Available online at www.sciencedirect.com
ScienceDirect
Geochimica et Cosmochimica Acta 71 (2007) 3391–3392
Geochimica et Cosmochimica Acta
www.elsevier.com/locate/gca

Response

Reply to Comment by Agrawal and Verma on "Tectonic classification of basalts with classification trees"

Pieter Vermeesch

ETH Zurich, Isotope Geology and Mineral Resources, Clausiusstrasse 25, NW C85, CH-8092 Zurich, Switzerland

Received 20 April 2007; accepted in revised form 27 April 2007; available online 3 May 2007

Agrawal and Verma (2007) allege six problems with the use of classification trees for the tectonic discrimination of oceanic basalts, as proposed by Vermeesch (2006). In the following, I will demonstrate that the first five of their points are false, whereas the sixth is partially correct but can easily be fixed.

The results of Vermeesch (2006) are said to be irreproducible because a large number of data points are "unclassifiable" due to the absence of all the primary and surrogate variables. However, Section 4.2 of Vermeesch (2006) explains that in such cases a "follow the majority" decision should be made. For example, in the absence of TiO_2, P_2O_5, and Zr (the primary and surrogate variables for the first split of the full tree; Figure 4 of Vermeesch, 2006), the sample should be sent to the "Yes"-side of the split, because this is where the majority (520/756) of the training data go. The "follow the majority" rule is an integral part of the classification tree method and all but obviates points (i)–(v) of the Comment by Agrawal and Verma (2007).

Thanks to the simple but effective way of dealing with missing data, the sparseness of the training data used by Vermeesch (2006) is not a problem. Agrawal and Verma (2007) remark that not even a single sample in the training set was analyzed for all the variables, and that only one MORB sample was analyzed for Sn. It is important to note that the variable Sn was used in neither of the two classification trees presented by Vermeesch (2006). The fact that classification trees are not hurt by sparse datasets should be seen as a positive feature and can hardly be considered a criticism of the method.

The sixth and final point of Agrawal and Verma (2007) is that geochemical analysis should consider only the relative and not the absolute values of its components. I welcome the opportunity to elaborate on this point here. It is true that the constant-sum constraint ("closure") of compositional data implies some degree of correlation between the components. Loss or gain of any component causes a change in the concentration of all the other components. This problem is well known in geochemistry and is generally solved by taking ratios. For a parametric method such as discriminant analysis, Aitchison (1986) advocates taking log-ratios. However, taking logarithms is not necessary for non-parametric tools such as the classification tree. The latter only considers the order of the split variables, which is not affected by taking logarithms.

For the sake of illustration, a ratio-based tree was built using the same dataset as Vermeesch (2006), but converting major oxide concentrations (in weight percent) to elemental concentrations (in parts per million). The following variables were used: La/Ti, Ce/Ti, Nd/Ti, Sm/Ti, Eu/Ti, Gd/Ti, Tb/Ti, Dy/Ti, Ho/Ti, Er/Ti, Tm/Ti, Yb/Ti, Lu/Ti, Sc/Ti, V/Ti, Sr/Ti, Y/Ti, Zr/Ti, Nb/Ti, Hf/Ti, Ta/Ti, Th/Ti, U/Ti, Sr/Zr, Zr/Nb, Nb/Th, La/Sm, La/Yb, Gd/Yb, Th/Ta, Nb/La, Th/Yb, Th/U, Nb/U and Nb/Ta (Fig. 1). Only 751 of the original 756 data were used for the tree construction, because one IAB and four OIBs lacked all the necessary variables. Using the entire dataset of 756 training data, the resubstitution error of the ratio-based tree is 14% and its 10-fold cross-validation error is 18%. Because the surrogate variables are also composed of ratios (Table 1), they are subject to some degree of spurious correlation (Chayes, 1971). There is no way around this, but the cross-validation error estimate suggests that it only affects the performance of the tree to a minor degree.

To illustrate once again the use of surrogate variables and the "follow the majority" rule, consider a sample with a Sr/Ti ratio of 0.01 and lacking all other variables. The primary split variable (Sr/Zr) is missing, so the first surrogate variable must be used (Table 1). Because $Sr/Ti = 0.01 < 0.02056053$, the sample is sent to the right side of the first node. We have now arrived at the third node, and the primary and surrogate variables are La/Yb, La/Sm and Yb/Ti, respectively. All these variables are missing, so

