Report on "**Particle size distributions in Earth Sciences: a review of techniques and a new procedure to match 2D and 3D analyses**" by Pizzati et al.

The authors present an article that is twofold: a critical review of the techniques used to calculate grain size distributions, and a proposed method for calculating a correction factor that allows the average apparent grain size to be compared with the real (3D) average grain size. In general, the text reads smoothly, the figures are illustrative, and the authors have taken great care to describe the methods in the main text and the supplementary materials and appendices. In addition, the dataset (manually outlined grain maps) provided by the authors is of great value in its own right.

As for the review part, I found it useful and Figure 1 is very informative. However, this part lacks a brief review of the methods that reconstruct the 3D structure of particles in consolidated materials (i.e. tomography), which I believe is mandatory if the authors want to present a fair review.

Regarding the new method, from my experience I am not keen on using this type of correction (or proportionality) factors to convert apparent grain size averages to real (3D) averages in most situations because this procedure is mathematically ill-posed and in certain situations can fail badly. However, I admit that occasionally there is no alternative and these simple stereological correction factors can provide useful estimates and, subject to certain limitations, their use may be acceptable. However, throughout the text, I have found a few errors or misconceptions that the authors need to address before publication, see details below.

Besides, mixing a general review, which should present the state of the art but not add anything new except its vision of the subject and outline future lines of research, with the introduction of a new method is, in my view, risky. First, because one of the two goals may be overshadowed by the other, and second, because it makes the article too long and the review part is not strictly necessary to understand the new method. In my opinion, I would have preferred to deal with the two things separately, but this is a decision to be taken by the authors with the agreement of the editor, and I will not complain if the authors finally decide to go down this "mixed" route.

In summary, I think this paper has potential for publication in Solid Earth, but in its current state, it has a few flaws in the analysis and needs further revision.

I hope you find this review useful and constructive.

Marco A. Lopez-Sanchez

2023-12-20

Major issues

If the authors intend to give a general overview of the techniques used today to measure grain size populations (Section 1.2), then a section presenting tomography techniques on aggregate materials cannot be missing. The only 3D reconstruction technique mentioned is serial sectioning (L241), and only in a vague way. It is not necessary to go into this in detail, but a very brief summary must be given, with some comments on its limitations and advantages compared

to other techniques (scope, availability, etc.). In recent years, there have been a large number of reviews on these tomographic techniques, which can be used as a starting point.

L469-470 "*Large datasets allowed…to minimize the effect of random grain sectioning*". This is misleading. The use of large datasets will not minimise the effect of sectioning but will make the population of apparent sizes more representative and the estimate of their average more precise (but not more accurate with respect to the 3D average). The difference between apparent (2D) and true (3D) particle size will not decrease as the sample size continues to increase it will remain the same, i.e., within the precision range established.

Section 6.1 issues

L832-833 "*We employed the very same grain size classes for both analytical methods to allow precise comparison of data distribution, without the bias induced by different bin sizes*" and figure 14 (volume-weighted grain size population plots). Several issues here. First, you can accurately compare populations even if you use different bin sizes by simply normalising the volume density of each interval by the bin size. There is no need to use the same bin size. Using the same bin size for all datasets is counterproductive and that is why you see these artefacts in the coarse size range when you plot the size populations in Figure 14. A better alternative is to use an appropriate bin size according to the population's features (e.g., using a rule of thumb) and then normalise for comparison. Another problem is the use of a solid line to represent the populations because, if I understand correctly, you are using the histogram behind the scenes. Why not show grain-size populations with bins then? Usually, continuous lines are used when you can calculate any value because you have an established model for the distribution, or when you use non-parametric techniques such as the 'kernel density estimation' instead of the histogram. Furthermore, it is not clear what criteria the authors used to draw the line (left edge, right edge or centre of the bin), which is key for the calculation of the mode.

L836-838. "*We preferred to prioritize and put emphasis on evidencing small differences between grain size classes rather than achieving volume frequency stability across the whole grain size distributions.*" Why? In most cases, these instabilities are a product of not measuring sufficient samples to accurately determine the grain size population to the bin size you want in that range of values (i.e., artefacts). I can find no compelling reason to justify this. If these instabilities in the coarse grain size range are real and not due to a statistical problem related to sample size, then there should be an explanation for the physical process that prevents certain grain sizes from being deposited in the sediment and favours others in this size range. Is this the case?

