Exploratory data analysis

Article · January 1981

CITATIONS
31

READS
8,584

2 authors, including:

Kelvyn Jones
University of Bristol
269 PUBLICATIONS **10,319** CITATIONS

SEE PROFILE

Chapter 13

Exploratory data analysis

115

N.J. Cox and K. Jones

In *exploratory data analysis*, attempts are made to identify the major features of a data set of interest and to generate ideas for further investigation, whereas in *confirmatory data analysis*, attention is focused on model specification, parameter estimation, hypothesis testing and firm decisions about data. This distinction, made by the statistician J.W. Tukey, is essentially that between *descriptive* and *inferential* statistics. However, confusion can easily arise because Tukey has recently produced many special techniques for exploratory work, yet placed these new methods at some distance from classical statistics, whether descriptive or inferential. Tukey's innovations are now explained in a variety of texts (Tukey, 1977; Mosteller and Tukey, 1977; McNeil, 1977; Erickson and Nosanchuk, 1977) and they have attracted the interest of some geographers as methods which may be used in teaching and research (Cox and Anderson, 1978; Cox, 1978). Other geographers (e.g. Mather, 1976) march under the banner of exploratory data analysis yet do not employ Tukey's new procedures. In this review we adopt the wider sense of 'exploratory data analysis' and do not confine attention to Tukey's innovations. Four themes are apparent in exploratory data analysis: displays, residuals, transformations and resistance (Hoaglin, 1977). None of these is novel, either in statistics or in geographical data analysis, yet from a survey of the field it seems that each deserves greater emphasis in future geographical work.

Tukey's new exploratory methods have met a variety of reactions, ranging from wild enthusiasm (e.g. Wainer, 1977) to outright condemnation (Ehrenberg, 1979a, b). In particular, Ehrenberg criticized exploratory data analysis, especially as presented in the text by Tukey (1977), as poorly explained and motivated, and as fundamentally mistaken in its implication that data analysts need to work without prior knowledge, which is usually available and should be considered (cf. Ehrenberg, 1975). However, those who find Tukey's text unduly cryptic and idiosyncratic may readily be directed to other accounts, while proponents of exploratory methods do not suggest that prior knowledge should be ignored in exploratory work (Cox and Anderson, 1980). It is to be hoped that geographers will avoid unfounded extreme reactions to exploratory data analysis: an attempt to place recent work in a larger context, statistical and geographical, should help in this respect.

In this review we direct attention to those attitudes and procedures which appear most fruitful in exploratory work, and consider the relationship between exploratory (descriptive) and confirmatory (inferential) approaches.

Displays

'Plot both your data and the results of data analysis' is one of the basic attitudes of exploratory data analysis. Displays reveal the major features of data, help in the production of ideas for further investigation, and are useful in checking assumptions (Anscombe, 1973). However, while 'graphicacy' has long been an educational concern among geographers (e.g. Balchin, 1976), and many texts explain a limited standard set of graphical techniques (e.g. histograms, pie diagrams and scatter diagrams), graphical display remains neglected to some extent in quantitative geography. Since graphical inspection allows the identification of outliers, nonlinearities, discontinuities, skewness and other characteristics of the data which may make or mar the analysis, it is sensible to supplement data analyses with appropriate plots. The increasing availability of flexible computer graphics systems makes it easier to do this as a matter of routine, but it should be noted that, for data sets of moderate size, some new plots may be produced manually in a short time.

More telling, perhaps, than any general exhortation is a simple example given by Anscombe (1973, pp. 19-20), who devised four very different data sets which have the same univariate means and the same least squares regression results. It is clear from scatter diagrams whether bivariate regressions are appropriate, but relying on calculated summary measures in model evaluation would produce quite misleading interpretations in three out of four cases.

Here we draw attention to some novel plots for univariate data and (in the next section on Residuals) to appropriate displays for analysis of residuals (especially those from regression models). Lack of space precludes discussion of other kinds of plot, notably those designed specifically for the exploration of multivariate data (Gnanadesikan, 1977; Everitt, 1978).

