even if these risks may not be entirely present. And if the authors are indeed certain that no changes have occurred, they could easily check this by measuring the different moulds and comparing whether the measurement on mould 1, 2 and 3 are all the same, which would validate the point they are trying to make.

Author's response to reviewer: The reviewer is correct that bioerosion can impact the measures collected here, which is why marks where bioerosive processes impacted depth were not included in the analysis. The specimens cited above had multiple modifications present, and those that were altered were excluded from our results.

Reviewer's reply: have the authors clearly stated this in the text? It would be very easy to add a sentence that clarifies this. If this is stated in the text, please can they indicate where, because I cannot find it.

Author's response to reviewer: We have added the information for the objective used in our analysis.

Reviewer's reply – thank you for including this information

Author's response to reviewer: ... For example, measuring a mark with a 20x lens would produce a point cloud that would exceed the memory capabilities of the S-Neox and would take hours instead of minutes to capture, if it were even possible...

Reviewer's reply: I won't go too much into this, but the idea that the point cloud exceeds the memory of the computer is interesting... I, and other colleagues, study cut marks, tooth marks and percussion marks at 20x, and have no problem at all with memory. Yes it takes longer, but science requires sacrifice, and while the work has been done on moulds, there is no issues with limited access to the samples which would imply that the authors could not have taken the time to scan all the marks at 20x. Maybe older versions of the same microscope present storage issues that newer version of the S-Neox do not present, if that is true then it might be necessary to state the year, version, or some indication about which edition of the microscope is being used, because I repeat, in our lab (University of Bordeaux), we have an S-Neox and have had no problem studying cut marks at 20x magnification.

Author's response to reviewer: the quantitative method used effectively discriminates between the processed marks so removing form is not problematic for our particular method and is necessary to capture the variables of interest.

Reviewer's reply: I'll have to take the author's word for it, however they still have not stated anywhere which polynomial is used to remove the form from these scans. This is an incredibly simple statement to include in the methods – A polynomial of X was used to remove the form.

Author's response to reviewer: While there is some literature that agrees with [QDA should not be used for ordination] it mostly relates to LDA and geometric morphometrics and not QDA on variables related to size rather than shape.

Reviewer's reply: There is a lot to unpack, even in this short paragraph. First a point of debate: "variables related to size rather than shape" - Size and shape can be, and are often, inherently related - unless you provide evidence to say otherwise, this statement can be very problematic. Because the objective of the paper is not to look at this question, I will brush past this statement, but I just would like to point out that care is needed here because this is not necessarily true, and in any other paper I would very strongly disagree with such a statement. Regarding the much more important point, I think the author's have missed the point with reference to the issues of using discriminant scores for ordination. I am not talking about geometric morphometrics (GMM), I am talking about mathematics. All the literature I can find on QDA and the corresponding plots still relate them to the calculation of canonical variates. In fact, QDA and LDA are built on some fundamentally identical algebraic calculations, that means they are completely related. Considering how QDA and LDA are heavily linked algebraically, this would in turn imply that all issues related with LDA, which are related to CVA, are related in turn to QDA as well. I'm talking exclusively about algebra, not about GMM. The reason I mention GMM is simply because this is where the majority of research into these algebraic properties lie. However, I would like to point out that authors such as Cardini and Bookstein do not even use landmark based data in their papers about this topic, they use simple multivariate normal distributions, and thus still prove my point – QDA should not be used for ordination. Likewise, the paper by Albrecht (that I provide below) talks mostly about linear measurements, not landmarks... therefore it is directly relatable to the present study.

It is interesting that the scores in the plots are also called canonical something. From this perspective how can you be so sure the plot is different from what has been criticised by other authors? Likewise, no indications or citations are included in the paper to clarify how calculations are performed, beyond the statement that JMP was used. JMP is a private licensed software, which I do not have a licence for (in my opinion this in itself is an issue) so I cannot perform any tests to check for replicability, or even evaluate the suitability and reliability of the algorithm in this specific context.

In sum; (1) if QDA is less sensitive to sample imbalance, I would like some references to support this, or simulations to prove this. (2) You still need to clarify if cross validation is done, and if it isn't performed, then justification as to why must be provided, because in a lot of circumstances the exclusion of CV would be considered completely problematic. (3) I am very happy to see a PCA is now included, which would generally resolve a number of the issues I had with the QDA plot, however, there are a number of issues related to the PCA. (3a) The PCA provided is very very hard to read, and I cannot see where the fossil points fall. Have they even been included? Please improve the presentation of this figure so it is clear to read and see how the archaeological samples compare with the experimental ones. This is something you did with the QDA plot (you already know my opinion about how reliable this plot is), but have refrained to do for the PCA; (3b) why have the confidence intervals been excluded for the PCA, but included for the (problematic) QDA? Please standardise the presentation of your results. (3c) Having red and orange points on the PCA makes the samples really difficult to distinguish. Also, is it just me, or do the colours not match up between the QDA and the PCA? Or even within the PCA at all. In the key there is blue, purple, green and purple, but in the PCA I cannot see blue, but instead see red, orange, green and purple. What is going on here?!

If the authors are still adamant that their use of QDA is appropriate, please provide justification. I provide justification here to highlight why it is not appropriate (once again, before the authors claim that literature originating from GMM is not relevant here, please read the methods sections of these papers to see that in most cases, the methods are not rooted in GMM analyses at all);

Albrecht, G. H. (1992). Assessing the affinities of fossils using canonical variates and generalized distances. *Journal of Human Evolution*, 7(4), 49–69.
Bookstein, F. L. (2019). Pathologies of between-groups principal components analysis in geometric morphometrics. *Evolutionary Biology*, 46, 271–302.
Cardini, A., & Polly, P. D. (2020). Cross-validated between Group PCA scatterplots: a solution to spurious group separation? *Evolutionary Biology*, 47, 85–95.
Cardini, A., O'Higgins, P., & Rohlf, F. J. (2019). Seeing distinct groups where there are none: spurious patterns from between-group PCA. *Evolutionary Biology*, 46, 303–316.
Mitteroecker, P., & Bookstein, F. (2011). Linear discrimination, ordination, and the visualization of selection gradients in modern morphometrics. *Evolutionary Biology*, 38, 100–114.
Rohlf, J. F. (2021). Why clusters and other patterns can seem to be found in analyses of high-dimensional data. *Evolutionary Biology*, 48, 1–16.

It's incredibly important to point out that in each of these referenced papers, the authors may talk a lot about GMM, however the data they are actually using to explore the algebraic and mathematical issues presented are simply multivariate distributions. This means the data used for these statistical and theoretical experiments are completely analogous with the present author's data. In particular I recommend the paper by Cardini and Polly (2020) for proof on the importance of cross-validating their discriminant function plots. For the importance of cross-validation in classification I could recommend a large number of papers and books, however would simply state that the book by Hastie, Tibshirani and Friedman (2009) *The Elements of Statistical Learning* is more than a good starting point.

Reply to the reviewer: We employ QDA here due to the unequal variance and size of the samples and we assess our model setting prior probabilities proportional to the occurrence of each mark to account for the differences in sample size.

Reviewer's reply: Ok then please clarify this in the text, and also state what prior probabilities were used. This can be easily included in a single sentence.

Reply to the reviewer: We have added an ROC curve as supplemental materials that reflects the requested information and shows the model is highly effective. Thanks for the suggestion! We will also report accuracy as it is a standard that more scientists are familiar with.

Reviewer's reply: While the authors have already corrected the presentation of results to include more than just accuracy (ROC curves are a great addition), I have to simply state, for the purpose of scientific clarification, that I strongly disagree with the term or concept that accuracy is a "standard that more scientists are familiar with" - I could not disagree more with the logic that just because everyone else does it means that it is right. For years it was general consensus that the world was flat, this doesn't mean that these individuals are the future of science. For years nobody thought anything was smaller than the atom, scientific research proved them wrong too (if anyone is paying attention they might catch a Friends reference in here). Science evolves and so should we, just because "scientists" (this should actually be archaeologists/palaeoanthropologist/palaeontologists) are more familiar with accuracy, that doesn't mean we shouldn't break the mould and develop our analyses to adopt correct practices. Moreover, if you actually consult the literature regarding classification in data science, the authors would see that (mostly) non-archaeological scientists publish far more than simply accuracy. Given this last point, I would argue that in data science reporting only accuracy is certainly not a standard, which would make this statement incorrect on a number of different levels.

Just to prove the point, and make sure that we go beyond the field of data science and into the realm of archaeology: imagine we have a new set of 100 marks; 80 cut, and 20 trampling. Imagine our algorithm claims that all marks are cut marks. This would result in 80% accuracy, however, it would also be misleading considering how the algorithm is actually allocating the label cut mark to everything (i.e. hasn't learnt to classify anything at all). Now, if we calculate sensitivity and specificity, at least one of these values would be 0, successfully helping us detect that our 80% accuracy is misleading, and the algorithm is in-fact awful, and should not be used in real case studies (EVER!). Now imagine that instead of archaeologists we are doctors and we have potentially killed 80 patients. Just because our patients have been dead for 1.9 million years (wow, what a date!), doesn't mean we should not be just as rigorous and strict with our diagnostics. I apologise for rambling, but I find this point fundamental to clarify, especially if the authors find more great finds in the future and have different reviewers who are unfamiliar with this concept.

Final statement - Once again, the reasoning for my lengthy review letter is to ensure clarity in my concerns, or open potential debate for future reference, and I hope the authors find these comments constructive (if my tone is taken at some points to be aggressive, I sincerely apologise, this is certainly not my intention). I strongly value and respect the work of all the authors, and think that these finds are incredible; I just want to ensure that there is no wiggle room or overlooked error that can be used against or to discredit this study. If I can find these little problems, then other (less amenable) researchers certainly can as well. I am incredibly grateful to be part of the review process, and look forward to seeing this paper published as soon as possible! Signed - Lloyd A. Courtenay.

Reviewer #3

(Remarks to the Author)

I have carefully reviewed the authors' responses to my comments, the revised version of the manuscript, and the supplementary information provided. I am grateful that they have considered my suggestions and that they are incorporated into the manuscript.

Reviewer #4

(Remarks to the Author)

Hominin presence in Eurasia by at least 1.95 million years ago, by Curran et al.

The authors have convincingly addressed most of my comments, and even when objecting, the answers given are reasonable and make sense. After going through the detailed rebuttal, it seems to me that the evidence provided on the three main critical aspects of work (i.e., human origin of the cut marks on bones; (ii) antiquity of the fossil material and (iii) palaeoenvironment reconstruction) are robust enough to support publication. I would especially like to thank the authors for dealing with the comments made by the reviewers in such a constructive way.

There are a couple of general comments I would like to make here after reading the revised manuscript, although they should not preclude the publication of this very interesting work that undoubtedly deserves to be shared with the community.

First, I must admit that I remain quite sceptical about the precision of the general chronostratigraphic framework of the ORV and Dacian basin. Yes, biochronology can give a rough chronological estimate of a given site, but this should be rather regarded as what it is, a rough estimation. Estimation because the inherent uncertainty typically associated to biochronology is usually minimized (or ignored). It may be related to (i) the taxonomic attribution of fossil specimens, which may vary depending on the specialist involved, (ii) the validity/robustness of large scale biostratigraphy correlations across a country or even a continent, which may often be questionable, and (iii) and the existing true independent age control available at other sites used for comparison (i.e., based on methods like magnetostratigraphy and numerical dating method), which is often limited and has an uncertainty on its own. This is why I personally have difficulties to accept precise age constraints based on biochronology only (e.g., here: 'This formation extends from the base of the Pleistocene to as young as 1 Ma; Biochronological assessments indicate Grăunceanu is Late Villafranchian (~2.2-1.9 Ma) and is attributed to mammalian biostratigraphic zones MN17/MmQ19-11). For example, in absence of numerical age results (and even though, the existing age errors are usually not negligible) or magnetostratigraphic evidence indicating the undisputed presence of Jaramillo, the minimum age constraint of 1 Ma proposed here is questionable. Let's say that an Early Pleistocene chronology may roughly be estimated for the Tetoiu Fm based on biostratigraphic evidence. Following a similar reasoning, there are many examples of localities that have been equally attributed to either the Middle, Late or Epivillafranchian biochrons depending on the authors. For example, I note that on the figure given in the rebuttal, the chronological window proposed for the site covers the latest Middle Villafranchian and earliest Late Villafranchian, whereas the text mentions a much finer age window ('Late Villafranchian (~2.2-1.9 Ma)'). Anyway, my point here is that I feel that the general chronostratigraphic context of the ORV and Dacian basin (and mainly the Tetoiu Fm) would deserve a dedicated section in Supplementary Material, as biostratigraphy provides a crucial support to the U-Pb dating results. This would help the reader unfamiliar with the area getting the bigger picture. In particular, it is unclear how the various sites mentioned in the text do relate (e.g., La Pietriș, Valea Roșcăi, Fântâna lui Mitilan). It is mentioned that some of them are coeval, while other are younger/older, but no data are provided to support this. I understand that this information might be found in various other papers, but it would be beneficial to the present work if a synthesis could be given in Supplementary Material. In particular the figure of the rebuttal showing the biochronological evaluation of the Grăunceanu should probably be added. It does provide useful information. The other localities from the ORV are that are mentioned in the text could be added to that figure too.

Finally, I would like to add that, in my view, there is still an inherent general uncertainty on the context of the site and of the fossil assemblage. I do get the explanations of the authors on this matter, and I accept them, but the fact that the site is no

longer accessible, and that most of the information related to the excavation is gone, naturally raises some questions on the integrity, extension and stratigraphy of the fossil assemblage. This is complicated by the fact that various successive campaigns have been performed over more than a decade, and we ignore how they relate. Moreover, dating analyses were not performed on the same bones that have the cut marks, or on the equid teeth used for palaeoenvironment reconstruction, so there is also an inherent uncertainty on whether all these remains are truly associated, i.e., come from the same excavation area and the same stratigraphic unit. In that regard, would it be possible to at least know from which excavation campaigns were taken all the specimens analysed here for cut marks, dating and paleoenvironment analyses? That said, I get the authors' explanations and especially regarding the taphonomy suggesting a single depositional event, and no doubt they did their best to narrow down this uncertainty, but this is still indirect evidence that cannot fully compensate for the absence of a proper stratigraphy and spatial location of the remains.

Specific comments

1. L.74-75: The chronostratigraphy of the Tetoiu Fm described from Andreescu et al. (2013) is not as robust and precise as suggested here. There is no independent age control available, and if I am not wrong the age constraints are only based on biostratigraphy. I'd rather see something like: 'An Early Pleistocene age may be considered for this Fm based on biostratigraphic evidence.'. I did not see evidence in Andreescu et al. (2013) for such a precise minimum age boundary of 1 Ma.
2. L.104-106: Please indicate that this relative chronology is based on biostratigraphic evidence.
3. Fig. 1: please check the spelling of 'El-Kherba'.
4. L.196-198: I understand these ages only refer to Grăunceanu. This should probably be clarified.
5. L.199, 'cluster tightly': maybe a number (e.g., coefficient of variation, or a time range between the youngest and oldest results) can be provided to support this statement.
6. Fig. 3: please indicate in caption that age errors are 2σ .
7. L. 231: VSMOW should probably be spelled out for its first occurrence in the text.
8. Table 1 and Supplementary Table S7: The various ages reported there are quite confusing, maybe this should be clarified in caption. Sometimes an age range is given, with either 1 or 2 decimal digits. Not sure a 2 decimal digit precision is appropriate for most sites (Fântâna lui Mitilan even shows 3 digits). It is also unclear whether the age range is taken from the errors on the numerical ages, from the combination of various ages, or simply correspond to various levels from the same site.
 - Not sure to understand how the age range of 2.3-2.0 Ma for Grăunceanu was obtained since the various U-Pb ages range from 1.87 to 2.01 Ma
 - Ain Hanech and El Kherba (Algeria) are usually regarded as coeval. Please see Duval et al. (2023) for the most recent age estimation.
 - Ain Boucherit (Algeria) is made of two levels, the lower level (Lw) dated to ~2.4 Ma and the upper level (Up) to ~1.9 Ma. Maybe these 2 different levels (usually referred as Ain Boucherit Lw and Ain Boucherit Up) should be differentiated, otherwise the age range given may suggest a significant uncertainty on the age of the site, which is not the case.
 - Pirro Nord: please note the most recent dating study on that site (Duval et al., 2024). Although a biochronological age of 1.3-1.6 Ma cannot be excluded, there is strong evidence suggesting that the site may be much younger.
 - Sima del Elefante : maybe indicate the layers, e.g., Sima del Elefante (TE7-9). Methods also include BS.
 - Fuente Nueva-3: a rounded age of '1.2 Ma' would probably be more appropriate given the associated uncertainty.
 - Barranco Leon: Methods also include MS and BS, in addition to ESR.
9. L.324-330: for a matter of consistency, please use the same number of decimal digits for all percentages.
10. L.326: Do the 3.32% include both Lower and Upper levels at Ain Boucherit? Please clarify.
11. L.335: maybe you can refer to Supplementary Table S8 here.
12. Ref 71 in SM: please note the name of the journal is 'L'Anthropologie'.
13. Supplementary Table S8: please indicate in caption the meaning of 'Full Taph?'. Note that the age reported for Grăunceanu is not consistent with that in Table S7. Same comment applies for Dmanisi.

References

- Duval, M., Arnold, L.J., Bahain, J.-J., Parés, J.M., Demuro, M., Falguères, C., Shao, Q., Voinchet, P., Arnaud, J., Berto, C., Berruti, G.L.F., Daffara, S., Sala, B., Arzarello, M. (2024). Dating the earliest evidence of human presence in western Europe: New results from Pirro Nord (Italy). *Quaternary Geochronology* 101519. [10.1016/j.quageo.2024.101519](https://doi.org/10.1016/j.quageo.2024.101519).
- Duval M., Sahnouni M., Parés J.M., Zhao J.-x., Grün, R., Abdessadok S., Pérez-González A., Derradji A., Harichane Z., Mazouni N., Boulaghraief K., Chelli Cheheb R., van der Made J. (2023). On the age of Ain Hanech Oldowan locality (Algeria): First numerical dating results. *Journal of Human Evolution* 180, 103371. <https://doi.org/10.1016/j.jhevol.2023.103371>.
- Sahnouni M., Parés J.M., Duval M., Cáceres I., Harichane Z., Van der Made J., Pérez-González A., Abdessadok S., Kandi N., Derradji A., Medig M., Boulaghraif K., Semaw S. (2018). 1.9-2.4 million-year-old artefacts and stone tool cutmarked bones from Ain Boucherit, Algeria. *Science* 362 (6420), pp. 1297-1301. DOI: 10.1126/science.aau0008.

Version 2:

Reviewer comments:

Reviewer #2

(Remarks to the Author)

Again it is a pleasure to be involved in the reviewing process of this paper, however I have rewritten this review a number of times to make an effort to be as clear as possible. While some of the comments made in response to my last review are debatable, they are not necessarily the focus of the study, therefore I will leave these points for now, and this can hopefully lead to interesting debates and lines of research in the future. Nevertheless, there is one point that has still not been corrected to a sufficient extent, and is arguably one of the most important points I have made. Without properly responding to this point, I do not feel comfortable accepting the metric analysis of the taphonomic marks in their current state. Once again I have no problem in saying that the marks are cut marks, from a traditional perspective, but my issue is that when using quantitative approaches, a certain number of statistical and mathematical requirements have to be met in order for the results to be considered robust, reliable, and not a misrepresentation of spurious separations between samples.

In my first and second review, I stated very clearly that it is fundamental that QDA results are cross-validated. This is important for both the evaluation of the algorithms used for classification, and can wildly change the quality of the ordination plots. When I brought it up again in the second round of reviews (and I provided a number of references to prove the importance of my point), the authors state that it has already been included in their manuscript. JMP is a private licenced software, so I have close to no way of testing some of the main issues, however, I have been able to obtain access to the non-pro version of JMP for a limited time in order to run some tests. Likewise, I have read in detail the documentation of the software to try and find the information I requested. Finally, in this review I will be providing R code to help clarify not only my points, but also provide a potential means of fixing some of the issues. I hope that the authors can see that the fact that I have taken the time to provide an R script is not only for the purpose of proving the importance of cross-validation, but in the hope that the authors may also find this code useful. If the authors wish to use their own data, then I have provided comments within the code to clearly indicate where their data needs to be inserted.

The authors state that in their original manuscript that they “clearly reported ... that a 25% validation set was extracted from the experimental sample of BSM to validate the model. We now call this a 25% cross-validation set for clarity”. I respectfully disagree, and I am not sure we are talking about the same thing here. First, there are many types of cross-validation algorithms, which one are you using (if you are using one at all)? In the case of classification, the most recommended form of performing cross-validation on a classification algorithm is k-fold cross-validation, with $k = 5$, or more commonly $k = 10$. Did the authors do this? Next, there is a distinct difference between a test set and a validation set, which is where I think some source of conclusion lies. Finally, how can 25% of the data be used to cross-validate the plot? In general cross-validation of the discriminant function plots is slightly different.

