

Accepted Manuscript

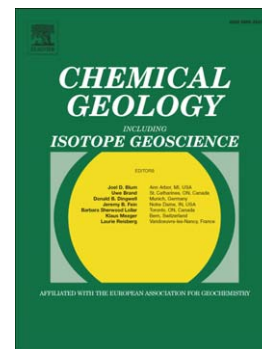
Making geological sense of 'Big Data' in sedimentary provenance analysis

Pieter Vermeesch, Eduardo Garzanti

PII: S0009-2541(15)00238-7
DOI: doi: [10.1016/j.chemgeo.2015.05.004](https://doi.org/10.1016/j.chemgeo.2015.05.004)
Reference: CHEMGE 17574

To appear in: *Chemical Geology*

Received date: 14 January 2015
Revised date: 30 April 2015
Accepted date: 5 May 2015



Please cite this article as: Vermeesch, Pieter, Garzanti, Eduardo, Making geological sense of 'Big Data' in sedimentary provenance analysis, *Chemical Geology* (2015), doi: [10.1016/j.chemgeo.2015.05.004](https://doi.org/10.1016/j.chemgeo.2015.05.004)

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Making geological sense of ‘Big Data’ in sedimentary provenance analysis

Pieter Vermeesch^{a*} and Eduardo Garzanti^b

^aLondon Geochronology Centre, University College London, United Kingdom

^bLaboratory for Provenance Studies, Università di Milano-Bicocca, Italy

April 30, 2015

Abstract

Sedimentary provenance studies increasingly apply multiple chemical, mineralogical and isotopic proxies to many samples. The resulting datasets are often so large (containing thousands of numerical values) and complex (comprising multiple dimensions) that it is warranted to use the internet-era term ‘Big Data’ to describe them. This paper introduces Multidimensional Scaling (MDS), Generalised Procrustes Analysis (GPA) and Individual Differences Scaling (INDSCAL, a type of ‘3-way MDS’ algorithm) as simple yet powerful tools to extract geological insights from ‘Big Data’ in a provenance context. Using a dataset from the Namib Sand Sea as a test case, we show how MDS can be used to visualise the similarities and differences between 16 fluvial and aeolian sand samples for five different provenance proxies, resulting in five different ‘configurations’. These configurations can be fed into a GPA algorithm, which translates, rotates, scales and reflects them to extract a ‘consensus view’ for all the data considered together. Alternatively, the five proxies can be jointly analysed by INDSCAL, which fits the data with not one but two sets of coordinates: the ‘group configuration’, which strongly resembles the graphical output produced by GPA, and the ‘source weights’, which can be used to attach geological meaning to the group configuration. For the Namib study, the three methods paint a detailed and self-consistent picture of a sediment routing system in which sand composition is determined by the combination of provenance and hydraulic sorting effects.

keywords: provenance – statistics – sediments – U-Pb – zircon – heavy minerals

1 Introduction

Some 65% of Earth’s surface is covered by siliclastic sediments and sedimentary rocks. Unravelling the provenance of these materials is of key importance to understanding modern sedimentary environments and their ancient counterparts, with important applications for geomorphology, paleotectonic and paleogeographic reconstructions, hydrocarbon exploration and reservoir characterization, and even forensic science (e.g., Pye, 2007; Vermeesch et al., 2010; Garzanti et al., 2012, 2014a,b; Stevens et al., 2013; Nie et al., 2014; Scott et al.,

*corresponding author, email: p.vermeesch@ucl.ac.uk, tel: +44 (0)20 7679 2428

29 2014). Over the years, thousands of studies have used a plethora of chemical, mineralogical and isotopic
30 indicators to trace sedimentary provenance. The complexity of the resulting datasets can be organised on a
31 number of hierarchical levels:

32 1. A single sample

33 Siliclastic sediments are made of grains, and on the most basic level, geological provenance analysis
34 extracts certain properties from these grains. These properties can either be categorical (e.g. mineral-
35 ogy) or continuous (e.g., age). In rare cases, analysing just a single grain can already yield important
36 insight into the provenance of a sediment. For example, a single grain of alluvial diamond confirms the
37 existence of kimberlitic lithologies in the hinterland. In general, however, provenance studies require
38 not just one but many grains to be analysed. The provenance information contained in a representative
39 collection of grains can be visualised with graphical aids such as histograms, pie charts or kernel density
40 estimates (Vermeesch, 2012).

41 2. Multiple samples

42 Subjective comparison of detrital zircon U-Pb age distributions or heavy mineral compositions reveals
43 the salient similarities and differences between two samples. Things become more complicated when
44 more than two samples need to be compared simultaneously. For example, a dataset comprising $n =$
45 10 age distributions presents the observer with $n(n-1)/2 = 45$ pairwise comparisons. If $n = 100$, this
46 increases quadratically to 4,950 pairwise comparisons, which is clearly too much for the human brain
47 to process. Multi-Dimensional Scaling (MDS) is a technique aimed to simplify this exercise (Section
48 3). Originating from the field of psychology, the method is commonly used in ecology (Kenkel and
49 Orlóci, 1986) and palaeontology (e.g., Dunkley Jones et al., 2008; Schneider et al., 2011). MDS was
50 introduced to the provenance community by Vermeesch (2013), and has instantly proved its value for
51 the interpretation of large datasets (e.g., Stevens et al., 2013; Nie et al., 2014).

52 3. Multiple methods

53 Several provenance methods are in use today which can be broadly categorised into two groups. Each
54 of these tells a different part of the provenance story:

- 55 (a) **Multi-mineral** techniques such as heavy mineral analysis and bulk geochemistry provide ar-
56 guably the richest source of provenance information, but are susceptible to hydraulic sorting
57 effects during deposition as well as chemical dissolution by diagenesis and weathering (Garzanti
58 et al., 2009; Andò et al., 2012). These effects obscure the provenance signal and can be hard to
59 correct.
- 60 (b) **Single mineral** techniques such as detrital zircon U-Pb geochronology are less sensitive to hy-
61 draulic sorting effects and, in the case of zircon, scarcely affected by secondary processes as well.
62 However, zircon is ‘blind’ to sediment sources such as mafic volcanic rocks and carbonates. Fur-
63 thermore, the robustness of zircon comes at a price, as it is difficult to account for the effect of
64 sediment recycling (Garzanti et al., 2013).

65 Great benefits arise when these two types of methods are used in tandem. A string of recent studies
66 combining conventional bulk and heavy mineral petrography techniques with detrital geochronology
67 have shown that this provides a very powerful way to trace provenance (e.g., Stevens et al., 2013;

68 Garzanti et al., 2012, 2014a,b). Combining multiple methods adds another level of complexity which
69 requires an additional layer of statistical simplification. The datasets resulting from these multi-sample,
70 multi-method studies are so large and complex that it is warranted to use the internet-era term ‘Big
71 Data’ to describe them. This paper introduces Procrustes analysis (Section 4) and 3-way MDS (Section
72 5) as valuable tools to help make geological sense of ‘Big Data’. These methods will be applied to a
73 large dataset from the Namib Sand Sea, which combines 16 samples analysed by 5 different methods
74 (Section 2). Although the use of some mathematical equations was inevitable in this paper, we have
75 made the text as accessible as possible by reducing the algorithms to their simplest possible form. The
76 formulas given in Sections 3-5 should therefore be considered as conceptual summaries rather than
77 practical recipes, with further implementational details deferred to the Appendices.