Section 6.2 issues

This section has the main methodological issue of the study. In Tables 2 and 3, and in the text of the section itself, the authors compare average values that do not make sense to compare directly. I will try to explain the issues as best as possible:

1. When one compares different types of averages from a common technique, be it laser granulometry as in Table 2 or 2D image analysis as in Table 3, no average is truer than another, they are all measures (models) of central tendency. Consequently, it makes no sense to calculate the deviation of the others by assuming that one of them is the true measure of central tendency (here the authors assumed that D_m and $D_{w\lambda}$ are the true ones). What you can do is study/compare which measure of central tendency performs better (i.e. needs a smaller sample size for a similar estimation error) and is

more robust (i.e. is less affected by the presence of outliers). That is all. You can do this by applying random sampling techniques to real datasets (you already have them, and of great quality!) or by creating synthetic datasets. You will find that some of them will be more or less optimal depending on the shape of the distribution, the expected number of outliers, etc, and in some cases, they will be comparable or not, depending on the symmetry or asymmetry type of the distribution. See for example Lopez-Sanchez (2020).

2. On the other hand, when comparing averages between techniques, it is fair to establish that the averages estimated by the laser granulometry methods are the real (3D) ones. However, because particle size distributions are asymmetric, you cannot compare a median with an arithmetic mean, a geometric mean, a mode, or any other combination you can think of. Nor can you compare a weighted average (by whatever, volume, area, aspect ratio, etc.) with an unweighted one. If you are interested in comparing different measures of central tendency between techniques, you can compare arithmetic means to arithmetic means, medians to medians, modes to modes, and so on. If you also want to test the effect of volume and aspect ratio weighting, you should weigh all the averages taken from apparent sections the same to evaluate the performance of each, as you do when comparing D_m to $D_{w\lambda}$.

Section 7 (concluding remarks) issues

Regarding point 2 (specifically in L1091-1094), the data you provide in the study do not support the claim that other averages (geometric mean, median, etc) are less reliable and more sensitive to the shape of the particle size distribution than the one proposed here. As a researcher who has worked in depth on this particular topic, the opposite is, in fact, true (e.g. Lopez-Sanchez 2020). Weighted averages perform poorly and are less robust compared to classical averages such as arithmetic and geometric means or medians, i.e. they require a much larger sample size to stabilise or to achieve a specific margin of error and are more prone to destabilise when there is data contamination. In fact, the sample sizes calculated to stabilise the averages at 5% and 1% error, illustrated in Figure 19, are extremely large compared to the typical sample sizes required to stabilise a common average without a stereological correction factor (i.e. the weight).

I guess that the authors here confuse the robustness and efficiency of a measure of central tendency with the validity of using a simple stereological model. In other words, one thing is that the stereological method you propose is the one that provides the closest value between the estimates based on image analysis and the laser granulometry (i.e. the accuracy), and quite another is the efficiency of a measure of central tendency (indeed, in section 6.4 you only tested the efficiency of Dwλ, not the others). I understand that your aim here is to find a simple method that allows you to compare grain sizes by both techniques, but this means you are assuming a ill-posed stereological model that may fail in some cases and inherit all the drawbacks of using a weighted average (worse overall performance and worse robustness to outliers). For example, if you need to compare grain size populations acquired exclusively from 2D sections, it is best to do so using an average that applies no weighting or correction factor at all (this is the case in many other areas of research).

Regarding point 4, It is strange that the authors did not compare the results of their proposed method with those obtained by more sophisticated stereological methods. Since one of the arguments given is that this method is much simpler than other stereological methods (e.g. Saltykov's method), this comparison is very relevant, because if the results obtained with the new

method are the same or better, there is no point in using more complicated procedures. Taking this into account would have added value to the study, but unfortunately, it was not done.

