Power-transform removal of skewness from large data sets

S. J. Mancey B.Sc.
R. J. Howarth B.Sc., Ph.D., M.I.M.M.
Applied Geochemistry Research Group, Imperial College, London

620.113:550.84(410)

Synopsis

Univariate geochemical data cannot usually be treated as normally distributed because of their asymmetrical distributions, which result in high skewness coefficients. A method for reducing skewness in large data sets by use of a power transform is described. The technique is quite general in its application, and its use is illustrated with a regional geochemical data set based on the analyses of *ca* 50 000 stream-sediment samples from England and Wales.

Both major and trace-element geochemical frequency distributions are often so positively skewed (with the usual exception of Si in major-element data) that they do not conform to the normal (Gaussian) distribution. Many transforms have been proposed for reducing this departure from normality in order to make subsequent use of parametric statistics more reliable. The log transform has been widely used for this purpose with geochemical data since Ahrens,[1] but it does not necessarily ensure that a more normal distribution will result. Indeed, the log-transformed data will often be little better than the original because a large original positive skewness may be replaced by a negative but generally smaller skewness.

Table 1 Determination of λ based on subsets of 2000 values*

Element	Subset 1	Subset 2	Average
Al	0.39	0.42	0.41
Ca	−0.27	−0.27	−0.27
Fe	0.22	0.18	0.20
K	0.46	0.31	0.38
Si	2.68	2.88	2.78
As	−0.11	−0.11	−0.11
Ba	−0.31	−0.33	−0.32
Cd	−0.34	−0.32	−0.33
Co	0.32	0.34	0.33
Cr	0.20	0.27	0.23
Cu	0.02	0.02	0.02
Ga	0.35	0.32	0.33
Li	0.45	0.41	0.43
Mn	−0.15	−0.19	−0.17
Mo	0.00	0.02	0.01
Ni	0.45	0.49	0.47
Pb	−0.22	−0.21	−0.22
Sc	0.59	0.58	0.58
Sn	−0.24	−0.27	−0.25
Sr	−0.10	−0.11	−0.11
V	0.52	0.48	0.50
Zn	−0.17	−0.18	−0.17

*Data from the England and Wales geochemical survey.[7]

Manuscript first received by the Institution of Mining and Metallurgy on 1 June, 1979; revised manuscript received on 2 January, 1980. Paper published in May, 1980.

Box and Cox[2] suggested a specific power transform to improve normality that would ensure an optimum result for a particular set of data in that it would be transformed to zero skewness. This process is referred to here as 'deskewing', since the aim is to make the transformed distribution symmetrical – this is usually accompanied by a closer approximation to a Gaussian distribution. Draper and Cox[3] showed that transformation can still help to make the data more tractable, even if normality is not achieved.

To avoid impossibly large data-transfer overheads because of the iterative nature of the calculations, all the observations need to be stored in the computer central memory. The available storage, therefore, limits the size of the data set that can be processed (for example, 10 000 samples for a large computer).

A method for deskewing large data sets, using a subset of the total data, is demonstrated by the transformation of 22 000 values derived from a large regional geochemical data set. The methodology is quite general and the technique discussed would be useful with any data.

Box–Cox power transform

Box and Cox[2] described a generalized power transform for improvement of the normality of univariate distributions, of the form

$$Z = \left\{ \begin{array}{ll} (x^{\lambda} - 1)/\lambda, & \lambda \neq 0 \\ \ln x, & \lambda = 0 \end{array} \right\} \quad x > 0$$

where Z is the set of transformed observations, x the set of original observations and λ the power coefficient. The back-transform is easily obtained if required by use of $(\ln(1+\lambda Z)/\lambda)$ or antilog $(\ln Z)$ as appropriate.