While a number of authors can be found throughout literature using the terms validation and test sets for different things, in general, the standard approach is to refer to the validation set as the data used by the algorithm itself to define the discriminant function, and the test set is what we separate and then use to evaluate how well this discriminant function is actually working (I think this is what you refer to as validation, given that JMP uses this term in the Analysis/Predictive Modelling/Make Validation Column menu). I suspect that the authors are calling the validation set what I (and many) would actually call the test set. This would imply that they are still not necessarily performing cross-validation, which is why the simple statement “25% cross-validation” does not actually or fully clarify anything.

Through playing with JMP, and the additional code I provide in R, I am pretty certain that the authors are not performing the cross-validation required. JMP appears to be able to do everything I will describe, however in the non-pro version of the software I cannot actually test this fully. Nevertheless, the documentation clearly states that K-fold cross validation is possible (<https://www.jmp.com/support/help/en/18.0/index.shtml#page/jmp/kfold-and-holdback-validation.shtml#>), which goes beyond simply separating a test set to evaluate model performance afterwards. This however requires a number of additional steps to be performed, however (see previous link).

From what I can tell by the table on this website, the algorithm is not doing k-fold cross validation automatically, I can only assume that using a number of other functions they can get around this. Otherwise the word “No” under “K-fold cross-validation” and “Discriminant Models” does not bode well for the use of this software for classification tasks using this algorithm. Likewise, I still do not see clearly whether the cross-validation approach will be taken into account when computing the discriminant scores as well. This needs to be checked and clarified if the authors are adamant about including a CV plot in their supplementary materials (which I have already suggested avoiding all together).

To validate, and additionally clarify the points I am trying to make, I will describe the experiments I have carried out in both R and JMP in continuation. If the authors are able to test that k-fold cross-validation is possible in JMP, and does indeed correct the plots, then this would be a good means of solving all issues I am about to describe.

In the first example, I took the dataset from <https://www.kaggle.com/datasets/muthuj7/weather-dataset?resource=download>. This is a known benchmark dataset, and in R I can show that a cross-validated QDA is able to reach a 94% accuracy. I'd like to strongly point out however that (as I already said in a previous review about how accuracy is a misleading metric), the QDA algorithm is unable to correctly classify any of the “null” class from this dataset, which leads to a kappa statistic of 0.65, i.e. the algorithm is not sufficient to correctly learn from this data. If I perform the same analysis in JMP without performing k-fold cross-validation the results are completely different. The accuracy now drops to 45.9%, with a kappa statistic of 0.124. I think this is already evidence enough to show that cross-validation is not only a valuable means of boosting model performance, enhancing discriminant function robustness, but how misleading accuracy results can be if cross-validation is not performed. Confusion matrices of both these calculations are included as Figures 1 and 2 in the file I provide alongside this review.

To provide sample sizes that are a little more analogous with archaeological datasets, however, I then took a subset of 30

individuals per sample from this dataset. I used a seed in my R code to ensure replicability of results, considering how random subsampling is performed. With this seed, you will see that one of the observations included in the PCA (Fig. 3) is a potential outlier (it is in fact not an outlier, as it does belong to the original distribution perfectly, a pretty cool example of how sampling strategies can make single extreme observations appear like outliers when they are in fact not). In a cross-validated canonical variates plot (Fig. 4) the point is forced closer to the rest of the sample, however, a cross-validated plot (Fig. 5) shows that the point is still an outlier, which is much truer to the actual nature of the distribution. Either way, the two plots are different, showing that cross-validation will produce a difference in plots and may force certain patterns to emerge that may not be an accurate representation of the underlying data. To show how this effects cross-validation performance for classification, if we do not do cross-validation, JMP calculates an accuracy of 56.5%, with a Kappa of 0.347, while the cross-validated version of the algorithm produces an accuracy of 56.5% and a Kappa of 0.3466. While the differences are minimal, they are there.

In the second example I increase the complexity of the dataset by adding 100 new dimensions of noise to the dataset, simply to provoke the 'curse of dimensionality' to occur, which again will lead to the spurious separation of samples mentioned by other authors. Once again note that without cross-validation, the plots are not comparable at all, and the cross-validated plot is much truer to the original distribution as shown in the PCA. Likewise, the outlier from the first experiment is once again still identified as such. In the non-cross-validated approach using JMP, the accuracy comes out at 43.5% but has a surprising kappa value of -0.049, indicating that the discriminant function is not a good classifier at all. R, on the other hand, has the courtesy of refusing to perform the analysis, as the sample size and complexity does not allow for the computation of a reliable discriminant function.

In the final example, I perform these experiments on purely synthetic data, that are designed to have 0 differences between the three samples. Once again, spurious separation occurs on non-cross-validated results, but does not appear on the cross-validated results. This is proof of the theorem that non cross-validated canonical variance plots can be incredibly difficult to interpret and are not always a necessarily reliable representation of the data. A non-cross-validated QDA in JMP reveals an accuracy of 47%, however once again R refuses to perform the computation.

To prove the fact that simply separating 75% of the data for the calculation of the discriminant function does not imply that cross-validation has been done, figures X, X, and X additionally show how using the final example, when using the validation set and when excluding the validation set, spurious separation occurs regardless.

It is important to point out that in some of the above cases, only limited differences are seen between the results of k-fold cross-validated algorithms, and discriminant functions fit without performing cross-validation. Regardless, the fact that even slight differences are there are important. In my previous reviews I provided a number of references to prove this. Seeing how this was insufficient in stressing my point, here I have provided code, and statistical proof that this is still an issue. Interestingly, R in a number of the above cases refused to calculate, or was unable to calculate the discriminant function, while JMP succeeded. Nevertheless, I refer the authors to JMP's documentation, where it clearly states that in some cases QDA can be problematic (<https://www.jmp.com/support/help/en/18.0/index.shtml#page/jmp/discriminant-method.shtml#>). The authors need to ensure that they report robust and reliable quantitative results, and even if repeating the statistics with a correct implementation of cross-validation produces the same results, at least then we can be sure that your results are accurate.

In sum, for Sup. Fig. 22 please either; (1) cross-validate the calculation of the CV scores in the plot, or (2) completely delete this figure. In my opinion, option 2 is the easiest approach and does not put into question the reliability of the study. For classification results, I insist that the authors must use cross-validation, preferably k-fold cross-validation with $k = 10$, as is standard procedure in most analyses of this type. This latter point is the most important and needs to be rectified, or at least clarified.

Signed – Lloyd A. Courtenay

PS - I may be mistaken but I also recall an Excel or csv file being included in an earlier version of this manuscript, which is now no longer available, if this was data it would have been nice to have access to this data again to run the same tests on your actual data.

Version 3:

Reviewer comments:

Reviewer #2

(Remarks to the Author)

The authors have finally acknowledged and made modifications in light of my concerns, which I am grateful for. I want to emphasize that my intention was never to be overly critical or harsh. Rather, I believe it is fundamental to evaluate quantitative analyses rigorously. My recommendations were aimed at reinforcing the contents of the paper, and my approach was intended to be constructive, not deconstructive. As I have always stated, I had no issue referring to the marks as cut marks. However, statistical analyses require specific criteria to be met, and I was simply trying to highlight those requirements.

I would like to thank the authors for all their hard work and patience. This is truly a fascinating and important study.

Kind regards

Lloyd

(Remarks on code availability)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Response to Reviewers

We greatly appreciate the time and effort the four reviewers put into their suggestions, which have improved this manuscript. In order to make a more readable revision, we have made a number of changes to align with Nature Communications specific format that were not tracked (formatting of headings, supplementary figure references, etc.). All changes to the manuscript (and supplemental documents) to address reviewer comments have been tracked.

Below we provide a point-by-point (bolded) response to the reviewers' (non-bolded) comments. Please do not hesitate to reach out with further questions and clarifications.

Reviewer #1 (Remarks to the Author):

Curran et al discuss evidence suggesting the presence of hominins in Romania approximately 2 million years ago, predating the widely accepted oldest Eurasian site of Dmanisi (Georgia). Their evidence includes cut-marked bones and new Uranium-Lead (U-Pb) dates of fossil dentine. The authors also use horse enamel isotope data for a coarse paleoclimate and paleoenvironmental reconstruction. Except for a few typos and awkward sentences, the manuscript is generally well-written. Figures need a little extra work (see below), but they are fairly well-crafted. Unfortunately, I must defer the evaluation of the hominin presence and U-Pb components of the paper to other reviewers, as they are outside my area of expertise. I can only say that their arguments and conclusions seem reasonable in light of the data to a non-expert. My main concern with this manuscript is the oversimplification of the interpretation of the isotope data. More specifically, here are the two main issues:

1) Interpretation of enamel oxygen isotope data. The authors interpret lower-than-present precipitation $\delta^{18}\text{O}$ (calculated from horse enamel $\delta^{18}\text{O}$) as an increase in winter precipitation component to aquifer recharge relative to the present. This is one of the many possible explanations. Alternative possibilities include a lower mean annual temperature, and a change in air mass trajectories. These alternative possibilities must be discussed here in depth.

--We have now clarified how we came to this conclusion by adding estimates of winter and summer precipitation amounts, which we summarize in the main text and then detail in the supplementary discussion. Regarding potential changes in air mass trajectory compared to the present, we do not have information about the Early Pleistocene setting in this region. At present, two main, isotopically different vapor sources are at play in Romania, namely the North Atlantic and the Eastern Mediterranean (which contributes water vapor with higher isotopic values due to increased evaporation at source). Local meteoric water lines in this region have equations similar to the global meteoric water line, indicating the North Atlantic as being the most important moisture source (Dragusin et al., 2017). We have added some text that mentions these possible influences but note that we either consider those possible interpretations unlikely due to the fauna at the site (lower mean annual temperatures), or unable to incorporate these sources into our analyses because of uncertainties regarding the fossil record (changes in air mass trajectories).

Furthermore, there is no evidence that the horses were drinking groundwater.

--We do not suggest that horses drank groundwater. In the Supplementary materials, we write: “Horses, which are obligate drinkers (water dependent), take most of their water from springs, rivers, or lakes, each of these sources being defined by their own isotopic drivers such as seasonality of recharge, evaporation, mixing, or recharge altitude. Rivers have lower amplitude of isotopic variability than rainfall, as they drain aquifers that are most of the time well mixed throughout the year.” We have clarified this in the main text.

Finally, the uncertainty in the average d18O value of precipitation is largely underestimated. The authors need to perform a rigorous error analysis and propagate the errors from ALL sources.

--We recognize that there are four potential sources of error when reconstructing $\delta^{18}\text{O}$ of meteoric water: 1) analytical measurement of samples, 2) the PO4-CO3 relationship, 3) diagenesis, and 4) meteoric water to enamel transfer functions. We have tried to clarify these complicating factors in the text and we address them here as well.

- Analytical errors are stated as 0.1‰ for CO3 at both labs, while it was 0.60‰, 0.37‰, 0.54‰, and 0.29‰ for PO4, for the four standards used.**
- We directly assessed the PO4-CO3 relationship for our M3 sample; this relationship is defined as $\delta^{18}\text{OPO4} = \delta^{18}\text{OCaCO3} \times 0.96 - 8.06$, which is very similar to that of Iacumin et al. (1996), $\delta^{18}\text{OPO4} = \delta^{18}\text{OCaCO3} \times 0.98 - 8.5$. The shift between our measured PO4-CO3 samples is on average 8.9‰, also similar to the Iacumin et al. (1996) value of 9.1‰, indicating little to no diagenesis. We therefore have high confidence that diagenetic changes are not meaningfully influencing our results.**
- Though we consider our original analyses to be robust, based on the reviewer’s concerns regarding sources of error we recalculated our $\delta^{18}\text{OPO4}$ estimates using the equation of Iacumin et al. (1996) instead of our own equation. Although the Iacumin et al. (1996) PO4-CO3 relationship is not provided with uncertainties, we decided to use it because it is based on a larger dataset, instead of using our own relationship derived from only five sample points. In addition, their equation also is likely to be more robust because they have smaller measurement uncertainties (0.15-0.20‰). In contrast our measurement run had higher values, which would render any error propagation from our equation meaningless. However, there are no errors to propagate from the Iacumin et al. (1996).**
- Of the three equations we then employed to reconstruct meteoric isotopic values (Delgado Huertas et al., 1995; Kohn and Cerlin, 2002; Amiot et al., 2002), only Amiot et al. incorporates error estimates. These errors associated with $\delta^{18}\text{O}$ in enamel PO4 and meteoric water are given as $\delta^{18}\text{Ow} = 1.1128 (\pm 0.0029) \times \delta^{18}\text{OPO4} - 26.4414 (\pm 0.0508)$, and integrating those uncertainties would affect only the second digit of the results, with no discernible change in our interpretation. Furthermore, all three equations (Delgado Huertas et al., 1995; Kohn and Cerling, 2002; Amiot et al., 2002) return mostly similar results and by averaging their results for the P2-M3 series we obtained the value of -10.72 (± 1.02 ‰), which is still below present-day values.**

- To conclude, our PO4-CO3 relationship resulting from a small dataset is similar to published values based on a much larger dataset of Iacumin et al. (1996) and shows no important signs of diagenesis. Out of three published transfer functions from PO4 to meteoric water only one is provided with uncertainties and those are very low, and all three return similar values, resulting in an average value with a standard deviation of 1.02‰. Thus, a full propagation of errors as outlined in Amiot et al. (2002) only has a minor effect on the final values. We therefore consider the standard deviation of 1.02‰ to be the most important statistical parameter for interpreting our results; we have added this standard deviation to Figure 4 to make this clearer. We have also edited the text to reflect the change in our use of Iacumin et al.'s equation.

2) Interpretation of enamel carbon isotope data. There are many issues with the calculation of MAP from enamel $\delta^{13}\text{C}$, including a) the possibility that the investigated taxon was selective in terms of its diet and habitat;

--The reviewer is correct that our investigated taxon, *Equus*, is highly selective in its diet and habitat, but that is precisely why we feel this is a powerful taxon to investigate. Modern *Equus* preferentially graze throughout the year, and all existing data suggest this has been the case for this genus since the Pleistocene. Because this taxon is so restricted in its diet, we have especially high confidence that we are actually capturing seasonal changes in mean annual precipitation (and/or migratory behaviors), rather than seasonal dietary shifts. Other analyses by our research team (Curran et al., 2021) have investigated isotopic dietary signals in ruminant taxa (i.e., deer), as well as meso- and microwear analyses of the teeth to get at dietary behaviors in species from Graunceanu. As we state in the supplementary materials discussion of the stable isotopes: "Plant tissue $\delta^{13}\text{C}$ reconstructed from enamel $\delta^{13}\text{C}$ values varies between $-28.9\text{‰} \pm 0.5\text{‰}$ and $-25.7\text{‰} \pm 0.5\text{‰}$ (VPDB), which is consistent with feeding in a woodland to woodland-mesic C3 grassland (as defined by Domingo (2013), and supports prior environmental reconstructions for Grăunceanu (Curran et al., 2021)."

b) the uncertainty in the enrichment factor between diet and enamel;

--The isotopic enrichment between diet and enamel is given by Cerling and Harris (1999) as $14.1 \pm 0.5\text{‰}$ and we integrated the uncertainty in the new calculation of the seasonal precipitation amounts. This results in precipitation amount estimates for the driest month of 6-22 mm/month, while for the wettest month they are between 90 and 148 mm/month. The text has been updated to include these new ranges of precipitation.

and c) the uncertainty in the relationship between the $\delta^{13}\text{C}$ of modern C3 plants and MAP. Like with $\delta^{18}\text{O}$, the authors largely underestimate the error in the reconstructed MAP. I am inclined to say that because of all these issues these MAP calculations and paleoenvironmental reconstructions have such a large uncertainty that they are almost meaningless, especially if only one taxon is used like in this case. The authors should consider eliminating the calculated MAP from the manuscript unless they provide a robust estimate for the uncertainty.

--We appreciate the reviewer's concerns regarding our MAP calculations and their uncertainties. We agree that our previous estimates of MAP had a broad range (868-1134 mm/year) even though they did not include the 0.5‰ uncertainty from Cerling and Harris

(1999); including those uncertainties would certainly increase those ranges in such a way that they are meaningless. Unfortunately, the uncertainty in the relationship between the $\delta^{13}\text{C}$ of modern C3 plants and MAP as given by the equation by Kohn (2010) is unknown since they did not provide uncertainties. Thus, instead of taking our original approach, we now instead discuss the reconstructions for just the driest (i.e., summer) and wettest (i.e., winter) months in order to define seasonal variability in precipitation. The text has been edited to reflect this change in focus.

These two issues must be fixed before the manuscript can be considered for publication in Nature Communications. Minor issues include hard to read figures and text. For instance, the text in Fig. 1 is hard to read (especially in the inset) and the “18” and “13” in Fig. 2 should be superscripts. Again, the manuscript also needs some minor editorial work to fix some typos and to rewrite some poorly structured sentences.

--Figure 1 will be submitted as a higher resolution image and Figure 4 (the isotope figure) have been edited.

--We also have gone back through the manuscript to correct any typos we found and checked for awkward sentences, though we cannot be certain which ones Reviewer 1 is referring to specifically.

Reviewer #2 (Remarks to the Author):

The paper by Curran and colleagues titled “Hominin presence in Eurasia by at least 1.95 million years ago” presents a highly important find that hominins were indeed present and active outside of Africa at least 1.95 million years ago. I have thoroughly enjoyed reviewing this paper, and certainly think it is relevant and falls within the scope of this journal. As the authors themselves point out, however, extraordinary claims require extraordinary evidence to avoid and answer possible scepticism. While I think the photographic evidence of the cut marks is very compelling, I will focus on assuring that the classifications and analyses are as rigorous as possible. For this reason, I will therefore dedicate my review solely to the taphonomic questions of the paper, as all other aspects are far from my area of expertise.

First of all, I think a terminological clarification is needed, as I find the use of the word score throughout the text to be quite problematic. The word “score” typically refers to carnivoran activity, as in “tooth score”, while I have very rarely seen people refer to cut marks as scores. It is fundamental that we are as clear as possible in the words we use to describe certain processes, so as to avoid confusion, and because of this I encourage the authors to use a different term that is less likely to be mistaken with other taphonomic alterations (we already have enough issues with taphonomic equifinality, we don’t want to add any “terminological equifinality” to the mix!). In the taphonomic atlas published by Fernández-Jalvo and Andrews, for example, the authors talk about a “linear” category, where both tooth scores and cut marks can be considered “linear marks” – they have a length that is more than twice their width. I think it is much more appropriate to use a term like this, than use a word that is already shared in the name of a different type of taphonomic alteration.

--We have changed the generalized form of “scores” to “linear marks” throughout the manuscript.

I also think more evidence about how the collections were obtained and how they are currently being curated is also very important, which ties in to my next comment about the moulding process. The authors state that at least two moulds were taken per cut mark, however a recent study by Valtierra et al. (2023, DOI: 10.1111/arcm.12938) have shown that successive extraction of moulds can have an effect on the morphology of the traces being studied. While they showed that statistically these changes are not too important, this work was carried out on the morphological study of a single profile along the BSM, at a very specific point. It is also important to point out that with each mould taken, the natural variability of metric variables decrease, which could have a potential impact on certain analyses. While some of the variables used in the quantitative analysis of the present study are similar and should not, therefore, change too much, I am pretty sure that there will be a change in the morphology of the trace after each mould is taken. This leads me to ask, which moulds were used for the quantitative analysis? The first mould taken? Or the last mould taken? How were the fossil specimens cleaned as well? When performing such high resolution analyses, and considering variables such as volume, these different processes could have an impact. While I don't think this would have too much of an impact, it is important to at least document and be aware of it.

--The molds measured by Pante were the second mold taken from each specimen.

Valtierra et al. 2023 used different methods of data collection and analysis than those that are applied here so their results are not comparable to this study. Their methods replicate those of Maté-González et al., 2015, which used a low resolution structured light scanner and only data extracted from profiles. They also tested the impact of multiple moldings on modern and not fossilized bones that have surfaces that are much more resistant to change due to their lithified states. We agree with the reviewer that our molding process likely did not impact our results, but we appreciate their research directed at determining the impact of molds on mark morphology.

Beyond overlying trephic modifications, I also think it is very important to consider the impact dendritic etchings and root marks may have had on the samples, especially when considering the methodology employed for quantitative BSM analysis and identification. For example, in VGr.0519, VGr.1557, VGr.1767, VGr.2275, VGr.2687 and VGr.2649 you can see overlapping biochemical alterations that would surely affect the calculation of BSM volume and other parameters, wouldn't these parameters now be unreliable for the identification of the traces? To what extent should the analysis therefore include these variables? What happens when you perform calculations on some of the samples including or excluding some of these parameters? Do the classifications change? It is very true that almost any methodology dedicated to the high-resolution analysis of BSMs would be conditioned by these alterations, however, calculations such as volume are more likely to be effected by such changes to the cortical surface.

--The reviewer is correct that bioerosion can impact the measures collected here, which is why marks where bioerosive processes impacted depth were not included in the analysis. The specimens cited above had multiple modifications present, and those that were altered were excluded from our results.

