Making sense of 'Big Data' in provenance studies

PIETER VERMEESCH[1]

[1]University College London, p.vermeesch@ucl.ac.uk

There is a lot to do on the Internet about the concept of 'Big Data', in which huge online databases are 'mined' to reveal previously hidden trends and relationships in society. One could argue that sedimentary geology has entered a similar era of 'Big Data', as modern provenance studies routinely use multiple proxies to dozens of samples, resulting in large multivariate datasets comprising thousands of data points. Just like the Internet, sedimentary geology now requires specialised statistical tools to visualise and interpret such large datasets. These can be organised on three distinct levels of progressively higher order:

1. A single sample: The most effective way to reveal the provenance information contained in a representative sample of detrital zircon U-Pb ages are probability density estimators such as histograms and kernel density estimates. The widely popular 'probability density plots' implemented in IsoPlot and AgeDisplay compound analytical uncertainty with geological scatter and are therefore invalid [1].

2. Several samples: Multi-panel diagrams comprising many detrital age distributions or compositional pie charts quickly become unwieldy and uninterpretable. For example, if there are N samples in a study, then the number of pairwise comparisons between samples increases quadratically as N(N-1)/2. This is simply too much information for the human eye to process. To solve this problem, it is necessary to (a) express the 'distance' between two samples as a simple scalar and (b) combine all N(N-1)/2 such values in a single two-dimensional 'map', grouping similar and pulling apart dissimilar samples. This can be easily achieved using simple statistics-based dissimilarity measures (e.g. the Kolmogorov-Smirnov statistic) and a standard statistical method called Multidimensional Scaling (MDS) [2].

3. Several methods: Suppose that we use four provenance proxies: bulk petrography, chemistry, heavy minerals and detrital geochronology. This will result in four MDS maps, each of which likely show slightly different trends and patterns. To deal with such cases, it may be useful to use a related technique called 'three way multidimensional scaling'. This results in two graphical outputs: an MDS map, and a map with 'weights' showing to what extent the different provenance proxies influence the horizontal and vertical axis of the MDS map. Thus, detrital data can not only inform the user about the provenance of sediments, but also about the causal relationships between the mineralogy, geochronology and chemistry.

[1] Vermeesch, 2012, Chem. Geol. v312, p190-194. [2] Vermeesch, 2013, Chem. Geol. v341, p140-146.