An initial value of λ is chosen and successively modified to reduce the asymmetry of the transformed distribution and, thus, improve its approach to a normal distribution. The criteria for the achievement of symmetry are either the attempted reduction of skewness to zero or the joint reduction of skewness to three, by a

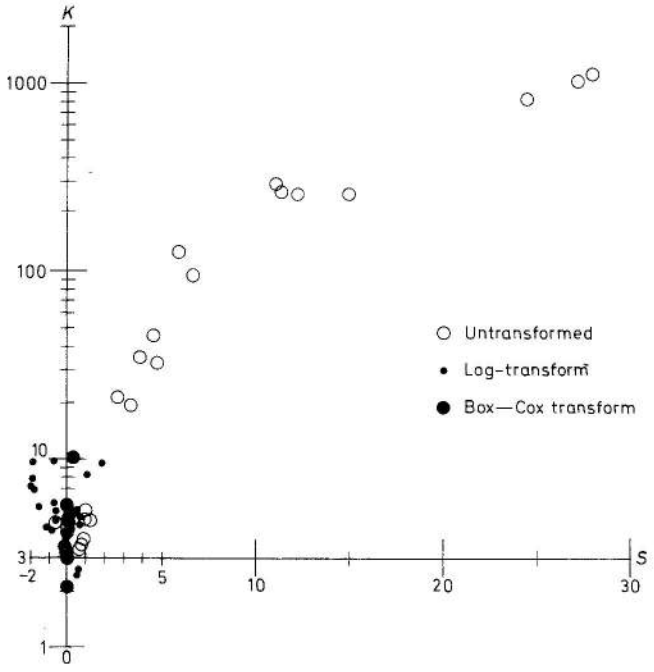

Fig. 1 Effect of data transformation on skewness (S) and kurtosis (K) of 22 elements in regional stream-sediment geochemical survey of England and Wales, based on ~ 22 000 values (see text and Table 2)

Table 2 Effect of transformation on skewness and kurtosis of England and Wales regional stream-sediment data*

Element	Skewness			Kurtosis		
	Untransformed	log	Box–Cox†	Untransformed	log	Box–Cox†
Al	0.68	−0.60	0.03	3.3	4.6	3.1
Ca	3.40	0.48	−0.01	19.1	2.4	2.1
Fe	0.77	−0.63	−0.09	3.6	9.9	3.9
K	0.97	−0.72	−0.01	5.4	5.2	3.6
Si	−0.88	−1.85	0.07	4.4	9.5	3.0
As	27.13	0.49	0.05	1009.0	5.2	4.1
Ba	12.16	1.13	−0.03	261.4	8.4	5.4
Cd	10.60	0.50	0.01	268.0	2.5	1.8
Co	6.73	−1.92	−0.04	96.4	7.7	4.6
Cr	27.89	−0.70	0.29	1135.0	5.6	10.3
Cu	24.40	−0.03	0.05	863.2	5.1	5.4
Ga	1.24	−0.82	0.07	4.8	4.4	3.0
Li	2.61	−1.84	−0.16	21.5	7.0	3.6
Mn	4.76	0.65	−0.05	32.3	4.7	5.5
Mo	3.85	−0.03	−0.60	35.3	3.8	3.8
Ni	5.88	−1.58	0.09	126.9	5.8	4.5
Pb	11.40	0.71	−0.04	266.3	4.9	4.2
Sc	1.00	−1.92	0.01	4.8	7.2	2.8
Sn	14.97	1.87	0.06	258.7	9.5	2.7
Sr	4.54	0.21	−0.12	46.2	4.2	4.5
V	0.76	−0.96	0.02	3.5	4.3	2.6
Zn	11.07	0.46	0.05	291.7	4.4	3.8

*Based on ~22 000 values for each element.
†Data transformed by use of average $\hat{\lambda}$-values from Table 1.

variety of optimization techniques (implemented in the computer program of Howarth and Earle[4]), of which that of Dunlap and Duffy,[5] which optimizes on skewness alone, is the fastest. The iterative calculation is repeated until no improvement in the chosen criterion results; as with all transformations the method is sensitive to the presence of outliers in the data.[6] λ is generally[4] in the range -1 – $+1$. The aim is to produce not a Gaussian distribution but an optimally deskewed distribution. Draper and Cox[3] showed that, even if transformation to normality is not achieved, the λ-transform could regularize the data, and that such a transformation has been found to be beneficial for a wide variety of data.[2,3,4]