With regards to data acquisition, much more information is needed regarding the microscopic analysis, including the precise FOV and numerical aperture of the lens used and the windows scanned; variables that have been proven in the past by authors such as Calandra et al. (2019, DOI: 10.1038/s41598-019-42713-w) to effect the replicability of analyses.

--We have added the information for the objective used in our analysis. The 5x lens has a numerical aperture of .15, a working distance of 23.5 mm, a field of view of 3400 um x 2837 um, a spatial sampling of 2.76 um, and an optical resolution of .93 um. However, we are not sure what the reviewer is referring to by “the window scanned”. It is important to note that cut mark analysis is not equivalent to microwear analysis as the resolution required for microwear is much higher and the measured variables are different.

Also, if I recall correctly, the S Neox has the option of obtaining 3D models using either Focus Variation or Confocal techniques, that would also condition the quality of the 3D model produced, so the modality must be stated.

--The S-Neox does not have a focus variation only mode, but it can combine focus variation with confocal data or use confocal data alone. The results are not impacted by the different techniques as the focus variation data is only incorporated to capture high slopes that are unmeasurable by the confocal method.

Why was only 5x magnification used? What colour light was used? My own experience with a similar microscope found 5x and 20x to produce different results for parameters such as roughness and volume when doing use-wear analysis. For this reason I'd also argue that comparing data obtained from the S Neox and the Nanovea ST400 could definitely be considered problematic for some of the quantitative variables included in the study (again, see the work of Ivan Calandra). The authors state that the results should not be impacted because of “coarser scales of measurement” (which would probably answer my question of why was only 5x used), however this is yet to be tested, and when you calculate variables such as volume and roughness, then multiple researchers, including myself, would in fact disagree with this statement.

--We understand the reviewer's concerns about the light and magnification, but it is important to keep in mind the scale of bone surface modification (BSM) analyses is much larger than that for microwear. Microwear analyses only sample a small area of a tooth's surface, which allows the use of higher magnification objectives. BSM's are many times larger than this and high magnification measurements are not possible on this scale due to the time required and limitations of the equipment used. For example, measuring a mark with a 20x lens would produce a point cloud that would exceed the memory capabilities of the S-Neox and would take hours instead of minutes to capture, if it were even possible. We chose the 5x lens because it is the closest in resolution and xy sampling rate to the previously used Nanovea st 400. Slight variations in measurements will occur in repeated measurements of marks even on the same device. This error range has been reported and explored in Pante et al., 2017 (which is cited in the text currently) and is on the order of a few microns for a variable like depth. Variations between measurements collected with different color lights or between devices are on a nanometer scale and only relevant when conducting high-resolution surface texture analysis. For example, in our sample, the variation between the mean values for depth between BSM are typically on the scale of tens to hundreds of microns so measurement variability between instruments and/or analysts equates to noise for our scales of analysis. This is even true for our measurements of Ra, which varies greatly between mark types. Ra measures the deviation of a surface from a mean line and the greatest deviation in a profile through a BSM is its overall depth. Therefore, our use of Ra is not in the traditional sense, which requires much higher

magnification to measure. However, the variable as measured here remains useful in discriminating between mark types.

Information about how many non-measured points during the scans are also required information, so as to properly assess the quality of the scans used for the analysis.

--Scans from molds generally result in 100% data capture. Occasionally missing data points occur, but because they are isolated and outside of the mark they do not impact the measured variables. Every study that employs confocal microscopy fills in missing points and the quality of scans can be better judged by the provided image captures of the 3-D models, which are provided in the Supplemental figures.

Regarding the post-processing of the 3D models obtained using the microscope, the author states that they removed form from the data, yet no indication about what polynomial was used to do this. I have to say that I think form removal for cut mark analysis is potentially very problematic, however I understand why it has been done considering some of the variables that have been measured. Form is removed in microwear studies because this directly has an impact on roughness parameters, and will also influence how easy it is to compute each traces volume, however when we are talking about the properties of the profile, form removal will have an effect on mark morphology.

--These methods are described in Pante et al., 2017. Importantly, the mark is excluded from this transformation, which is detailed in Pante et al., 2017. Regardless, the quantitative method used effectively discriminates between the processed marks so removing form is not problematic for our particular method and is necessary to capture the variables of interest.

With regards to the statistical analysis of the data, and classification tasks performed, I also have some concerns that need addressing. First of all, Quadratic Discriminant Analysis (QDA) should not be used for ordination, as has been done for supplementary figure 22. While use of QDA algorithms is perfectly acceptable for classification, it should not be used for ordination for a number of mathematical and epistemological reasons. QDA is a classification algorithm, so of course there will be separation between samples in the constructed and visualised feature spaces, because that's what it is designed to do – force separation to exist. There is also a huge issue with these types of constructed feature spaces due to issues with sample size and imbalance. While the individual to variable ratio is acceptable for the dataset provided, the sample imbalance is large enough to be problematic, in particular between cut marks and percussion marks

--While there is some literature that agrees with this it mostly relates to LDA and geometric morphometrics and not QDA on variables related to size rather than shape. We employ QDA here due to the unequal variance and size of the samples and we assess our model setting prior probabilities proportional to the occurrence of each mark to account for the differences in sample size.

(side note – why are these two compared in general, are there people who consider equifinality to exist between percussion and cut marks?).

--Percussion marks inflicted by sharp edged anvils can resemble cut marks in our experience. Short and shallow cut marks can also resemble microstriations associated with

percussion. The term equifinality can be misleading as it is a function of analyst experience with control collections or the methods employed.

What the authors should use instead, if they wish to visualise the multivariate differences between samples, is perform and visualise a PCA with a biplot, or simply present boxplots for each variable separately. If they want to use the QDA figure, then it must be used as a complementary representation of the data. It's also important to ensure that the feature space is constructed using cross-validation, otherwise separation is much more likely to be spurious in nature. For an example, see <https://doi.org/10.1007/s11692-020-09494-x> - although this paper was done for the purpose of discussing use of similar techniques in geometric morphometrics, the algebraic properties of the ordination technique are very similar, while the input to the figures in this paper are not landmark data but instead standard multivariate distributions and are thus completely comparable to the present case study. Finally, an ordination technique like these ones should not be presented without including in the axes labels the percentage of informative that is captured by each dimension, otherwise they are very hard to interpret.

--We have added label percentages to the figure caption (Canonical 1 is 86.9, Canonical 2 is 7.9%) and a PCA biplot as supplemental material.

For the evaluation of how well a classification algorithm works before being applied to archaeological data (presented in the first paragraph of page 22), the “accuracy” metric is often misleading, especially in cases where the imbalance is notable. Instead the authors must present other metrics such as precision, recall, sensitivity or specificity, that are more informative of other parts of the confusion matrix

--We have added an ROC curve as supplemental materials that reflects the requested information and shows the model is highly effective. Thanks for the suggestion! We will also report accuracy as it is a standard that more scientists are familiar with.

Some minor comments:

In the results section of the main article the authors state that a reduced number of bones presenting polishing is proof of limited transport. Polish does not necessarily imply movement, it's the action of water-born abrasive particles altering the overall surface of the bone, this could occur with a low flowing course of water over a prolonged period of time without moving the bones at all. In the supplementary materials the authors mention rounding as well. I think it's more informative to simply present all of the water-based alterations to cortical surfaces in the main text as well.

--We have changed the term “transport” to “water-based alterations”

In the last paragraph of the taphonomy section in the results the authors mention that some specimens have “more degraded surfaces”, it would be nice if the nature of this degradation be clarified – is it due to fluvial processes, for example?

--We have added the clarification to specify that we mean root-etching and surface exfoliation.

Have percussion marks or evidence of anthropogenic fracture patterns been identified in the site?

--Percussion marks were amongst the BSMs we were looking for, though no evidence of them was found. We have updated the text to reflect this. We also recognize that breakage patterns are useful for taphonomic analysis, but due to our time constraints (we only had 25 day total to examine 5,000+ specimens), we prioritize bone surface modifications in our analysis. We acknowledge this and suggest it as a future study in our full taphonomic analysis manuscript.

On line 227 the authors include some citations of papers that support the notion that sites exist with identified cut marks but no associated lithic assemblages. Among these cited papers, they include BK from Olduvai, however BK does contain a huge amount of lithics published as far back as the Leakeys, and is even mentioned in the abstract of the very paper they cite. I can assure you BK has lithics, while Mary Leakey even noted the presence of handaxes from this site. You also cite Dikika, which is very problematic, and many would even argue that not only are there no lithics, but no cut marks either

--The reviewer is corrected and we have removed the reference to BK from Olduvai and also clarified our language to indicate that these are published reports of cut marks.

In summary, I once again wish to thank the authors for their work, I truly have no problem at all with the validity of their finds, I just hope to ensure that the presentation of results are as rigorous, robust, and transparent, as possible. I have no problem disclosing my name as a reviewer - Lloyd A. Courtenay

--We appreciate the positive comments and thoughtful suggestions. Thank you.

Reviewer #3 (Remarks to the Author):

The paper presented by Curran and colleagues, entitled "Hominin presence in Eurasia by at least 1.95 million years ago," presents clear evidence of ancient anthropogenic activity in the form of butchery marks from several sites in the Olteț River Valley in Romania. The remains come mainly from a single Grăunceanu assemblage, although they also provide data from some other penecontemporaneous sites. A list of U-Pb dates establishes an age of 1.95 Ma. This work offers highly relevant information for understanding human dispersal prior to 2 Ma and provides palaeoenvironmental data, through isotopic analysis, suggesting conditions conducive to the development of such human movements.

The authors present bone remains with unquestionable cut marks, which have been analyzed using both quantitative and qualitative methods. Nevertheless, it is surprising that the authors state that the fossils were examined under strong, low-angled light for evidence of bone surface modifications. It would have been more appropriate to use microscopic observational methods (stereomicroscope, ESEM, digital microscopes, etc.). It seems plausible to suggest that this is the reason why the analysis method of Rodrigo et al., 2009, has been modified, with the aim of avoiding the microscopic variables of the analysis. It is likely that had a microscopic approach been employed, a greater number of marks would have been identified. The marks attributed to cut marks are clearly visible to the naked eye, and therefore it is reasonable to hypothesize that a microscopic analysis would have identified a greater number of evidence, as well as helped to attribute some of the probable marks to high-confidence marks.

--We agree with the reviewer that the use of ESEM and/or stereomicroscopy would likely enable us to identify additional cut marks. Using macroscopic methods, hand-lenses, and dino-lites, we were able to identify 20 cut marks. Our team was constrained by time (and funding) and used qualitative and quantitative methods to maximize our data collection on the 5,000+ specimens in the time we had (~25 days total).

The authors propose that some of the striae are attributable to high-confidence cut marks, while others may be associated with probable or unknown marks. Although they employ a quantitative methodology, they utilize a similarity-by-association approach to ascertain the origin of the marks as either anthropogenic or non-anthropogenic. It would be preferable for the authors to not use association to attribute the marks, but rather to only apply the proposed analytical principles. It is unclear why the teeth were excluded from the taphonomical analysis. If one of the objectives was to identify butchering marks, the teeth may show evidence of cuts related to the cheek defleshing and/or tongue removal. In addition, some of them could have been dated, obtaining direct dates from remains modified by lithic tools.

--We looked at any dentition that was still in the mandible or maxilla, but did not analyze isolated teeth, which we have clarified in the manuscript.

In the text (third paragraph of the Taphonomy section), the number of remains with high-confidence cut marks is initially indicated as seven. However, later in the text, it is stated that there are eight remains, and these are also the number of fossils that appear in Table S2.

Therefore, it is necessary to correct the text.

--Thank you for catching this- you are correct (there are seven high-confidence cut marks from VGr. and one from FM). We have made this clearer in the text.

In some instances, the observed marks appear to be clearly anthropic in origin. However, the analysis suggests otherwise. This discrepancy may be a consequence of the macroscopic analysis developed by the authors and the omission of certain microscopic variables that could have biased the results. A good example is the metopodial VGr.0429. Upon examination of the image (Fig. S12), it appears to be a clear anthropic disarticulation mark. However, the analysis suggests that it is, in fact, a tooth mark. It is recommended that these probable cutmarks be reanalyzed under microscopic tools (a stereomicroscope, at a minimum).

--We agree with the reviewer that this mark is anthropic in origin. The mark was relatively deep and on the edge of the condyle, which is the likely reason that the quantitative analysis assigned it to “tooth mark”. Our qualitative approach identified the mark as a cut mark and thus it is unlikely that further magnification would alter our conclusion. This combination of analytical procedures (qualitative and quantitative) was used for all marked specimens, thus we are confident in our assessment of the marks described here, recognizing that it is conservative.

VGr.1490 is described as a cutmark with a high probability and a high confidence, yet it appears both in the text (page 6 of SI) and in table S2 as a probable cut mark. Could this be an error? In the SI description, the authors even assign the activity carried out by hominins, tendon extraction.

--VGr.1490 is identified as a “probable cut mark”. The quantitative analysis identified this as a cut mark (with high posterior probability), however given its curved trajectory, we

assigned the more conservative identification of “probable”, though we agree that is it most likely anthropic in origin.

Why do the authors attribute the mark observed on specimen VGr.2052 (Fig. S18) to trampling if the analysis indicates that it is a cut mark with a high probability? This suggests that the authors sometimes adhere to the result of the quantitative analysis, but at other times, the observation of the researchers takes precedence.

--We have edited our description of this mark and have clarified our final assessment procedure, in which we did give slightly higher priority to the qualitative analysis done by human observers, who can implement a more holistic approach than the quantitative analysis can alone.

Given the paucity of information on the stratigraphic context of these ancient collections, it is understandable that the authors refrain from offering interpretations regarding the type of occupation or the mode of access to the resources. However, it is crucial to clarify the authorship of the marks and their age.

--We agree that these questions are important. However, our data cannot identify an author of the cut marks beyond hominin. We discussed the ages for the deposits in this manuscript.

Tables 1 and S7 are invaluable for the compilation of all ancient sites with evidence of hominids or anthropogenic activity (lithics and butchery marks) and the corresponding bibliography. It is recommended to include additional bibliographical references concerning published extended taphonomic studies of the cited sites:

--We appreciate the reviewer’s compliment and have added these references.

- For the Ain Boucherit site (Algeria):

Cáceres, I., Chelli Cheheb, R., van der Made, J., Harichane, Z., Boulaghraief, K., Sahnouni, M., 2023. Assessing the subsistence strategies of the earliest North African inhabitants: evidence from the Early Pleistocene site of Ain Boucherit (Algeria). *Archaeological and Anthropological Sciences* 15, 87. <https://doi.org/10.1007/s12520-023-01783-8>

For the Sima del Elefante site (Spain):

Blasco, R., Blain, H.A., Rosell, J., Díez, J.C., Huguet, R., Rodríguez, J., Arsuaga, J.L., Bermúdez De Castro, J.M., Carbonell, E., 2011. Earliest evidence for human consumption of tortoises in the European Early Pleistocene from Sima del Elefante, Sierra de Atapuerca, Spain. *Journal of Human Evolution* 61, 503–509. <https://doi.org/10.1016/j.jhevol.2011.06.002>

Huguet, R., Saladié, P., Cáceres, I., Díez, C., Rosell, J., Bennàsar, M., Blasco, R., Esteban-Nadal, M., Gabucio, M.J., Rodríguez-Hidalgo, A., Carbonell, E., 2013. Successful subsistence strategies of the first humans in south-western Europe. *Quaternary International* 295, 168–182. <https://doi.org/10.1016/j.quaint.2012.11.015>.

The evidence presented is of significant interest and should be published. The utilization of microscopy to analyze the attributes of the cut marks is recommended in order to identify

additional marks, correctly assign probable ones, and enhance the images presented. It is also recommended to add some high-resolution images of examples of trampling or carnivore toothmarks.

--We thank the reviewer for their insightful comments. We will be including further BSM images in our other (detailed taphonomy) manuscript, which is currently under review at the Journal of Human Evolution.

Reviewer #4 (Remarks to the Author):

Hominin presence in Eurasia by at least 1.95 million years ago, by Curran et al.

This is a very interesting work with some exciting results suggesting early evidence of human presence in East Europe by > 2 Ma, i.e. much older than thought until very recently. Please note that the evaluation of the cut marks on bones and whether there are human made is beyond my area of expertise. I will mostly focus my review on the chronology and general context of the fossils.

As a general comment, I would say that the manuscript is overall well written and easy to follow. The results and data presented apparently support the interpretation proposed by the authors. Actually, the main issues are in my opinion not related to what is shown in the work, but rather to what is NOT shown. In that regard, the absence of a detailed contextualization of the fossils is my major concern. I get that the fossils are coming from old excavations and that the various sites are most likely no longer accessible, but we have virtually no idea about their stratigraphic position, and whether the fossils represent one single or various assemblages. In other words, the integrity of the fossil assemblages is simply unknown. This major question mark about the integrity of the fossil assemblage, and whether all fossils correspond to a single depositional event, is actually not addressed in the present work, since the fossils dated with U-Pb are not those that have cut marks. While one may assume that they have all the same age, we actually don't know, and this should be verified by running additional analyses on the cut-marked bones.

--We have added further text to describe the depositional context, about which we have a full manuscript under review at the Journal of Human Evolution. Our detailed taphonomic analysis strongly suggests that these remains are from a single depositional context due to uniform surface appearance, lack of weathering, and overall consistent preservation. In sum, the remains are likely to have been deposited in a low-energy overbank flood. We therefore have high confidence that these remains were deposited in a relatively short period of time and the U-Pb for Graunceanu also reflect the ages of the cut marked bones. Our U-Pb results for Fantana lui Mitilan and Valea Roscai also support the dating for Graunceanu since Fantana lui Mitilan has younger fauna and is dated to a younger age and Valea Roscai has been compared to Graunceanu biochronologically and also has a similar age.

The same comment applies to the stable isotope analyses performed on the equid teeth. Direct U-series dating should be performed in order to confirm its antiquity and make sure it can be closely associated to the other fossils.

--Given the short time frame and lack of funding for further analyses, additional dating is not possible at this time. Nevertheless, as discussed in the previous comment, the taphonomic preservation of the VGr. assemblage is very uniform and thus, we consider it to be a single deposit.

In that regard, I genuinely wonder whether Rare Earth Elements could also help demonstrate that all fossils have the same origin and come from the same stratigraphic unit.

--While we agree that rare earth elements (REE) may potentially be useful, in this context the very large range of REE contents observed within a single tooth (as a result of variable postmortem uptake) we see no easy way that REE could be used in this way to distinguish different stratigraphic sections.

The whole context is briefly summarized in Methods, but remains (too) vague. The authors mention that the location of various localities have been 'reidentified', therefore it should be possible to get further information about the context, and establish a proper stratigraphic column, description of the sediment, with some additional information about the formation of the fossil horizon and the depositional environment for the 3 localities. A geological or geomorphological map of the ORV with the stratigraphic correlations of the various sections would also be very useful.

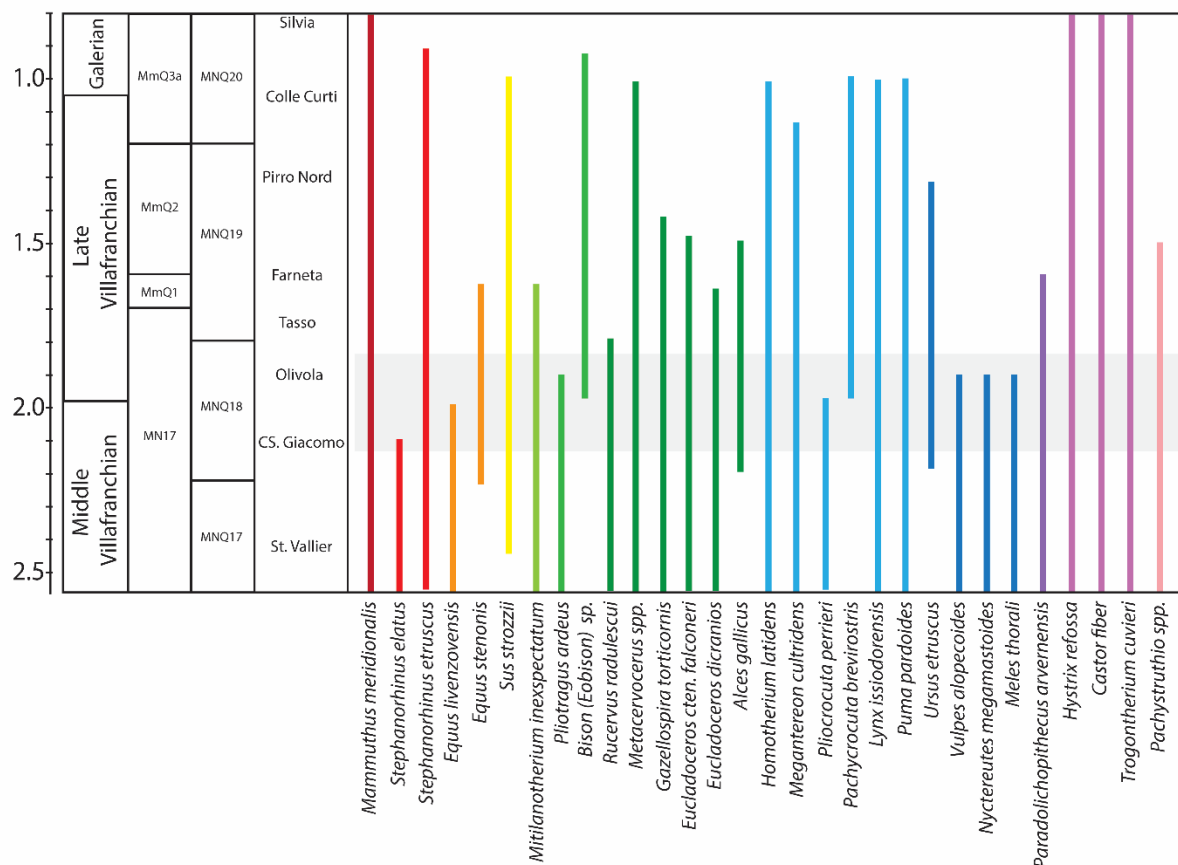
--Though we have reidentified the site locations, the original excavations completely exhausted the fossil deposits. Nevertheless, the original reports for Graunceanu (the main focus of this paper), describe the site as a single deposit (a "bone nest" or bed) and our taphonomic analysis verifies this, as specimen surface condition was found to be extremely consistent across the entire Graunceanu assemblage.