Traditionally the histogram has pride of place for presenting univariate frequency distributions, and it will continue to be very useful. Three new kinds of displays deserve consideration, however, for this task. In a *rootogram*, not class frequencies but roots of class frequencies are plotted as ordinates, on the grounds that a square root transformation tends to stabilize variation in counts (Tukey, 1970, pp. 163-5; 1972, pp. 312-5; 1977, Ch. 17). In a *box plot*, minimum and maximum values are marked by point symbols and median and quartiles are marked by bars joined in a box. Thus range and interquartile range (or midspread, to use a Tukey term) are represented by distances between symbols. Further information can be added if desired and box plots for different sets of data juxtaposed for comparison (Tukey, 1972, pp. 301-3; 1977, Ch. 2; McNeil, 1977, Chs 1-2; Erickson and Nosanchuk, 1977, Ch. 4; McGill *et al.*, 1978: note variations in terminology). An example is given in Figure 13.1. Box plots are related to the dispersion diagrams once popular in geography, particularly for climatic data (e.g. Crowe, 1933; Gregory, 1978, pp. 147-50).

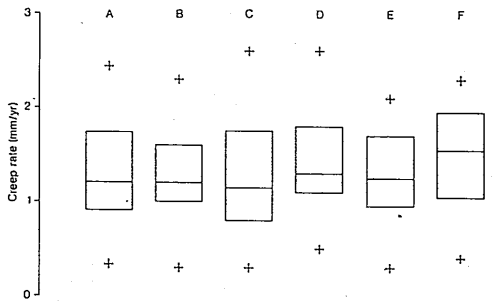

Figure 13.1 Multiple box plot of soil creep rates at Rookhope, upper Weardale, measured by A inclinometer pegs, B Anderson's tubes, C aluminium pillars, D Young's pits, E dowelling pillars, F Cassidy's tubes (Anderson and Cox, 1978)

In a *stem-and-leaf display*, values are represented by the combination of a stem (coarse) and a leaf (fine): the number of leaves on each stem corresponds to the class frequency of a histogram (Tukey, 1972, pp. 295-6; 1977, Ch. 1; McNeil, 1977, Ch. 1; Erickson and Nosanchuk, 1977, Ch. 2). As a simple case consider these annual rainfall figures (in mm) for Durham in the period 1952 to 1976 (Cox and Anderson, 1978, pp. 33-4): 575, 521, 778, 484, 628, 551, 615, 440, 782, 708, 574, 672, 515, 814, 756, 718, 742, 756, 562, 543, 563, 503, 589, 562, 683. The first five are plotted on the stem-and-leaf display below:

```
7 | 7
6 | 2
5 | 72
4 | 8
```

The numbers on the left of the vertical line are the stems: in this case 7, 6, 5, 4 (for 700, . . . , 400). The numbers on the right are the leaves, the leading digits after the stems. In this case the remaining digits have been dropped, although it would naturally be possible to round to the nearest digit. Adding the other figures to the display we obtain:

```
8 | 1
7 | 7805145
6 | 2178
5 | 72571646086
4 | 84
```

and the display can be tidied up by placing leaves on each stem in ascending order.

```
8 | 1
7 | 0145578
6 | 1278
5 | 01245666778
4 | 48
```

Hence numerical ordering produces a simple display which shows the form of the frequency distribution, costs less effort than a histogram yet contains more information, and helps when calculating measures based on ordered values, such as the median or midspread.

Residuals

Residuals are the remainders left after a model (any kind of summary description, from something simple like a measure of level to something more complex like a multiple regression) has been fitted to data. Usually we have a basic partition

$$\text{data} = \text{fit} + \text{residual}.$$

One common strategy is to assume that the model first thought of is so good that the residuals can be set aside safely as the amount unexplained, expressed indirectly as a standard error, a coefficient of determination or some other gross summary statistic.

A more realistic strategy, fundamental to exploratory data analysis and worth wider adoption in quantitative geography, is to doubt whether the model first thought of really is a good summary and to scrutinize the residuals carefully for any pattern which should be reflected in a revised model. The general approach is one of 'summarizing by fit and exposing by residuals' (Tukey and Wilk, 1966, p. 698).

Graphical display is the most useful weapon available for analysis of residuals. Some of the most valuable kinds of plots will be considered briefly in the specific context of regression analysis.

A general 'catch-all' plot shows the residuals e_i against the fitted values $\hat{Y}_i$ of the response or dependent variable (see, e.g., Chatterjee and Price, 1977, Ch. 2). Since a correctly specified regression model would account for all the systematic variation in the response, the corresponding residual plot would show no discernible pattern (e.g. Figure 13.2 (a)). Clear patterns of any kind indicate, however, that the model might need reformulation. A curved band of residuals (Figure 13.2 (b)) might be tackled by adding a square or a higher-order term or a cross-product term. A wedge shaped pattern (Figure 13.2 (c)) shows heteroscedastic variation suggesting the use of weighted least-squares or an appropriate transformation. A solitary point (Figure 13.2 (d)) indicates that the data set contains an outlier which needs further consideration.

