Journal of Structural Geology, Vol. 11, No. 6, pp. 775 to 778, 1989
Printed in Great Britain

0191-8141/89 $03.00 + 0.00

Smoothing a sample of circular data

N. I. FISHER

CSIRO Division of Mathematics and Statistics, Lindfield, NSW 2070, Australia

(*Received* 4 *January* 1989; *accepted* 14 *March* 1989)

Abstract—The purpose of this note is to demonstrate certain pitfalls associated with the use of circular histograms or rose diagrams for displaying two-dimensional orientation data, and to recommend an alternative method of summarizing and displaying distributional features of the data.

INTRODUCTION

AN ESSENTIAL aspect of good statistical analysis of a sample of data (whether linear or circular) is Exploratory Data Analysis (Tukey 1977). On the basis of some suitably-chosen graphical displays and, possibly, some simple summary quantities and 'smoothing' of the data, the data analyst can form an initial appreciation of a suitable statistical model for the data (e.g. uniform, unimodal or multimodal) and note the characteristics (e.g. outliers), which might help determine the choice of method in the next, more formal, stage of statistical analysis. For two-dimensional orientation data, the most-widely used method of graphical display is a variant of an angular histogram known as a rose diagram. Rose diagrams are simple to draw, and can convey useful information about general features of the data.

A rose diagram is constructed by specifying a grouping interval or *bin width* W (e.g. W = 5°), and specifying one of the *bin boundaries* (e.g. 0°, so that the intervals are 0°–5°, 5°–10°, . . .), with a direction like 10° falling into the 10°–15° bin. Sectors are then constructed with radii proportional to the amount of data or, preferably, the square root of the amount of data in the various bins. The bin width W controls the amount that the data are *smoothed*, and such a smoothing parameter is a key feature of any smoothing procedure. However, the need to specify a *bin boundary* is an undesirable property of all histograms because it introduces an unfortunate artefact: the choice of bin boundary can have a considerable effect on the shape of the histogram.

Figure 1 is an example of this boundary effect for linear data. The precise nature of the data is unimportant, but their purpose is not: it was desired to infer that the distribution was bimodal. Figure 1(b) shows the data set shifted *en masse* by adding 25 to each datum, so that the bins of the histogram are the same width as in Fig. 1(a) but differently located with respect to the data. On the basis of this figure, there is little evidence to support the hypothesis of bimodality.

Similar difficulties arise with circular data. Figure 2 shows a sample of azimuths and four associated rose diagrams from a fracture analysis at Wallsend Borehole Colliery (Enever *et al.* 1980, Set 17). Note that, as the data are axial, both ends of each axis have been shown in the raw data plot. The rose diagrams have been drawn with the same 10° bin widths but different locations for bin boundaries. The differences, although not as striking as in the previous example, are enough to cause concern. (In fact, this difficulty arose some time ago with these data when they were plotted with respect to Magnetic North and subsequently re-plotted after adjustment to Grid North.) Deciding on the correct bin width—that is,

