Chemical Geology 341 (2013) 140–146

Contents lists available at SciVerse ScienceDirect

Chemical Geology

journal homepage: www.elsevier.com/locate/chemgeo

Multi-sample comparison of detrital age distributions

Pieter Vermeesch *

Department of Earth Sciences, University College London, Gower Street, London WC1E 6BT, UK

ARTICLE INFO

Article history:
Received 4 October 2012
Received in revised form 15 January 2013
Accepted 18 January 2013
Available online 28 January 2013

Editor: K. Mezger

Keywords:
Zircon
Provenance
U–Pb dating
Geochronology
Statistics

ABSTRACT

The petrography and geochronology of detrital minerals form rich archives of information pertaining to the provenance of siliclastic sediments. The composition and age spectra of multi-sample datasets can be used to trace the flow of sediments through modern and ancient sediment routing systems. Such studies often involve dozens of samples comprising thousands of measurements. Objective interpretation of such large datasets can be challenging and greatly benefits from dimension-reducing exploratory data analysis tools. Principal components analysis (PCA) is a proven method that has been widely used in the context of compositional data analysis and traditional heavy mineral studies. Unfortunately, PCA cannot be readily applied to geochronological data, which are rapidly overtaking petrographic techniques as the method of choice for large scale provenance studies. This paper proposes another standard statistical technique called multidimensional scaling (MDS) as an appropriate tool to fill this void. MDS is a robust and flexible superset of PCA which makes fewer assumptions about the data. Given a table of pairwise 'dissimilarities' between samples, MDS produces a 'map' of points on which 'similar' samples cluster closely together, and 'dissimilar' samples plot far apart. It is shown that the statistical effect size of the Kolmogorov–Smirnov test is a viable dissimilarity measure. This is not the case for the p-values of this and other tests. To aid in the adoption of the method by the geochronological community, this paper includes some simple code using the statistical programming language R. More extensive software tools are provided on http://mudisc.london-geochron.com.

1. Introduction

Ever since the development of single grain U–Pb dating by (ion and laser) microprobe analysis, the method has been applied to detrital zircon (DZ) as a means of reconstructing the provenance of siliciclastic rocks. Initially, DZ geochronology was primarily used to trace the provenance of such rocks back to individual 'protosources' or source terranes (Gehrels et al., 1995; Pell et al., 1997). But in recent years, the ever-increasing throughput and ever decreasing cost of DZ geochronology have enabled a more sophisticated kind of applications, in which the U–Pb age distributions of multiple samples are used as a characteristic 'fingerprint' to trace the flow of zircon grains through the sediment routing system.

This paper introduces methods that make the interpretation of such datasets more objective, using a recently published provenance study from China as an example. Stevens et al. (in press) present a dataset comprising ten sand(stone) samples from the Mu Us desert, a Quaternary loess sample, a modern fluvial sand sample from the Yellow River, and a dataset of DZ ages from the Tibetan headwaters of the Yellow River taken from Pullen et al. (2011). The degree of similarity between these samples can be assessed on a qualitative basis by jointly plotting their respective age spectra (Fig. 1). Another commonly used visual aid is the so-called 'QQ plot', in which various quantiles of the samples are plotted against each other, the idea being that two samples follow an identical distribution if and only if their quantiles plot on a 1:1-line (Fig. 2).

Both the QQ plots and the age spectra can become unwieldy if they contain more than a dozen or so samples. For example, Fig. 1 contains $n = 13$ kernel density estimates (KDEs, Vermeesch, 2012) showing the probability distributions of 2025 single grain age estimates, while the QQ-plots in Fig. 2 form an upper triangular matrix with $n(n-1)/2 = 78$ pairwise comparisons. This is simply too much information for the human eye to process. To solve this problem, we need a 'filter' removing the redundant features of the individual distributions while preserving and amplifying the significant differences between them. This paper makes the case that a standard statistical technique called multidimensional scaling (MDS) can be used effectively for this purpose (Sections 3 and 4).

In addition to the DZ ages, all but one (T) of the samples in the Chinese study were subjected to heavy mineral (HM) analysis. With the exception of samples 1 and 8, the HM analyses were performed on separate aliquots from the U–Pb measurements. For samples 1 and 8, the HM mounts were prepared by mixing leftover mineral separates from the DZ study. Between 201 and 419 grains were counted in the 63–250 μm size fraction of each sample, resulting in an additional 2901 datapoints. Part of the aim of this paper is to treat these categorical data on an equal footing with

* Tel.: +44 20 7679 2418.
E-mail address: p.vermeesch@ucl.ac.uk.