Power transformation of large data sets

As each iteration requires the transformation of all the observations and the calculation of one of the skewness/kurtosis criteria, the total computation time increased linearly as the data set became larger (e.g. 10 sec CDC Cyber 174 central processor time per 6500 samples) to the limit of the available storage for the observations. If the data had to be read sequentially for each iteration, an exponential increase in computational overheads could be expected, and the calculation would become impractical for very large data sets. A solution to this problem is to use a subset of the data (held in central memory) for the computation of $\hat{\lambda}$ – an estimation of the true value of λ –

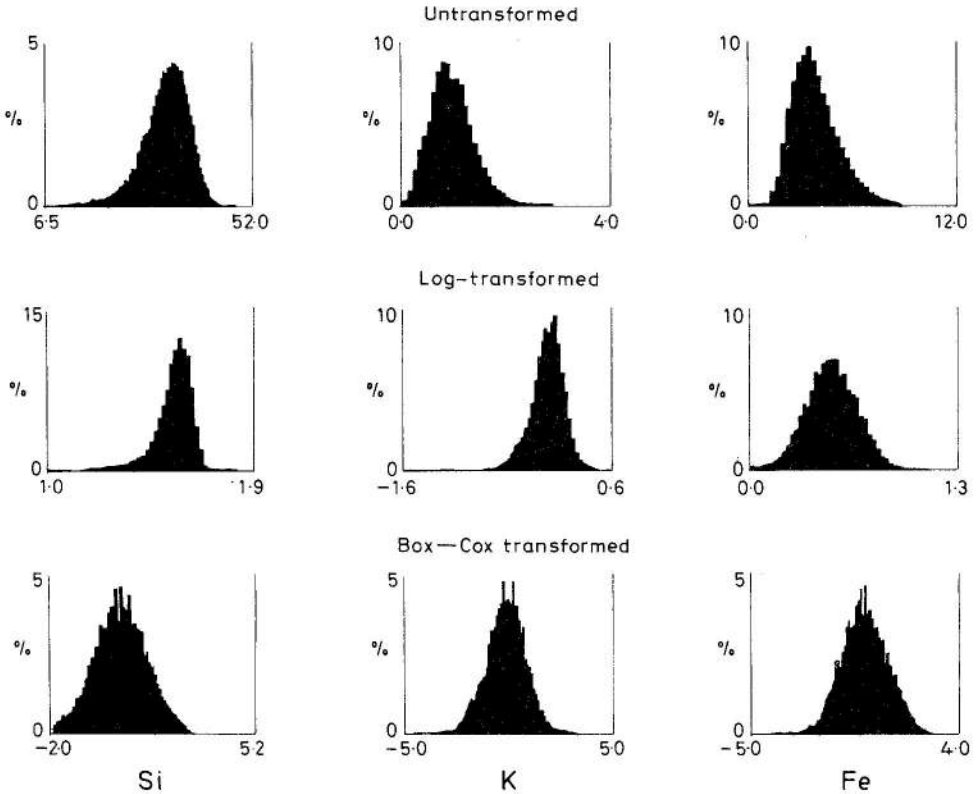

Fig. 2 Effect of data transformation on frequency distribution shape: untransformed data, %

