Copy 3
1972/123

COMMONWEALTH OF AUSTRALIA

DEPARTMENT OF NATIONAL DEVELOPMENT

BUREAU OF MINERAL RESOURCES, GEOLOGY AND GEOPHYSICS

RECORD NO. 1972/123

PLANE REPRESENTATION OF MULTIVARIATE DATA STRUCTURE

by

S. Henley

The information contained in this report has been obtained by the Department of National Development as part of the policy of the Commonwealth Government to assist in the exploration and development of ineral resources. It may not be published in any form or used in a company prospectus or statement ithout the permission in writing of the Director, Bureau of Mineral Resources, Geology & Geophysics.

RECORD NO. 1972/123

PLANE REPRESENTATION OF
MULTIVARIATE DATA STRUCTURE

by

S. Henley

PLANE REPRESENTATION OF MULTIVARIATE DATA STRUCTURE

A method of dimensionality reduction described by Sammon[1] (which he called Nonlinear Mapping) operates by the generation from a two or three dimensional array, of a triangular distance matrix as close as possible to the Euclidean distance matrix computed from the multidimensional data.

It is possible to extend this principle to use any distance measure, as pointed out by Sammon; alternatively any measure of similarity may be used. The same principle may also be applied in the R-mode, to study relationships among variables, by using any measure of correlation.

The method of solution adopted by Sammon for Nonlinear Mapping, using a steepest descent algorithm, may be applied in any case where the distance, similarity, or correlation function is differentiable with respect to the data values.

An R-mode variant of the method of potentially wide application is taken as an illustration: given a data matrix comprising N observations of L variables, with values x_{pq}, p=1, ... ,L and q=2, ...,N , let us define a corresponding set of only d observations ($d \ll N$; commonly d=2) of L random variables, with values y_{pq}, p=1, ... ,L and q=1, ... ,d.

Given that the values x_{pq} are standardised to zero mean and unit variance, the product-moment correlation coefficient between the ith and jth variables is

$$r^*_{ij} = \frac{1}{N-1} \sum_{k=1}^{N} x_{ik}x_{jk},$$

and similarly, in the d-space, the correlation coefficient between the ith and jth variable is

$$r_{ij} = \frac{1}{d - 1} \sum_{k=1}^{d} y_{ik}y_{jk}.$$

The error E which is to be minimised may now be defined as

$$E = \frac{1}{L(L - 1)} \quad . \quad \sum_{i<j}^{L} (r^*_{ij} - r_{ij})^2$$

and the square root of this is the mean error per off-diagonal element of the matrix.

The iterative procedure adjusts the values of y_{pq}; the new d-space configuration after the mth iteration is given by

$$y_{pq}(m) = y_{pq}(m-1) - MF.\Delta_{pq}(m-1)$$

where
$$\Delta_{pq} = \frac{\partial E}{\partial y_{pq}} \Big/ \left| \frac{\partial^2 E}{\partial y_{pq}{}^2} \right|$$

and MF is a stepping coefficient or 'magic factor' empirically found to give best results at a value of about 0.35.

The partial derivatives are

$$\frac{\partial E}{\partial y_{pq}} = 2 \sum_{i\neq p}^{L} y_{iq} (r_{ip} - r^*_{ip})$$

and
$$\frac{\partial^2 E}{\partial y_{pq}{}^2} = 2 \sum_{i\neq p}^{L} y^2_{iq}/(d - 1).$$

In order to ensure that the second derivative takes a non-zero value, the starting configuration must contain at least two non-zero values of y_{pq} for each of the d observations; the configuration need not, however, be random, and a starting condition with all $y_{pq}=1$ is perfectly adequate.

Results of the method give plots (Fig. 1) very similar in appearance to those from principal components analysis (Fig. 2) in which the first two components are plotted; differences result from the method presenting maximum 'structure' rather than maximum variance.

References

[1] Sammon, J.W., IEEE Trans. on Computing, C-18, 401-409, (1969).

[2] Henley, S., PhD thesis, Nottingham, (1970).

S. HENLEY

Bureau of Mineral Resources, Geology and Geophysics
Canberra, A.C.T.
Australia.

CAPTIONS TO THE FIGURES

FIGURE 1. R-mode Nonlinear Mapping of the product-moment correlation relationships among 35 geochemical variables measured on 96 Devonian sedimentary rocks from Cornwall, England[2].

FIGURE 2. Plot of the first two principal components computed from the same set of data as used to construct Figure 1.

