Chemical identification of metamorphic protoliths using machine learning methods: A manual

D. Hasterok

June 19, 2019

1 Introduction

In many cases, a metamorphic protolith class, igneous or sedimentary, can be reliably determined by field observations, analysis of zircon morphology, details of the zircon spectra, or perhaps some other means. However, there are cases where these data are inconclusive, contradictory, or not immediately clear. For these cases and to add weight to interpretations, we have trained a classifier to predict the origin as igneous or sedimentary on the basis of major element chemistry.

The method and codes detailed here use a classifier trained selected from a range of machine learning methods included in the MATLAB® Classification App. The full details of the training can be found in (Hasterok et al., revised). Some methods perform better among igneous or sedimentary rocks, but our preferred method, RUSBoost ensemble decision trees with 30 splits and 1000 learners, was balances the performance of each while still performing quite well overall.

2 Performance

The classifier was trained on a global dataset of 497401 meta+igneous and 35959 meta+sedimentary geochemical samples from a version of a global geochemical dataset, now updated by Gard et al. (2019). The database represents a combination of exisiting academic geochemical databases, governmental reports individual peer-reviewed publications, and theses/dissertations. The classifier trained is trained on a random selection of 90% of the data used for training and ~10% of the data held in reserve for post-training validation. The performance is given in Table 1 below.

Table 1: Protolith classifier overall performance.

	true	predicted protolith			
		igneous		sedimentary	
	N	*N*	%	*N*	%
		training dataset			
igneous	447669	428440	95.7	19229	4.3
sedimentary	32355	3258	10.1	29097	89.9
		validation dataset			
igneous	49732	47475	95.5	2257	4.5
sedimentary	3604	530	14.7	3074	85.3

Although the performance is quite high, it can vary considerably by rock type as shown in Table 2. Mafic and higher alkaline performance is very high. The classifier performance is good, but less accurate, for igneous granites and granodiorite and sedimentary arkose, wacke and iron-rich shales. The classifier performance is very poor for very high silica igneous rocks (e.g. quartz veins and pegmatites) and carbonatites, which can be difficult to distinguish from marbles.

3 Protolith Prediction

The protolith classifier is relatively simple to use. A suite of MATLAB codes are provided at . If you are interested in the dataset used to train the classifier, you can find it at .

3.1 Input Data Format

The input of data is relatively simple, requiring an Excel worksheet (*.xls or *.xlsx). The spreadsheet format requires a table of geochemistry data; an example is provided in protolith_template.xlsx. Leave no blank rows at the top. Each column should be single column, with the first column being a description of the contents of each row. At a minimum, the table must include a sample name, SiO_2, TiO_2, Al_2O_3, MgO, CaO, Na_2O, K_2O, P_2O_5 and FeO and/or Fe_2O_3. All iron as Fe2+ or Fe3+ will be converted to total iron, FeO before processing. Units for major element oxides should be in weight percent.

All other elements or oxides are not used, but will not affect the classifier; though, if additional elements or metadata are included they can be used with some of the other plotting and processing codes provided through github. Note all elements should be given in ppm, though if ppb are included in the table, append (ppb) to the row name, e.g. 'Au (ppb)', as the code will automatically convert these data to ppm. For platinum group elements sometimes ppt is used, again (ppt) will be converted to ppm. It is common for S or C to be given in wt.%. for these data, append (%) to the row name. If no (units) is provided, the assumed units for individual elements is assumed to be ppm.

3.2 Running the Code

There are a number of functions and scripts, but only protolith_predictor.m needs to be run within MATLAB to predict protoliths from the input file. The protolith classification function can be run in two formats: either with the Excel filename chosen through a dialog once the function is called/run or by explicity calling the function with the filename, *filename.xlsx*.

To run, make sure you have downloaded the files from github. Make sure MATLAB's current working directory is set to the folder containing the protolith classification codes, or the directory has been added to the MATLAB path using *addpath* or by selecting the setpath under the environment tab and adding the path.