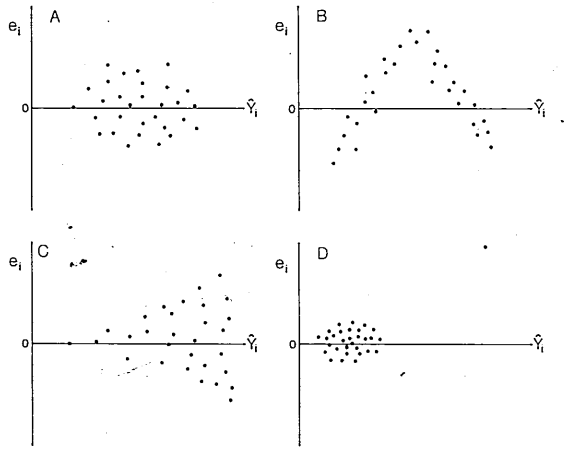

Figure 13.2 Residual plots — residuals v. predicted values of the dependent variable

Even if scatter diagrams were not used before regression, such residual plots would highlight the anomalies in Anscombe's (1973) data sets. While in one case there is no obvious pattern in the residuals, in the other instances a curvilinear pattern and definite outliers are very clear (Figure 13.3). The idiosyncrasies of these different data sets, hidden behind identical numerical summaries, are thus evident from graphical displays.

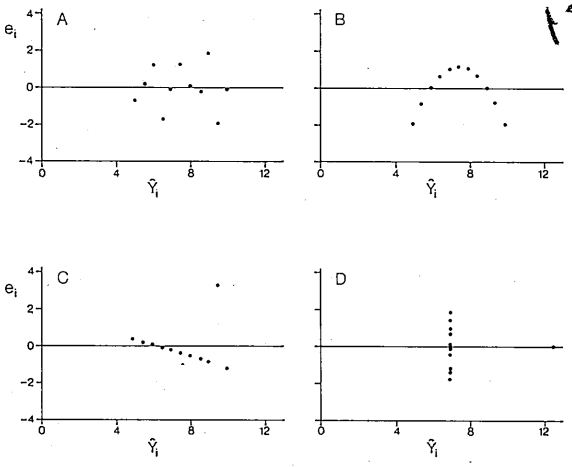

Figure 13.3 Residual plots for Anscombe's (1973) data sets

There are as yet only a few isolated examples of the use of residual plots with geographical data. Crewe and Payne (1971, 1976) used residual plots in their studies of voting behaviour in British elections. A first simple model regressed percentage Labour vote against percentage of manual workers. Fifteen constituencies (including the twelve Northern Ireland seats) were identified as true outliers on the basis of residual plots and substantive reasoning. The fifty largest positive and negative residuals remaining were used to derive a classification of constituencies and to suggest various explanatory variables that might be included in a revised model. After considerable trial and error, a final model was produced which accounted for nearly 90 per cent of the variation in the percentage Labour vote. Moreover, when the residuals from this model were examined, the apparent pattern was one of local effects which could not be reflected in a national model. Barnard (1978) attempted to develop a model of the distribution of the elderly in south-west Hampshire in terms of various dwelling and accessibility measures. A residual plot for the original regression model revealed a substantial outlier, found on investigation to be an Enumeration District in which nearly all the residents were military personnel renting Ministry of Defence accommodation. This outlier was subsequently deleted as anomalous.

Several other kinds of plots may be useful. Partial residual plots (Larsen and McCleary, 1972) essentially show the relationship between the response and a particular explanatory variable after the effects of other explanatory variables have been removed. They have been used by Jones (1980) in a study of geographical variations in mortality for tackling the difficult problem of identifying appropriate functional relationships between variables. It may also be worth plotting the frequency distri-

bution of residuals, or residuals in series (as a graph or a map), or residuals against other variables (Tukey and Wilk, 1966, p. 699; Box *et al.*, 1978, pp. 182-7; Silk, 1979, pp. 245-7).