78 2 The Namib dataset

79 The statistical methods introduced in this paper will be illustrated with a large dataset from Namibia. The
80 dataset comprises fourteen aeolian samples from the Namib Sand Sea and two fluvial samples from the
81 Orange River (Figure 1). These samples were analysed using five different analytical methods:

- 82 1. Geochronology: ~100 zircon U-Pb ages were obtained per sample by LA-ICP-MS. For samples N1-N13,
83 this was done using methods described by Vermeesch et al. (2010). N14, T8 and T13 are new samples
84 which were analysed at the London Geochronology Centre using an Agilent 7700x ICP-MS coupled to
85 a New Wave NWR193 excimer laser with standard two volume ablation cell.
- 86 2. Heavy minerals: a full description of samples N1-N14 was given by Garzanti et al. (2012). Samples T8
87 and T13 were reported (as samples S4328 and S4332) by Garzanti et al. (2014a,b).
- 88 3. Bulk petrography: is also taken from Garzanti et al. (2012, 2014a,b).
- 89 4. Major element composition: 10 major elements were measured by acid dissolution (Aqua Regia) ICP-
90 ES at AcmeLabs Inc. in Vancouver, Canada (protocol 4A/B).
- 91 5. Trace element composition: 27 trace elements were measured by acid dissolution (Aqua Regia) ICP-ES
92 and ICP-MS at AcmeLabs (protocol 4A/B).

93 The complete dataset is available as an Online Supplement in a tabular form that can be imported into
94 the software discussed later in this paper. Taken altogether, the entire dataset contains 16,125 physical
95 measurements covering a variety of ordinal and compositional spaces. This is a prime example of ‘Big Data’
96 in a provenance context. A lot can be learned by a simple qualitative analysis of the measurements. For
97 example, the zircon age distributions reveal prominent peaks at ~600 and ~1,000 Ma, consistent with a
98 hinterland affected by Damara and Namaqua orogenesis, while the widespread occurrence of pyroxene and
99 basaltic rock fragments indicates the existence of a volcanic sediment source (Garzanti et al., 2012, 2014a).
100 But it is difficult to go beyond these general observations without statistical assistance because there is
101 simply ‘too much’ data. In the following sections, we will follow the hierarchical organisation of Section 1
102 to gain a better understanding of the multivariate dataset in different steps. First, we will integrate the
103 different age distributions and compositions into five MDS maps (Section 3). Then, we will integrate these
104 MDS maps into a single ‘Procrustes analysis’ (Section 4). Finally, we will jointly analyse the five datasets
105 using ‘3-way MDS’ to gain further insight into the sediment routing system (Section 5).

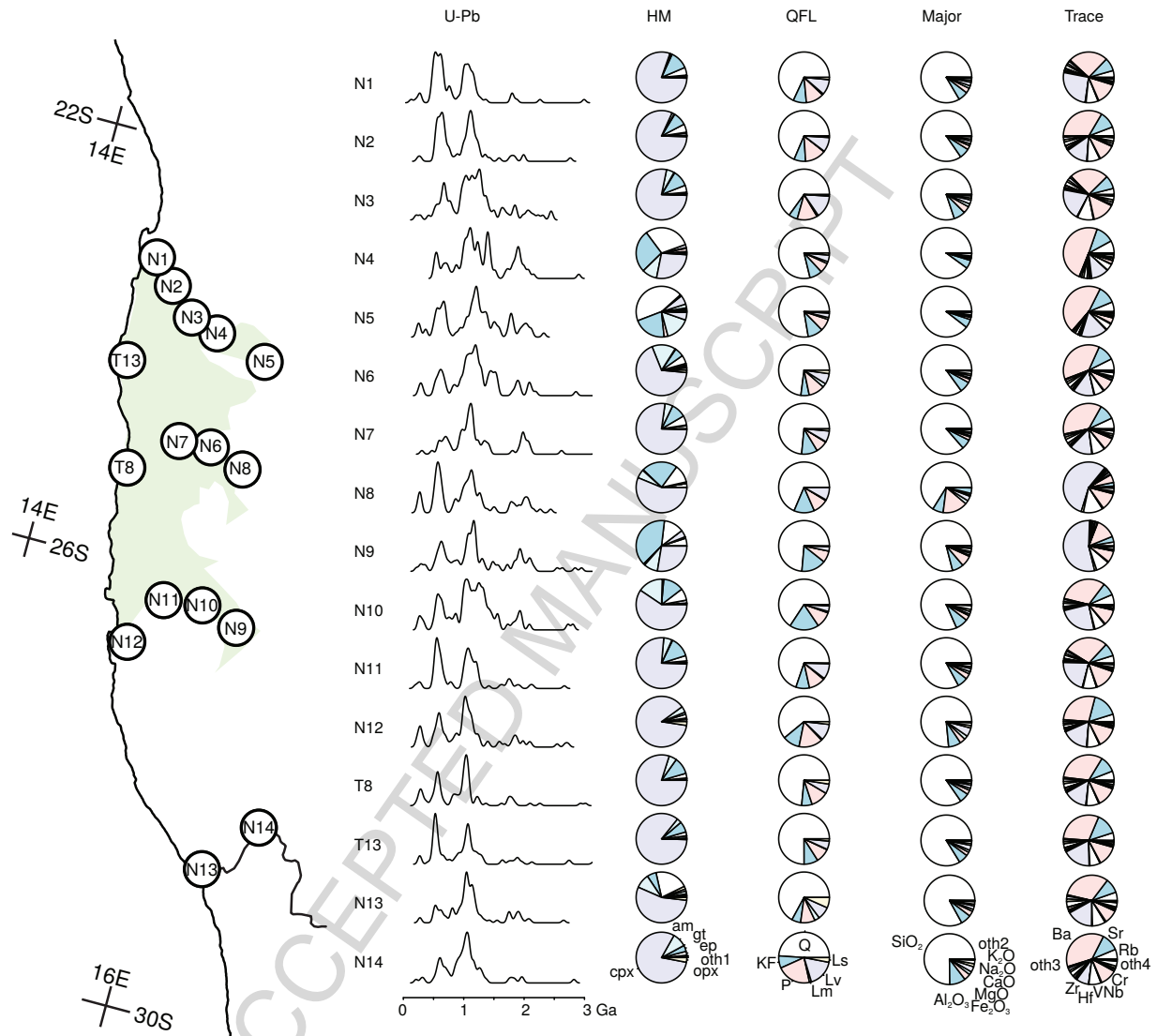


Figure 1: The Namib dataset comprises of 1,533 detrital zircon U-Pb ages (shown as kernel density estimates with a bandwidth of 30 Ma, Vermeesch, 2012), 3,600 heavy mineral counts ('HM'), 6,400 petrographic point counts ('QFL'), and chemical concentration measurements for 10 major and 27 trace elements. 'opx' = orthopyroxene, 'cpx' = clinopyroxene, 'am' = amphibole, 'gt' = garnet, 'ep' = epidote, 'oth1' = zircon + tourmaline + rutile + Ti-oxides + sphene + apatite + staurolite; 'Q' = quartz, 'KF' = K-feldspar, 'P' = plagioclase, 'Lm', 'Lv' and 'Ls' are lithic fragments of metamorphic, volcanic and sedimentary origin, respectively; 'oth2' = $\text{TiO}_2 + \text{P}_2\text{O}_5 + \text{MnO}$; 'oth3' = Sc + Y + La + Ce + Pr + Nd + Sm + Gd + Dy + Er + Yb + Th + U, and 'oth4' = Cr + Co + Ni + Cu + Zn + Ga + Pb. This figure makes the point that an objective interpretation of a large database like this is impossible without the help of statistical aids.