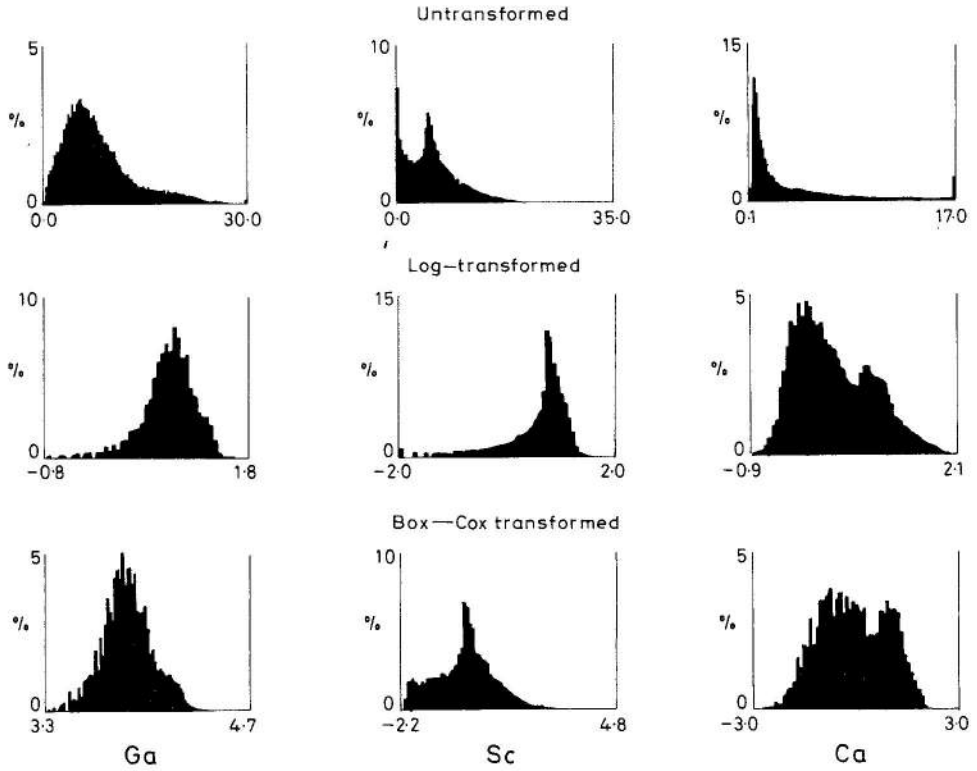

Fig. 3 Effect of data transformation on frequency distribution shape: untransformed Ga and Sc, ppm; Ca, %

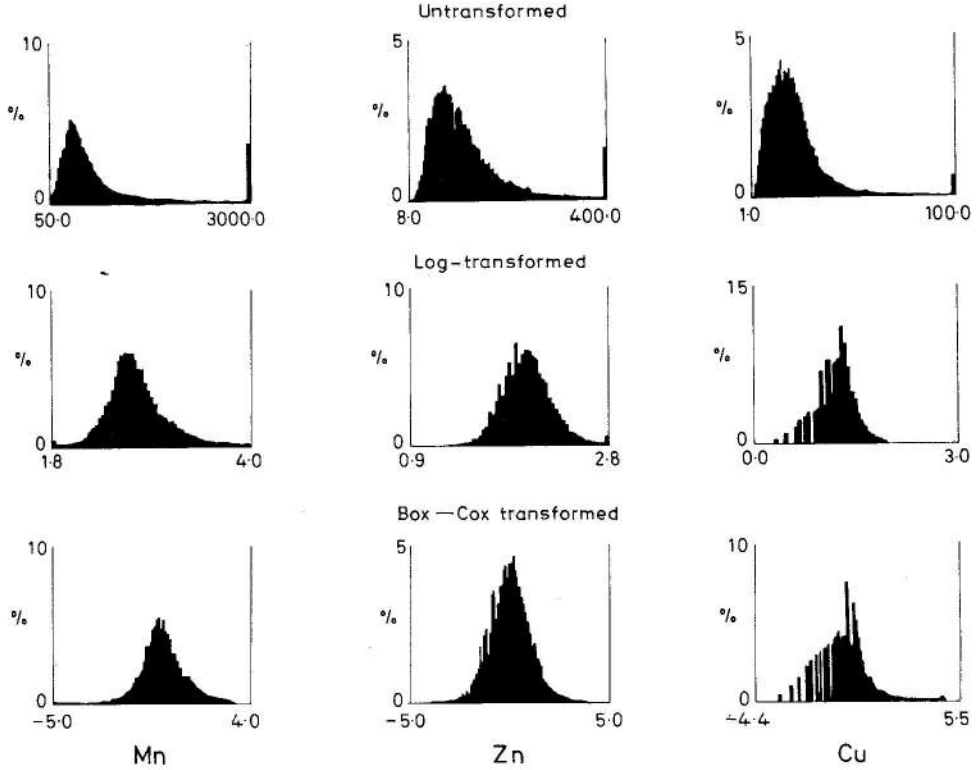

Fig. 4 Effect of data transformation on frequency distribution shape: untransformed data, ppm

that could be used to transform the complete data set in one pass through the data file. The appropriate selection of a representative subset would be necessary and the normal rules for the choice of subsample size would apply; it had been found that a 10% sample size would be generally satisfactory.

