

Supervised Machine Learning for Columbia River Basalt Group Classification

A. E. STEINER^{1*} AND J. A. WOLFF¹

¹Washington State University, Pullman, WA 99164, USA

(*correspondence: ashley.steiner@wsu.edu)

(jawolff@wsu.edu)

The Columbia River Basalt Group (CRBG) is a large igneous province located in the Pacific Northwest, USA that has a complex stratigraphy of ~210,000 km³ of basalts and basaltic andesites divided into half a dozen formations, more than 40 members and as many as 30 distinct flows distributed among a subset of those members. Many of these members can be distinguished from one another based upon the bulk rock geochemistry as determined by X-ray fluorescence (XRF) spectrometry and inductively-coupled plasma mass spectrometry (ICP-MS).

The practice of identifying CRBG lavas based upon XRF-determined geochemistry has been utilized at the Peter Hooper GeoAnalytical Lab for the last ~40 years to great effect. Classification of unknown basalts within the group has historically been performed for academic and commercial uses by painstakingly ‘eyeballing’ bivariate plot after bivariate plot of major and trace data.

In this study, we have compiled our database of labeled CRBG XRF & ICP-MS geochemical analyses and created a preliminary pipeline of supervised machine learning models that use a multiclass logistic regression classifier to classify unknown lavas of the CRBG into formations, members and flows, respectively. The model was developed using the open-source Python module Scikit-learn v0.20.2 [1]. The logistic regression classifier utilizes the limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) solver and yielded a 97% accuracy score on a 20% test split of the 3700-sample CRBG geochemical dataset for formation classification. The computed log loss score for the four main formations are as follows: Grande Ronde 0.038838, Imnaha 0.014756, Saddle Mountains 0.005578, and Wanapum 0.042380. These low values support the assertion that these formations are geochemically distinct in 27-dimension space (10 major oxides and 17 trace elements). Furthermore, applying this classification model to geographically separated but geochemically identical lavas may provide insights into common petrogenetic processes.

[1] Pedregosa *et al.* (2011), *Journal of Machine Learning Research* 12, 2825-2830.