106 3 Multidimensional Scaling

107 The Namib study contains 16 samples, which can be visualised as kernel density estimates (for the U-Pb
108 data) or pie charts/histograms (for the compositional datasets). For each of the five provenance proxies,

109 we have $16 \times 15 / 2 = 120$ pairwise comparisons, which is clearly too much to handle for an unaided human
 110 observer (Figure 1). Multidimensional Scaling (MDS) is a technique aimed to simplify the interpretation of
 111 such large datasets by producing a simple two-dimensional map in which ‘similar’ samples plot close together
 112 and ‘dissimilar’ samples plot far apart. The technique is rooted in the field of psychology, in which human
 113 observers are frequently asked to make a subjective assessment of the dissimilarity between ‘stimuli’ such as
 114 shapes, sounds, flavours etc. A classic example of this is the colour-vision experiment of Helm (1964), which
 115 recorded the perceived differences between 10 colours by a human observer, resulting in a 9×9 dissimilarity
 116 matrix. Let $\delta_{i,j}$ be the ‘dissimilarity’ between two colours i and j (‘red’ and ‘blue’, say). Then MDS aims to
 117 find a monotone ‘disparity transformation’ f

$$f(\delta_{ij}) = \delta'_{ij} \quad (1)$$

118 and a configuration¹ X

$$X = \begin{bmatrix} x_1 & x_2 & \cdots & x_i & \cdots & x_j & \cdots & x_n \\ y_1 & y_2 & \cdots & y_i & \cdots & y_j & \cdots & y_n \end{bmatrix} \quad (2)$$

119 so as to minimise the (‘raw’) stress S

$$S = \sum_{i < j} \left[\delta'_{ij} - \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \right]^2 \quad (3)$$

120 The (x,y)-coordinates resulting from Equation 2 can be plotted as a map which, in the case of the Helm
 121 (1964) dataset, reveals the well-known colour circle (Figure 2a). Exactly the same principle can be used
 122 for geological data with, of course, dissimilarities not based on subjective perceptions but analytical data.
 123 There is a rich literature documenting ways to quantify the dissimilarity between petrographic or geochem-
 124 ical datasets. Further details about this are provided in Appendix A.

125
 126 Applying these methods to the Namib dataset, we can convert the raw input data (Figure 1) into five
 127 dissimilarity matrices. For the purpose of this exercise, we have used the Kolmogorov-Smirnov statistic for
 128 the U-Pb data, the Bray-Curtis dissimilarity for the heavy mineral and bulk petrography data, and the
 129 Aitchison distance for the major and trace element compositions (see Appendix A for a justification of these
 130 choices). Each of the resulting dissimilarity matrices can then be fed into an MDS algorithm to produce
 131 five configurations (Figure 2). Note that, because the Bray-Curtis dissimilarity does not fulfil the triangle
 132 inequality, the petrographic and heavy mineral datasets cannot be analysed by means of classical MDS
 133 (Vermeesch, 2013). The MDS maps of Figure 2 were therefore constructed using a nonmetric algorithm (see
 134 Kruskal and Wish, 1978; Borg and Groenen, 2005; Vermeesch, 2013, for further details). It is important
 135 to note that nonmetric MDS merely aims to reproduce the ‘rank order’ of the input data, rather than the
 136 actual dissimilarities themselves (Kruskal, 1964; Borg and Groenen, 2005). Bearing this in mind, the five
 137 MDS maps representing the Namib dataset reveal some clear trends in the data.

138
 139 A first observation is that the coastal samples (N1, N2, N11, N12, T8 and T13) plot close together in
 140 all five MDS maps, with the easternmost samples (N4, N5, N8 and N9) plotting elsewhere. Second, the

¹In this paper we will only consider two-dimensional solutions, which simplifies the notation and interpretation. It is easy to generalise the equations to more than two dimensions.

141 Orange River samples (N13 and N14) tend to plot closer to the coastal samples than to the inland samples.
 142 And third, within the eastern group, the northern samples (N4 and N5) are generally found in a different
 143 direction from the southern samples (N8 and N9), relative to the coastal group. But in addition to these
 144 commonalities, there also exist notable differences between the five maps. Specific examples of this are the
 145 odd position of N14 in the bulk petrography configuration (Figure 2d), the different orientation of the major
 146 and trace element configurations (Figures 2e and 2f) and countless other minor differences in the absolute
 147 and relative inter-sample distances. Also note that not all five datasets fit their respective MDS configuration
 148 equally well. A ‘goodness of fit’ measure called ‘Stress-1’ can be obtained by normalising the ‘raw’ stress
 149 (Equation 3) to the sum of the squared fitted distances (Kruskal, 1964; Kruskal and Wish, 1978). The
 150 resulting Stress-1 values range from 0.02 to 0.07, indicating ‘excellent’ fits to some and ‘fair’ fits to other
 151 datasets (Figure 2b-f). The five MDS maps, then, present us with a multi-comparison problem similar to
 152 the one presented by Figure 1, with the only difference being that it does not involve multiple KDEs or pie
 153 charts, but multiple MDS maps. Making this multi-sample comparison more objective requires an additional
 154 layer of statistical simplification, in which all the data are pooled to produce a ‘consensus’ view.

155 4 Procrustes analysis

156 According to Greek mythology, Procrustes was an inn keeper who managed to fit all travellers to a single
 157 bed, regardless of their size or length, by stretching or amputation. Similarly, in a statistical context,
 158 a Procrustes arrangement can be found that resembles each of several MDS maps by a combination of
 159 stretching, translation, reflection and rotation. In mathematical terms, Generalised Procrustes Analysis
 160 (GPA, Gower, 1975; Gower and Dijksterhuis, 2004; Borg and Groenen, 2005) proceeds in a similar manner
 161 to the method laid out for MDS in Section 3. Given K sets of two-dimensional MDS configurations X_k (for
 162 $1 \leq k \leq K$)

$$X_k = \begin{bmatrix} x_{1k} & x_{2k} & \cdots & x_{ik} & \cdots & x_{nk} \\ y_{1k} & y_{2k} & \cdots & y_{ik} & \cdots & y_{nk} \end{bmatrix} \quad (4)$$

163 GPA aims to find a transformation g constituting of a combination of scale factors s_k , orthonormal
 164 transformation matrices T_k and translation matrices t_k (Borg and Groenen, 2005):

$$g(X_k) = s_k X_k T_k + t_k = X'_k = \begin{bmatrix} x'_{1k} & x'_{2k} & \cdots & x'_{ik} & \cdots & x'_{nk} \\ y'_{1k} & y'_{2k} & \cdots & y'_{ik} & \cdots & y'_{nk} \end{bmatrix} \quad (5)$$