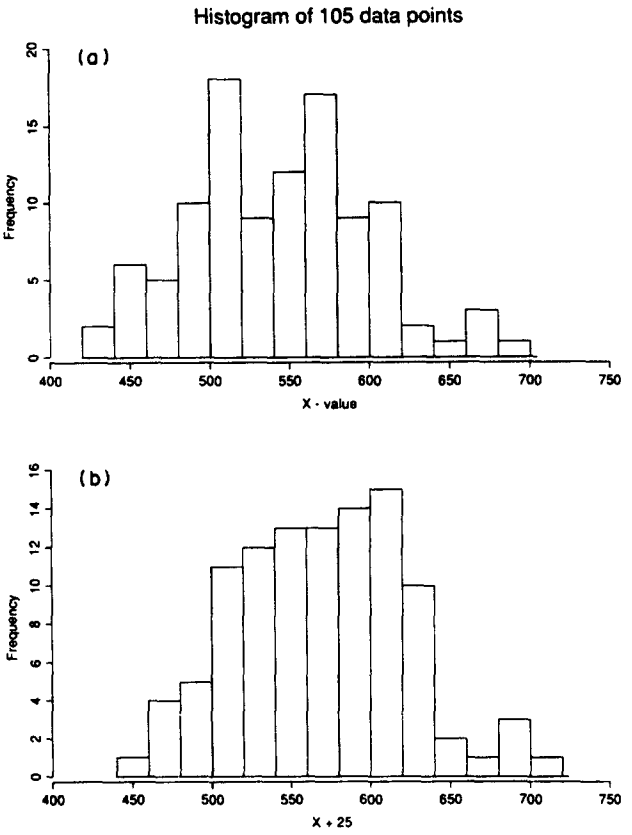

Fig. 1. Histograms of 105 data points (a) data unshifted; (b) data shifted by 25 units.

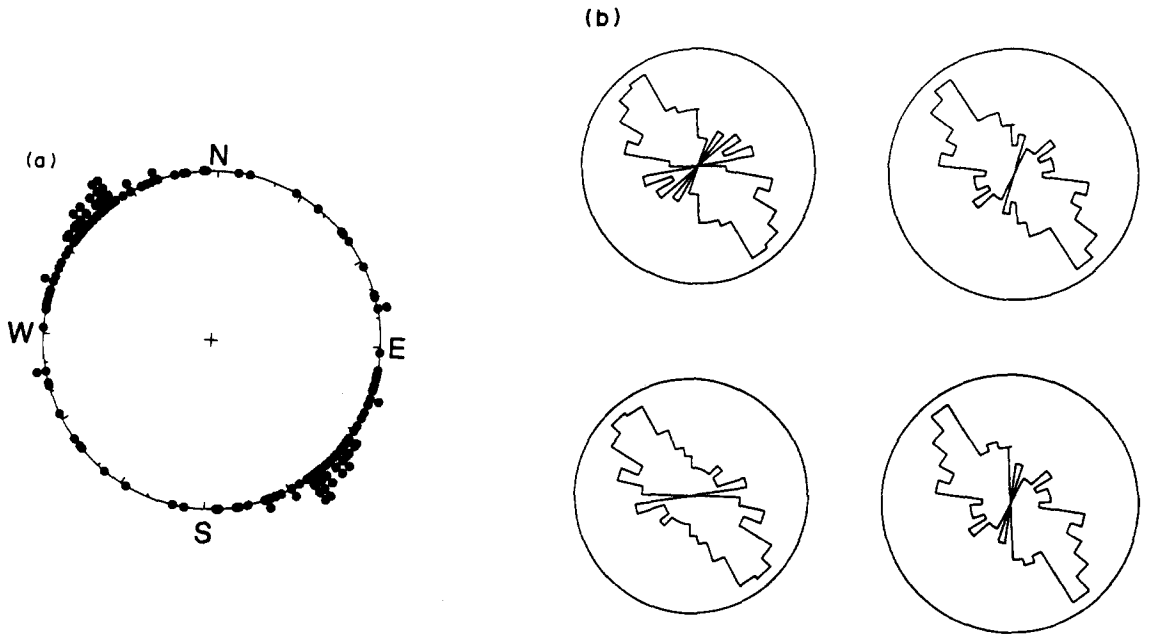

Fig. 2. (a) 75 fracture measurements; data are undirected axes, so both ends of each axis have been plotted. (b) Four rose diagrams for data in (a), with bin boundaries shifted.

on the correct amount of smoothing—is a sufficiently difficult problem without having to worry about where to position the bins.

An alternative to the histogram and rose diagram is a 'non-parametric density estimate'. This is a method of smoothing the data which has the advantage over histograms and rose diagrams of avoiding an arbitrary choice of cell boundary. However, in common with these latter displays it does require a choice of smoothing parameter. Silverman (1986) gives a general account of the subject.

One form of non-parametric density estimate is described in the next section, and examples of its use and further discussion are given for the data of Fig. 2, and for a more complex data set, in the final section.