E-mail address: pvermees@erdw.ethz.ch

doi:10.1016/j.gca.2007.04.021

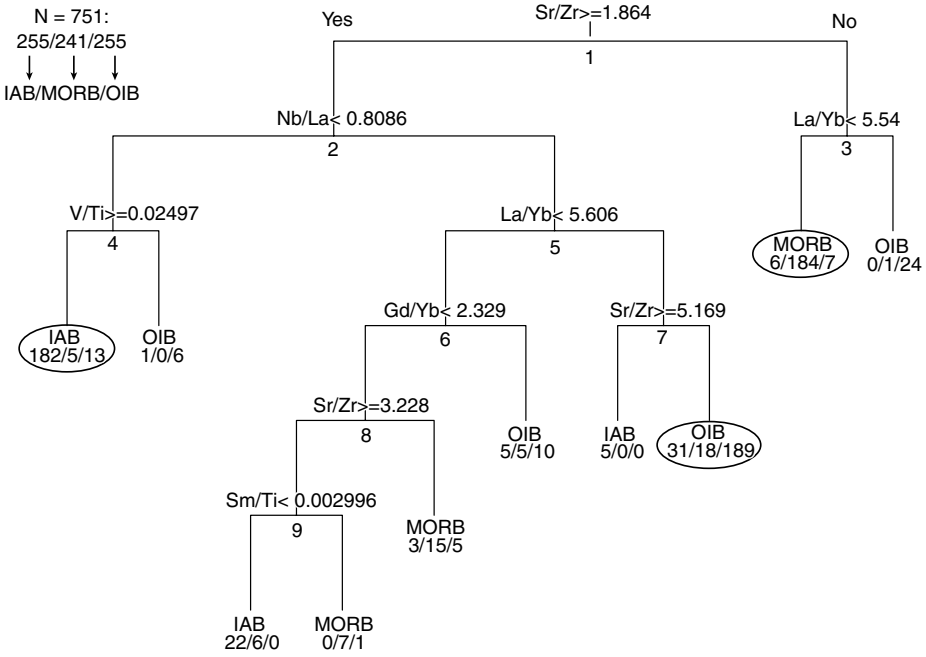

Fig. 1. Example of a ratio-based tree. The "heaviest" nodes are encircled as in Vermeesch (2006).

Table 1
Surrogate splits for the ratio-based tree (Fig. 1)

Split number	IAB/MORB/OIB	Primary split	Surrogate 1	Surrogate 2	Surrogate 3
1	255/241/255	Sr/Zr $\geqslant$ 1.86375	Sr/Ti $\geqslant$ 0.02056053	—	—
2	249/56/224	Nb/La $<$ 0.8085888	Zr/Nb $\geqslant$ 13.66719	Nb/Th $<$ 4.770525	Sr/Zr $\geqslant$ 3.527116
3	6/185/31	La/Yb $<$ 5.539706	La/Sm $<$ 2.778826	Yb/Ti $\geqslant$ 0.0002107271	—
4	183/5/19	V/Ti $\geqslant$ 0.02497367	—	—	—
5	66/51/205	La/Yb $<$ 5.606413	La/Sm $<$ 2.480184	La/Ti $<$ 0.0007888696	Ce/Ti $<$ 0.002014049
6	30/33/16	Gd/Yb $<$ 2.328818	La/Yb $<$ 3.675245	—	—
7	36/18/189	Sr/Zr $\geqslant$ 5.168944	—	—	—
8	25/28/6	Sr/Zr $\geqslant$ 3.228164	Sr/Ti $\geqslant$ 0.03181145	Zr/Ti $<$ 0.008364239	La/Sm $\geqslant$ 1.078327
9	22/13/1	Sm/Ti $<$ 0.002996233	Eu/Ti $<$ 0.001157938	Yb/Ti $<$ 0.00315352	La/Ti $<$ 0.002535417

If all primary and surrogate splits are missing, a "follow the majority" rule is used.

we must use the "follow the majority" rule. Because 197 out of the 222 training data that arrived at node 3 were sent to the left, our sample is classified as MORB. The misclassification rate of samples with missing data is worse than that of samples which were analyzed for all the components. However, provided that the sample of unknown tectonic affinity and the training data are comparatively sparse, the cross-validation error of 18% should be a reasonably accurate estimate of the true misclassification rate. For this reason, comparing the performance of a classification tree with that of a discriminant analysis lacking any missing data, as done by Verma et al. (2006) is fundamentally unfair.

APPENDIX A. SUPPLEMENTARY DATA

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gca.2007.04.021.

REFERENCES

Agrawal S., and Verma S. P. (2007) Comment on "Tectonic classification of basalts with classification trees" by Pieter Vermeesch (2006). *Geochim. CosmoChim. Acta* **71**, 3388–3390.

Aitchison J. (1986) *The Statistical Analysis of Compositional Data*. Chapman and Hall.

Chayes F. (1971) *Ratio Correlation; A Manual for Students of Petrology and Geochemistry*. Chicago University Press.

Verma S. P., Guevara M., and Agrawal S. (2006) Discriminating four tectonic settings: five new geochemical diagrams for basic and ultrabasic volcanic rocks based on log-ratio transformation of major-element data. *J. Earth Sys. Sci.* **115**, 485–528.

Vermeesch P. (2006) Tectonic discrimination of basalts with classification trees. *Geochim. CosmoChim. Acta* **70**, 1839–1848.

Associate editor: Richard J. Walker