The lack of independent age control is also another major concern. There is some circular reasoning in using the biochronology (which is based on independently dated assemblages) to support the numerical ages obtained here. The uncertainty on biochronological inferences is usually minored and not really taken into account, and especially on large mammals. I doubt very much that the precision on the biochronological age is sufficient to undoubtedly position the site around 2.2-1.9 Ma, within a small 0.3 Ma age window. This would need further explanation with a dedicated section in SM, and in particular about the robustness of the age control on the localities used as chronological tie points. Rodents may possibly help refining the chronology of the assemblage (e.g. Maul & Markova QI, 2007), but seems like from Table S1 that there are too few remains. As mentioned above, the absence of independent age control adds a non-negligible uncertainty to the dating results, especially when knowing that dental tissues behave as open system, and may show U leaching, which would artificially age the sample. I had a quick look at Andreescu et al. QI (2013), and I didn't find the existing age control on the Tetoiu Fm very convincing to securely confirm its chronological attribution to the Early Pleistocene.

--Our research team has previously published a detailed biochronological assessment of the fauna at these fossil localities (Terhune et al., 2020); we have also published several additional papers reviewing specific taxonomic groups (Werdelin et al., 2023; Croitor et al., 2024) that also discuss the biochronology of these fauna. Though many species from these localities are not useful from a chronological perspective, others (*Sus strozzi*, *Bison* (*Eobison*), *Alces*, *Nyctereutes megamastoides*, *Vulpes alopecoides*, *Meles thoralis*, *Pliocrocota*, *Canis*) serve as important biomarkers for the Early Pleistocene and consistently point

toward Graunceanu being dated to the 2.2-1.9 Ma window we originally suggested in our 2020 paper.

The figure below represents this biochronological assessment from Terhune et al. (2020) (though has been updated slightly to account for minor changes in taxonomy, corresponding to Table S1). The gray box represents our approximate suggested time period of deposition for Graunceanu based on the fauna.



The U-Pb dates provide additional support to this estimate and are also consistent with biochronology from the two other localities analyzed, Fantana lui Mitilan and Valea Roscai. While we very much wish we had micromammal remains with which to independently test this dating, these fossils are lacking from the ORV localities. We count ourselves lucky that we in fact have two relatively robust independent dating results for the ORV localities that are consistent with one another; as the reviewer likely already knows based on their last comment for our Table S7 (see below), this is relatively rare. Indeed, many early Eurasian sites have been dated solely on the basis of biochronology or other relative dating techniques like magnetostratigraphy (see updated Table S7, which now includes dating methods for each site).

On an unrelated matter, I see in Table S5 that several samples are in secular equilibrium, while other do show U-leaching (i.e., beyond equilibrium values). For example, this is the case for sample VGr.2256-2. My first reaction would be that this outcome may seriously question the reliability of the U-Pb age of 1.95 Ma obtained on that same tooth. I think this aspect would deserve some additional discussion in the manuscript. From the information given in SM, I understand that both solution and laser ablation analyses were carried out on the teeth. Some samples were microdrilled. I am not sure to understand why, and this could be clarified in the text perhaps.

--We apologise for any confusion that may have arisen here. To clarify- all the U-Th determinations were performed on micro-drilled samples and all the U-Pb analyses were performed via laser ablation. While we thought the former was clear from the existing supplementary text, we have now added the words, ‘laser ablation’ to the main text U-Pb section to further clarify.

It is also unclear to me how LA analyses were performed, either by drilling vertically within the dentine, or through mapping an area of the occlusal surface? If so, can we see the maps with the spatial distribution of the various ratios? Please add some extra explanations in SM. The exact location of the samples analyzed for U-series (given in Table S5) within the teeth could also be given. Fossil teeth are known to be highly heterogeneous in terms of spatial distribution of radioelement, and especially old teeth (e.g., Duval et al. QG, 2011) and it is unclear to me how this uncertainty was evaluated. Some domains within the dentine or enamel may locally show very distinct U-series data, which would bias the age results. It would be strongly advisable to sample other parts of each tooth to evaluate this variability and help identify more recent uranium uptake or leaching overprints.

--Table S3 in the Supplementary information indicates that the laser sampling mode was ‘static spot’ analysis. As a result there are no maps.

--As we discussed in the SM, U mobility affecting U-Pb laser spot isochrons, and U mobility affecting ^{234}U and ^{230}Th ratios operate at different spatial and temporal scales in fossil dentine and are diagnosed in two different ways. We clarify this below but feel that the text we currently provide in the supplementary discussion is sufficient.

--The ultimate test for U-Pb isochron dating, particularly with such large spot populations is, “do we have a clean, clear U-Pb isochron?” If the answer is “yes,” then we can infer with confidence that an “early uptake” model is applicable, with relatively rapid post-burial U uptake from the surrounding environment and no subsequent U mobility on a scale that might significantly affect the equilibrium U-Pb age calculation. In this manuscript we present U-Pb isochron data for a sub-set of the investigated teeth where this was found to be the case.

--We have used a simple reliability test for our $^{234}\text{U}/^{238}\text{U}$ measurements, by pairing them with $^{230}\text{Th}/^{238}\text{U}$ measurements as for U-Th disequilibrium dating. Given the clearly greater than 1 Ma age of these samples from their U-Pb isochrons, closed-system evolution from any plausible initial $^{234}\text{U}/^{238}\text{U}$ value must be so close to the infinite ^{238}U - ^{234}U - ^{230}Th isochron so as to be analytically indistinguishable from it. Recent minor U gain or

loss (even where too minor to affect the U-Pb isochron age) will drive a sample away from the infinite isochron, towards younger apparent U-Th ages (U gain) or to a clearly open-system position (U loss). Dentine ^{238}U - ^{234}U - ^{230}Th measurements closer to the infinite isochron than their uncertainty indicate no recent U mobility at the point analysed and give a reliable measurement of undisturbed $^{234}\text{U}/^{238}\text{U}$ for that tooth.

--We measured $^{234}\text{U}/^{238}\text{U}$ on drilled dentine samples of ~ 2.5 mg and ~ 75 μg . The subset of the 75 μg analyses which had not gained or lost U gave a very similar mean $^{234}\text{U}/^{238}\text{U}$ to that of the ~ 2.5 mg samples, $\sim 30\times$ larger, indicating that recent dentine U mobility of these teeth (which had all passed the U-Pb isochron test above) is fine-scale, and likely internal. Any of the following would all lead to scatter in (and rejection of) the U-Pb isochron: greater U mobility, U mobility persisting over much of the burial time of the tooth, or significant heterogeneity of initial $^{234}\text{U}/^{238}\text{U}$ in the tooth. There is no expectation that the tooth reported by Duval et al. (2011) would have passed this isochron test (had we been able to undertake this at the time the paper was written), as that tooth had clearly been subject to very significant recent U uptake unlike any of the teeth considered here and would have been rejected from this study.

--A successful (i.e., low scatter, low MSWD) U-Pb isochron indicates any U mobility is at a level which does not affect the U-Pb age. Minor, relatively recent U mobility can affect measured $^{234}\text{U}/^{238}\text{U}$ ratios so we have applied a simple test for this using the ^{238}U - ^{234}U - ^{230}Th isochron space. Some of the $^{234}\text{U}/^{238}\text{U}$ sub-samples were found not to have been measurably affected by U mobility, and the tight clustering of their values allows us to apply their mean (with sufficient uncertainty to cover their scatter) to all of our disequilibrium corrected U-Pb isochrons.

Specific comments

1. Main text, 'The current best': What does 'best' mean here? Please be more specific.

--We have updated this wording.

2. What is the definition of 'ephemeral traces' here? There several occurrences of this term throughout the manuscript and I guess this should be explained somewhere.

-- By ephemeral we are referring to the fleeting traces of hominins represented by either small or less-well documented localities that hint at a larger presence of hominins throughout Eurasia in the Early Pleistocene, but that is not as robustly demonstrated as sites such as Dmanisi. We have tried to clarify this in the introduction.

3. U-Pb dating, ' 2.01 ± 0.20 Ma to 1.87 ± 0.16 '. Please indicate at least once in the main text whether errors given are 1 or 2σ . 'Ma' missing after ' 1.87 ± 0.16 '

--Values reported in the main text are $\pm 2\sigma$. This text has been updated and the missing Ma added.

4. U-Pb dating, 'which is ubiquitous in fossil teeth and bone samples': Do you mean at this specific site? Please clarify the sentence. There are actually many works showing the timing of

the U uptake may significantly vary within a given dental tissue or bone (e.g., Grün and Taylor Ancient TL 1996; Grün, RM 2009; Duval et al., QG, 2011), between the various tissues of a given tooth, and between fossils from the same site/horizon.

--Our statement refers to the fact that almost all fossil teeth and bone samples – from any site – will experience some form of post mortem U-uptake. We are, of course, aware of the variabilities in timing both between different sites and across individual samples but we do not actually refer to timing at all here and so believe that our generic statement remains accurate.

5. U-Pb dating, ‘Grăunceanu took place >1.95 Ma and potentially $>2.01\pm0.20$ Ma’: actually these two ages mentioned here are virtually the same given the associated error. There is no need in my opinion to mention them both, one is sufficient.

--We agree that from a statistical perspective these are the same, though we feel it’s important to emphasize that deposition of the site might have occurred well before 2.0 Ma. We have edited this text slightly to reflect this nuance.

6. U-Pb dating, ‘These results make these sites the most securely dated localities from Early Pleistocene Europe’: the concept of ‘securely dated’ is very relative. The absence of independent age control and the presence of minimum ages on fossils do not make the sites better constrained than others in my view.

--We have updated this wording to now say “These U-Pb results, coupled with the biochronology for these localities¹¹⁻¹⁴, lend high confidence to our age of deposition for Grăunceanu of >1.95 Ma and potentially prior to 2.0 Ma, and indicate that Grăunceanu is the earliest locality in Europe to show evidence for hominin activity.”

7. Discussion, ‘Our paleoenvironmental data for Grăunceanu indicate a more open, arid environment with some nearby woodlands and water sources’: do you mean more open and arid than today? Please specify. Additionally, the estimated MAP for Grăunceanu is higher than current conditions, which seems to be in contradiction with the statement above on a more arid environment.

--We have changed our manuscript from reporting MAP to reporting seasonal data. That is, we estimated precipitation amounts for the driest (summer) and wettest (winter) months. Thus, the driest months would have been in summer, resulting in more arid conditions than today. We also removed “more” from that sentence.

8. Discussion, ‘While it is clear that the hominin presence in Eurasia at this time was not extensive’: Why is it clear? And what do you mean by extensive? There is actually early evidence of human presence in Europe from Ukraine to Spain.

--We have edited this text to say “While it is clear that the hominin presence in Eurasia at this time was likely geographically and temporally discontinuous, the preponderance of ephemeral traces for hominins in this region can no longer be ignored.”

9. Fig. 3 caption: please specify whether errors are displayed at 1 or 2σ .

--Values are reported as $\pm$ Ma in Fig. 3.

10. Fig. S24: I am not sure to understand the interest of showing elemental concentrations in dental tissues? What kind of information does it bring?

--We generated this image in our pilot studies and agree that it is not necessary for the manuscript and have removed this image. Please note that this deletion does not show up as a tracked change to allow for a more readable manuscript revision.

11. Fig. S24: Metal concentration or elemental concentration?

--Figure removed.

12. Please would it be possible to show instead the maps of U concentrations, U234/U-238 and apparent age for each of the dated samples?

-- U-Th analyses were conducted by micro-drilling with very few sampling locations per sample and so maps cannot be constructed. The U-Pb dating method requires the use of all laser ablation spots from a given sample in order to determine an age via the ‘isochron’ construction. Once again, therefore, maps cannot be constructed using individual spots.

13. Fig. S25: How was the present-day U234/U-238 measured?

-- The determination of this parameter is explained in detail on page 8 of the Supplementary information.

14. Table S1: Does the composition of the fossil assemblage come from previous studies (if so, please cite the relevant references), or from the direct study of the specimens by the authors? If the data given for this table are coming from a new taxonomic revision of the fossil assemblage, then a dedicated section including a proper systematic paleontology study should be included.

--This updated table incorporates our prior work describing the fauna from the Oltet River Valley (Terhune et al., 2020, 2021; Croitor et al. 2024; Werdelin et al., 2023). These references are already provided in the main text where we cite this table, but we have also added them to the caption for Table S1.

15. Table S4: which U-234/U-238 values were used for the disequilibrium corrected age? This may be clarified, as Table S5 shows that there are several values available for each sample.

-- As noted above, page 8 of the Supplementary information contains a detailed explanation of how the value of 1.0039 ± 0.0035 , which is used for correction of all U-Pb ages, was determined.

16. Table S5:

- Please clarify whether MXF and UMF samples have all been measured through solution analyses, or UMF samples have been analyzed using Laser Ablation?

-- As indicated in the SM text and table, all $^{238}\text{U}/^{234}\text{U}$ measurements are by solution, per Hellstrom (2003). As noted in the table caption, MXF/UMF indicate the two sample digestion procedures described in the SM text.

- Where (=which lab) was Th-230/Th-232 ratio measured?

-- The $^{230}\text{Th}/^{232}\text{Th}$ was missing a superscript “b” to indicate measurement at the University of Melbourne after Hellstrom (2003).

- Not sure I understand well the U mobility column. Do ‘Closed’ mean ‘secular equilibrium’, and ‘U gain’ mean that the ratio are still below secular equilibrium?

-- As noted in the table caption and SM text, and rephrased above, uranium mobility is indicated by measurable departure from the infinite age isochron. These samples cannot, however, be considered to be at secular equilibrium until $^{234}\text{U}/^{238}\text{U}$ is within its uncertainty of unity, which will occur hundreds of thousands of years into the future.

- The apparent U-Th age can be added to this table. Out of curiosity, how can we explain that the same tooth may provide an apparent U-Th age of <500 ka and a U-Pb age of > 1.9 Ma. This would probably deserve some additional comments in SM.

-- We respectfully disagree with the reviewer here as we feel that the U-Th ages to this table would be misleading. Where a U-Th age can be calculated, it simply indicates recent U gain for the sub-sample analysed and has no meaning beyond that. Regarding the difference between calculated U-Pb age, and apparent U-Th age, we dedicated text in the SM to discussing this, also rephrased above. The U-Th age calculation, where it is possible to make one, is not meaningful as it merely indicates recent micro-scale uranium gain.

17. I have my doubts on the reliability of the age of many localities mentioned in that table. Can the authors at least add one column mentioning the dating methods employed to constrain the age of the site (e.g., U-Pb, palaeomagnetism, biochronology, ESR, TCN, etc.)?

--We have added the dating methods in a new column in Table S7.

Response to Reviewer 1

I have read the revised manuscript by Curran et al., titled “Hominin presence in Eurasia by at least 1.95 million years ago.” As I mentioned in my previous review, the manuscript is generally well-written and of broad interest. However, I am unable to comment on the robustness of the evidence for hominin presence or the U-Pb ages, as these techniques fall outside my expertise.

My main concern with the original version related to the interpretation of the stable isotope data. The authors have provided a thorough and detailed response to my comments, and I am overall satisfied with the revisions, which include:

- a) Calculations of the oxygen isotope composition of meteoric precipitation 1.95 Ma using three different transfer functions, along with an uncertainty estimate.
- b) A brief discussion justifying the conclusion of a higher winter precipitation component to aquifer recharge relative to today.
- c) Estimates for precipitation amounts in the driest and wettest months, rather than an average for the whole year.

I would only recommend that the authors clarify in the text that the Kohn et al. equation was developed for mean annual precipitation (MAP), and its validity for reconstructing seasonal precipitation is untested. Additionally, while Kohn (2010) did not provide uncertainties for his equation, these are discussed in Kohn and McKay (2012, *Paleo* 3).

Overall, I am satisfied with this revised version and recommend the manuscript for publication in *Nature Communications*.

--We thank the reviewer for their support of our paper. We have edited the methods and supplemental text to include a discussion of the uncertainties in the seasonal range estimates, and added the Kohn and McKay (2012) reference.

Response to Reviewer 2

Author's response to reviewer: The molds measured by Pante were the second mold taken from each specimen.

Reply: Please include this in the paper, there is no reason why this very simple sentence cannot be included, thus responding to the reviewer's (my) concerns.

--We have added text to the methods section indicating that the second set of molds were employed in the quantitative analysis.

Author's response to reviewer: Valiterra et al. 2023 used different methods of data collection and analysis than those that are applied here so their results are not comparable to this study. Their methods replicate those of Maté-González et al., 2015, which used a low resolution structured light scanner and only data extracted from profiles.

I debate: The authors don't think there is a correlation between depth, width, and volume. This is a potentially mathematically flawed statement; please take a triangle, change its width and height, and see how any minute change will also have an impact on the area. Regardless of whether the measurements taken are not precisely the same, if other authors note that there is a metric change to cut marks due to the moulding process, then one cannot completely disregard the study because slightly different measurements are being taken by another study. Each of these metrics are inherently dependent on the global properties of a mark, even if measured on only a single profile, or the entire mark. Also, if other authors were able to observe changes in the metric properties of cut marks at such a 'low resolution', then wouldn't these changes be more pronounced at a higher resolution? Half the time, lower resolution techniques are not able to differentiate between slight variations in variables, but high resolution techniques are, so if a low resolution technique is able to detect something, then it would be reasonable to assume that the high resolution technique is more likely to detect such a pattern. This can be simply acknowledged in the revised version of the text.

--As noted above we have clarified in the text which set of molds were used in our analysis, and in response to the comments below we have clarified in the text that all bones are fully mineralized and unlikely to be affected by molding.

Author's response to reviewer: They also tested the impact of multiple moldings on modern and not fossilized bones that have surfaces that are much more resistant to change due to their lithified states. We agree with the reviewer that our molding process likely did not impact our results, but we appreciate their research directed at determining the impact of molds on mark morphology.

Reviewer's reply: I understand the reply, however the authors have not clarified anything in the text to support this. I have studied materials at 3 Ma that present no sign of mineralization at all, and are equally as fragile as a modern bone, while on the contrary I also currently study materials that are only 20 ka that are completely mineralized. (1) Just because bones have fossilized, doesn't make their cortical surface more or less fragile, while (2) not all sites present the same degree of mineralization. The authors haven't mentioned at any point the mineralization of the bones in their paper. At this moment in time, we have to take the authors word that the moulding process has not damaged the bone. Transparency is needed and the authors should at least clarify some of these points. Also, if they are certain that the two moulds are the same and no changes have occurred, then the first and the second mould of the same mark should produce the same metric results, could this not be tested, just in case?

--We have added text in the methods section to clarify that all of the fossils analyzed as part of this study are heavily mineralized and that we did not observe any damage to the bones during the

molding process. We therefore have high confidence that the fossils specimens were not damaged by the molding process and that results from different molds are comparable.

Overall reviewer's reply – The authors reply with some valid points to me, the reviewer, but these points are still left unattended in the text, leaving all other readers of this body of work in the dark. They could easily clarify these points, in most cases with a single sentence. I repeat – if these concerns and questions could have occurred to me, they can probably occur to somebody else. We need to be as transparent as possible in our analyses, especially in the case of such extraordinary findings. Including the sentence “The molds measured were the second mold taken from each specimen”, is a very simple statement to include, with an additional acknowledgement about the potential risks, even if these risks may not be entirely present. And if the authors are indeed certain that no changes have occurred, they could easily check this by measuring the different moulds and comparing whether the measurement on mould 1, 2 and 3 are all the same, which would validate the point they are trying to make.

--Please see comments to individual points above. At this point we feel that an additional analysis focused on error related to molding individual marks is beyond the scope of this study. We also feel it likely that any variation among molds is small in comparison to error associated with different techniques or observers.

Author's response to reviewer: The reviewer is correct that bioerosion can impact the measures collected here, which is why marks where bioerosive processes impacted depth were not included in the analysis. The specimens cited above had multiple modifications present, and those that were altered were excluded from our results.