165 and a ‘group configuration’ \bar{X}

$$\bar{X} = \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_i & \cdots & \bar{x}_j & \cdots & \bar{x}_n \\ \bar{y}_1 & \bar{y}_2 & \cdots & \bar{y}_i & \cdots & \bar{y}_j & \cdots & \bar{y}_n \end{bmatrix} \quad (6)$$

166 so as to minimise the least squares misfit SS :

$$SS = \sum_{k=1}^K \sum_{i=1}^n (x'_{ik} - \bar{x}_i)^2 + (y'_{ik} - \bar{y}_i)^2 \quad (7)$$

167 Applying this method to the five (i.e., $K=5$) Namib MDS maps of Figure 2 produces a Procrustes map
 168 (Figure 3) confirming the salient points raised in Section 3. The GPA analysis shows the dichotomy between

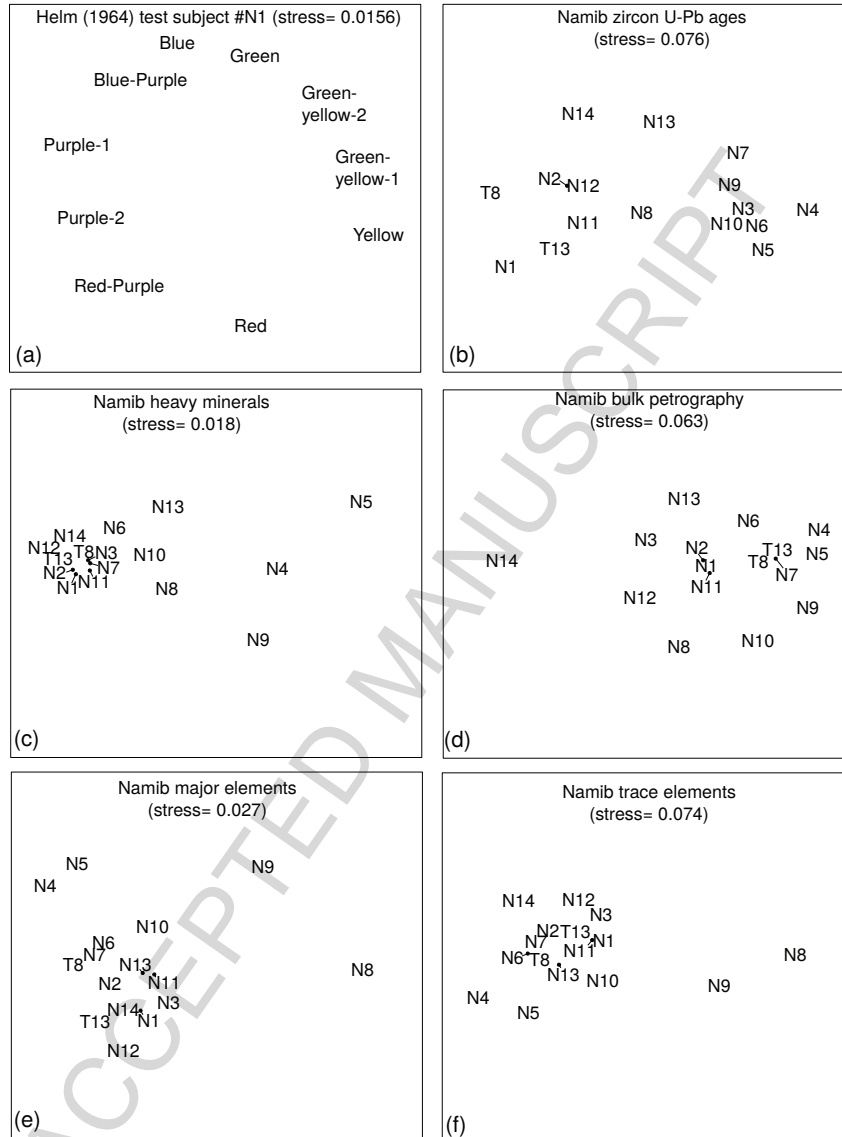


Figure 2: Nonmetric (2-way) MDS analyses of (a) Helm (1964)'s colour-vision data (for a single observer, 'N1') and (b)-(f) the five Namib datasets. The 'stress' values indicate 'excellent' (< 0.025) to 'fair' (< 0.1) fits (Kruskal and Wish, 1978; Vermeesch, 2013). Axes are plotted on a one-to-one scale with omitted labels to reflect the fact that non-metric MDS aims to preserve the ranks rather than the values of the dissimilarities. The MDS maps for the Namib dataset all paint a consistent picture in which (i) the coastal dune and Orange river samples (N1, N2, N11, N12, T8 and T13) plot close together and the inland samples (N4, N5, N8 and N9) plot elsewhere; and (ii) the northeastern samples (N4 and N5) are generally found in a different direction from the southeastern samples (N8 and N9), relative to the coastal group. However, there are also some distinct differences between the five configurations. The Procrustes and 3-way MDS analysis presented in Figures 3 and 4 make an abstraction of these differences.

169 the coastal and eastern sands, as well as the similarity of the coastal sands with the Orange River, and it
 170 does so more clearly than any of the five original MDS maps (Figure 2). It also emphasises the significance
 171 of the differences between the northeastern and southeastern samples, which plot at right angles from each
 172 other relative to the coastal samples. The GPA map, then, paints a detailed picture of the sediment routing
 173 system in the Namib Sand Sea, which would have been difficult to obtain from a simple visual inspection
 174 of the raw data. However, GPA weighs all five MDS configurations equally and does not readily take into
 175 account the significant differences in ‘goodness of fit’ (‘Stress-1, Section 3) between them. Also, although the
 176 trends and groupings among samples are clear from the GPA map, the underlying reasons for these features
 177 are not. The next section introduces a method aiming to solve this problem and thus yields additional insight
 178 into the sediment routing system of Namibia.

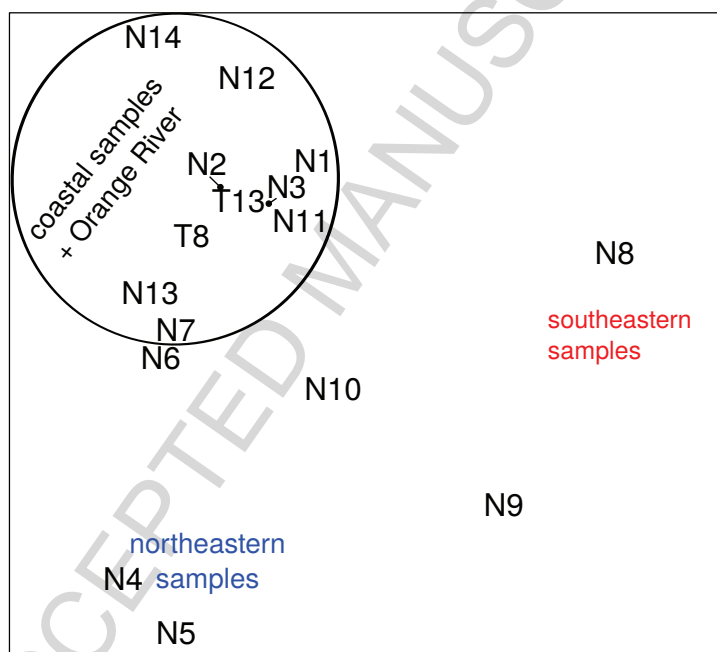


Figure 3: Generalised Procrustes Analysis (GPA) of the Namib dataset, pooling together all five MDS maps of Figure 2 into a single ‘average’ configuration. This confirms the strong similarities between sand samples collected along the Atlantic coast (N1, N2, N11, N12, T8, T13) and the Orange River (N13 and N14) as opposed to samples collected further inland (N4 through N10).

