Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC2
https://doi.org/10.5194/gmd-2020-391-RC2, 2021

Comment on gmd-2020-391

Ignacio Fuentes (Referee)

Referee comment on "*dh2loop* 1.0: an open-source Python library for automated processing and classification of geological logs" by Ranee Joshi et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2020-391-RC2, 2021

In general, the article is well written and presents some valuable results. I'm not totally sure how it is related to the topic of the journal (modeling), but that's a problem for the editor. There are several parts in the manuscript where results were presented including some methodological descriptions. Therefore, I suggest the authors a minor revision of the article, and to move some of the methodological parts in the results section to the materials and methods.

Here are some minor details:

Abstract:

Line 23: what is an extraction rate of 865? What units? Or is it a typo and should be 86.5%?

Introduction:

Line 43: I'm not sure if lithological drill core logging is "inevitably" subjective. In my impression, a lack of standardization in the lithological descriptions makes it subjective, but the subjectivity might be reduced through a standard procedure of description. However, it is not clear that such standardization is what we want.

Materials and methods

Line 146: ”The module was re-written into python to be make it more compact”... The grammar there sounds funny.

Line 195: Figure 2 seems to be wrongly enumerated in the text. The study area is referring to Fig 2, but it corresponds to Fig 3.

Figure 4: Regarding volcaniclastic rocks, they are classified as igneous and sedimentary. Is it ok to have the same subgroup in two lithological groups?

Line 325: it should be EPSG:4326 for WGS84

Line 326: Relative level with respect to the sea level? Does the relative level refers to any reference level or is it a standard level? Because if it refers to any reference level, there is no way to know the real location unless it is corrected using a DEM and assuming the collar at the surface of the terrain.

Results:

Line 498: you specified 820,612 entries for lithology, and 273,684 matched records with the thesaurus. What happened with the remaining 546,819 entries (66.6% of the total entries)?

Can you give a simple example of entries not matched?

Line 507: Does it mean that in about 40,000 records you had a ratio() score from the fuzzy string matching) lower than 80? Maybe you could be more explicit in this? Additionally, you defined the score threshold based on the exact match. But, might it be a kind of balance between the number of matched records and the exact match percentage? I’m just wondering because it seems that by defining that threshold you lose a lot of entries to be matched (83.5%). Could you give a simple example of records not matched?

Lines 542 - 560: These are more materials and methods than results.

Line 549: Couldn’t you get the rest of the hierarchical categories based on the lower hierarchy defined?

Table 2. It gives an example of unmatched cases, so disregard that part of previous comments.

Lines 663 - 678. Accuracy metrics should be included in the materials and methods section and not in the results.

Discussion

Lines 812: Did you tried to replace the “same as above” with the descriptions?

Lines 824: Is there any way to automatise the building of a thesaurus given new advances in NLP and machine learning?

Line 831: “extraction rate of 16% from the Comments is not bad at all” This sounds too subjective, how do you define what is a good or a bad extraction rate?

Lines 852-853: “For sedimentary rocks, the lack of a standard syntax as to how comments are recorded impacts the classification. “ You see, imagine if such standardization is achieved, wouldn’t it reduce the subjectivity

Line 863: grammar error “ Soils are technically are not rocks”