For data in which there are reasonable spatial correlations, a suitable subset would be one that covers the sampled geographical area evenly. If the data are stored in the computer in a sequential manner, a regular sample can be obtained by extracting every nth value, where n is chosen to give a subset of manageable size. The alternative would be to select the subsample from the total data file at random, but this would often be more complicated to implement, and would be more likely to result in uneven spatial representation of the geographical units. Several subsets could be taken and the values compared to check the reliability of $\hat{\lambda}$.

Fig. 5 Spatial distribution of two Ca populations and their overlap region (separated by decomposition of the Box–Cox transformed values): 0 – 1.34, 1.35 – 2.99, ⩾ 3.00%. Regional stream-sediment geochemistry of England and Wales

Example

Regional geochemical data from England and Wales

The data set used to illustrate this method is the result of the analysis of ~ 50 000 stream-sediment samples for 22 elements. The samples were collected for a regional geochemical survey of England and Wales[7] at an average sample density of 1/2.6 km^2. All the elements were determined by direct-reading emission spectrometer, except for zinc (atomic absorption) and arsenic, cadmium, molybdenum and tin (colorimetric). Full details of the survey were given by Webb *et al.*[7]

The values at individual sample sites were averaged into square 6.25-km^2 map cells, resulting in a map grid of 254 rows and 209 columns. The small number of values below the statistical detection limit were nonzero positive numbers; it was preferred that they be retained for this analysis rather than set to some arbitrary value – such as half the detection limit. These data were then smoothed by use of a local 3 × 3 cell moving-average filter and some isolated blank cells were infilled, resulting in a final set of ~ 22 000 occupied map-cell element concentration values – the remainder mainly corresponded to grid points over sea areas. To facilitate further statistical treatment it was decided to transform the occupied map-cell data rather than the original point-source values so that the results would be directly comparable with the earlier published moving-average maps.[7] This does not affect the comparisons made here, since the original distribution is the same in all cases. Emphasis throughout the atlas study of England and Wales is on broad-scale regional patterns of variation.

Selection of subsets

The grid of map-cell values for each element was sampled by taking every 10th occupied cell; this was repeated with a second grid, offset from the first by five cells, giving two sets of data with ~ 2000 values in each. Howarth and Earle (Fig. 7)[4] and other workers in the Applied Geochemistry Research Group have shown that performance is quite stable down to small sample sizes; a 10% subsample has been used for convenience.