Resistance

Much of the theory behind statistical methods is concerned with optimal procedures, which are best according to specific criteria if particular assumptions about generating processes are satisfied. However, assumptions that relationships are additive and linear, that variables are identically and independently distributed and Gaussian, or that generating processes are stationary and isotropic - to name but a few prominent examples - are chosen at least in part for their mathematical convenience. They do indeed allow the derivation of many elegant and rigorous theorems. It is nevertheless rare in practice that such assumptions are exactly satisfied. Hence it is desirable that procedures be *robust*, and work well under a variety of conditions, not just under idealized specific conditions for which they have been proved optimal.

Robustness is especially important in exploratory work since it would be foolish to be confident that an unexamined set of data satisfies specific assumptions about behaviour. One kind of robustness particularly valuable in geography is *resistance* to data drawn from distributions which are longer-tailed than the Gaussian, and particularly to *outliers*, isolated values which are detached from the main body of data (cf. Wainer, 1976; Mosteller and Tukey, 1977, for introductions to resistance). This arises from the fact that while many of the procedures favoured by geographers (e.g. correlation, regression, analysis of variance, Student's t test) are most appropriate, if not optimal, for Gaussian variables, many distributions of geographical interest are longer-tailed and include outliers. These outliers are frequently genuine values, and not merely products of observational or experimental blunders. Dublin often appears as an outlier on plots of socio-economic data for the Irish Republic because it is genuinely different from the rest of the country! Such individual outliers can have a great distorting influence on any fitted model, and thus have both geographical and statistical importance.

It would be misleading to imply that resistant methods are the only means of dealing with outliers, which can be accommodated, incorporated, identified or rejected (Barnett, 1978), as seems appropriate. In general, graphical display is the most valuable way of checking for outliers, although this becomes more difficult as the number of data and the number of variables increase, and so resistant methods assume greater value (Cox and Hinkley, 1974, pp. 270-1; Wainer, 1976). If outliers can be identified, parallel 'with' and 'without' analyses can be conducted. Haggett *et al.* (1977, pp. 364-5), for example, repeated a regression of retail sales against personal income for Irish counties without values for Counties Cork and Dublin. It is also good practice to couple 'resistant' and 'unresistant' (or 'robust' and 'fragile') analyses (e.g. Wainer, 1976; Erickson and Nosanchuk, 1977).

The problem of robust estimation of the centre of a symmetric distribution has been the subject of extensive analytical and simulation studies by statisticians (e.g. Andrews *et al.*, 1972). It is now abundantly clear that the mean can perform very poorly with long-tailed distributions; that the median, among other estimators, is better in the presence of long tails or outliers; and that simple estimators exist which perform well under a wide range of conditions. Much of this has long been known to geographers, yet the discussions given in some geographical texts follow a comment about the resistance of the median with a dismissal on spurious grounds. Norcliffe (1977, p. 54) asserted generally that the mean is more efficient than the median, but this is not universally true (cf. Andrews *et al.*, 1972); Gregory (1978, pp. 25-6) unfavourably compared the median, possessing 'no real mathematical qualities' with the mean, 'based on sound mathematics', an invocation of mathematical respectability quite without foundation.

More positively, means and medians can be seen as limiting cases of the family of 'trimmed means' (see, e.g., Wainer, 1976, for further explanation). A p% trimmed mean is calculated by setting aside the p % largest values and the p % smallest values and taking the mean of the (100 - 2p) % of values which remain. p = 0 produces a mean, p = 50 a median and p = 25 a midmean. In principle p can be chosen according to the degree of resistance required and the character of the data, although there is little reason for using merely one value of p. One example of the use of trimmed means with air pollution data was given by Cleveland and Guarino (1976).

Moment-based measures of spread, asymmetry and tailedness (e.g. standard deviation and classical skewness and kurtosis) are generally unresistant, and there is much to be said for greater use of quantile-based measures (cf. Tukey, 1977; McNeil, 1977; Erickson and Nosanchuk, 1977). The interquartile range or 'midspread', for example, is more useful as a measure of spread than many geographers allow: common objections to it often boil down to prejudices that it is old-fashioned and not totally respectable.

Resistance is an important property not only for summary measures but also for other methods of data analysis. For example, two-way tables are often analysed using a model of the form

$$\text{data} = \text{level} + \text{row effect} + \text{column effect} + \text{error}.$$

The parameters are usually estimated via table mean, row means and column means. Tukey has devised a resistant method of iterative estimation for such

tables known as median polish (Tukey, 1977, Chs 10 and 11; McNeil, 1977, Ch. 5; Erickson and Nosanchuk, 1977, Ch. 15). Anderson and Cox (1978) used median polish in a comparison of different instruments for measuring soil creep, and found a clear picture emerging from a fairly messy set of data, which was supported by an independent and more conventional analysis.