Reviewer's reply: have the authors clearly stated this in the text? It would be very easy to add a sentence that clarifies this. If this is stated in the text, please can they indicate where, because I cannot find it.

--We have added text to the methods section that states “in a few instances marks that were molded correctly were subsequently removed from analysis due to bioerosion.”

Author's response to reviewer: ... For example, measuring a mark with a 20x lens would produce a point cloud that would exceed the memory capabilities of the S-Neox and would take hours instead of minutes to capture, if it were even possible...

Reviewer's reply: I won't go too much into this, but the idea that the point cloud exceeds the memory of the computer is interesting... I, and other colleagues, study cut marks, tooth marks and percussion marks at 20x, and have no problem at all with memory. Yes it takes longer, but science requires sacrifice, and while the work has been done on moulds, there is no issues with limited access to the samples which would imply that the authors could not have taken the time to scan all the marks at 20x. Maybe older versions of the same microscope present storage issues that newer version of the S-Neox do not present, if that is true then it might be necessary to state the year, version, or some indication about which edition of the microscope is being used, because I repeat, in our lab (University of Bordeaux), we

have an S-Neox and have had no problem studying cut marks at 20x magnification.

--The model of the S Neox scanner used in this study was manufactured in 2018 (now specified in the text). Thus, the reviewer is correct that we used a slightly older version of the scanner that the reviewer also has access to; memory on this older version is limited in comparison to newer models. We have added the manufacture date to the text. Given that our analyses with the 5x objective have clearly demonstrated anthropogenic alterations to the ORV remains, we feel that to reanalyze everything, including the sample of 898 experimentally made BSM, at a higher resolution would not provide a return on the time investment that reanalysis would require. Our QDA model based on the data collected with the 5x objective is shown through multiple statistical methods of assessment to be highly effective at classifying BSM.

Author's response to reviewer: the quantitative method used effectively discriminates between the processed marks so removing form is not problematic for our particular method and is necessary to capture the variables of interest.

Reviewer's reply: I'll have to take the author's word for it, however they still have not stated anywhere which polynomial is used to remove the form from these scans. This is an incredibly simple statement to include in the methods – A polynomial of X was used to remove the form.

--Methods for the removal of mark form are described in Pante et al. 2017 (which is already cited in the methods text). Because the underlying shape of every bone is different, a different polynomial equation is used for each. We now specify that readers should see Pante et al., 2017 for further detail.

Author's response to reviewer: While there is some literature that agrees with [QDA should not be used for ordination] it mostly relates to LDA and geometric morphometrics and not QDA on variables related to size rather than shape.

Reviewer's reply: There is a lot to unpack, even in this short paragraph. First a point of debate: "variables related to size rather than shape" - Size and shape can be, and are often, inherently related - unless you provide evidence to say otherwise, this statement can be very problematic. Because the objective of the paper is not to look at this question, I will brush past this statement, but I just would like to point out that care is needed here because this is not necessarily true, and in any other paper I would very strongly disagree with such a statement. Regarding the much more important point, I think the author's have missed the point with reference to the issues of using discriminant scores for ordination. I am not talking about geometric morphometrics (GMM), I am talking about mathematics. All the literature I can find on QDA and the corresponding plots still relate them to the calculation of canonical variates. In fact, QDA and LDA are built on some fundamentally identical algebraic calculations, that means they are completely related. Considering how QDA and LDA are heavily linked algebraically, this would in turn imply that all issues related with LDA, which are related to CVA, are related in turn to QDA as well. I'm talking exclusively about algebra, not about GMM. The reason I mention GMM is simply because this is where the majority of research into these algebraic properties lie. However, I would like to point out that authors such as Cardini and Bookstein do not even use landmark based data in their

papers about this topic, they use simple multivariate normal distributions, and thus still prove my point – QDA should not be used for ordination. Likewise, the paper by Albrecht (that I provide below) talks mostly about linear measurements, not landmarks... therefore it is directly relatable to the present study.

It is interesting that the scores in the plots are also called canonical something. From this perspective how can you be so sure the plot is different from what has been criticised by other authors? Likewise, no indications or citations are included in the paper to clarify how calculations are performed, beyond the statement that JMP was used. JMP is a private licensed software, which I do not have a licence for (in my opinion this in itself is an issue) so I cannot perform any tests to check for replicability, or even evaluate the suitability and reliability of the algorithm in this specific context.

In sum; (1) if QDA is less sensitive to sample imbalance, I would like some references to support this, or simulations to prove this. (2) You still need to clarify if cross validation is done, and if it isn't performed, then justification as to why must be provided, because in a lot of circumstances the exclusion of CV would be considered completely problematic. (3) I am very happy to see a PCA is now included, which would generally resolve a number of the issues I had with the QDA plot, however, there are a number of issues related to the PCA. (3a) The PCA provided is very very hard to read, and I cannot see where the fossil points fall. Have they even been included? Please improve the presentation of this figure so it is clear to read and see how the archaeological samples compare with the experimental ones. This is something you did with the QDA plot (you already know my opinion about how reliable this plot is), but have refrained to do for the PCA; (3b) why have the confidence intervals been excluded for the PCA, but included for the (problematic) QDA? Please standardise the presentation of your results. (3c) Having red and orange points on the PCA makes the samples really difficult to distinguish. Also, is it just me, or do the colours not match up between the QDA and the PCA? Or even within the PCA at all. In the key there is blue, purple, green and purple, but in the PCA I cannot see blue, but instead see red, orange, green and purple. What is going on here?!

If the authors are still adamant that their use of QDA is appropriate, please provide justification. I provide justification here to highlight why it is not appropriate (once again, before the authors claim that literature originating from GMM is not relevant here, please read the methods sections of these papers to see that in most cases, the methods are not rooted in GMM analyses at all);

Albrecht, G. H. (1992). Assessing the affinities of fossils using canonical variates and generalized distances. *Journal of Human Evolution*, 7(4), 49–69.

Bookstein, F. L. (2019). Pathologies of between-groups principal components analysis in geometric morphometrics. *Evolutionary Biology*, 46, 271–302.

Cardini, A., & Polly, P. D. (2020). Cross-validated between Group PCA scatterplots: a solution to spurious group separation? *Evolutionary Biology*, 47, 85–95.

Cardini, A., O'Higgins, P., & Rohlf, F. J. (2019). Seeing distinct groups where there are none: spurious patterns from between-group PCA. *Evolutionary Biology*, 46, 303–316.

Mitteroecker, P., & Bookstein, F. (2011). Linear discrimination, ordination, and the visualization of selection gradients in modern morphometrics. *Evolutionary Biology*, 38, 100–114.

Rohlf, J. F. (2021). Why clusters and other patterns can seem to be found in analyses of high-

dimensional data. *Evolutionary Biology*, 48, 1–16.

It's incredibly important to point out that in each of these referenced papers, the authors may talk a lot about GMM, however the data they are actually using to explore the algebraic and mathematical issues presented are simply multivariate distributions. This means the data used for these statistical and theoretical experiments are completely analogous with the present author's data. In particular I recommend the paper by Cardini and Poly (2020) for proof on the importance of cross-validating their discriminant function plots. For the importance of cross-validation in classification I could recommend a large number of papers and books, however would simply state that the book by Hastie, Tibshirani and Friedman (2009) *The Elements of Statistical Learning* is more than a good starting point.

--We appreciate the reviewer's detailed response regarding QDA, LDA, CVA, and PCA. As we think is demonstrated by the reviewer's own text, there is considerable debate in the literature regarding the appropriate use of these methods, both in geometric morphometric analyses and beyond. Because of this we feel it useful to provide both a CVA and PCA plot in our supplemental data, which will provide readers with as much data as possible (as this reviewer advocates for throughout their response) with which to make informed interpretations of the data. Please also note that our CVA plot is an accurate representation of the posterior probabilities that are used to assess the confidence of the QDA model in its classification of each mark. Marks that have posterior probabilities for multiple mark types appear in areas where the confidence intervals of these mark types overlap in the CVA plot. We do appreciate the reviewer's comments about making sure that these two plots are consistent in their presentation; indeed we had thought the colors were comparable but some error must have occurred in the creation of the PCA plot that caused the problems noted. We have updated the figure and now hope that it meets the expectations of the reviewer.

The reviewer also questioned whether we used cross-validation, which was clearly reported in our original submission where we specified that a 25% validation set was extracted from the experimental sample of BSM to validate the model. We now call this a 25% cross-validation set for clarity.

Reply to the reviewer: We employ QDA here due to the unequal variance and size of the samples and we assess our model setting prior probabilities proportional to the occurrence of each mark to account for the differences in sample size.

Reviewer's reply: Ok then please clarify this in the text, and also state what prior probabilities were used. This can be easily included in a single sentence.

--Our existing text already states "Prior probabilities were set proportional to the occurrence of each mark type in the dataset to offset the disproportionate representations of each mark type in the experimental sample."

Reply to the reviewer: We have added an ROC curve as supplemental materials that reflects the requested information and shows the model is highly effective. Thanks for the suggestion! We will also report accuracy as it is a standard that more scientists are familiar with.

Reviewer's reply: While the authors have already corrected the presentation of results to include more than just accuracy (ROC curves are a great addition), I have to simply state, for the purpose of scientific clarification, that I strongly disagree with the term or concept that accuracy is a "standard that more scientists are familiar with" - I could not disagree more with the logic that just because everyone else does it means that it is right. For years it was general consensus that the world was flat, this doesn't mean that these individuals are the future of science. For years nobody thought anything was smaller than the atom, scientific research proved them wrong too (if anyone is paying attention they might catch a Friends reference in here). Science evolves and so should we, just because "scientists" (this should actually be archaeologists/palaeoanthropologist/palaeontologists) are more familiar with accuracy, that doesn't mean we shouldn't break the mould and develop our analyses to adopt correct practices. Moreover, if you actually consult the literature regarding classification in data science, the authors would see that (mostly) non-archaeological scientists publish far more than simply accuracy. Given this last point, I would argue that in data science reporting only accuracy is certainly not a standard, which would make this statement incorrect on a number of different levels.

Just to prove the point, and make sure that we go beyond the field of data science and into the realm of archaeology: imagine we have a new set of 100 marks; 80 cut, and 20 trampling. Imagine our algorithm claims that all marks are cut marks. This would result in 80% accuracy, however, it would also be misleading considering how the algorithm is actually allocating the label cut mark to everything (i.e. hasn't learnt to classify anything at all). Now, if we calculate sensitivity and specificity, at least one of these values would be 0, successfully helping us detect that our 80% accuracy is misleading, and the algorithm is in-fact awful, and should not be used in real case studies (EVER!). Now imagine that instead of archaeologists we are doctors and we have potentially killed 80 patients. Just because our patients have been dead for 1.9 million years (wow, what a date!), doesn't mean we should not be just as rigorous and strict with our diagnostics. I apologise for rambling, but I find this point fundamental to clarify, especially if the authors find more great finds in the future and have different reviewers who are unfamiliar with this concept.

--We have addressed these comments in our original revision by providing an ROC curve. We maintain that classification accuracy is a valid measure for assessing the strength of discriminant models and also a more familiar statistic to paleoanthropologists than an ROC curve.

Response to Reviewer 4

The authors have convincingly addressed most of my comments, and even when objecting, the answers given are reasonable and make sense. After going through the detailed rebuttal, it seems to me that the evidence provided on the three main critical aspects of work (i.e., human origin of the cut marks on bones; (ii) antiquity of the fossil material and (iii) palaeoenvironment reconstruction) are robust enough to support publication. I would especially like to thank the authors for dealing with the comments made by the reviewers in such a constructive way.

There are a couple of general comments I would like to make here after reading the revised manuscript, although they should not preclude the publication of this very interesting work that undoubtedly deserves to be shared with the community.

-- We thank the reviewer for their continued support of this manuscript and appreciate their professional and constructive feedback.

First, I must admit that I remain quite sceptical about the precision of the general chronostratigraphic framework of the ORV and Dacian basin. Yes, biochronology can give a rough chronological estimate of a given site, but this should be rather regarded as what it is, a rough estimation. Estimation because the inherent uncertainty typically associated to biochronology is usually minimized (or ignored). It may be related to (i) the taxonomic attribution of fossil specimens, which may vary depending on the specialist involved, (ii) the validity/robustness of large scale biostratigraphy correlations across a country or even a continent, which may often be questionable, and (iii) and the existing true independent age control available at other sites used for comparison (i.e., based on methods like magnetostratigraphy and numerical dating method), which is often limited and has an uncertainty on its own. This is why I personally have difficulties to accept precise age constraints based on biochronology only (e.g., here: 'This formation extends from the base of the Pleistocene to as young as 1 Ma; Biochronological assessments indicate Grăunceanu is Late Villafranchian (~2.2-1.9 Ma) and is attributed to mammalian biostratigraphic zones MN17/MmQ19-11). For example, in absence of numerical age results (and even though, the existing age errors are usually not negligible) or magnetostratigraphic evidence indicating the undisputed presence of Jaramillo, the minimum age constraint of 1 Ma proposed here is questionable. Let's say that an Early Pleistocene chronology may roughly be estimated for the Tetoiu Fm based on biostratigraphic evidence. Following a similar reasoning, there are many examples of localities that have been equally attributed to either the Middle, Late or Epivillafranchian biochrons depending on the authors. For example, I note that on the figure given in the rebuttal, the chronological window proposed for the site covers the latest Middle Villafranchian and earliest Late Villafranchian, whereas the text mentions a much finer age window ('Late Villafranchian (~2.2-1.9 Ma)'). Anyway, my point here is that I feel that the general chronostratigraphic context of the ORV and Dacian basin (and mainly the Tetoiu Fm) would deserve a dedicated section in Supplementary Material, as biostratigraphy provides a crucial support to the U-Pb dating results. This would help the reader unfamiliar with the area getting the bigger picture. In particular, it is unclear how the various sites mentioned in the text do relate (e.g., La Pietriș, Valea Roșcăi, Fântâna lui Mitilan). It is mentioned that some of them are coeval, while others are younger/older, but no data are provided to support this. I understand that this information might be found in various other papers, but it would be beneficial to the present work if a synthesis could be given in Supplementary Material. In particular the figure of the rebuttal showing the biochronological evaluation of the Grăunceanu should probably be added. It does provide useful information. The other localities from the ORV are that are mentioned in the text could be added to that figure too.

--We agree completely that biochronological estimates of ages for fossil sites are challenging, and we had the same concerns that the reviewer outlines here when we first described the biochronology of these localities in our original publication (Terhune et al., 2020). This is in fact what spurred us on to seek the radiometric dates presented in this paper. Happily, our U-Pb dates and all prior biochronological estimates, both by our team and previous researchers, are roughly consistent with

one another, and they are consistent with previous descriptions of the faunal horizons of the ORV (which include the sites discussed here) as described by Radulescu and Samson (1990). We agree with the reviewer that more clarity of the positions and relationships of the sites and faunal horizons is valuable and we have added a section in this regard to the Supplemental Materials. However, we have decided not to include a detailed discussion of the biochronology of the fauna from the ORV sites, as it is our opinion that this would replicate previous work and take away from the focus of the paper, which is on the U-Pb dating and cut marks. We do agree with the reviewer that more precision regarding the upper limit of the Tetoiu Formation is necessary, and have also updated this in the main text and supplemental to 1.3 Ma to be consistent with descriptions by Andreescu et al.

Finally, I would like to add that, in my view, there is still an inherent general uncertainty on the context of the site and of the fossil assemblage. I do get the explanations of the authors on this matter, and I accept them, but the fact that the site is no longer accessible, and that most of the information related to the excavation is gone, naturally raises some questions on the integrity, extension and stratigraphy of the fossil assemblage. This is complicated by the fact that various successive campaigns have been performed over more than a decade, and we ignore how they relate. Moreover, dating analyses were not performed on the same bones that have the cut marks, or on the equid teeth used for palaeoenvironment reconstruction, so there is also an inherent uncertainty on whether all these remains are truly associated, i.e., come from the same excavation area and the same stratigraphic unit. In that regard, would it be possible to at least know from which excavation campaigns were taken all the specimens analysed here for cut marks, dating and paleoenvironment analyses? That said, I get the authors' explanations and especially regarding the taphonomy suggesting a single depositional event, and no doubt they did their best to narrow down this uncertainty, but this is still indirect evidence that cannot fully compensate for the absence of a proper stratigraphy and spatial location of the remains.

--We appreciate the reviewer's comments here and understand where this confusion comes from. All of the materials examined for this study from Graunceanu come from the original excavations that took place in the 1960s. Subsequent investigations have not revealed additional faunal accumulations at Graunceanu, since it seems likely that the entirety of the bone bed was exhausted in the original excavations. Our taphonomic analyses (as well as original images and descriptions of these excavations) strongly suggest that the faunal assemblage from Graunceanu stems from a single depositional event. We very much wish we had a clearer picture of the internal stratigraphy of the site, but in the absence of that we still have high confidence in the depositional integrity of the locality. This is further supported by a continuous chain of custody of these remains from the original excavators; one of those original excavators, Constantin Radulescu, was the PhD supervisor of Alex Petculescu, who is a coauthor on this study and currently manages the Bucharest collection. In the known history of the collection there is no mention of anything other than the homogeneous site that was excavated in the 1960s. Though the collection at present is split between two museums, it is very clear from markings on the bones that they were excavated at the same time and then subsequently split up in the intervening decades (no one seems to know why this split occurred).

--We agree it would have been ideal to directly date the cut marked specimens; this certainly would give us even higher confidence in the dates of any anthropogenic activity. However, methodologically it was not possible for this to happen, since the U-Pb dating requires dental samples and the vast majority of the specimens exhibiting cut marks are from the postcrania. The one mandibular specimen with cut marks (from Fantana lui Mitilan) is a nearly complete cervid mandible and dentition which is important both taxonomically and taphonomically, and therefore in our opinion should not be destructively sampled. While we did destructively sample the horse dentition for isotopic study, this analysis was conducted after the U-series dating campaign was concluded.

--We have edited the methods in response to both of these comments and hope these edits sufficiently alleviate the reviewer's concerns.

Specific comments

1. L.74-75: The chronostratigraphy of the Tetoiu Fm described from Andreescu et al. (2013) is not as robust and precise as suggested here. There is no independent age control available, and if I am not wrong the age constraints are only based on biostratigraphy. I'd rather see something like: 'An Early Pleistocene age may be considered for this Fm based on biostratigraphic evidence.'. I did not see evidence in Andreescu et al. (2013) for such a precise minimum age boundary of 1 Ma.

--see comments above.

2. L.104-106: Please indicate that this relative chronology is based on biostratigraphic evidence.

--done

3. Fig. 1: please check the spelling of 'El-Kherba'.

--Thanks for catching this error. The figure has been updated.

4. L.196-198: I understand these ages only refer to Grăunceanu. This should probably be clarified.

--Done

5. L.199, 'cluster tightly': maybe a number (e.g., coefficient of variation, or a time range between the youngest and oldest results) can be provided to support this statement.

--A time range between the oldest and the youngest dates are already provided in the preceding sentence; we feel that adding anything else here might be more confusing than helpful.

6. Fig. 3: please indicate in caption that age errors are 2σ .

--Done

7. L. 231: VSMOW should probably be spelled out for its first occurrence in the text.

--Done

8. Table 1 and Supplementary Table S7: The various ages reported there are quite confusing, maybe this should be clarified in caption. Sometimes an age range is given, with either 1 or 2 decimal digits. Not sure a 2 decimal digit precision is appropriate for most sites (Fântâna lui Mitilan even shows 3 digits). It is also unclear whether the age range is taken from the errors on the numerical ages, from the combination of various ages, or simply correspond to various levels from the same site.

--Thank you for this comment. It is very difficult to collate data from so many different localities and ensure that the information presented is comparable. We have done our best in this regard but feel it is important to leave the ages of the sites as close to how they are reported in the original publications as possible. This means that some dates are reported with more precision than others, and others provide age ranges instead of specific numeric dates. We have tried to clarify this in the caption for Supplemental Table 7. We have changed the date reported for Fantana lui Mitilan to more closely reflect what is presented in our main text, however.

- Not sure to understand how the age range of 2.3-2.0 Ma for Grăunceanu was obtained since the various U-Pb ages range from 1.87 to 2.01 Ma

--Thank you for catching this error. We have edited this to now say >1.95 Ma.

- Ain Hanech and El Kherba (Algeria) are usually regarded as coeval. Please see Duval et al. (2023) for the most recent age estimation.

--Thank you for pointing us to this publication. We have updated the ages for both sites in our tables.

- Ain Boucherit (Algeria) is made of two levels, the lower level (Lw) dated to ~2.4 Ma and the upper level (Up) to ~1.9 Ma. Maybe these 2 different levels (usually referred as Ain Boucherit Lw and Ain Boucherit Up) should be differentiated, otherwise the age range given may suggest a significant uncertainty on the age of the site, which is not the case.

--We have separated these two layers in both the table in the main text and Supplemental tables 7 and 8.

- Pirro Nord: please note the most recent dating study on that site (Duval et al., 2024). Although a biochronological age of 1.3-1.6 Ma cannot be excluded, there is strong evidence suggesting that the site may be much younger.

