Earth Syst. Sci. Data Discuss., referee comment RC1
https://doi.org/10.5194/essd-2022-23-RC1, 2022

Comment on essd-2022-23

Anonymous Referee #1

Referee comment on "*Artemisia* pollen dataset for exploring the potential ecological indicators in deep time" by Li-Li Lu et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2022-23-RC1, 2022

The paper by Lu et al presents an *Artemisia* pollen morphometric dataset, with a view to using these data to differentiate *Artemisia* species and link these to different environments. The authors have clearly put some effort into sampling the herbarium specimens, taking photographs and generating the data, and the data are all fully accessible and approriately curated. However, I have a number of concerns about the extent of the data and how it has been presented and analysed, that to me compromise the results of the study and the potential for the dataset to be employed in further research. There are also issues with how the data generating process is documented in the paper, and given that this is the main point of a paper in ESSD (i.e. to support the publication of a dataset, including presenting how the data have been generated) I find this particularly problematic. I'll detail each of these points, one at a time:

1. Extent of the data: the dataset comprises six morphological characters measured for the *Artemisia* pollen, plus three ratios based on these and a further character ('Pertoration spacing' in Table 1 – should this be Perforation spacing?) that is only relevant for the outgroup taxa. Two of the six measured characters – polar length and equatorial width – can be measured by light microscopy (LM, which most palynologists routinely use) and four – diameter of spinule base, spinule height, granule spacing, and spinule spacing – need to be measured using SEM. However, a number of characters that could readily be measured using LM are left out, most obviously including exine thickness at the equator and poles, colpi length and pore dimensions. These would have been easy to include, and may well be useful for morphometric analysis and classification. I therefore think that these characters should be added in (on line 127 the authors state that colporate pattern was measured, so I assume this was considered at some point).

Also, while the authors include 36 species, with 20 pollen grains measured per species, there is no consideration of plant-to-plant (i.e. intraspecific) variability. Therefore, it is not clear how much differences between species are really just differences between plants, and if this dataset is to be truly valuable as a resource then 3 to 5 plants per species would need to be sampled.

I also note that only 33 species are from *Artemisia*, while the other three are 'outgroup' taxa. It's not clear to me what the point of the outgroup taxa is – this is not a phylogenetic analysis, there is no requirement to have outgroups when using cluster analysis or other exploratory multivariate methods, and given that the pollen from these other taxa looks quite different to *Artemisia*, with an additional character measured that is not relevant for the main taxa being studied, it is inevitable that the outgroup taxa will cluster separately to the 33 *Artemisia* species. This therefore seems like an odd addition that could just as easily be removed.

The other data presented, for example distribution and climate data, are valuable but have simply been downloaded from GBIF and WorldClim, and are thus not original data generated by the authors but rather are publicly accessible data that anyone could already find and download.

2. Documentation of the data generating process: on line 126 the authors state that for each species 20 grains were measured using LM, and 5 using SEM. However, there are 20 measurements for the characters relating to sculpture size and distribution, which if I understand it right were measured using SEM. So where did these come from? Were 20 grains measured using SEM? This really needs to be clear so that people know what they are dealing with.

And in the cluster analysis part (lines 135 to 137) the authors state that five main clusters were distinguished. But why five, when three main types of *Artemisia* pollen are considered elsewhere in the text?

3. Potential for re-use: The lack of measured characters, and the lack of within-species replication, really limits how useful the data will be for future work. This is compounded by the fact that the authors have already carried out an analysis of the data and shown that there is a lot of morphological overlap among taxa, which more or less answers the question of whether these sorts of measurements could be used for classification of individual pollen grains. The reliance on measurements of small sculptural elements that require SEM images also limits how useful this dataset will be for routine use, because not only is LM still the main way palynologists study their samples, even if SEMs are available there is only so much picking of pollen from samples that palynologists are going to do to generate further data to analyse using the dataset that the authors are presenting.

It is possible that the images that the authors have made available along with the data will be useful for further analysis, i.e. with deep learning based classification, but the authors do not flag this up so the whole resource might be missed, and again the lack of intraspecific replication probably limits how useful these photos are for classification attempts.

4. Data analysis: a brief comment on this, but it would be useful to accompany the cluster analysis with an ordination technique such as PCA, because with this you can see how

much your different groups overlap in the morphospace, and therefore how distinct they are morphologically (cluster analysis imposes a hierarchical structure whether one is there or not, while ordinations reveal both gradients and groupings in the data, and give a more visually intuitive understanding of how similar or different the various taxa are).

So, with all of this said, what could the authors do to fix this? As it stands I think the paper would be a better fit for a regular journal, where the aim is to answer a scientific question rather than present a dataset. Something specialist like Palynology, Grana or Review of Palynology and Palaeobotany would be suitable here, although the authors would have to make clear that the lack of sampling might limit the strength of their conclusions. The data would be very appropriate as supplementary material to such a paper though, including the data downloaded from GBIF and WorldClim that would support the analyses in the paper very well.

If the authors do want this to be published in ESSD, then I think there is nothing for it but to generate more data, using more characters and more plant specimens. In the paper itself I suggest much more emphasis on careful documentation of how the data were generated (which at the moment only covers two paragraphs of the whole paper – lines 111 to 129), and much less on the analysis that comes with it. More detail on different contexts in which the data and images could be re-used would also be worthwhile, given that this is one of the main points of the publication.