179 5 3-way MDS

180 As we saw in Section 4, Procrustes analysis is a two-step process. First, the various datasets are analysed by
 181 MDS. Then, the resulting MDS configurations are amalgamated into a single Procrustes map. The question
 182 then arises whether it is possible to skip the first step and go straight from the input data to a ‘group
 183 configuration’. Such methods exist under the umbrella of ‘3-way MDS’. In this paper, we will discuss the
 184 oldest and still most widely used technique of this kind, which is known as INdividual Differences SCALing
 185 (INDSCAL, Carroll and Chang, 1970). The method is formulated as a natural extension of the basic MDS

186 model outlined in Section 3. Given K dissimilarity matrices $\delta_{ij,k}$ ($1 \leq i, j \leq n$ and $1 \leq k \leq K$), INDSCAL
 187 aims to find K disparity transformations f_k

$$\delta'_{ij,k} = f_k(\delta_{ij,k}) \text{ (with } \sum_{i < j} \delta'^2_{ij,k} = \text{constant } \forall k), \quad (8)$$

188 a group configuration \bar{X} (defined as in Equation 6), and a set of dimension weights W

$$W = \begin{bmatrix} w_{x1} & w_{x2} & \cdots & w_{xk} & \cdots & w_{xK} \\ w_{y1} & w_{y2} & \cdots & w_{yk} & \cdots & w_{yK} \end{bmatrix} \quad (9)$$

189 so as to minimise a modified stress parameter S'

$$S' = \sum_{k=1}^K \sum_{i < j} \left[\delta'_{ij,k} - \sqrt{w_{xk}(x_i - x_j)^2 + w_{yk}(y_i - y_j)^2} \right]^2 \quad (10)$$

190 To illustrate the application of INDSCAL to real data, it is instructive to revisit the colour-vision exam-
 191 ple of Section 3. In addition to test subject ‘N1’ shown in Figure 2a, the study by Helm (1964) involved
 192 thirteen more participants. Each of these people produced one (or two, for subjects N6 and CD2) dissim-
 193 ilarity matrix(es), resulting in a total of sixteen MDS maps, which could in principle be subjected to a
 194 Procrustes analysis (Section 4). Alternatively, the sixteen dissimilarity matrices can also be fed into the
 195 INDSCAL algorithm. The resulting ‘group configuration’ (\bar{X}) is a map that fits the perceived differences of
 196 all fourteen observers by stretching and shrinking (but not rotating) in the x- and y-direction (Figure 4.a).
 197 The degree of stretching or shrinking associated with each observer is given by the ‘source weights’ (W),
 198 which can be plotted as a second piece of graphical output (Figure 4.b). For the colour-vision experiment,
 199 the group configuration shows the familiar colour circle, and the source weights express the degree to which
 200 this colour circle is distorted in the perception of the colour deficient test subjects (prefix ‘CD’) relative to
 201 those subjects with normal colour vision (prefix ‘N’). The latter all plot together in the northwest quadrant
 202 of the diagram, whereas the former plot in the southeast quadrant. Multiplying the x-y coordinates of the
 203 group configuration with the respective dimensions of the source weights yields sixteen ‘private spaces’, which
 204 are approximate MDS maps for each test subject. For the colour deficient subjects, these private spaces
 205 will have an oblate shape, emphasising the reduced sensitivity of the colour deficient test subjects to the
 206 red-green colour axis relative to the blue-yellow axis. In summary, whereas an ordinary MDS configuration
 207 can be rotated by an arbitrary angle without loss of information, this is not the case for an INDSCAL group
 208 configuration. The principal axes of the latter generally have an interpretive meaning, which is one of the
 209 most appealing aspects of the method (Arabie et al., 1987; Borg and Groenen, 2005).

210
 211 The five datasets of the Namibian study can be analysed in exactly the same manner as Helm (1964)’s
 212 colour data, producing the same two pieces of graphical output as before. The resulting ‘group configuration’
 213 (Figure 4c) looks remarkably similar to the GPA map of Figure 3. It shows the same separation between
 214 samples collected from the eastern and western parts of the desert, and the same 90° angle between the
 215 northeastern and southeastern sampling locations. But whereas the GPA map did not offer any explanation
 216 for these observations, the source weights of the INDSCAL analysis do provide some important clues (Figure
 217 4d). The provenance proxies based on the analysis of bulk materials (chemistry and petrography) attach
 218 stronger weights to the horizontal dimension. The proxies based on density separates (U-Pb ages and heavy
 219 minerals), on other hand, weigh the vertical dimension more heavily. Because the former proxies are more

220 sensitive to hydraulic sorting effects and comparatively less sensitive to provenance than the latter proxies
221 (see Section 1), this observation leads to the interpretation that hydraulic sorting (predominantly) separates
222 samples along the x-dimension, whereas the provenance signal (predominantly) separates samples along the
223 y-dimension.

224 6 Discussion, caveats and conclusions

225 Until recently, large multi-proxy provenance studies like the Namib case study presented in this paper were
226 prohibitively expensive and time consuming. However, continued technological advances in mass spectrometry
227 (Frei and Gerdes, 2009) and petrography/geochemistry (Allen et al., 2012) promise to change this picture.
228 In anticipation of the impending flood of provenance data resulting from these advances, this paper borrowed
229 some simple yet powerful ‘data mining’ techniques from other scientific disciplines, which help to make geo-
230 logical sense of complex datasets. Some readers will be familiar with Principal Components Analysis (PCA),
231 which is a dimension-reducing procedure that is commonly used to interpret geochemical, petrographic and
232 other compositional data (Aitchison, 1983; Vermeesch, 2013). Multidimensional Scaling is a flexible and
233 powerful superset of PCA which allows geologists to extend PCA-like interpretation to isotopic data such as
234 U-Pb ages (Vermeesch, 2013). Generalised Procrustes Analysis and Individual Differences Scaling are higher
235 order supersets of MDS which can be used to integrate multiple proxies in a single comprehensive analysis.

236
237 The application to the Namib Sand Sea has yielded results that are broadly consistent with previous
238 interpretations by visual inspection of the age distributions, petrographic diagrams etc. The statistical tools
239 presented in this paper offer two key advantages over the traditional approach. First, they are far more
240 objective and easy to use. Expert knowledge of mineralogy, petrography and isotope geochemistry, while
241 still desirable, becomes less crucial because the statistical tools automatically extract geologically meaningful
242 differences between the datasets. Second, the methods introduced in this paper provide a way to compare
243 datasets of very different nature in a common framework. Thus the new approach to data interpretation
244 makes it much easier to combine petrographic and isotopic provenance proxies.

