EGUsphere, referee comment RC1
https://doi.org/10.5194/egusphere-2022-633-RC1, 2022

Comment on egusphere-2022-633

Guillaume Duclaux (Referee)

Referee comment on "Clustering has a meaning: optimization of angular similarity to detect 3D geometric anomalies in geological terrains" by Michał P. Michalak et al., EGUsphere, https://doi.org/10.5194/egusphere-2022-633-RC1, 2022

Review of "Clustering has a meaning: optimization of angular similarity to detect 3D geometric anomalies in geological terrains" by MichaÅ□ Michalak, LesÅ□aw Teper, Florian Wellmann, Jerzy Å»aba, Krzysztof Gaidzik, Marcin Kostur, Yuriy Maystrenko, and Paulina Leonowicz.

This manuscript introduces a new workflow for dealing with geological-surface mapping using sparse subsurface data. In particular, this work develops and investigates two new features for geological mapping using unsupervised machine-learning : 1) the role of structural data representations (as normal and dips vectors) on clustering results, and 2) the characterisation of Voronoi diagrams to explain the meaning of the boundaries between obtained clusters. The potential of these two methods are illustrated through applications to a couple of examples focusing at the very large scale on clustering regular data for the bottom Jurassic surface of the Central European Basin System, and at a smaller scale on clustering of irregular data for a middle Jurassic interface within the Krakow-Silesian Homocline in South-Central Poland.

Now, I am definitely not an expert in either machine-learning, nor clustering methods... so I've reviewed this manuscript from the perspectives of a structural geologist to whom such methods could be very useful for interpreting subsurface geology and structures.

Overall, the manuscript is well written and organised, and seems well suited for EGUsphere readership. The application of the unsupervised clustering method is presented, tested and analysed for different k-means and different vector representations in numerous figures (that still require some editing and clarifications). Limitations are appropriately discussed which keeps the contribution very honest. Such new machine-learning approach will potentially provide opportunities for geologists to (re)interpret subsurface structures in regions with either available geological surfaces, or dense boreholes coverage. Based on my review - as a structural geologist - I would recommend accepting this manuscript after moderate revisions of the figures and minor revisions of the text.

I present below a few key points for which I have some questions/concerns followed by a list of minor comments.

1) Choice of the optimum number of clusters: I have some trouble understanding the k-means choices based on the elbow method the authors have employed to determine the optimal number of clusters in their case studies... First, I wonder whether Figure 7 is flawed? Why y the y-axes values so variable between the normal and dip representations for a single dataset? Shouldn't the numbers be similar between a and b (CEBS), and c and d (KSH)? Each structural data provide both dip and normal values, for CEBS there should be 236380 data points. I might miss something related to the y-axis for each figure: what is "tot_withinss"? Why does Figure 7B can either have 2 or 4 optimum number of clusters (Line 285-286)?
The number of clusters will be very important for determining the data clustering pattern based on the cluster centers analysis, so I reckon this section should be strengthened.

2) Stereographic representations: This applies for figure 2c, 8, 9, 10, 11, 12, 13 and 14.
- On Figure 2c it isn't clear which hemisphere is displayed for the data + we don't know where the North is located on figures 2a and 2b.
- Fig 8 to 14: There a lower and an upper hemisphere half globes shown next to the stereonets in all these figures. For the lower hemisphere the whole Stereonet is shown, not for the upper hemisphere... what is the grid spacing on these stereo? I may have missed it but it isn't clear to me what projection is being used (I suppose an azimuthal polar stereographic projection, is that correct?) - structural geologist would classically use equal-area or equal-angle stereonets. Please clarify this in the caption of Figure 8 where the stereos first appear.

3) Direct screenshots from Paraview are hard to read. I'm thinking of Figures 1, 2 and mostly 4 and 6. The scale are not always meaningful, or hard to read. For example in Figure 1 the color scale legend is scalars... I guess elevation would be more adequate. The bounding boxes scale units in figure 2 are nor readable either and are totally missing in Figure 4 and 6. I believe redrafting Figure 4 would massively improve its readability. The moiré pattern visible in Figure b and c is just terrible. Units should also be added to the scale bars, and finally the scale bars' min and max values should be written in the same encoding than the rest of the values (-7.2e+03 and 1.6e+03 should be -7200 and 1600; 1.6e-04 should be 0; 6.1e+01 should be 61; 2.2e-03 and 3.6e+02 should be 0 and 360). Same goes for Figure 6.

Minor comments:
+ Equation 2 page 4: why is each line for Eq. (2) development given a different number? This is the same equation and as such should be only referred to Eq (2).
+ Line 144: please remove the second "a" in this line: "[...] whether a specific 2D point p [...]"
+ Line 234: "en échelon" is missing its accent.
+ Line 240: Genus and species for *Strenocera subfurcatum* should be written in italic font.
+ Line 272: Please revise the reference for the Anon borehole database citation. I

understand it is not published, though.
+ Line 350-351: Theorem 1 --> Do you mean Eq 1?
+ Line 373-374: How would you differentiate between a graben structure and an "antithetic shear with hanging walls dipping against the main fault"?
+ Line 440: missing s in "this result suggests"
+ Figures and captions in general: the figures use lower case letter (a, b, c...) while in the captions and the text upper case letters are used (A, B, C...). Please harmonise between the figures and the text/captions.

Nice, 24/08/2022
Guillaume Duclaux