Classification and Prediction of Lithologic Assemblages based on Soil Geochemistry and Geospatial Relationships

E.C. Grunsky[1], J. M. McKinley[2], and U.A. Mueller[3]
[1] Department of Earth and Environmental Sciences, University of Waterloo, Waterloo, Ontario, Canada, N2L 3G1.
egrunsky@gmail.com
[2]School of the Natural and Built Environment, Queen's University Belfast, BT7 1NN, UK
[3] School of Science, Edith Cowan University, Joondalup, Western Australia, WA 6027

Abstract

Regional scale soil geochemistry data have been used to characterize, classify and predict regional geological assemblages comprised of mixtures of igneous, metamorphic and sedimentary lithologies ranging from the Neoproterozoic to Cenozoic eras in Northern Ireland. Each of these eras have distinct lithologic assemblages comprised of clastic sediments, sedimentary carbonates, metamorphosed clastic sediments, basalts and granitoid rocks. Within the Phanerozoic eon, there is a significant amount of overlap of lithologies (sedimentary carbonates, wackes, sandstones). Despite this overlap, the stratigraphic assemblages of Northern Ireland can be subdivided into Age Brackets, which can be used as a basis for characterization and classification of the regional soil geochemical data.

The Northern Ireland soil geochemical dataset was collected with a spatial sampling density of 2 km^2 with continuous geospatial coverage across the region. The regional soil survey consists of 5,659 sample sites from which the following elements were evaluated: As, Au, B, Ba, Ca, Ce, Co, Cr, Cu, Fe, Ga, K, La, Mg, Mn, Mo, Nb, Ni, P, Pb, Pd, Pt, Rb, Se, Sn, Sr, Th, Ti, U, V, Y, Zn, Zr.

These data are compositional in nature and. as a result the individual elements are not independent. Logratio transforms were applied to the data to create a data space by which standard statistical methods can be applied. The centred logratio transform was applied to the data followed by a principal component analysis (PCA) and minimum/maximum autocorrelation factor analysis (MAF), as part of a method for characterizing and recognizing processes within the data ("process discovery") by which the unique multi-element assemblages define the Age Bracket package lithologies. Classification methods including linear discriminant analysis (LDA) and random forests, using cross-validation, were used to validate these Age Bracket classes ("process validation"). The method of LDA, using PCA and MAF showed an overall classification accuracy of 75%. Age Bracket classification using the method of random forests applied to the MAF scores, yield a higher classification accuracy that is close to 80%. High classification accuracies are associated with volcanic and granitic lithologies, while lower classification accuracies and higher confusion is associated within and between the Phanerozoic clastic assemblages. Co-kriging of the posterior probabilities derived from the application of linear discriminant analysis and kriging of random forest votes for each Age Bracket demonstrate the geospatial coherence of the classifications. This study confirms that the unique association of multi-element soil geochemistry, within a compositional framework, represents distinct lithologic assemblages that are geospatially coherent.