245
246 Despite the intuitive appeal of INDSCAL and its apparent success in the Namib study, it is important
247 to mention a few caveats. Whereas the group configuration is quite robust (as exemplified by the similar-
248 ity of Figures 3 and 4d), the same cannot be said about the source weights. Consider, for example, the
249 INDSCAL analysis of the Namib data, which used a combination of Kolmogorov-Smirnov (for the U-Pb
250 data), Bray-Curtis (for the mineralogical data) and Aitchison (for the bulk chemistry) measures. Replacing
251 the Kolmogorov-Smirnov statistic with (Sircombe and Hazelton, 2004)’s L2-norm, say, results in a similar
252 group configuration but in significantly different source weights with a less clear interpretation (although the
253 bulk and density separated proxies still plot in opposite corners). The instability of the source weights may
254 easily lead to over-interpretation, causing some (e.g., Borg and Groenen, 2005) to recommend abandoning
255 INDSCAL in favour of GPA or similar techniques.

256
257 Thanks to the widespread acceptance of MDS, GPA and INDSCAL in other fields of science, several
258 software options are available (see Appendix B for details). These tools can be combined with other types of
259 inferential techniques such as cluster analysis, regression, bootstrapping etc. This paper barely scratches the
260 surface of the vast field of MDS-related research. We refer the user to the reference works by Arabie et al.

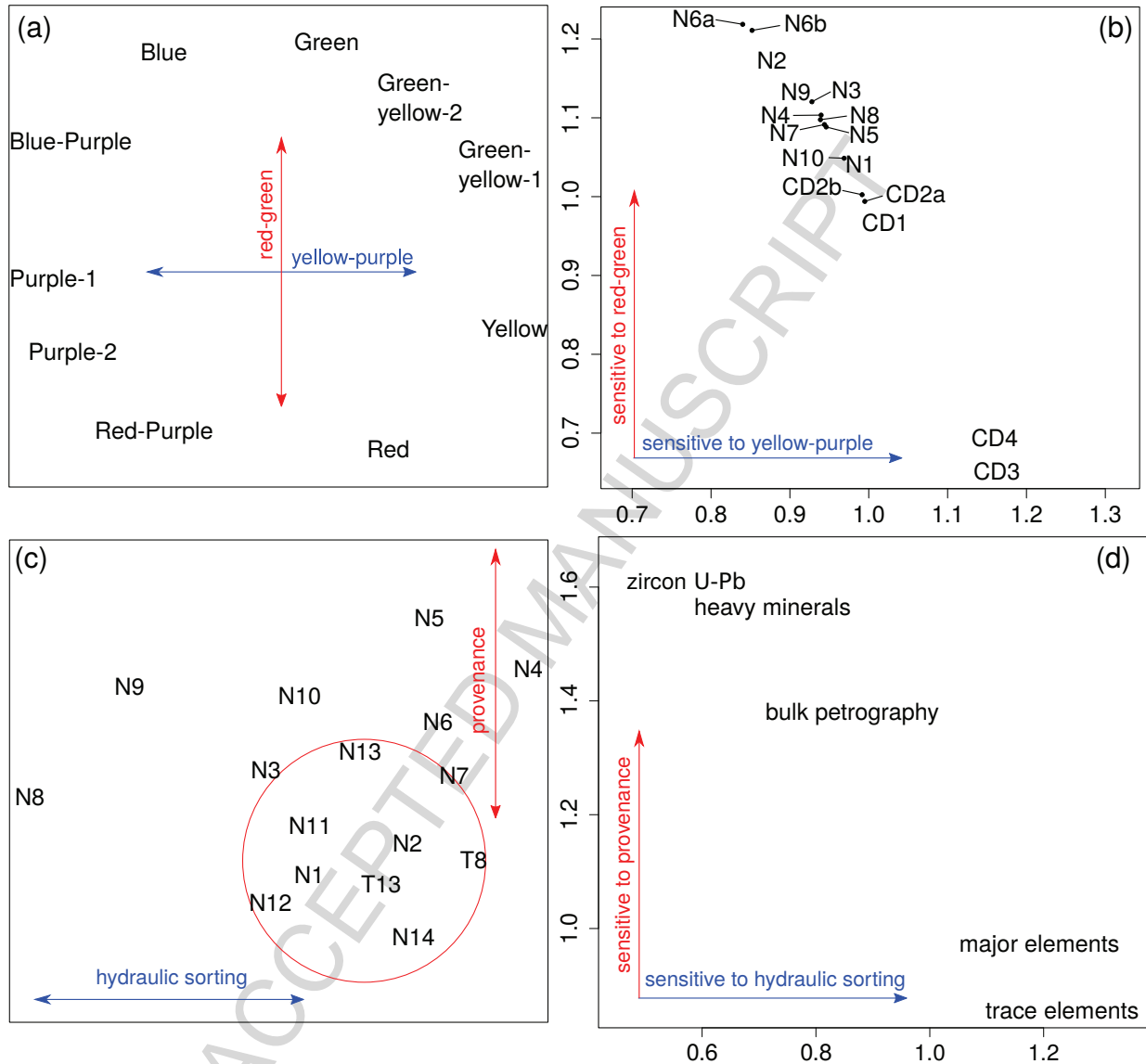


Figure 4: 3-way MDS analysis of the colour-vision experiment by Helm (1964, (a)-(b)) and the Namib dataset [(c)-(d)]. The left panels [(a) and (c)] show the ‘group configurations’, whereas the right panels [(b) and (e)] show the ‘source weights’. For the Namib dataset, the former shows essentially the same picture as the Procrustes analysis (Figure 3). The map of ‘source weights’ (d) shows the degree of importance each of the five proxies attach to the horizontal and vertical dimension of the group configuration. An intuitive interpretation of these two dimensions suggests that the y-axis shows the provenance signal (which dominates the proxies based on density separates, see Section 1), whereas the hydraulic sorting effect dominates the x-axis (and the bulk analysis proxies).

261 (1987); Borg and Groenen (2005); Borg et al. (2012); Gower and Dijksterhuis (2004) for further details and
 262 ideas and hope that our paper will encourage others to explore these extension in order to address a new

263 class of geological problems.

264 Acknowledgments

265 We would like to thank Ingwer Borg, Jan de Leeuw, Patrick Mair, Patrick Groenen, Christian Hennig and
 266 two anonymous reviewers for feedback and statistical advice. This research was funded by NERC grant
 267 #NE/1009248/1 and ERC grant 259505 ('KArSD').

