A new statistical method to identify geochemical data structure

*Hikaru Iwamori[1], Kenta Yoshida[1], Hitomi Nakamura[1,2,3], Tatsu Kuwatani[1], Morihisa Hamada[1], Satoru Haraguchi[1], Kenta Ueki[4]

1. Solid Earth Geochemistry, Japan Agency for Marine-Earth Science and Technology, 2. Department of Earth and Planetary Sciences, Tokyo Institute of Technology, 3. Ocean Resources Research Center for Next Generation, Chiba Institute of Technology, 4. Earthquake Research Institute, The University of Tokyo

Identifying the data structure including trends and groups/clusters in geochemical problems is essential to discuss the origin of sources and processes from the observed variability of data. A rapidly increasing number and high dimensionality of recent geochemical data require efficient and accurate methods for capturing the data structure. For example, the two databases of GEOROC and PetDB contain ~382,000 sets of data in total. Jenner and O’ Neil [2012] provided analysis of 60 elements in 616 ocean floor basaltic glasses. The structure including trends and groups of these data cannot be identified by graphical methods (e.g., Harker diagrams and identifying trends/groups based on them). As will be demonstrated, even 2-dimensional data may be misinterpreted by graphical methods.

Here we propose a new multivariate statistical method that combines three conventional but powerful methods to capture the true structure of multivariate data [Iwamori et al., 2017, doi:10.1002/2016gc006663]; they are k-means cluster analysis (KCA), principal component analysis (PCA), and independent component analysis (ICA). The reasons for selecting the three methods are (i) KCA and PCA are probably the most fundamental yet powerful tools for multivariate analyses; (ii) ICA is not as common as PCA but is a unique tool for identifying hidden independent structures; and (iii) the three methods are newly found to be closely related and can be integrated to analyze the data effectively. In this study, we first describe the relationship of these three methods to elucidate the entire data structure based mainly on synthetic data. We apply this to a natural data set of isotopic compositions of basalts for which ICA has been performed. On the basis of the results, an effective combination of the methods is clarified, for which we provide an Excel program “KCA” at both doi:10.1002/2016gc006663 and http://dsap.jamstec.go.jp/ to allow readers to test and apply the program to individual problems.

Keywords: multivariate statistical analysis, cluster analysis, principal component analysis, independent component analysis, geochemical data