A KERNEL DENSITY ESTIMATE FOR CIRCULAR DATA

A density estimate which is simple to implement is the kernel estimate, essentially a moving average of the data. Suppose that the sample comprises n measurements $\theta_1, \ldots, \theta_n$ transformed to the range $(0, 2\pi)$. The general form of the estimated density $f(\theta)$ in some direction θ is (e.g. Silverman 1986, p. 15)

$$f(\theta) = \frac{1}{nh} \sum_{i=1}^{n} W\left(\frac{\theta - \theta_i}{h}\right),$$

where W is the so-called kernel (or weighting) function and h the smoothing parameter. Generally speaking, the precise choice of W is less critical than the choice of h: the larger the value of h, the more smoothing results. The amount of smoothing should be related to the sample size and to the dispersion of the data, with larger sample sizes or greater concentration in the data corresponding to smaller values of h, and conversely.

The procedure outlined below is an adaptation to circular data of linear data methods in Silverman (1986), using a quartic kernel function $W(\theta) = 0.9375(1 - \theta^2)^2$ for $-1 \le \theta \le 1$, and zero otherwise (see also Silverman 1986, pp. 4–5, 31). Suppose first that the data are in a single clump.

Step 1. Calculate the mean resultant length $\bar{R}$ of the data:

$$C = \sum_{i=1}^{n} \cos\theta_i, \quad S = \sum_{i=1}^{n} \sin\theta_i, \quad \bar{R} = (C^2 + S^2)^{1/2}/n.$$

Step 2. Calculate

$$K = \begin{cases} 2\bar{R} + \bar{R}^3 + 5\bar{R}^5/6 & \bar{R} < 0.53 \\ -0.4 + 1.39\bar{R} + 0.43/(1 - \bar{R}) & 0.53 \le \bar{R} \le 0.85 \\ (\bar{R}^3 - 4\bar{R}^2 + 3\bar{R})^{-1} & \bar{R} > 0.85 \end{cases}$$

and $\sigma = 1/K^{1/2}$. (K is an estimate of the von Mises concentration parameter—see Best & Fisher 1981.)

Step 3. Calculate $h_0 = 7^{1/2}\sigma/n^{1/5}$. (For a single group of data, this parameter is designed for use with a quartic kernel.)

Step 4. For any given direction θ and smoothing parameter h, calculate the density estimate $f(\theta)$ using the following algorithm:

4.1. $i = 0$

sum = 0.

4.2. $i = i + 1$

$$d_i = |\theta - \theta_i|$$
$$e_i = \min(d_i, 2\pi - d_i).$$

4.3. If $e_i \geq h$ go to 4.2.

4.4. Sum = sum + $(1 - e_i^2/h^2)^2$.

4.5. If $i < n$ go to 4.2.

4.6. $f(\theta) = 0.9375 \times \text{sum}/(n \times h)$.

4.7. Repeat this for, say, 100 values of θ equally spaced between 0 and 2π, yielding

$$(\theta_1^*, f_1), \ldots, (\theta_{100}^*, f_{100}),$$

where

$$f_j = f(\theta_j^*), \quad j = 1, \ldots, 100.$$

4.8. To plot the density around a circle of unit radius, normalize f_j by

$$g_j = f_j/\max(f_1, \ldots, f_{100})$$

and let

$$f_j^* = (1 + g_j)^{1/2} - 1, \quad j = 1, \ldots, 100.$$

Join up the points

$$(\theta_1^*, f_1^*), \ldots, (\theta_{100}^*, f_{100}^*), (\theta_1^*, f_1^*).$$

$f_1^*, \ldots, f_{100}^*$ are used instead of $f_1, \ldots, f_{100}$ to avoid distorting the plot: use of $f_1, \ldots, f_{100}$ results in overemphasis of large peaks at the expense of small ones.

Step 5. It is usually helpful to look at the density estimates corresponding to values h in the range $0.25h_0$, $1.5h_0$.

More care is required when the data appear to have two or more modal groups or clumps. Whilst it is difficult to formulate a simple general procedure, satisfactory results can often be obtained by estimating $\bar{R}$ in Step 1 just from the data in the largest clump.

More effective algorithms can be based on the Fast Fourier Transform; see Silverman (1986) for details.