268 Appendix A: dissimilarity measures

269 This section provides a few examples of dissimilarity measures to compare two sediment samples (A and B,
 270 say). Let us first consider the case of categorical data ($A = \{A_1, A_2, \dots, A_n\}$ and $B = \{B_1, B_2, \dots, B_n\}$,
 271 where A_i represents the number of observations of class i , etc.) such as heavy mineral counts. Vermeesch
 272 (2013) used Aitchisons central logratio distance:

$$\delta_{AB}^{ait} = \sqrt{\sum_{i=1}^n \left[\ln\left(\frac{A_i}{g(A)}\right) - \ln\left(\frac{B_i}{g(B)}\right) \right]^2} \quad (11)$$

273 where 'g(x) stands for 'the geometric mean of x (Aitchison, 1986; Vermeesch, 2013). Note that the same
 274 distance is obtained irrespective of whether the input data are expressed as fractions or percents. The
 275 Aitchison distance breaks down for datasets comprising 'zero counts' ($A_i = 0$ or $B_i = 0$ for any i). This
 276 problem can be solved by pooling several categories together, or by using a different dissimilarity measure
 277 such as the Bray-Curtis dissimilarity:

$$\delta_{AB}^{bc} = \frac{\sum_{i=1}^n |A_i - B_i|}{\sum_{i=1}^n (A_i + B_i)} \quad (12)$$

278 where $|\cdot|$ stands for the absolute value. Note that the Bray-Curtis dissimilarity does not fulfil the triangle
 279 inequality. It can therefore not be used for 'classical' MDS (in which the disparity transformation is the
 280 identity matrix, Vermeesch, 2013). However, this is not an issue for nonmetric MDS (as well as certain
 281 classes of metric MDS). For ordinal data such as U-Pb ages, it is useful to define the empirical cumulative
 282 distribution functions (CDFs):

$$F_A(t) = \frac{1}{n}(\#a_i \leq t) \text{ and } F_B(t) = \frac{1}{m}(\#b_i \leq t) \quad (13)$$

283 where n and m are the sample sizes of A and B, respectively and ' $\#x \leq t$ ' stands for "the number of items
 284 in x that are smaller than or equal to t ". The simplest CDF-based statistic was developed by Kolmogorov
 285 and Smirnov and uses the maximum absolute difference between $F_A(t)$ and $F_B(t)$ (Feller, 1948):

$$\delta_{AB}^{ks} = \max_t |F_A(t) - F_B(t)| \quad (14)$$

286 The Kolmogorov-Smirnov (KS) statistic takes on discrete values in steps of $|\frac{1}{n} - \frac{1}{m}|$ and may therefore yield
 287 dissimilarity measures with duplicate values, which in turn may cause problems in certain MDS algorithms.
 288 Furthermore, the KS-statistic is most sensitive to the region near the modes of the sample distribution, and

289 less sensitive to the tails. Finally, when $F_A(t)$ and $F_B(t)$ cross each other multiple times, the maximum
 290 deviation between them is reduced. Therefore, the KS-statistic (or variants thereof such as the Kuiper
 291 statistic) cannot ‘see’ the difference between a uniform distribution and a ‘comb’-like distribution. Although
 292 alternative statistics such as Cramér-von Mises and Anderson-Darling solve any or all of these problems,
 293 they generally exhibit an undesirable dependence on sample size. One promising alternative which does not
 294 suffer from this problem is the L2-norm proposed by Sircombe and Hazelton (2004). This measure explicitly
 295 takes into account the analytical uncertainties and may therefore be the preferred option when combining
 296 samples from different analytical sources.

297 Appendix B: software

298 The methods introduced in this paper are widely used in a variety of research fields, and several software
 299 options are available, including Matlab (Trendafilov, 2012), SPSS (PROXSCAL, Busing et al., 1997), PAST
 300 (Hammer and Harper, 2008) and R (De Leeuw and Mair, 2011). This section contains the shortest workable
 301 example of R code needed to reproduce the figures in this paper. The `BigData.Rdata` input file and a more
 302 general purpose code can be downloaded from <http://mudisc.london-geochron.com>.

```

303 library(MASS)           # performs nonmetric MDS
304 library(smacof)        # performs INDSCAL
305 library(shapes)        # performs GPA
306 library(robCompositions) # supplies the Aitchison distance
307 library(vegan)         # supplies the Bray-Curtis distance
308
309 load("BigData.Rdata")  # load the raw input data (DZ, HM, QFL, Major and Trace)
310 snames <- names(d$DZ)  # extract the list of sample names
311 n <- length(snames)    # n = the number of samples
312 m <- length(d)         # m = the number of datasets
313
314 # this function calculates the dissimilarity between age distributions
315 getDZdist <- function(dat,labels=smnames) {
316   n <- length(dat)
317   diss <- matrix(nrow=n,ncol=n,dimnames=list(snames,smnames))
318   for (i in 1:n){ for (j in 1:n){ # loop through the rows and columns
319     diss[i,j] <- ks.test(dat[[i]],dat[[j]])$statistic }}
320   return (as.dist(diss))          # convert to a 'distance' object
321 }
322
323 # calculate the dissimilarity matrices for each of the five datasets
324 DZdist <- getDZdist(d$DZ,labels=smnames)      # U-Pb data: KS statistic
325 QFLdist <- vegdist(d$QFL,'bray',labels=smnames) # bulk petrography: Bray-Curtis
326 HMdist <- vegdist(d$HM,'bray',labels=smnames) # heavy minerals: Bray-Curtis
327 MajorDist <- dist(cenLR(d$Major)$x.clr) # major elements: Aitchison distance
328 TraceDist <- dist(cenLR(d$Trace)$x.clr) # trace elements: Aitchison distance

```

```

329 distlist <- list(DZ=DZdist,QFL=QFLdist,HM=HMdist,Major=MajorDist,Trace=TraceDist)
330
331 # the following lines produce a GPA map
332 X <- array(dim=c(n,2,m)) # initialise the 3-way matrix of MDS configurations
333 for (i in 1:m) { # loop through all the datasets
334   X[, ,i] <- isoMDS(distlist[[i]],k=2)$points} # perform a nonmetric MDS analysis
335 pfit <- procGPA(X) # perform a GPA analysis
336 xp <- pfit$mshape[,1] # x-coordinates of the procrustes configuration
337 yp <- pfit$mshape[,2] # y-coordinates of the procrustes configuration
338 plot(xp,yp,type="n",asp=1) # create an empty plot (replace "n" with "p" to show points)
339 text(xp,yp,snames) # plot the procrustes configuration
340
341 # perform an INDSCAL analysis
342 ifit <- smacofIndDiff(distlist, constraint="indscal", type="ordinal")
343 dev.new() # open a new graphics window for the group configuration
344 plot(ifit,plot.type="confplot",asp=1) # plot the group configuration
345 dev.new() # open a new graphics window for the source weights
346 weights <- unlist(ifit$cweights) # extract the source weights
347 xw <- weights[4*seq(m)-3] # weights of the horizontal axis
348 yw <- weights[4*seq(m)] # weights of the vertical axis
349 plot(xw,yw,type="n",asp=1) # create an empty plot
350 text(xw,yw,names(d)) # plot the source weights

```

References

- 352 Aitchison, J., 1983. Principal component analysis of compositional data. *Biometrika* 70, 57–65.
353 doi:10.1093/biomet/70.1.57.
- 354 Aitchison, J., 1986. *The statistical analysis of compositional data*. London, Chapman and Hall.
- 355 Allen, J.L., Johnson, C.L., Heumann, M.J., Gooley, J., Gallin, W., 2012. New technology and methodology
356 for assessing sandstone composition: A preliminary case study using a quantitative electron microscope
357 scanner (QEMScan). *Geological Society of America Special Papers* 487, 177–194.
- 358 Andò, S., Garzanti, E., Padoan, M., Limonta, M., 2012. Corrosion of heavy minerals during weathering and
359 diagenesis: A catalog for optical analysis. *Sedimentary geology* 280, 165–178.
- 360 Arabie, P., Carroll, J.D., DeSarbo, W.S., 1987. *Three Way Scaling: A Guide to Multidimensional Scaling*
361 and Clustering. volume 65. Sage.
- 362 Borg, I., Groenen, P.J., 2005. *Modern multidimensional scaling: Theory and applications*. Springer.
- 363 Borg, I., Groenen, P.J., Mair, P., 2012. *Applied multidimensional scaling*. Springer.
- 364 Busing, F., Commandeur, J.J., Heiser, W.J., Bandilla, W., Faulbaum, F., 1997. Proxscal: A multidimen-
365 sional scaling program for individual differences scaling with constraints. *Softstat* 97, 67–74.