Perhaps the greatest need for resistant methods is in applications of correlation, regression and related multivariate analyses. Many resistant procedures have been devised by statisticians, and they deserve close attention from geographers. For example, Wainer (1976) outlined a method of estimating standard deviation, correlation and slope resistantly, and suggested that principal components and factor analyses be based on variance-covariance matrices derived in this way. There are simple methods for line fitting from bivariate medians of thirds of data sets (Erickson and Nosanchuk, 1977, Chs 11 and 12; McNeil, 1977, Ch. 3), while biweight estimation is an elegant resistant alternative to least squares (Mosteller and Tukey, 1977, Chs 10 and 14; McNeil, 1977, Ch. 7), used on air pollution data by Cleveland and Guarino (1976).

Transformations

Thus far we have used themes identified in exploratory data analysis by Hoaglin (1977) as headings for this review, and considered displays, residuals and resistance in turn. The remaining theme of transformations probably needs least emphasis for a geographical audience. Standard texts intended for geographers (e.g. Haggett *et al.*, 1977; Norcliffe, 1977; Taylor, 1977; Gregory, 1978) include sections on transformations, and transformations of variables of geographical interest are commonplace, at both elementary and research levels. Most of the basic transformations employed by statisticians (cf. Hoyle, 1973) are known to geographical data analysts: not only the common power, root and logarithmic transformations, but also transformations useful for categorical data such as the logit, probit and various others based on inverse trigonometric and hyperbolic functions.

Transformation of geographical data has usually been motivated by an inferential approach, and particularly by the idea that hypothesis testing requires data drawn from Gaussian (normal) distributions. One common aim of re-expressing variables has hence been normalization of frequency distributions. The degree of success of any transformation may be assessed by inspection of histograms or probability plots, calculation of skewness and kurtosis or performance of some distribution-free test such as chi-square or Kolmogorov-Smirnov. This approach has often been successful in its own terms. For example, many variables encountered by geographers are right-skewed (but not highly irregular) in distribution, and thus behave quite respectably if logarithms or square roots are taken. Investigators have then proceeded joyfully to inferential procedures; unfortunately they have frequently overlooked the fact that these may be inappropriate or irrelevant on other grounds (see below). Indeed it is common to find among geographers a notion that lack of normality is the basic statistical problem, and thus that paradise has been attained once a semblance of normality has been produced, whether by subterfuge or by honest means. This notion ignores the standard principle that other assumptions about data (especially mutual independence) are often crucial. The very term 'normal', still used by an overwhelming majority of geographers, continues to act as a misleading influence.

Some geographers have hoped that a common 'blanket' transformation will simultaneously normalize a variety of variables (perhaps a haphazard mixture of attributes about to be offered as ritual sacrifice in a principal components or factor analysis). No doubt data analysis would be simpler if life were easier, but there seem to be many empirical and statistical grounds for the contrary idea that each variable should be treated individually. Since right-skewness is a common kind of departure from normality, taking logarithms (for example) will usually improve a majority of variables, but neither logarithms nor any other simple transformation can serve as a panacea for non-normality (see, for example, results of Gardiner, 1973).

The view dominant in exploratory data analysis is that transformation of variables need be justified only by convenience. It need not be motivated by any specific inferential assumption, but merely by the aim of easier and more effective description. If pressed too far, however, this distinction of aims turns into a false antithesis. Transformations have been used frequently to achieve approximately linear relationships. It is well known, for example, that logarithmic transformation of one or both variables brings power function and exponential relationships into the family of linear relationships (cf. Tufte, 1974, pp. 108-31, for an especially lucid account). Here the aim of easier and more effective description is coupled with the hope of using the well-developed inferential machinery of the linear model.

The exploratory data analysis texts of Tukey (1977), Mosteller and Tukey (1977) and McNeil (1977) include not only many practical examples of a more liberal view of transformations, but also much valuable advice about specific issues such as the handling of zeros and the use of folded transformations. It is striking to find that Tukey regards additivity of effects and constancy of spread as more important in practice than normality (or even symmetry) of distribution, although any order of priority is confused by the happy circumstance that these three conditions frequently occur together

when they do exist (Tukey and Wilk, 1966, p. 702).