When the data are axial, or undirected, such as those in Fig. 2, the following adjustments are required:

Step 0. Convert the data to *vectors*, or directed lines, by doubling them.

Replace Step 4.8 by

Step 4.8*. Join up the points.

$$(\tfrac{1}{2}\theta_1^*, f_1^*), \ldots, (\tfrac{1}{2}\theta_{100}^*, f_{100}^*), (\tfrac{1}{2}\theta_1^* + \pi, f_1^*), \ldots, (\tfrac{1}{2}\theta_{100}^* + \pi, f_{100}^*), (\tfrac{1}{2}\theta_{100}^*, f_1^*),$$

where $f_1^*, \ldots, f_{100}^*$ are calculated as described in Step 4.8.

EXAMPLES AND FURTHER COMMENTS

Returning to the data of Fig. 2, a simple calculation leads to the value $h = 0.45$ (from Steps 1–3 of the algorithm). The corresponding density estimate is shown in Fig. 3. There is little evidence, from the raw data, of more than one modal group being present.

A more complex data set, comprising measurements of long-axis orientation of 148 feldspar laths in basalt (Smith 1988, Set 1-7-2 co.prn) is displayed in Fig. 4(a). The two main clumps of data have very similar dispersions, so that a smoothing value of $h = 0.26$ is appropriate to either. The densities shown in Figs. 4(b)–(d) were computed using $0.25h$, h and $1.5h$, respectively.

Clearly, the density in Fig. 4(b) is grossly undersmoothed. The density in Fig. 4(c) appears to be a reasonable representation of the data, with two dominant modal groups and a smaller N–S mode; the density

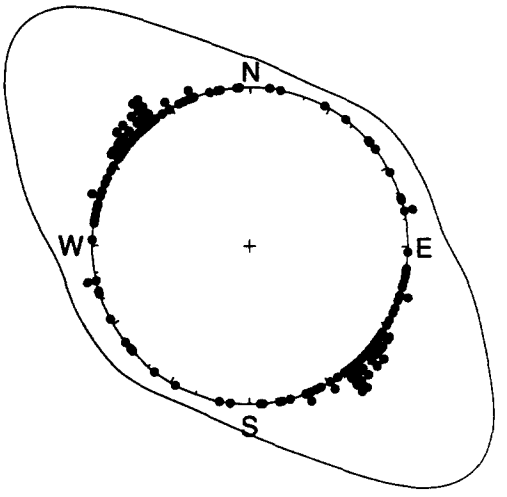

Fig. 3. Non-parametric density estimate for data in Fig. 2(a).

in Fig. 4(d) is probably oversmoothed, with the N–S mode virtually eliminated. Thus, some degree of subjectivity remains, which can only be resolved with a formal statistical test (the subject of a current investigation). Meanwhile, examination of two or three different degrees of data smoothing, as in this example, is strongly recommended.

In conclusion, it is recommended that a raw data plot, in conjunction with one or more non-parametric estimates of density, be regarded as the primary method of initial examination of a sample of orientations.

REFERENCES

Best, D. J. & Fisher, N. I. 1981. The bias of the maximum likelihood estimators of the von Mises–Fisher concentration parameters. *Commun. Statist.-Simulat. Comput.* **B10**, 493–502.

Enever, J. R., Shepherd, J., Cook, C. E., Creasey, J. W., Rixon, L. K., Crawford, G., Dean, A. & White, A. S. 1980. Analysis of factors influencing roof stability at Wallsend Borehole Colliery. CSIRO Division of Applied Geomechanics Report No. 15, CSIRO, Mount Waverley, Victoria, Australia.