In the command window, type

```
protolith_predictor
```

which will bring up a dialog from which the Excel input file can be selected. Alternatively,

```
protolith_predictor(filename.xls*)
```

can be used to open the file directly without a dialog.

3.3 Interpreting the Results

The code will create an output file *filename_classified.csv* with three columns: sample name, predicted class, and score. The predicted class will be listed as either igneous or sedimentary. The score is a value the varies from -1 to 1 with negative values indicating a predicted sedimentary class and positive values indicating a predicted igneous class. Values with a higher absolute value indicate a high confidence in the prediction and a values near zero have poor confidence. There is no clear point to determine a cutoff for using scores because it is possible that scores will indicate high confidence and yet be classified poorly; though, a absolute value of 0.5 will capture most high confidence data correctly.

4 Feedback

If you use this code to classify protoliths, I would like to know how it works for you—good or bad. Only through its success or failure can I improve the method further or make it simpler for one to use. What rock types are poorly classified? If you have multiple samples from the same unit, does it yield the same result on all sample, right or wrong?

References:

Gard, M., Hasterok, D., Halpin, J., 2019. Global whole-rock geochemical database compilation. Earth System Science Data Discussions , 1–23doi:10.5194/essd-2019-50.

Hasterok, D., Gard, M., Bishop, C., Kelsey, D., revised. Chemical identification of metamorphic protoliths using machine learning methods. Computers & Geosciences .

Table 2: Protolith classifier performance for individual rock types.

	training dataset			validation dataset		
rock type[a]	true positives	false negatives	% FN	true positives	false negatives	% FN
true igneous samples						
quartzolite	36	563	94	5	66	93
granite	62077	5037	7.5	6704	573	7.9
granodiorite	31549	5538	14.9	3553	627	15
diorite	33692	1826	5.1	3659	206	5.3
gabbroic diorite	56262	610	1.1	6209	76	1.2
subalkalic gabbro	99156	306	0.3	11090	35	0.3
peridotgabbro	2626	67	2.5	313	4	1.3
crustal peridotite	621	26	4	62	2	3.1
syenite	7883	222	2.7	921	35	3.7
quartz monzonite	14400	822	5.4	1568	103	6.2
monzonite	15443	1010	6.1	1736	113	6.1
monzodiorite	18304	385	2.1	2020	39	1.9
monzogabbro	14227	111	0.8	1546	17	1.1
alkalic gabbro	27075	144	0.5	3077	19	0.6
foid syenite	3376	57	1.7	350	5	1.4
foid monzosyenite	1802	73	3.9	214	10	4.5
foid monzodiorite	2729	54	1.9	320	11	3.3
foid gabbro	12975	143	1.1	1498	19	1.3
ultra-high alkali igneous	280	7	2.4	36	4	10
foidolite	3377	91	2.6	399	12	2.9
sanukitoid	1931	84	4.2	190	8	4
picrite/alkali picrite	3021	45	1.5	313	6	1.9
komatiite/meimechite	3697	36	1	381	8	2.1
mantle peridotite/pyroxenite	2627	2	0.1	303	3	1
carbonatite	665	433	39.4	81	69	46
silicocarbonatite	924	284	23.5	87	36	29
true sedimentary samples						
quartzite	3087	57	1.8	328	10	3
quartz arenite	147	0	0	24	0	0
litharenite	1300	10	0.8	156	2	1.3
sublitharenite	171	0	0	16	0	0
arkose	1953	455	18.9	222	72	24.5
subarkose	261	0	0	26	0	0
wacke	6136	478	7.2	639	84	11.6
shale	7129	617	8	754	96	11.3
iron-rich shale	3167	1469	31.7	304	219	41.9
iron-rich sand	1464	16	1.1	148	4	2.6
laterite/bauxite	308	1	0.3	37	2	5.1
limestone	2539	3	0.1	276	7	2.5
dolomite	1433	151	9.5	144	34	19.1

Only plutonic names for igneous rocks.