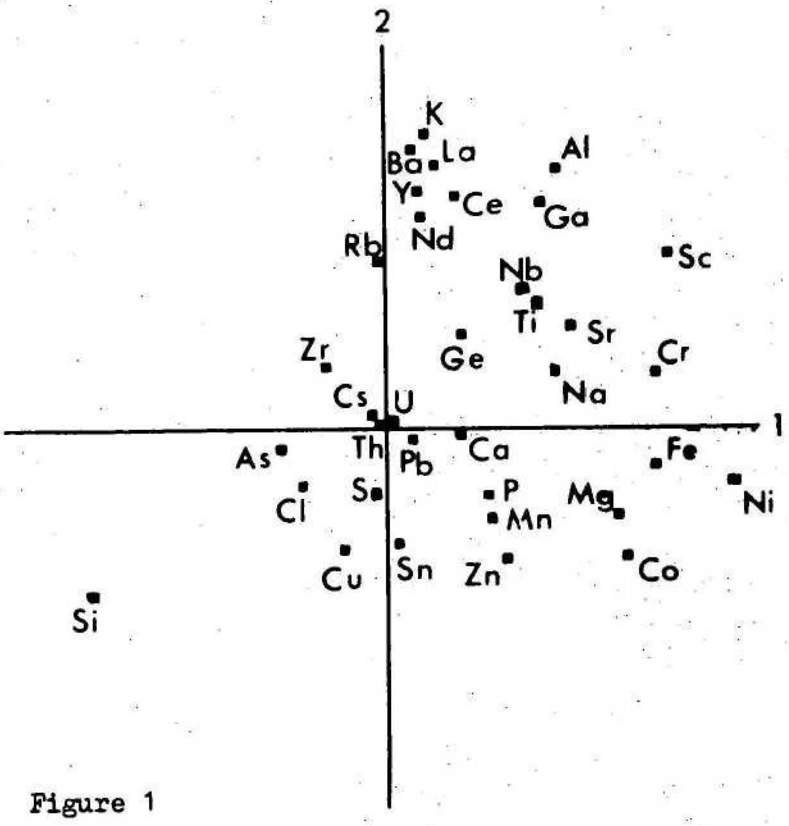

Figure 1

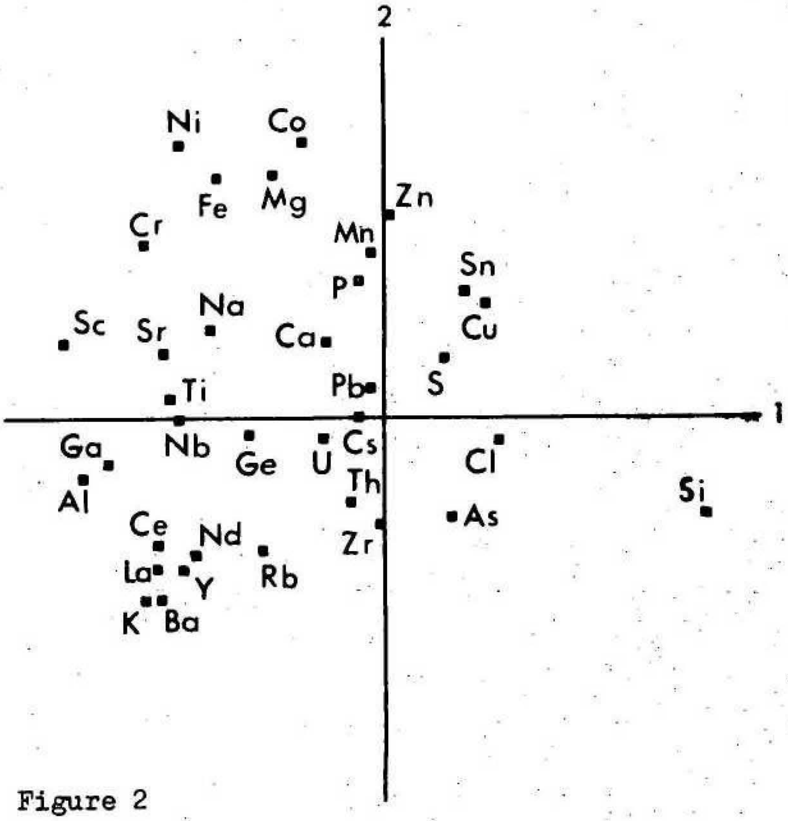

Figure 2