Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC1
https://doi.org/10.5194/gmd-2020-391-RC1, 2021

Comment on gmd-2020-391

Anonymous Referee #1

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-RC1, 2021

General comments

The paper is an interesting and innovative application of methods to extract meaningful quantified standardised information from highly heterogeneous drill log data that are typical of legacy mineral exploration. Even within company-imposed data structures, the logging geologists have historically recorded subjective data inconsistently and this hinders bulk aggregation of these data. The classification methods in this paper using open source scripts and libraries show potential to efficiently find similarities between logged lithological terms, codes and descriptive comments using hierarchical semantic relationships.

The paper is too long and dense in terminology and nuanced meaning. The authors have endeavoured to symbolise some of the different and complicated database terminology although possibly more is needed. The flow is logical and the writing is generally understandable although laboured in places. The figures and tables are appropriate and informative in most places; some minor improvements have been suggested. I have not checked references or URL links exhaustively.

Specific comments

Introduction, section 1: replace first 2 sentences with 'Drilling is a process of penetrating through the ground that is capable of extracting information about rocks from various depths below the surface. This is useful for establishing the geology beneath. Drill core or cuttings can be collected thus providing samples for description, interpretation and analysis.'

Introduction, section 3: The legacy data described seem to be hardcopy forms

subsequently digitised. Legacy digital data also suffer from lack of standardisation, inconsistency.

Introduction, line 78: These data are not 'unstructured' but they may not conform to standards or be consistently applied/described.

Material and Methods, Conventions: workflows need their own distinct convention (font). Later they are confusingly rendered as combinations of database table fields.

Thesauri: Some of the so-called 'synonyms' actually have distinct meanings from each other, even the listed example elevation vs relative (reduced) level. Maybe qualify 'synonym' as meaning 'nearly the same' or a 'close match' for their general intent is similar e.g the elevation terms all are recording a vertical height.

Thesauri, line 245: Rather than 'The opposite is true as well' suggest you explain specifically that more than one code may refer to the same lithology. Basically there is a many-to-many relationship between code and lithology.

Thesauri, line 252: The CGI vocabularies support the GeoSciML and EarthResourceML (note singular) geology data models but potentially other applications. Suggest you rephrase as '...the CGI-IUGS geoscience vocabularies accessible at http://geosciml.org/resource/def/voc/'.

Thesauri, Lithology Hierarchical Thesaurus: The 3-level hierarchy is highly simplified compared to CGI's Simple Lithology. Many of the 'Lithology_Subgroups' listed have parent-child relationships e.g. 'mafic_fine_grained_crystalline' is a child of 'mafic'. This should be mentioned, presumably some simplification and pragmatism is needed for your analysis.

Data Extraction, Collar Extraction: I wondered why this section needs to be here at all. The collar extraction isn't central to your paper focus on lithology. You don't utilise collar location in a spatial analysis or context - its only function in this paper seems to be a pre-filter for data quality. The method itself is good and useful, and ultimately important for data mining where spatial understanding is needed, just not essential for the lithology-driven analysis presented here.

Data Extraction, Survey Extraction: Ditto, I wondered why this section needs to be here at all. If it is retained then suggest the 4th field should be ‘Inclination’ not ‘Dip’. Dip is a measurement of the slope of a planar surface feature whereas inclination refers to the plunge of a linear feature. Additionally, how consistent are WAMEX records around positive inclinations meaning upwards-directed drill holes?

Data Extraction, Survey Extraction: The 'Calculated X, Y, Z values' are not particularly helpful or necessary in this Survey table i.e. only recording collar and the end of hole locations. Survey tables more typically describe changing azimuths and inclinations with 'depth' (i.e. account for curved drill holes) – does WAMEX not do this?

Data Extraction, Lithology Extraction: The fuzzywuzzy algorithm appears to be repeating pre-processing already mentioned in the previous paragraph (line 419-424).

Data Extraction, Line 432-433: What does 'Since the sorted intersection component of token_set(), will result in an exact match…' mean? Elaborate or explain more clearly.

Data Extraction, Line 439: What is an 'intersection token'?

Data Extraction, Table 1: The column of ticks and crosses is unexplained. In two cases the lower score is ticked implying it is the preferred result?