0009-2541/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemgeo.2013.01.010

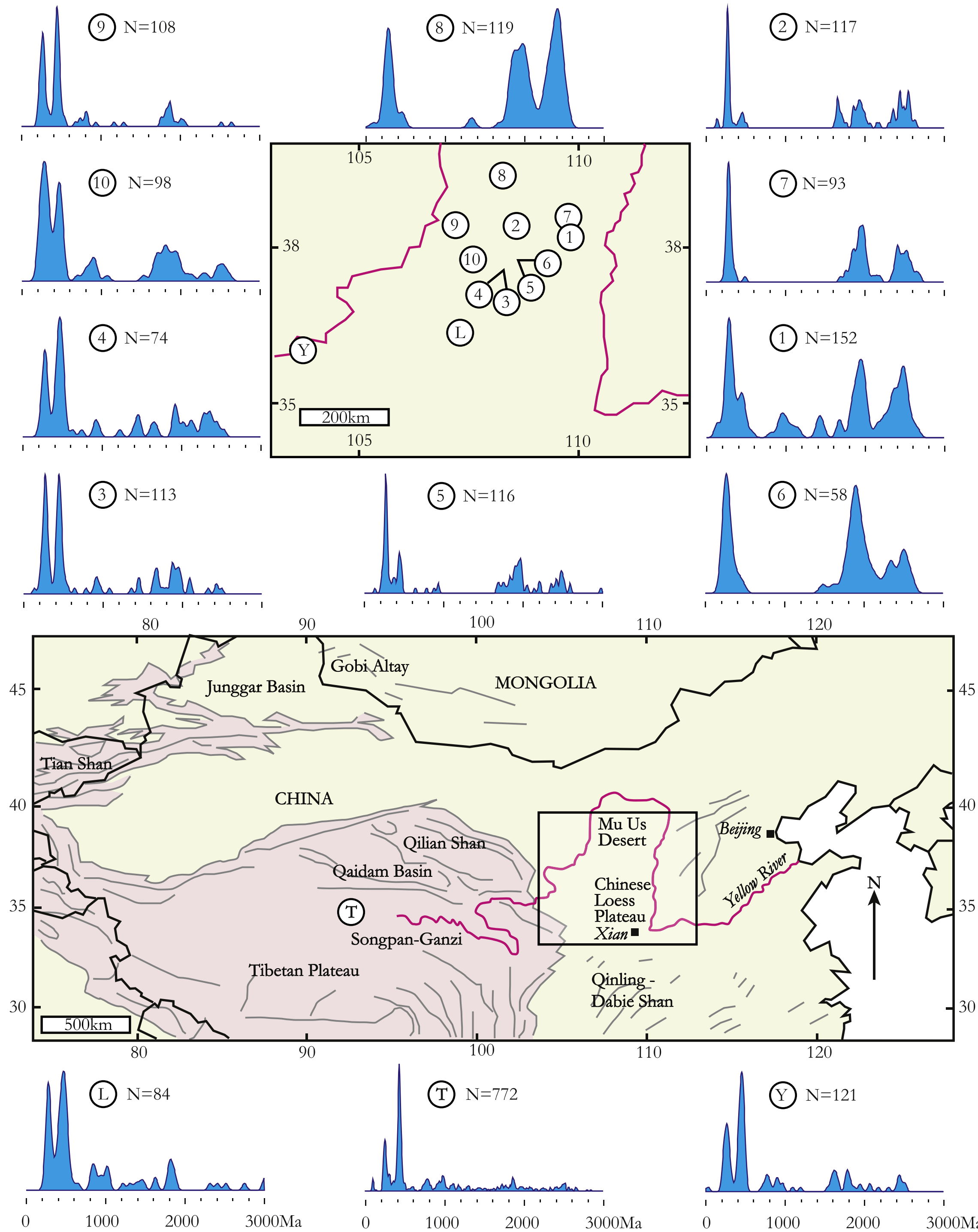

Fig. 1. Sample locations and kernel density estimates (KDEs, Vermeesch, 2012) of the Chinese loess study (N = number of concordant ages). Samples 1–5, 7 and 9–10 — dune sand; samples 6 and 8 — Cretaceous sandstone; sample L — Quaternary loess; sample Y — fluvial sand from the Yellow River; sample T — compilation of proposed sources in Yellow River headwaters from Pullen et al. (2011).

the continuous age data in a consistent mathematical framework (Section 5). All the analyses presented in this paper can be reproduced using the software discussed in Section 6 and made available on http://mudisc.london-geochron.com.

2. Measuring the dissimilarity between two samples

Before discussing multi-sample comparisons, it is useful to review the underlying principles for measuring the dissimilarity ($\delta_{i,j}$) between two samples (i and j, say). It is desirable for any dissimilarity measure to fulfil the following four requirements:

$$\delta_{i,j} \text{ should independent of sample size } N \tag{1}$$

$$\delta_{i,j} = 0 \text{ if } i = j \text{ and } \delta_{i,j} > 0 \text{ otherwise (nonnegativity)} \tag{2}$$

$$\delta_{i,j} = \delta_{j,i} \quad \text{(symmetry)} \tag{3}$$