--Thank you for pointing us to this publication. We have retained the original dates of 1.3-1.6 Ma, but have also included a reference to this paper in the supplemental table.

- Sima del Elefante : maybe indicate the layers, e.g., Sima del Elefante (TE7-9). Methods also include BS.

--We have opted not to include the level for this site since we generally do not provide level information for other sites (or at least not those with multiple distinct levels for which data are provided). We have added BS as one of the dating methods.

- Fuente Nueva-3: a rounded age of '1.2 Ma' would probably be more appropriate given the associated uncertainty.

--While we agree there is likely uncertainty here, we retain 1.19 Ma as the age since this is what is reported in the literature (see our comment above regarding the presentation of the age data for the fossil sites).

- Barranco Leon: Methods also include MS and BS, in addition to ESR.

--added

9. L.324-330: for a matter of consistency, please use the same number of decimal digits for all percentages.

--fixed

10. L.326: Do the 3.32% include both Lower and Upper levels at Ain Boucherit? Please clarify.

--That percentage did include both levels but we have now separated them out

11. L.335: maybe you can refer to Supplementary Table S8 here.

--added

12. Ref 71 in SM: please note the name of the journal is 'L'Anthropologie'.

--fixed

13. Supplementary Table S8: please indicate in caption the meaning of 'Full Taph?'. Note that the age reported for Grăunceanu is not consistent with that in Table S7. Same comment applies for Dmanisi.

--dates for Graunceanu and Dmanisi have been updated; thank you for catching those inconsistencies. Full Taph refers to whether a full taphonomic assessment of the entire assemblage for that site has been conducted and published. We have clarified that in the table.

Response to Reviewer 2 (provided below original comment text)

Again it is a pleasure to be involved in the reviewing process of this paper, however I have rewritten this review a number of times to make an effort to be as clear as possible. While some of the comments made in response to my last review are debatable, they are not necessarily the focus of the study, therefore I will leave these points for now, and this can hopefully lead to interesting debates and lines of research in the future. Nevertheless, there is one point that has still not been corrected to a sufficient extent, and is arguably one of the most important points I have made. Without properly responding to this point, I do not feel comfortable accepting the metric analysis of the taphonomic marks in their current state. Once again I have no problem in saying that the marks are cut marks, from a traditional perspective, but my issue is that when using quantitative approaches, a certain number of statistical and mathematical requirements have to be met in order for the results to be considered robust, reliable, and not a misrepresentation of spurious separations between samples.

In my first and second review, I stated very clearly that it is fundamental that QDA results are cross-validated. This is important for both the evaluation of the algorithms used for classification, and can wildly change the quality of the ordination plots. When I brought it up again in the second round of reviews (and I provided a number of references to prove the importance of my point), the authors state that it has already been included in their manuscript. JMP is a private licenced software, so I have close to no way of testing some of the main issues, however, I have been able to obtain access to the non-pro version of JMP for a limited time in order to run some tests. Likewise, I have read in detail the documentation of the software to try and find the information I requested. Finally, in this review I will be providing R code to help clarify not only my points, but also provide a potential means of fixing some of the issues. I hope that the authors can see that the fact that I have taken the time to provide an R script is not only for the purpose of proving the importance of cross-validation, but in the hope that the authors may also find this code useful. If the authors wish to use their own data, then I have provided comments within the code to clearly indicate where their data needs to be inserted.

The authors state that in their original manuscript that they “clearly reported ... that a 25% validation set was extracted from the experimental sample of BSM to validate the model. We now call this a 25% cross-validation set for clarity”. I respectfully disagree, and I am not sure we are talking about the same thing here. First, there are many types of cross-validation algorithms, which one are you using (if you are using one at all)? In the case of classification, the most recommended form of performing cross-validation on a classification algorithm is k-fold cross-validation, with $k = 5$, or more commonly $k = 10$. Did the authors do this? Next, there is a distinct difference between a test set and a validation set, which is where I think some source of conclusion lies. Finally, how can 25% of the data be used to cross-validate the plot? In general cross-validation of the discriminant function plots is slightly different.

While a number of authors can be found throughout literature using the terms validation and test sets for different things, in general, the standard approach is to refer to the validation set as the data used by the algorithm itself to define the discriminant function, and the test set is what we separate and then use to evaluate how well this discriminant function is actually working (I think this is what you refer to as validation, given that JMP uses this term in the Analysis/Predictive Modelling/Make Validation Column menu). I suspect that the authors are calling the validation set what I (and many) would actually call the test set. This would imply that they are still not necessarily performing cross-validation, which is why the simple statement “25% cross-validation” does not

actually or fully clarify anything.

Through playing with JMP, and the additional code I provide in R, I am pretty certain that the authors are not performing the cross-validation required. JMP appears to be able to do everything I will describe, however in the non-pro version of the software I cannot actually test this fully. Nevertheless, the documentation clearly states that K-fold cross validation is possible (<https://www.jmp.com/support/help/en/18.0/index.shtml#page/jmp/kfold-and-holdback-validation.shtml#>), which goes beyond simply separating a test set to evaluate model performance afterwards. This however requires a number of additional steps to be performed, however (see previous link).

From what I can tell by the table on this website, the algorithm is not doing k-fold cross validation automatically, I can only assume that using a number of other functions they can get around this. Otherwise the word “No” under “K-fold cross-validation” and “Discriminant Models” does not bode well for the use of this software for classification tasks using this algorithm. Likewise, I still do not see clearly whether the cross-validation approach will be taken into account when computing the discriminant scores as well. This needs to be checked and clarified if the authors are adamant about including a CV plot in their supplementary materials (which I have already suggested avoiding all together).

To validate, and additionally clarify the points I am trying to make, I will describe the experiments I have carried out in both R and JMP in continuation. If the authors are able to test that k-fold cross-validation is possible in JMP, and does indeed correct the plots, then this would be a good means of solving all issues I am about to describe.

In the first example, I took the dataset from <https://www.kaggle.com/datasets/muthuj7/weather-dataset?resource=download>. This is a known benchmark dataset, and in R I can show that a cross-validated QDA is able to reach a 94% accuracy. I’d like to strongly point out however that (as I already said in a previous review about how accuracy is a misleading metric), the QDA algorithm is unable to correctly classify any of the “null” class from this dataset, which leads to a kappa statistic of 0.65, i.e. the algorithm is not sufficient to correctly learn from this data. If I perform the same analysis in JMP without performing k-fold cross-validation the results are completely different. The accuracy now drops to 45.9%, with a kappa statistic of 0.124. I think this is already evidence enough to show that cross-validation is not only a valuable means of boosting model performance, enhancing discriminant function robustness, but how misleading accuracy results can be if cross-validation is not performed. Confusion matrices of both these calculations are included as Figures 1 and 2 in the file I provide alongside this review.

To provide sample sizes that are a little more analogous with archaeological datasets, however, I then took a subset of 30 individuals per sample from this dataset. I used a seed in my R code to ensure replicability of results, considering how random subsampling is performed. With this seed, you will see that one of the observations included in the PCA (Fig. 3) is a potential outlier (it is in fact not an outlier, as it does belong to the original distribution perfectly, a pretty cool example of how sampling strategies can make single extreme observations appear like outliers when they are in fact not). In a cross-validated canonical variates plot (Fig. 4) the point is forced closer to the rest of the sample, however, a cross-validated plot (Fig. 5) shows that the point is still an outlier, which is much truer to the actual nature of the distribution. Either way, the two plots are different, showing that cross-validation will produce a difference in plots and may force certain patterns to emerge

that may not be an accurate representation of the underlying data. To show how this effects cross-validation performance for classification, if we do not do cross-validation, JMP calculates an accuracy of 56.5%, with a Kappa of 0.347, while the cross-validated version of the algorithm produces an accuracy of 56.5% and a Kappa of 0.3466. While the differences are minimal, they are there.

In the second example I increase the complexity of the dataset by adding 100 new dimensions of noise to the dataset, simply to provoke the 'curse of dimensionality' to occur, which again will lead to the spurious separation of samples mentioned by other authors. Once again note that without cross-validation, the plots are not comparable at all, and the cross-validated plot is much truer to the original distribution as shown in the PCA. Likewise, the outlier from the first experiment is once again still identified as such. In the non-cross-validated approach using JMP, the accuracy comes out at 43.5% but has a surprising kappa value of -0.049, indicating that the discriminant function is not a good classifier at all. R, on the other hand, has the courtesy of refusing to perform the analysis, as the sample size and complexity does not allow for the computation of a reliable discriminant function.

In the final example, I perform these experiments on purely synthetic data, that are designed to have 0 differences between the three samples. Once again, spurious separation occurs on non-cross-validated results, but does not appear on the cross-validated results. This is proof of the theorem that non cross-validated canonical variance plots can be incredibly difficult to interpret and are not always a necessarily reliable representation of the data. A non-cross-validated QDA in JMP reveals an accuracy of 47%, however once again R refuses to perform the computation. To prove the fact that simply separating 75% of the data for the calculation of the discriminant function does not imply that cross-validation has been done, figures X, X, and X additionally show how using the final example, when using the validation set and when excluding the validation set, spurious separation occurs regardless.

It is important to point out that in some of the above cases, only limited differences are seen between the results of k-fold cross-validated algorithms, and discriminant functions fit without performing cross-validation. Regardless, the fact that even slight differences are there are important. In my previous reviews I provided a number of references to prove this. Seeing how this was insufficient in stressing my point, here I have provided code, and statistical proof that this is still an issue. Interestingly, R in a number of the above cases refused to calculate, or was unable to calculate the discriminant function, while JMP succeeded. Nevertheless, I refer the authors to JMP's documentation, where it clearly states that in some cases QDA can be problematic (<https://www.jmp.com/support/help/en/18.0/index.shtml#page/jmp/discriminant-method.shtml#>). The authors need to ensure that they report robust and reliable quantitative results, and even if repeating the statistics with a correct implementation of cross-validation produces the same results, at least then we can be sure that your results are accurate.

In sum, for Sup. Fig. 22 please either; (1) cross-validate the calculation of the CV scores in the plot, or (2) completely delete this figure. In my opinion, option 2 is the easiest approach and does not put into question the reliability of the study. For classification results, I insist that the authors must use cross-validation, preferably k-fold cross-validation with $k = 10$, as is standard procedure in most analyses of this type. This latter point is the most important and needs to be rectified, or at least clarified.

Signed – Lloyd A. Courtenay

PS - I may be mistaken but I also recall an Excel or csv file being included in an earlier version of this manuscript, which is now no longer available, if this was data it would have been nice to have access to this data again to run the same tests on your actual data.

--We appreciate the reviewer's continued support of our manuscript and their thoughtful comments on the statistical methods. It is clear from their comments that they feel strongly about these methods, and we have done our best to address their concerns in as thorough a manner as possible. We agree that some of the confusion here may have been related to differences in terminology, and we agree that a robust cross-validation of the model is critical for ensuring proper interpretation of the results.

To that end, we have revised the text to make our analyses as clear as possible, and now include an additional author who performed updated analyses. As per the reviewer's suggestion we now present both leave-one-out and 10-fold cross-validation of the QDA results. We have added a new Supplementary Table 3 that provides the confusion matrices, and have added the R code to run these analyses in the supplementary information (though note this requires the raw experimental data, which are not yet published). Both cross-validation methods return very similar results (81.4% and 82%, respectively), demonstrating the effectiveness of the QDA. In the course of these updates we noted a small error (n=3) in the counts of bone surface modifications in the experimental dataset and have subsequently updated Supplementary Figures 22 and 23, Supplementary Table 2, and all posterior probabilities listed in the text. These changes are all very minor and do not change the outcomes of our interpretations; in fact, they strengthen them.

All changes made have either been tracked in the text or sections with major changes are highlighted in yellow. Again, we appreciate the reviewer's attention to our statistics and support of our work, and we hope these updates satisfy their concerns regarding the quantitative assessments of the bone surface modifications we present here.

Reviewer 2 attachments:

```
# Example code for the cross-validation of discriminant plots and the
training of classification algorithms
# code by L.A.Courtenay, University of Bordeaux - [20 / 09 / 2024] to [24 /
09 / 2024]
# for queries or help contact ladc1995@gmail.com

# Raw Example of the Weather History Dataset -----

library(caret) # for cross-validation algorithms
library(MASS) # for QDA functions
library(Morpho) # the only package i know that offers cross-validated
canonical variate plots
library(ggplot2) # to visualise results

set.seed(123456789)

# !!!! Load here your own data, here i use a common known dataset just for
an example !!!!
# the dataset i used can be downloaded from
https://www.kaggle.com/datasets/muthuj7/weather-dataset?resource=download

dataset <- read.table("weatherHistory.csv", head = TRUE, sep = ",")
dataset <- dataset[,c(3, 4, 6, 7, 9, 11)] # remove some of the useless
variables that are not relevant for classification

# !!!! Load here your own data, here i use a common known dataset just for
an example !!!!

# extract numeric values and labels

numeric_values <- dataset[,2:6] # extract only numeric variables
labels <- as.factor(dataset[,1]) # extract the labels

# perform a PCA

numeric_values_prima <- scale(numeric_values)
pca_results <- prcomp(numeric_values_prima)

variances <- pca_results$sdev ^ 2
explained_variance <- variances / sum(variances)

pc1_variance <- explained_variance[1] * 100
pc2_variance <- explained_variance[2] * 100

plot_data <- data.frame(
  pc1 = pca_results$x[,1],
  pc2 = pca_results$x[,2],
  Sample = labels
)

ggplot(data = plot_data,
       aes(x = pc1, y = pc2, colour = Sample)) +
  geom_point(size = 1) +
  stat_ellipse(size = 1) +
```

```

xlab(paste("PC1  (", round(pc1_variance, 2), "%)")) +
ylab(paste("PC2  (", round(pc2_variance, 2), "%)")) +
theme_bw() +
coord_fixed()

# because the dataset is huge, CVA will take an incredibly long time to
compute,
# therefore for this example on the raw data I have refrained from doing so
and
# go straight onto calculating the benchmark accuracy using cross-validated
QDA

# seperating test data for evaluation

test_split <- 0.75

train_x <- numeric_values[1:(ceiling(nrow(dataset) * test_split)),]
test_x <- numeric_values[ceiling(nrow(dataset) *
test_split):nrow(dataset),]

train_y <- as.factor(labels[1:(ceiling(nrow(dataset) * test_split))])
test_y <- as.factor(labels[ceiling(nrow(dataset) *
test_split):nrow(dataset)])

train_data <- data.frame(train_x, Sample = train_y)
test_data <- data.frame(test_x, Sample = test_y)

# save the data so that it can be later imported into JMP

write.table(
  rbind(data.frame(train_data, Validation = "Train"),
        data.frame(test_data, Validation = "Test")
  ),
  "Full_JMP_data_for_QDA_experiments.csv",
  col.names = TRUE, row.names = FALSE, sep = ",",
)

# cross-validated QDA model

cross_validation_algorithm <- trainControl(method = "cv", number = 10)
QDAmodel_CV <- train(
  Sample ~ .,
  data = train_data,
  method = "qda",
  trControl = cross_validation_algorithm
)

# evaluate model

test_predictions <- predict(QDAmodel_CV, test_data)
confusionMatrix(test_predictions, test_y)

#

# Example 1 - Weather History dataset -----

```

```

library(caret) # for cross-validation algorithms
library(MASS) # for QDA functions
library(Morpho) # the only package i know that offers cross-validated
canonical variate plots
library(ggplot2) # to visualise results

set.seed(123456789)

# !!!! Load here your own data, here i use a common known dataset just for
an example !!!!
# the dataset i used can be downloaded from
https://www.kaggle.com/datasets/muthuj7/weather-dataset?resource=download

dataset <- read.table("weatherHistory.csv", head = TRUE, sep = ",")
dataset <- dataset[,c(3, 4, 6, 7, 9, 11)] # remove some of the useless
variables that are not relevant for classification

# because the dataset is huge i will also subsample 30 random examples from
each of the classes

split_data <- split(dataset, as.factor(dataset[,1]))
subsampled_data <- lapply(split_data, function(subset) {
  subset[sample(nrow(subset), 30, replace = FALSE),]
})
dataset <- do.call(rbind, subsampled_data)

# !!!! Load here your own data, here i use a common known dataset just for
an example !!!!

# shuffle data

dataset <- dataset[sample(nrow(dataset), replace = FALSE),]

# extract numeric values and labels

numeric_values <- dataset[,2:6] # extract only numeric variables
labels <- as.factor(dataset[,1]) # extract the labels

# perform a PCA

numeric_values_prima <- scale(numeric_values)
pca_results <- prcomp(numeric_values_prima)

variances <- pca_results$sdev ^ 2
explained_variance <- variances / sum(variances)

pc1_variance <- explained_variance[1] * 100
pc2_variance <- explained_variance[2] * 100

plot_data <- data.frame(
  pc1 = pca_results$x[,1],
  pc2 = pca_results$x[,2],
  Sample = labels
)

ggplot(data = plot_data,

```

```

    aes(x = pc1, y = pc2, colour = Sample)) +
  geom_point(size = 4) +
  stat_ellipse(size = 1) +
  xlab(paste("PC1 (", round(pc1_variance, 2), "%)")) +
  ylab(paste("PC2 (", round(pc2_variance, 2), "%)")) +
  theme_bw() +
  coord_fixed()

# non-cross-validated Canonical Variates plot

cva_results <- CVA(numeric_values, labels, cv = FALSE)

cv1_variance <- cva_results$Var[1,2][[1]]
cv2_variance <- cva_results$Var[2,2][[1]]

plot_data <- data.frame(
  cv1 = cva_results$CVscores[,1],
  cv2 = cva_results$CVscores[,2],
  Sample = labels
)

ggplot(data = plot_data,
       aes(x = cv1, y = cv2, colour = Sample)) +
  geom_point(size = 4) +
  stat_ellipse(size = 1) +
  xlab(paste("CV1 (", round(cv1_variance, 2), "%)")) +
  ylab(paste("CV2 (", round(cv2_variance, 2), "%)")) +
  theme_bw() +
  coord_fixed()

# cross-validated Canoncial Variates plot

cva_results_CV <- CVA(numeric_values, as.factor(labels), cv = TRUE)

cv1_variance <- cva_results_CV$Var[1,2][[1]]
cv2_variance <- cva_results_CV$Var[2,2][[1]]

plot_data <- data.frame(
  cv1 = cva_results_CV$CVcv[,1],
  cv2 = cva_results_CV$CVcv[,2],
  Sample = labels
)

ggplot(data = plot_data,
       aes(x = cv1, y = cv2, colour = Sample)) +
  geom_point(size = 4) +
  stat_ellipse(size = 1) +
  xlab(paste("Cross Validated CV1 (", round(cv1_variance, 2), "%)")) +
  ylab(paste("Cross Validated CV2 (", round(cv2_variance, 2), "%)")) +
  theme_bw() +
  coord_fixed()

# classification using QDA

# seperating test data for evaluation

```

```

test_split <- 0.75

train_x <- numeric_values[1:(ceiling(nrow(dataset) * test_split)),]
test_x <- numeric_values[ceiling(nrow(dataset) *
test_split):nrow(dataset),]

train_y <- as.factor(labels[1:(ceiling(nrow(dataset) * test_split))])
test_y <- as.factor(labels[ceiling(nrow(dataset) *
test_split):nrow(dataset)])

train_data <- data.frame(train_x, Sample = train_y)
test_data <- data.frame(test_x, Sample = test_y)

# save the data so that it can be later imported into JMP

write.table(
  rbind(data.frame(train_data, Validation = "Train"),
        data.frame(test_data, Validation = "Test")
  ),
  "JMP_data_for_QDA_experiments.csv",
  col.names = TRUE, row.names = FALSE, sep = ",",
)

# cross-validated QDA model

cross_validation_algorithm <- trainControl(method = "cv", number = 10)
QDAmodel_CV <- train(
  Sample ~ .,
  data = train_data,
  method = "qda",
  trControl = cross_validation_algorithm
)

# evaluate model

test_predictions <- predict(QDAmodel_CV, test_data)
confusionMatrix(test_predictions, test_y)

#

# Example 2 - Weather History dataset -----

library(caret) # for cross-validation algorithms
library(MASS) # for QDA functions
library(Morpho) # the only package i know that offers cross-validated
canonical variate plots
library(ggplot2) # to visualise results

set.seed(12345678)

# !!!! Load here your own data, here i use a common known dataset just for
an example !!!!
# the dataset i used can be downloaded from
https://www.kaggle.com/datasets/muthuj7/weather-dataset?resource=download

dataset <- read.table("weatherHistory.csv", head = TRUE, sep = ",")