Data Extraction, Line 453: 'Andesitic basalt' is an unfortunate example since 'basaltic andesite' is an established volcanic rock name. Would basaltic andesite wrongly revert to andesite in this classification process?

Data Extraction, Figure 8: I struggled to understand this graph. How can data with 100% Exact Match score only 80%?

Data Extraction Results, Unique Lithology Code Results: Database table field names seem to be inconsistent e.g. Company_LithoCode vs Company_Lithology vs Lithology_Code. Suggest careful check to ensure consistency of use otherwise confusing for the reader.

Data Extraction Results, Unique Lithology Code Results: Workflows such as 'Lithology_Code Detailed_Lithology' need to be distinctly symbolised. At the moment they look like co-joined database table names without an obvious algorithm progression between them. Suggest also where these are mentioned you mention 'workflow' or 'workflows'

Data Extraction Results, Table 2: Struggling to understand why row 3 is a Close Match when the almost identical row 5 is a Broad Match? If anything the 'basic volcanic rock', not being a recognised Lithology_SubGroup member, is broader rather than closer than 'mafic

fine grained crystalline'.

Data Extraction Results, Fuzzy String Matching Results: Mentions of 'comments' should be 'Comments' in most cases, possibly with special font.

Data Extraction Results, lines 611, 622: These results are suboptimal. You discuss this later but it seems your method is sometimes picking a subordinate lithology rather than the dominant lithology.

Data Extraction Results, line 653: I wasn't clear what 'the limitation' is – processing?

Discussion, Assessment of String Matching Results (line 844): Need to qualify that the 'classification of structures and textures and metamorphic rocks' has higher confidence in the study area dataset, not necessarily in others. I'm sure there will metamorphic-dominated terranes where the subordinate igneous rocks will be classified with higher confidence.

Technical corrections

line 44: delete ', particularly as it is likely to have been conducted by tens to hundreds of geologists...' with something like 'as all logging geologists have their own personal biases.

Line 49-50: delete 'even detection of'

line 57: The semi-automatic methods also are poor at describing textural characteristics (foliation, banding, grainsize variation)

line 70" Delete 'Elizabeth'?

line 105: limitations -> limitation

line 198: replace 'that occurred' with 'were emplaced'

line 199: replace ‘ultramafic mafic’ with ‘ultramafic to mafic’ and ‘local centres’ with ‘local eruptive centres’.

line 201: replace ‘volcanoclastic’ with volcaniclastic’

line 203: delete ‘profiles’ and delete ‘both’. Break sentence after ‘bedrock’ and start next with ‘Regolith...’

line 209: suggest replacing ‘complexity’ with ‘diversity’

Figure 3 needs a unit to describe dill hole density e.g. per square kilometre

line 254: Insert after ‘Added records’ ‘with examples’

line 270: Replace ‘GeoSciML’ with ‘the CGI-IUGS Simple Lithology vocabulary http://resource.geosciml.org/classifier/cgi/lithology’

line 276: Suggest deleting second half of sentence i.e. after ‘dictionary’

line 294: Delete orphan ‘And’

Figure 4: Lighten purple shade (or whiten text within)

line 358: Replace ‘Dip: it is the inclination angle perpendicular to the azimuth…’ with ‘Inclination: the plunge angle of the drill hole relative to horizontal…’.

line 358-360: Replace sentence ‘A positive value indicates an upward-directed drill hole and a negative value indicates a drill hole directed downwards.’

line 411: Replace ‘The string followed by key phrases such as…’ with ‘The string preceded by key phrases such as…’

line 415: Does ‘tokens with less than three characters’ mean or include short words?

Line 434: Insert ‘method’ after ‘ratio()’.

Line 511: Font change in ‘Company_Lithology’.

Line 570: Where is the ‘brown text’ in Table 2?

Figure 10: Lighten purple shade (or whiten text within).

Line 598: replace ‘take a look’ with ‘looked’

Line 663: replace ‘couple of’ with ‘four’

Line 671: replace ‘trumps’ with ‘trump’

Line 833: What ‘information being fed itself’ mean?

Line 887: delete ‘a couple of’