Multivariate statistics for hydrogeology: moving forward from "the present is the key to the past"

Irawan, D.E.1,a) and Rachmi, C.N. 2

1Applied Geology Research Division,

Faculty of Earth Sciences and Technology,

Institut Teknologi Bandung   
2Department of Child and Adolescent Health,

University of Sydney Clinical School, Australia

a)Corresponding author: dasaptaerwin@outlook.co.id

**Abstract.** Geology is one of the oldest science in the world. Originated from natural science, it grows from the observation of sea shells to the sophisticated interpretation of the earth interior. On recent development geological approach need to be more quantitative, related to the needs prediction and simulation. Geology has shifted from ‘the present is the key to the past’ towards ‘the present is the key to the past and also the future’. Hydrogeology is one of the promising branch of geology that relies more to quantitative analysis. Multivariate statistics is one of the most frequently used resources in this field.

We did some literature search and web scraping to analyze current situation and future trend of multivariate statistics application for geological synthesis. We used key words ‘(all in title) multivariate statistics (and) hydrogeology’ on Google Scholar and Crossref database. The final result was 118 papers. We used VosViewer and Zotero to do some text mining operations.

Based on the analysis we can draw some results. Cluster analysis and principal component analysis are still the most frequently used method in hydrogeology. Both are mostly used to extract hydrochemical and isotope data to analyze the hydrogeological nature of groundwater flow. More machine learning methods have been introduced in the last five years in hydrogeological science. `Random forest` and `decision tree` technique are used extensively to learn the from physical and chemical properties of groundwater. Open source tools have also shifted the use of major statistical or programming language such as: SAS and Matlab. Python and R programming are the two famous open source applications in this field. We also note the increase of papers to discuss hydrogeology and public health sector. Therefore such methods are also being used to analyze open demographic data like DHS (demographic health survey) and FLS (Family Life Survey). Strong community of programmer makes the exponential development of both languages, via platform like Github. This has become the future of hydrogeology.

**Keywords:** Multivariate statistics, Python, R, machine learning.