```

```

dataset <- dataset[,c(3, 4, 6, 7, 9, 11)] # remove some of the useless
variables that are not relevant for classification

# because the dataset is huge i will also subsample 30 random examples from
each of the classes

split_data <- split(dataset, as.factor(dataset[,1]))
subsampled_data <- lapply(split_data, function(subset) {
  subset[sample(nrow(subset), 30, replace = FALSE),]
})
dataset <- do.call(rbind, subsampled_data)

# add some artificial noise to make the dataset more interesting at higher
dimensions

for (i in 1:100) {

  dataset <- cbind(dataset, rnorm(nrow(dataset), runif(1, 0, 10), runif(1,
0, 10)))

}

# !!!! Load here your own data, here i use a common known dataset just for
an example !!!!

# shuffle data

dataset <- dataset[sample(nrow(dataset), replace = FALSE),]

# extract numeric values and labels

numeric_values <- dataset[,2:ncol(dataset)] # extract only numeric
variables
labels <- as.factor(dataset[,1]) # extract the labels

# perform a PCA

numeric_values_prima <- scale(numeric_values)
pca_results <- prcomp(numeric_values_prima)

variances <- pca_results$sdev ^ 2
explained_variance <- variances / sum(variances)

pc1_variance <- explained_variance[1] * 100
pc2_variance <- explained_variance[2] * 100

plot_data <- data.frame(
  pc1 = pca_results$x[,1],
  pc2 = pca_results$x[,2],
  Sample = labels
)

ggplot(data = plot_data,
  aes(x = pc1, y = pc2, colour = Sample)) +
  geom_point(size = 4) +
  stat_ellipse(size = 1) +

```

```

xlab(paste("PC1  (", round(pcl_variance, 2), "%)") +
ylab(paste("PC2  (", round(pc2_variance, 2), "%)") +
theme_bw() +
coord_fixed()

# non-cross-validated Canonical Variates plot

cva_results <- CVA(numeric_values, labels, cv = FALSE)

cv1_variance <- cva_results$Var[1,2][[1]]
cv2_variance <- cva_results$Var[2,2][[1]]

plot_data <- data.frame(
  cv1 = cva_results$CVscores[,1],
  cv2 = cva_results$CVscores[,2],
  Sample = labels
)

ggplot(data = plot_data,
  aes(x = cv1, y = cv2, colour = Sample)) +
  geom_point(size = 4) +
  stat_ellipse(size = 1) +
  xlab(paste("CV1  (", round(cv1_variance, 2), "%)") +
  ylab(paste("CV2  (", round(cv2_variance, 2), "%)") +
  theme_bw() +
  coord_fixed()

# cross-validated Canoncial Variates plot

cva_results_CV <- CVA(numeric_values, as.factor(labels), cv = TRUE)

cv1_variance <- cva_results_CV$Var[1,2][[1]]
cv2_variance <- cva_results_CV$Var[2,2][[1]]

plot_data <- data.frame(
  cv1 = cva_results_CV$CVcv[,1],
  cv2 = cva_results_CV$CVcv[,2],
  Sample = labels
)

ggplot(data = plot_data,
  aes(x = cv1, y = cv2, colour = Sample)) +
  geom_point(size = 4) +
  stat_ellipse(size = 1) +
  xlab(paste("Cross Validated CV1  (", round(cv1_variance, 2), "%)") +
  ylab(paste("Cross Validated CV2  (", round(cv2_variance, 2), "%)") +
  theme_bw() +
  coord_fixed()

#

# seperating test data for evaluation

test_split <- 0.75

train_x <- numeric_values[1:(ceiling(nrow(dataset) * test_split)),]

```

```

test_x <- numeric_values[ceiling(nrow(dataset) *
test_split):nrow(dataset),]

train_y <- as.factor(labels[1:(ceiling(nrow(dataset) * test_split))])
test_y <- as.factor(labels[ceiling(nrow(dataset) *
test_split):nrow(dataset)])

train_data <- data.frame(train_x, Sample = train_y)
test_data <- data.frame(test_x, Sample = test_y)

# save the data so that it can be later imported into JMP

write.table(
  rbind(data.frame(train_data, Validation = "Train"),
        data.frame(test_data, Validation = "Test")
  ),
  "JMP_noisy_data_for_QDA_experiments.csv",
  col.names = TRUE, row.names = FALSE, sep = ",",
)

# cross-validated QDA model

cross_validation_algorithm <- trainControl(method = "repeatedcv", repeats =
10)
QDAmodel_CV <- train(
  Sample ~ .,
  data = train_data,
  method = "qda",
  trControl = cross_validation_algorithm
)

# note that the algorithm gives an error because it is unable to compute a
discriminant function

#

# Synthetic Data -----

library(caret) # for cross-validation algorithms
library(MASS) # for QDA functions
library(Morpho) # the only package i know that offers cross-validated
canonical variate plots
library(ggplot2) # to visualise results
library(abind) # for tensor manipulation operations

set.seed(123456789)

# !!!! Load here your own data, here i use a common known dataset just for
an example !!!!

# for this example lets use purely synthetic data

dataset <- array(numeric(), dim = c(90, 0))

for (i in 1:100) {

```

```

dataset <- abind(dataset, rnorm(nrow(dataset), runif(1, 0, 10), runif(1,
0, 10)),
                along = 2)

}
dataset <- data.frame(labels = c(rep(1, 30), rep(2, 30), rep(3, 30)),
dataset)

# !!!! Load here your own data, here i use a common known dataset just for
an example !!!!!

# shuffle data

dataset <- dataset[sample(nrow(dataset), replace = FALSE),]

# extract numeric values and labels

numeric_values <- dataset[,2:ncol(dataset)] # extract only numeric
variables
labels <- as.factor(dataset[,1]) # extract the labels

# perform a PCA

numeric_values_prima <- scale(numeric_values)
pca_results <- prcomp(numeric_values_prima)

variances <- pca_results$sdev ^ 2
explained_variance <- variances / sum(variances)

pc1_variance <- explained_variance[1] * 100
pc2_variance <- explained_variance[2] * 100

plot_data <- data.frame(
  pc1 = pca_results$x[,1],
  pc2 = pca_results$x[,2],
  Sample = labels
)

ggplot(data = plot_data,
       aes(x = pc1, y = pc2, colour = Sample)) +
  geom_point(size = 4) +
  stat_ellipse(size = 1) +
  xlab(paste("PC1 (", round(pc1_variance, 2), "%)")) +
  ylab(paste("PC2 (", round(pc2_variance, 2), "%)")) +
  theme_bw() +
  coord_fixed()

# non-cross-validated Canonical Variates plot

cva_results <- CVA(numeric_values, labels, cv = FALSE)

cv1_variance <- cva_results$Var[1,2][[1]]
cv2_variance <- cva_results$Var[2,2][[1]]

plot_data <- data.frame(
  cv1 = cva_results$CVscores[,1],

```

```

    cv2 = cva_results$CVscores[,2],
    Sample = labels
)

ggplot(data = plot_data,
       aes(x = cv1, y = cv2, colour = Sample)) +
  geom_point(size = 4) +
  stat_ellipse(size = 1) +
  xlab(paste("CV1 (", round(cv1_variance, 2), "%)")) +
  ylab(paste("CV2 (", round(cv2_variance, 2), "%)")) +
  theme_bw() +
  coord_fixed()

# cross-validated Canoncial Variates plot

cva_results_CV <- CVA(numeric_values, as.factor(labels), cv = TRUE)

cv1_variance <- cva_results_CV$Var[1,2][[1]]
cv2_variance <- cva_results_CV$Var[2,2][[1]]

plot_data <- data.frame(
  cv1 = cva_results_CV$CVcv[,1],
  cv2 = cva_results_CV$CVcv[,2],
  Sample = labels
)

ggplot(data = plot_data,
       aes(x = cv1, y = cv2, colour = Sample)) +
  geom_point(size = 4) +
  stat_ellipse(size = 1) +
  xlab(paste("Cross Validated CV1 (", round(cv1_variance, 2), "%)")) +
  ylab(paste("Cross Validated CV2 (", round(cv2_variance, 2), "%)")) +
  theme_bw() +
  coord_fixed()

#

# seperating test data for evaluation

test_split <- 0.75

train_x <- numeric_values[1:(ceiling(nrow(dataset) * test_split)),]
test_x <- numeric_values[ceiling(nrow(dataset) *
test_split):nrow(dataset),]

train_y <- as.factor(labels[1:(ceiling(nrow(dataset) * test_split))])
test_y <- as.factor(labels[ceiling(nrow(dataset) *
test_split):nrow(dataset)])

train_data <- data.frame(train_x, Sample = train_y)
test_data <- data.frame(test_x, Sample = test_y)

# export for JMP

write.table(
  rbind(data.frame(train_data, Validation = "Train"),

```

```

    data.frame(test_data, Validation = "Test")
  ),
  "JMP_noisy_data_for_QDA_experiments.csv",
  col.names = TRUE, row.names = FALSE, sep = ",",
)

# note that R will not calculate the QDA for this data, therefore has been
excluded here

# Classifying archaeological samples -----

# if the user wishes to calculate the labels for new specimens, such as the
archaeological traces
# you need to load the data using the read.table function, or have them
separated at the begining
# then use the predict() function, e.g. predict(QDAmodel,
archaeological_data),
# to predict the new labels of each of the observations.
# Prediction probabilities can be obtained by including the parameter type
= "prob"
# in the predict() function, e.g. predict(QDAmodel, archaeological_data,
type = "prob")

# for ROC curves, you can use the pROC or ROCR packages, however ROC curves
are only
# useful for comparing binary classification tasks. For categorical
classification,
# you would need to calculate a curve for each of the pairs of labels. the
roc()
# function can be used to calculate the curve, and the auc() funciton will
then calculate
# the area under the ROC curve. If sample imbalance is present, however,
please note
# that ROC may not be the best approach for model evaluation. use help(roc)
to
# consult R's official documentation for this function.

```

Figure 1 – A PCA of the dataset used for some of the experiments described in this review. Note that in a previous review the authors stated that because they do not use GMM data the comments made about the reliability of cross-validation were not applicable to their own case study. In this experiment I have not used, in any way, shape, or form, morphological data. The variables included are variables relating to the weather, humidity, and speed of the wind.

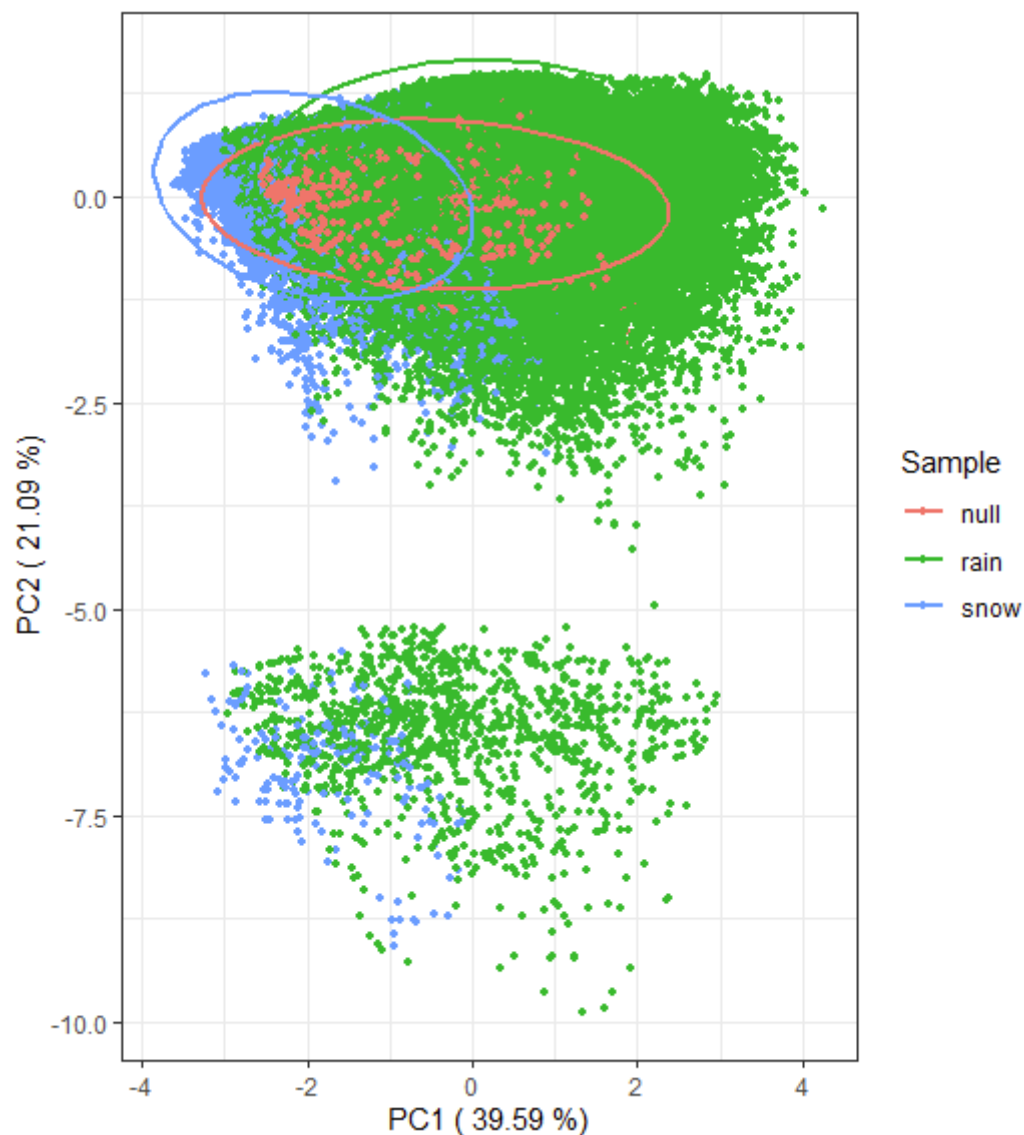


Figure 1 – A screen shot of the confusion matrix produced by JMP when used to classify the dataset used for the first couple of experiments. For this confusion matrix, QDA was trained on 75% of the data, and the confusion matrix was computed using the remaining 25% of the data for testing.

Real Sample	Conteo predicho		
	null	rain	snow
null	98	233	43
rain	10055	9332	2690
snow	2	12	1649

Fig. 2 – A screen shot of the confusion matrix produced by R when using a k-fold cross-validated (k = 10) QDA algorithm. Once again the confusion matrix was computed using 25% of the data for testing.

Prediction	Reference		
	null	rain	snow
null	0	0	0
rain	363	21289	209
snow	11	788	1454

Fig. 3 – A PCA presenting a subset of the original weather dataset, used for the first primary experiment. If you consult Figure 1, you will see that the apparent outlier is not, in fact, an outlier, yet simply a product of sampling bias.

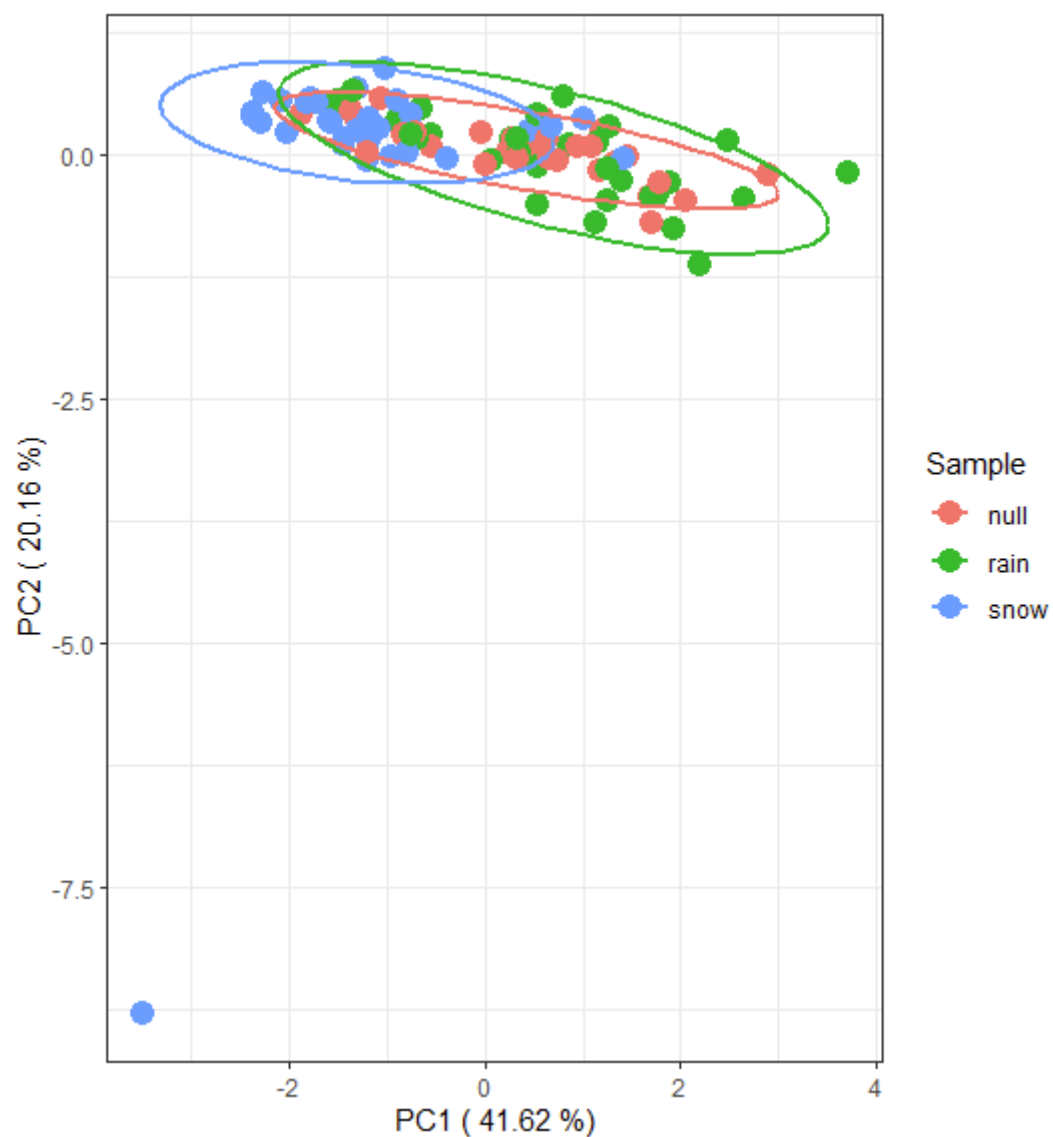


Fig. 4 – A Canonical Variance plot describing the separation of samples without the use of cross-validation.