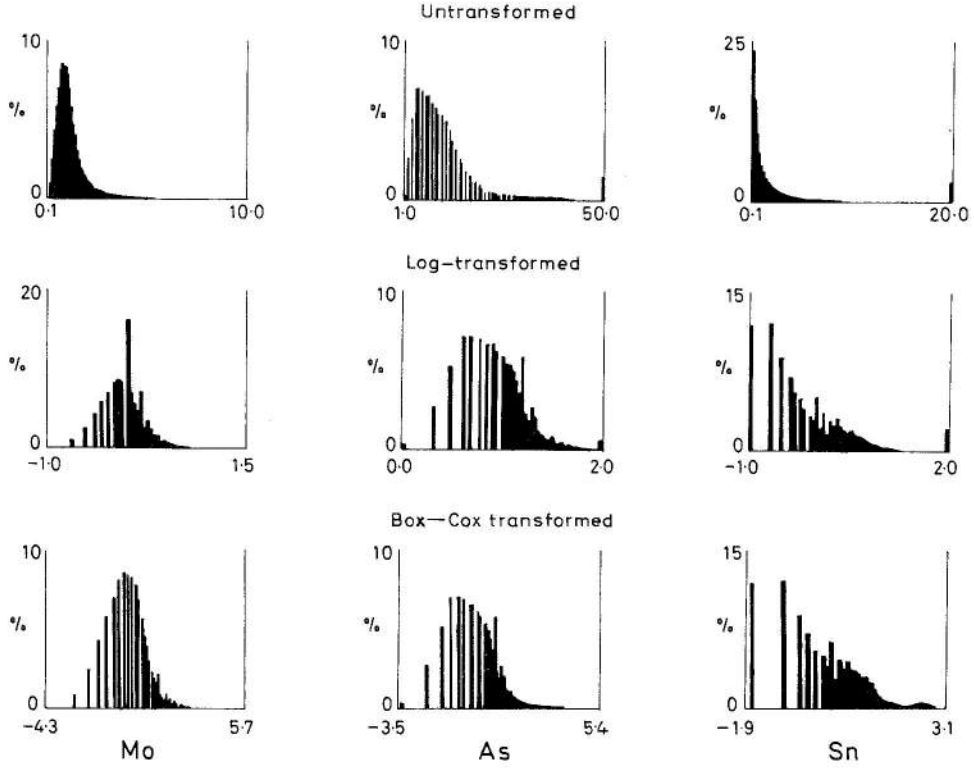

Fig. 6 Effect of data transformation on frequency distribution shape: untransformed values, ppm

$\hat{\lambda}$ was calculated for each variable in each subset by use of the Dunlap and Duffy[5] algorithm (modified to allow for data with an original negative skewness). The values for $\hat{\lambda}$ obtained from the two subsets were typically within 0.05 of each other (Table 1) and the mean value of $\hat{\lambda}$ for each element was used to transform the entire data set. The transformed data were subsequently standardized, to make the values comparable, by subtracting the mean value for each element and dividing the result by the corresponding standard deviation.

Results

The optimum $\hat{\lambda}$ values for each element are given in Table 1. The effectiveness of the power transform for production of a more symmetrical distribution than either the original or log-transformed data is demonstrated by the plot (Fig. 1) of skewness against kurtosis for all 22 elements (a detailed comparison is given in Table 2). Similar improvements in a smaller data set were illustrated by Howarth and Earle.[4]

Potassium and iron (Fig. 2) and also aluminium have moderately positively skewed distributions; of the other major elements, silicon (Fig. 2) is negatively skewed and calcium (Fig. 3) was bimodally distributed (discussed below). The effect of the log transform on silicon was to increase its negative skewness (from −0.88 to −1.85), but for the other elements skewness is reduced, although less effectively than with the power transform. Fig. 4 shows the distributions of three typical trace elements, manganese, zinc and copper. Barium, chromium, lead, strontium and vanadium (not shown) are broadly similar with unimodal positively skewed distributions in the untransformed data. The effect of recording data as integer ppm is apparent in the discontinuous nature of values below the mean in both the log- and λ-transformed copper frequency distributions; this effect is also shown by chromium, lead, strontium and vanadium.

Some distributions exhibit bimodality and still have a high overall skewness — for example, gallium and scandium (Fig. 3). Cobalt and lithium are similarly distributed. (The apparent bimodality of scandium may be an artefact caused by instrumental effects close to the detection limit.) In these cases the log-transform appears as a negatively skewed unimodal distribution and the λ-transform is very much more symmetrical, but still does not attain a normal distribution (although the skewness and kurtosis values may lie close to those expected, that is 0 and 3, respectively). The cutoff, caused by values that fall below the detection limit, is also apparent in these transformed frequency distributions. Calcium has a particularly interesting distribution since the untransformed data are suggestive of a unimodal lognormal distribution (Fig. 4). Both the log- and λ-transformed data show strongly bimodal distributions, the latter being more symmetrical. By the usual graphical technique for decomposition of a multimodal frequency distribution[8] based on dissection of the cumulative distribution, two overlapping populations were obtained. A map of these, together with their overlap region, clearly corresponds to spatially distinct groups of samples related to areas of broadly calcareous or non-calcareous rocks (Fig. 5). Blank areas in the map (plotted as computer-generated microfilm) correspond to sampling gaps, caused by major conurbations or areas with no tributary drainage.