- 366 Carroll, J.D., Chang, J.J., 1970. Analysis of individual differences in multidimensional scaling via an N-way
367 generalization of Eckart-Young decomposition. *Psychometrika* 35, 283–319.
- 368 De Leeuw, J., Mair, P., 2011. Multidimensional scaling using majorization: SMACOF in R. Department of
369 Statistics, UCLA .
- 370 Dunkley Jones, T., Bown, P.R., Pearson, P.N., Wade, B.S., Coxall, H.K., Lear, C.H., 2008. Major shifts
371 in calcareous phytoplankton assemblages through the Eocene-Oligocene transition of Tanzania and their
372 implications for low-latitude primary production. *Paleoceanography* 23.
- 373 Feller, W., 1948. On the Kolmogorov-Smirnov limit theorems for empirical distributions. *The Annals of*
374 *Mathematical Statistics* 19, 177–189.
- 375 Frei, D., Gerdes, A., 2009. Precise and accurate *in situ* U–Pb dating of zircon with high sample throughput
376 by automated LA-SF-ICP-MS. *Chemical Geology* 261, 261–270.
- 377 Garzanti, E., Andò, S., Vezzoli, G., 2009. Grain-size dependence of sediment composition and
378 environmental bias in provenance studies. *Earth and Planetary Science Letters* 277, 422–432.
379 doi:10.1016/j.epsl.2008.11.007.
- 380 Garzanti, E., Andò, S., Vezzoli, G., Lustrino, M., Boni, M., Vermeesch, P., 2012. Petrology of the Namib
381 Sand Sea: Long-distance transport and compositional variability in the wind-displaced Orange Delta.
382 *Earth-Science Reviews* 112, 173 – 189. doi:10.1016/j.earscirev.2012.02.008.
- 383 Garzanti, E., Resentini, A., Andò, S., Vezzoli, G., Pereira, A., Vermeesch, P., 2014a. Physical controls on
384 sand composition and relative durability of detrital minerals during ultra-long distance littoral and aeolian
385 transport (Namibia and southern Angola). *Sedimentology* doi:10.1111/sed.12169.
- 386 Garzanti, E., Vermeesch, P., Andò, S., Lustrino, M., Padoan, M., Vezzoli, G., 2014b. Ultra-long distance
387 littoral transport of Orange sand and provenance of the Skeleton Coast Erg (Namibia). *Marine Geology*
388 357, 25–36.
- 389 Garzanti, E., Vermeesch, P., Andò, S., Vezzoli, G., Valagussa, M., Allen, K., Kadi, K.A., Al-Juboury, A.I.,
390 2013. Provenance and recycling of Arabian desert sand. *Earth-Science Reviews* .
- 391 Gower, J.C., 1975. Generalized procrustes analysis. *Psychometrika* 40, 33–51.
- 392 Gower, J.C., Dijksterhuis, G.B., 2004. Procrustes problems. volume 3. Oxford University Press Oxford.
- 393 Hammer, Ø., Harper, D.A., 2008. Paleontological data analysis. John Wiley & Sons.
- 394 Helm, C.E., 1964. Multidimensional ratio scaling analysis of perceived color relations. *JOSA* 54, 256–260.
- 395 Kenkel, N.C., Orłóci, L., 1986. Applying metric and nonmetric multidimensional scaling to ecological studies:
396 some new results. *Ecology* , 919–928.
- 397 Kruskal, J., 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psy-*
398 *chometrika* 29, 1–27.
- 399 Kruskal, J.B., Wish, M., 1978. Multidimensional scaling. volume 07-011 of *Sage University Paper series on*
400 *Quantitative Application in the Social Sciences*. Sage Publications, Beverly Hills and London.

- 401 Nie, J., Peng, W., Möller, A., Song, Y., Stockli, D.F., Stevens, T., Horton, B.K., Liu, S., Bird,
402 A., Oalmann, J., Gong, H., Fang, X., 2014. Provenance of the upper Miocene-Pliocene Red
403 Clay deposits of the Chinese loess plateau. *Earth and Planetary Science Letters* 407, 35 – 47.
404 doi:<http://dx.doi.org/10.1016/j.epsl.2014.09.026>.
- 405 Pye, K., 2007. *Geological and soil evidence: Forensic applications*. CRC Press.
- 406 Schneider, L.J., Bralower, T.J., Kump, L.R., 2011. Response of nanoplankton to early Eocene ocean
407 de-stratification. *Palaeogeography, Palaeoclimatology, Palaeoecology* 310, 152–162.
- 408 Scott, R.A., Smyth, H.R., Morton, A.C., Richardson, N. (Eds.), 2014. *Sediment Provenance Studies in*
409 *Hydrocarbon Exploration and Production*. volume 386 of *Geological Society, London, Special Publications*.
410 Geological Society of London.
- 411 Sircombe, K.N., Hazelton, M.L., 2004. Comparison of detrital zircon age distributions by kernel functional
412 estimation. *Sedimentary Geology* 171, 91–111. doi:[10.1016/j.sedgeo.2004.05.012](https://doi.org/10.1016/j.sedgeo.2004.05.012).
- 413 Stevens, T., Carter, A., Watson, T., Vermeesch, P., Andò, S., Bird, A., Lu, H., Garzanti, E., Cottam, M.,
414 Sevastjanova, I., 2013. Genetic linkage between the yellow River, the Mu Us desert and the Chinese Loess
415 Plateau. *Quaternary Science Reviews* 78, 355–368.
- 416 Trendafilov, N.T., 2012. Dindscal: direct INDSCAL. *Statistics and Computing* 22, 445–454.
- 417 Vermeesch, P., 2012. On the visualisation of detrital age distributions. *Chemical Geology* 312-313, 190–194.
418 doi:[10.1016/j.chemgeo.2012.04.021](https://doi.org/10.1016/j.chemgeo.2012.04.021).
- 419 Vermeesch, P., 2013. Multi-sample comparison of detrital age distributions. *Chemical Geology* 341, 140–146.
- 420 Vermeesch, P., Fenton, C.R., Kober, F., Wiggs, G.F.S., Bristow, C.S., Xu, S., 2010. Sand residence times
421 of one million years in the Namib Sand Sea from cosmogenic nuclides. *Nature Geoscience* 3, 862–865.
422 doi:[10.1038/ngeo985](https://doi.org/10.1038/ngeo985).

Highlights:

- This paper introduces statistical aids for interpreting multi-proxy provenance data.
- Multi-Dimensional Scaling (MDS) visualises multiple samples in one 'configuration'.
- Generalised Procrustes Analysis (GPA) pools several MDS configurations together.
- Individual Difference SCALing (INDSCAL) adds geological meaning to MDS maps.
- These methods were applied to a 5-proxy dataset from the Namib Sand Sea.