Finally, there is a limitation with the methodology used, which is that the particles are embedded in a matrix without interacting with each other. This model is more similar to an unconsolidated sediment than to a consolidated rock where the grains are in contact with each other. It is difficult to predict the effect this has on the study, but it should be noted as a potential limitation.

Other comments

L44. The term "*texture*" has different meanings depending on the discipline (e.g. crystal preferred orientation vs size and shapes of grains). It is best to state clearly what is meant here.

L65. "*…and to define plastic deformation styles…*". The average grain size in plastically deformed rocks is mainly used as a paleopiezometer (stress estimate). Full grain size populations are typically used to quantify the volume fraction below a certain size threshold. The meaning of "*plastic deformation styles*", although I think I know what you mean, is ambiguous. Perhaps you could say that in the field of plastic deformation, grain size statistics are used for (i) palaeopiezometry or (ii) prediction of the dominant deformation mechanism.

L118. "*…extract 3D average grain diameters from 2D granulometric distributions*." To be more precise, consider changing "*extract*" to "*predict*", as it is not possible to calculate this value exactly, only to approximate it.

L130. *Subsection 1.2*. For the sake of clarity, consider breaking this section down into:

- Disaggregated materials
- 2D methods in aggregated materials
- 3D methods (serial sectioning and tomography) in aggregated materials

Or something similar.

L169. Consider changing "Matlab" to "MATLAB". Also, MATLAB is not a statistical program per se, but a *programming platform designed for scientists and engineers*. Maybe refer to "*…using MATLAB*" or a specific statistical MATLAB toolbox.

L232-236. Stereology is a set of mathematical methods relating 3D parameters defining the structure to 2D measurements obtainable on sections of the structure (Baddeley and Vedel Jensen, 2005). I believe that all the different methods listed here fall under the umbrella of stereology. Note that even the method proposed in this study is stereological by this definition.

L287 Figure 2. Consider adding the name of each sample next to it, as this makes the figure easier to read.

L382. Typo: *De Brouker* should read *De Broukere*

L436-437 Odd phrasing in "*Photomosaics were imported and calibrated in ImageJ image analysis, open-source software…*" Consider simplifying it to "*Photomosaics were imported and calibrated with the free open-source ImageJ software…"* or similar (if you used a particular ImageJ plugin for this, refer to it too). Also, indicate the ImageJ version used.

L589 Figure 9. Consider including the name of each sample directly in the figure, rather than referring to them as (a), (b), etc., as this will greatly simplify the figure caption and make it easier to read the figure without having to keep looking at the figure caption for reference. (Same applies to figures 14, 15 and 18)

L651 Table 1., Indicate at what sigma level is it calculated the error in *Dm*. Does this error take into account that the crystalline parameters of quartz and a spherical particle shape were assumed?

L762-763. "*…and the intercepts with X-axis are almost coincident*." In theory, the apparent (2D) population should show a lower x-axis intercept due to the sectioning effect. If this is not the case, it means that what you are observing is a limit imposed by the image resolution, i.e. you are not able to resolve apparent grain sizes below this limit. It makes no sense to have the same intercept value with both techniques if both are able to measure at the same resolution limit.

L820-821. "*The overall shape, skew (asymmetry) and modal peak position is equal in both the adopted methodologies (Fig. 14).*" This is a qualitative assessment and plots can be misleading. For comparison, you need to provide quantitative measures of population shape, skewness and mode, within the plot in Figure 14 if you wish. Also typo: …are equal…

L908. "*The mode can approximate the average diameter only in the case of weakly skewed grain size distributions*." Although it is true that in symmetric or quasi-symmetric populations the mode and different types of means (arithmetic, geometric, harmonic, etc.) become similar, the mode is just another type of average! There is no real average, only different types of averages.

L954-958. Make sure that all the methods cited here used equivalent circular diameters (ECDs) to measure the apparent grain size and set the conversion factor. In the past, it was more common to use linear intercepts (LIs), callipers or other methods and the 2D to 3D conversions are not directly applicable to ECDs, in fact, there is a conversion factor between LIs and ECDs.

L1036. Consider changing "*to describe*" to "*to summarise*" because an average only summarises a population. To describe a full population you need more than one statistic by definition.