Silverman, B. W. 1986. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

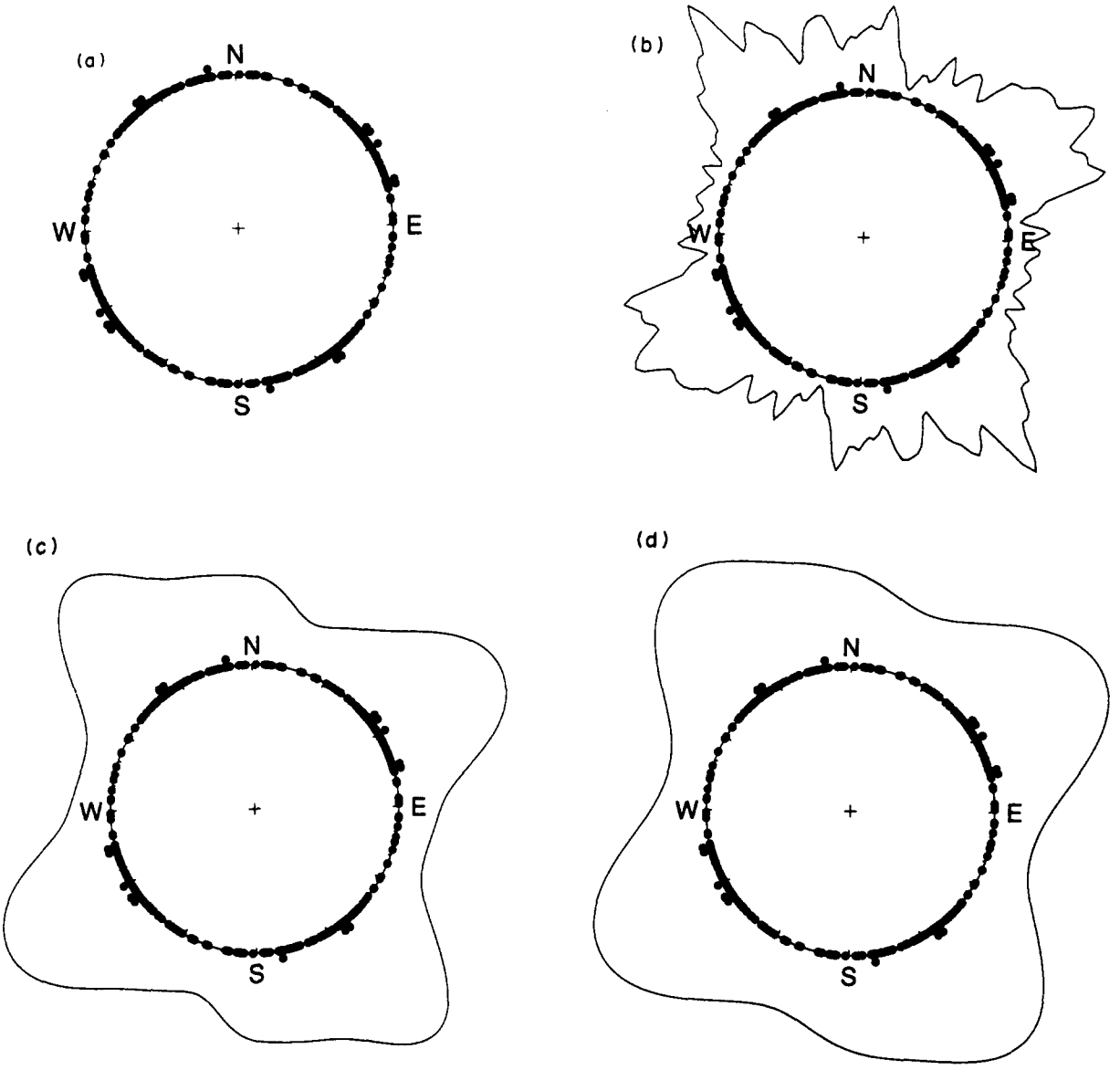

Fig. 4. (a) Orientations of principal axes of 148 feldspar laths; (b) non-parametric density estimate ($h = 0.045$); (c) non-parametric density estimate ($h = 0.25$); (d) non-parametric density estimate ($h = 0.39$).

Smith, N. M. 1988. Reconstruction of the Tertiary drainage systems in the Inverell region. Unpublished B.Sc. (Hon.) thesis, Department of Geography, University of Sydney, Australia.

Tukey, J. W. 1977. *Exploratory Data Analysis*. Addison Wesley, Reading, Massachusetts, U.S.A.