Fig. 6 (molybdenum, arsenic and tin) illustrates the effect of finite recording intervals on the transformation of unimodal positively skewed distributions. The distribution of cadmium (not shown) is similar to that of tin, and these two are the only unimodal distributions that cannot be made symmetrical either by the log- or by the λ-transformations.

It may be noted that these results also demonstrate that values of skewness and kurtosis near 0 and 3, respectively, do not necessarily imply that the distribution is normal (but it may well be symmetrical), although this assumption has been used on occasion to justify the assumption of a normal distribution.[9] For example, the clearly non-normal λ-transformed distribution of tin has a skewness of 0.06 and kurtosis of 2.7. It is not the aim here to obtain distributions that are necessarily Gaussian but

optimally symmetrical distributions (generally prior to further data analysis). Log-transformation has traditionally been used to normalize data in much geochemical work, but we have shown that the results may often be as deleterious for some elements as they are beneficial for others if it is indiscriminately applied. In contrast, the power transform (of which the log-transform is a special case) is always an improvement. The problem of the effect of wild values on the statistical analysis of a data set always applies irrespective of the nature of the transform, and such values should preferably be identified and removed from the data prior to serious data analysis.[6]

Conclusion

The Box–Cox power transform would appear to be a more powerful tool than the traditional log-transform for deskewing data prior to further statistical treatment. The method described here would allow its application to be extended to large data sets that would otherwise be untreatable.

Deskewing data is, naturally, only one of the first steps in its statistical analysis. Analysis of principal components based on the λ-transformed data for England and Wales has already proved to be extremely useful.[10]

Acknowledgement

The authors thank Dr. C. Y. Chork for useful discussions on the subject of this paper. The work carried out by S. J. M. was supported by a Natural Environment Research Council studentship. The England and Wales geochemical survey was undertaken with the aid of a grant from the Wolfson Foundation to Professor J. S. Webb. The calculations were carried out on the CDC 6400/174 facility of the Imperial College Computer Centre, and the results were plotted on the Calcomp 1670 microfilm plotter at the University of London Computer Centre.

References

1. Ahrens L. H. The lognormal distribution of the elements. *Geochim. cosmochim. Acta*, **5**, 1954, 49–73; **6**, 1954, 121–31.
2. Box G. E. P. and Cox D. R. An analysis of transformations. *J. R. statist. Soc.*, **B26**, 1964, 211–43.
3. Draper N. R. and Cox D. R. On distributions and their transformation to normality. *J. R. statist. Soc.*, **B31**, 1969, 472–6.
4. Howarth R. J. and Earle S. A. M. Application of a generalised power transformation to geological data. *Math. Geol.*, **11**, 1979, 45–62.
5. Dunlap W. P. and Duffy J. A. A computer program for determining optimal data transformations minimizing skew. *Behav. Res. Meth. Instrument.*, **6**, 1974, 46–8.
6. Andrews D. F. A note on the selection of data transforms. *Biometrika*, **58**, 1971, 249–54.
7. Applied Geochemistry Research Group, Imperial College of Science and Technology, London. *The Wolfson geochemical atlas of England and Wales* (Oxford: Oxford University Press, 1978), 74 p.
8. Sinclair A. J. Selection of threshold values in geochemical data using probability graphs. *J. geochem. Explor.*, **3**, 1974, 129–49.
9. Lister B. Second inter-laboratory survey of the accuracy of ore analysis. *Trans. Instn Min. Metall. (Sect. B: Appl. earth sci.)*, **86**, 1977, 133–48.
10. Mancey S. J. and Howarth R. J. *Factor score maps of regional geochemical data from England and Wales* (London: Applied Geochemistry Research Group, Imperial College of Science and Technology, 1978), 2 sheets.