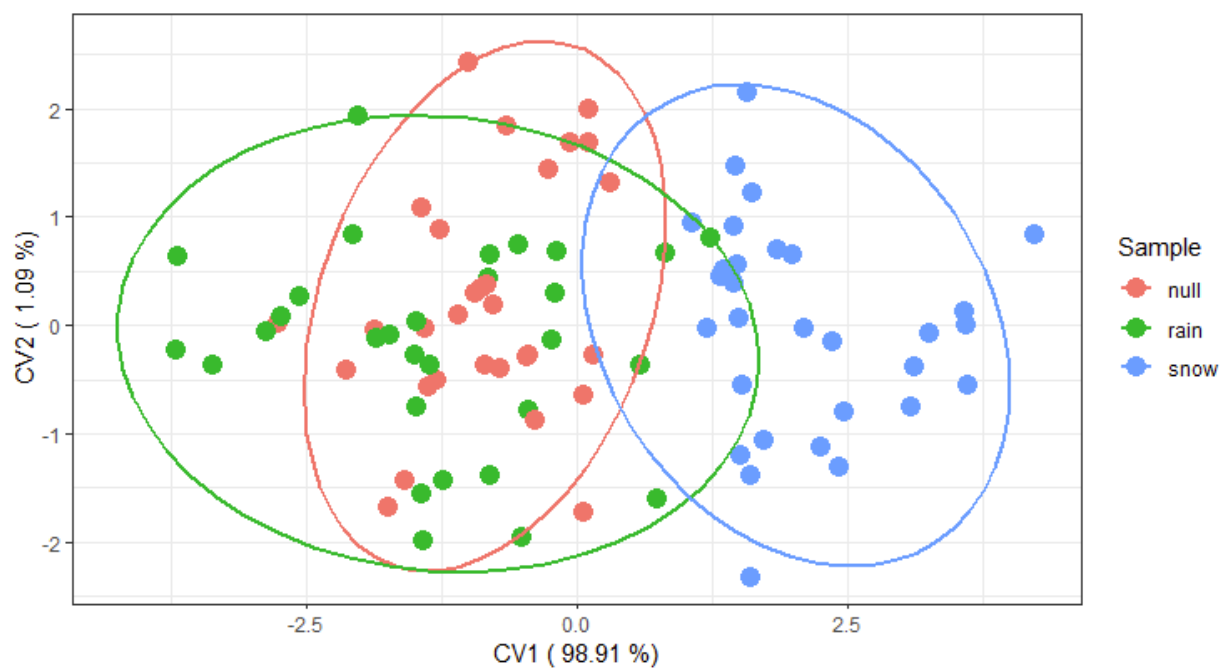


Fig. 5 – A cross-validated Canonical Variance plot describing the separation of samples – note that the outlier is still represented as such, implying that the plot is a more realistic representation of the actual underlying structure and nature of the data

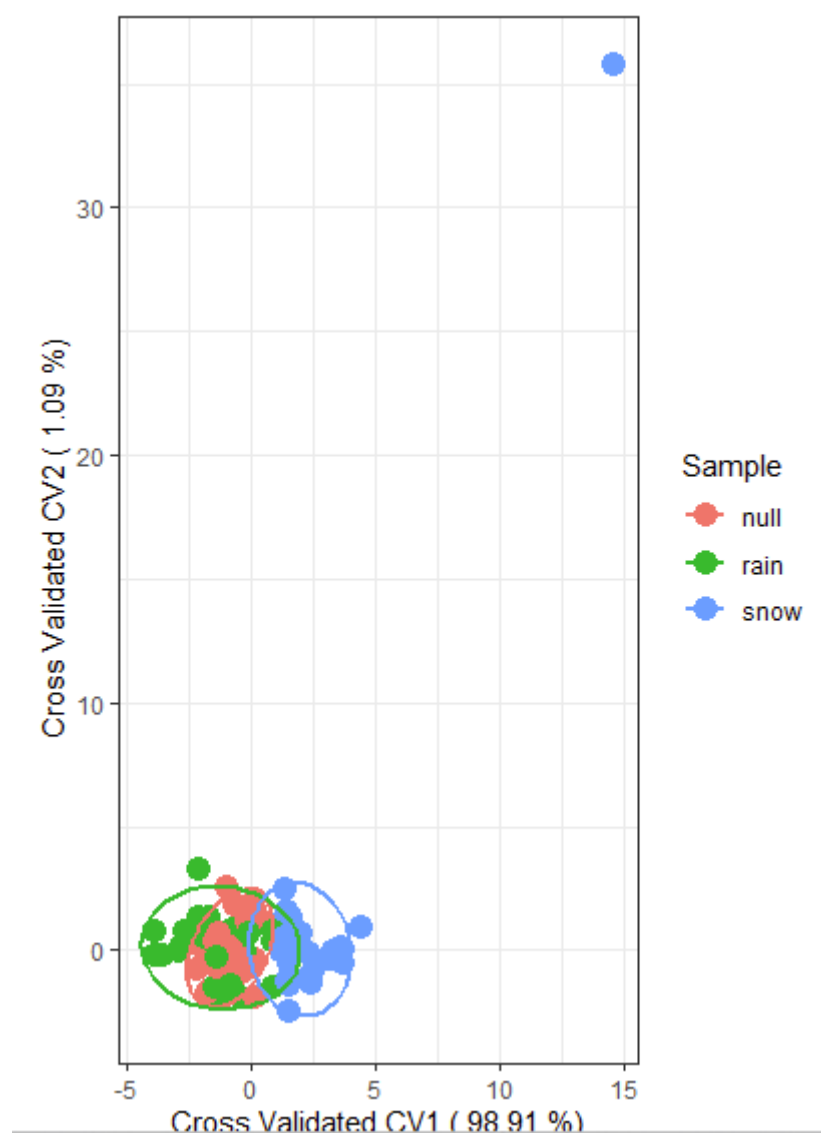
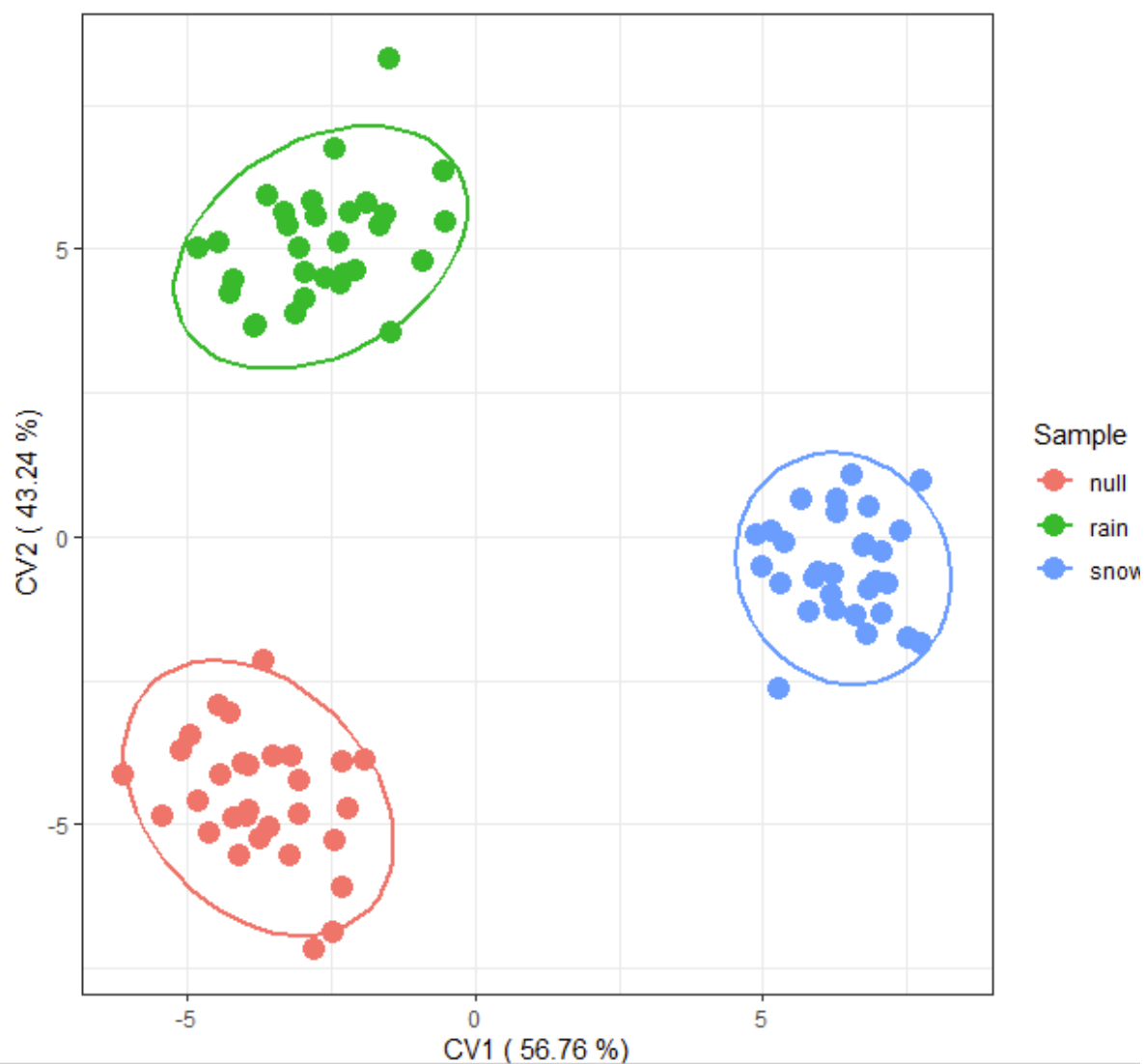


Fig. 6 – A simple Canonical Variance plot describing the separation of the same dataset from figures 3, 4, and 5, however 100 dimensions of noise have been added to provoke the spurious separation of samples, proving that non-cross-validated plots are not necessarily a robust or reliable representation of the data.



A scatter plot showing the relationship between Cross Validated CV1 (56.76 %) on the x-axis and Cross Validated CV2 (43.24 %) on the y-axis. The data points are categorized by sample type: null (red circles), rain (green circles), and snow (blue circles). Three concentric ellipses are drawn around the data points, representing the distribution of each sample type. The null ellipse is the innermost and most compact, followed by the snow ellipse, and the rain ellipse is the outermost and most elongated. The axes range from -30 to 30 on the y-axis and -20 to 20 on the x-axis.

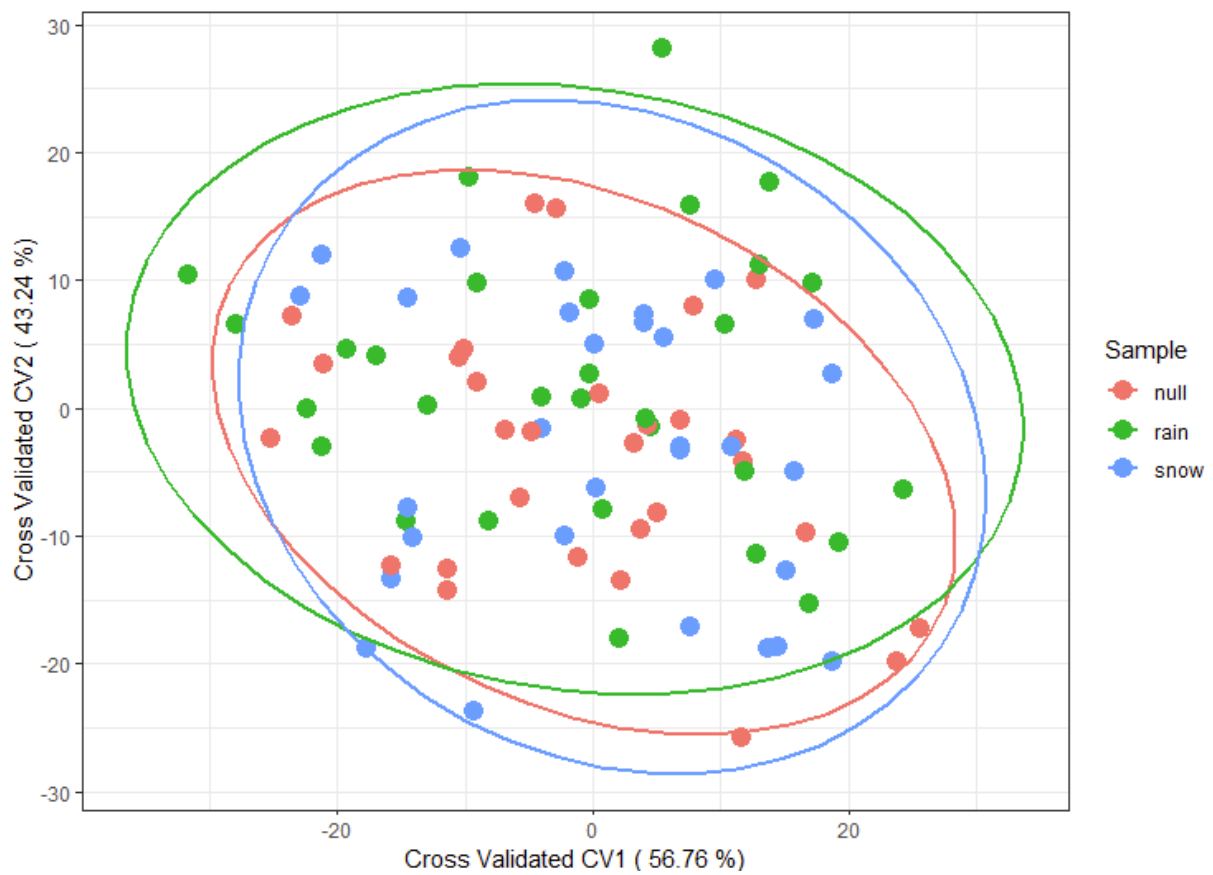


Fig. 8 - A simple Canonical Variance plot describing the separation of a purely synthetic 100 dimensional dataset where absolutely no differences between samples is supposed to exist. All three samples come from the exact same probability distribution. This is theoretical proof that non-cross-validated plots are not a reliable representation of these distributions, as they are supposed to all be the same. While this is product of an inflated variable to individuo ratio, it is an exaggerated means of proving the Marchenko-Pastur theorem.

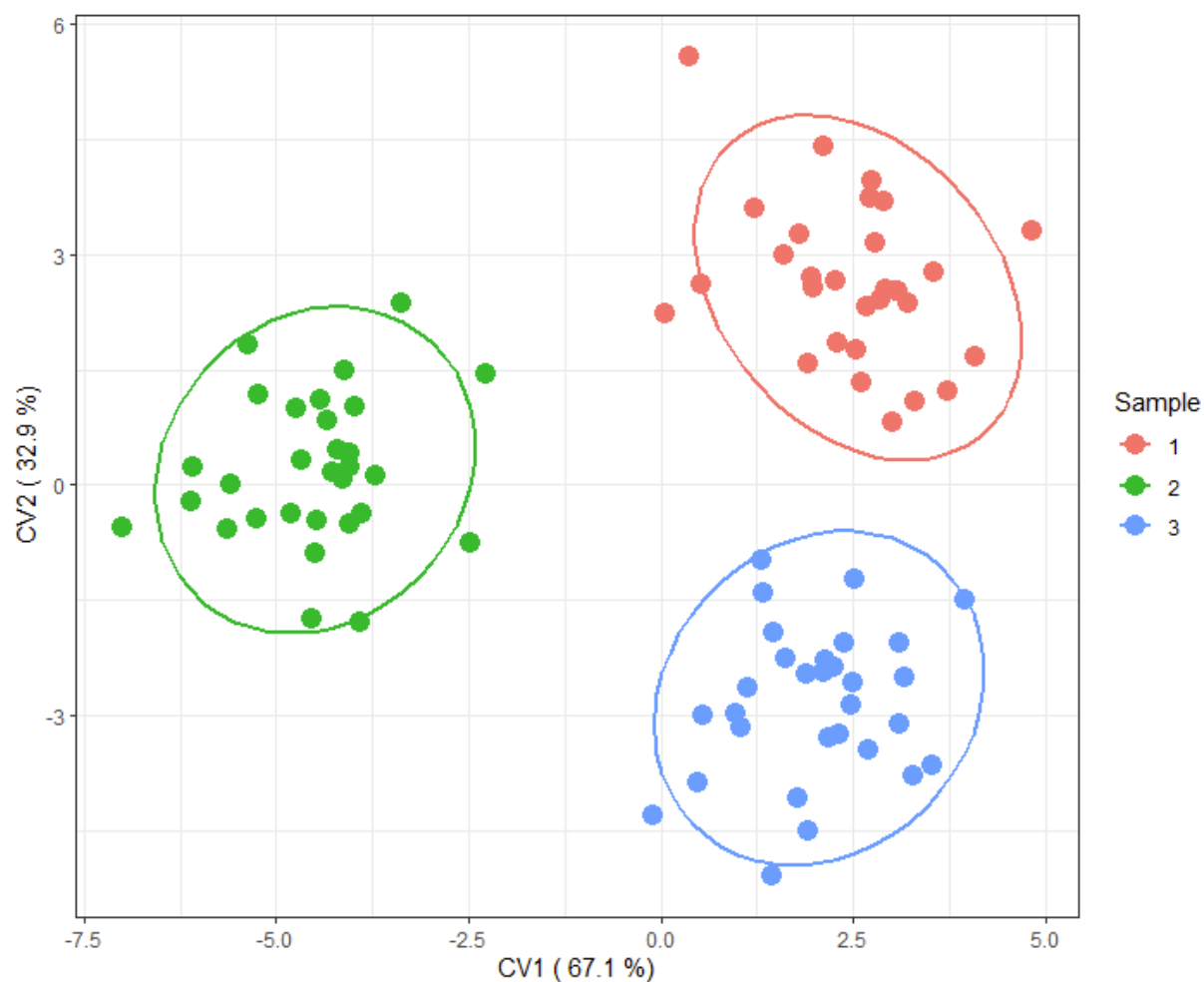


Fig. 9 - A cross-validated canonical variance plot of the data presented in figure 7. Note that the plots are not at all comparable.

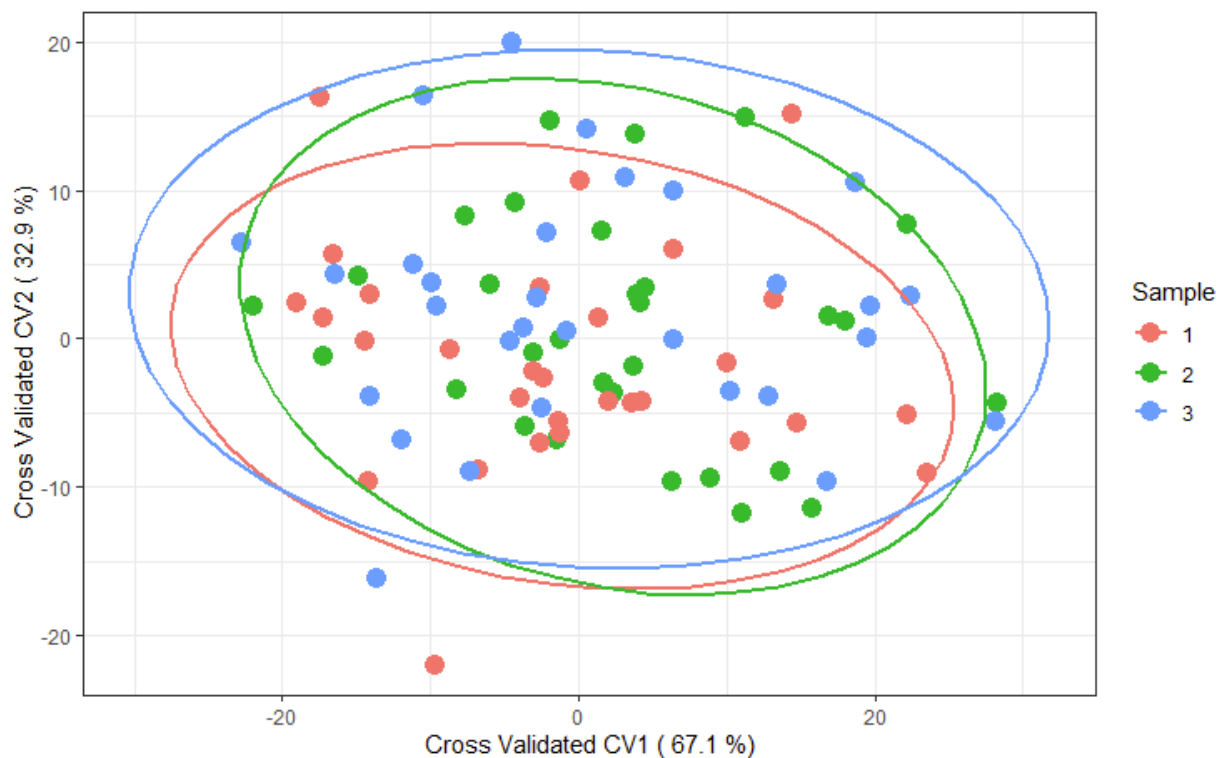


Fig. 10 – An example of calculating CV plots on the data from Figure 7 without performing the validation split. This proves that JMP is not necessarily cross-validating the computation of canonical variance scores, and thus the distributions presented in the plot are not reliable.

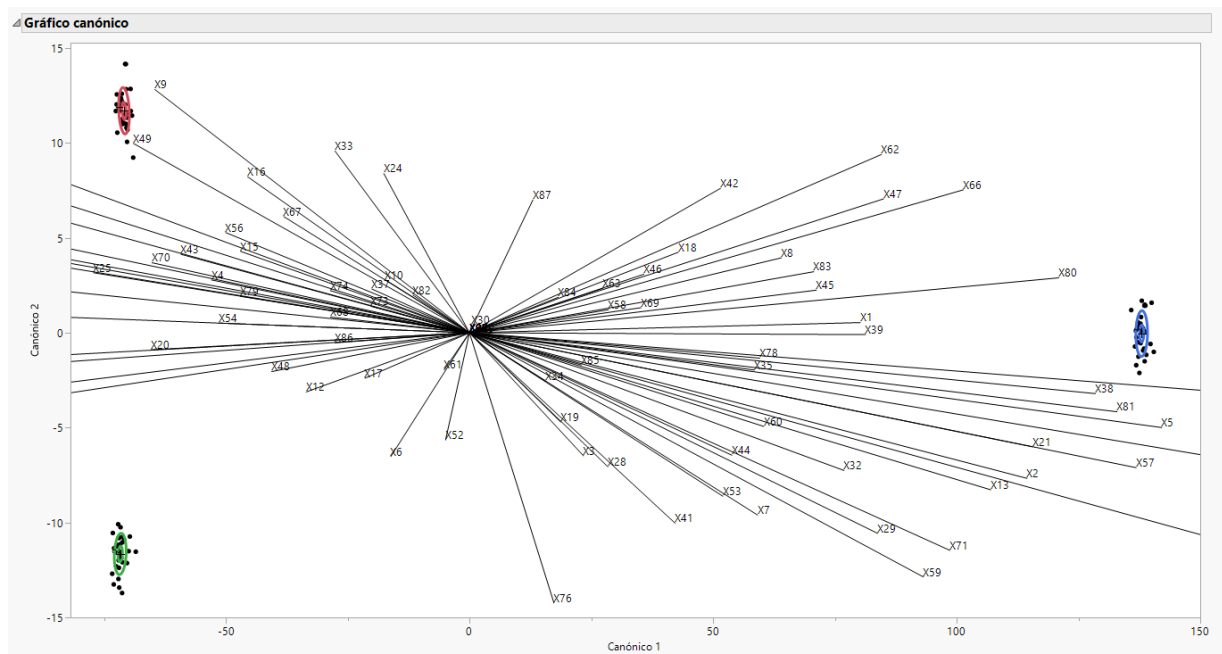


Fig. 10 – An example of calculating CV plots on the data from Figure 7 while using the 75% training split. The only reason this plot and the plot from Figure 10 are different is simply because the number of data points is lower, considering how 25% of the data has been excluded for “validation” (or testing). This is further proof that simply using the analysis/Predictive Modelling/Make Validation Column does not solve any of the issues or points I have been trying to make since my very first review letter.