Any reluctance by geographers to embark on transformation (e.g. Gould, 1970, pp. 442-3) is usually on one or both of two grounds: a feeling that transformed scales are unnatural, and an unease that transformation involves an unacceptable element of adhoc-ery, if indeed it does not verge upon statistical cheating. These objections are both exaggerated. The often cited case of pH, a logarithmic transformation of hydrogen ion concentration long accepted as a useful measure, serves as a reminder that 'naturalness' may reflect convention as much as reality. The thought that square roots of counts are less natural than raw counts is understandable, and accounts for reluctance to draw rootograms rather than histograms, but the feeling is reduced once preliminary rooting has been shown to produce a clearer picture in a few analyses. The appearance of adhoc-ery can also be avoided, and the choice of transformations made more systematic in a variety of ways. The standard power, root and logarithmic transformations can be seen to be members of a 'ladder of re-expressions' (Mosteller and Tukey, 1977, pp. 79-81): the analyst moves up and down the ladder as appropriate. The commutative property of such monotonic transformations

transform of quantile of raw data
= quantile of transform of raw data

can be used to reduce the work in choosing transformations from this ladder. (This property is an advantage of quantiles not possessed by means, and should be set against the fact that the additive property of means which allows them to be combined is not satisfied by quantiles: e.g. Ehrenberg, 1975, pp. 173-4.)

Smoothing

Smoothing is an approach used in exploratory data analysis of special interest to geographers, who are commonly faced with data in the form of spatial or time series. It is appropriate if data series may be regarded as a mixture of smooth and rough components, whereby

data = smooth + rough.

For example, data may be regarded as 'signal' mixed with 'noise', 'true values' mixed with 'measurement error', 'long-term trend' mixed with 'short-term fluctuation', or as 'regional trend' mixed with 'local deviation'.

Popular smoothing methods used in quantitative geography can be classified as either linear function fitting or local linear smoothing (Cox, 1979a). While such approaches are often useful and appropriate, their limitations (stressed here) justify interest in nonlinear smoothers introduced by Tukey as an alternative approach.

In linear function fitting we adopt a precisely specified model of the general form

data = linear function + stochastic error.

The linear function is usually a polynomial in the map coordinates (trend surface analysis). However, we thereby invoke particular assumptions about the structure of the data which may be unjustified (or unjustifiable) in exploratory work. Unless the linear function has some interpretation in terms of geographical processes, we might be making such interpretation more difficult. This approach leans heavily on the idea of linearity, whereas there are many grounds for expecting nonlinear behaviour of geographical responses. Usually least squares estimation is used: it is well known that this performs poorly in the presence of outliers or long-tailed data. The linear function is generally fitted globally to all the data: it may be more realistic not to assume that a single model is valid for every part of the data, but to move to local fitting.

In local linear smoothing, we compute a weighted average of the values in the neighbourhood of any particular value. One major advantage is then that weights can be chosen to suppress or magnify variations in particular frequency bands, at least if data are regularly spaced. Moving to local operations alleviates any doubts about the appropriateness of a global approach, and may reduce other difficulties, but we still lean heavily on linearity and (implicitly) on least squares.

Nonlinear smoothers based on running medians have been far less popular than linear smoothers, but they have been advocated in recent years by some statisticians (Beaton and Tukey, 1974; Tukey, 1977, Chs 7 and 16; McNeil, 1977, Ch. 6; Velleman, 1977) and geophysicists (Claerbout and Muir, 1973; Claerbout, 1976). When smoothing data series running medians tend to be more resistant than moving averages. A local median will be relatively uninfluenced by high or low spikes which usually cannot be regarded as part of the smooth, whereas a local mean will tend to mix such spikes in with the smooth.

Methods proposed by Tukey for smoothing one-dimensional series are based on taking running medians of successive trios of data values, often repeated until convergence and followed by Hanning, a linear smoother with weights ¼:½:¼. These methods can be generalized readily for cases of two-dimensional series, whether or not data are regularly spaced (Besag and McNeil, 1976; Cox, 1979a), but the one-dimensional methods have been far more widely used and, indeed, have been rather more successful.

This family of methods proposed by Tukey is designed for exploratory work. The attitude most fruitful in practice is to experiment with a variety of smoothers (setting aside any idea that one particular smoother might be 'best'), and to examine the resulting series of smooth and rough. This approach

is easiest to implement when a computer library of smoothing routines is coupled with a cathode ray tube display (for immediate scanning) and a pen plotter or some other hard copy device.

In continuing work Cox (1979a, b) has employed non-linear smoothers on various data series of geographical interest: socio-economic data for the Irish Republic (Cliff and Ord, 1973); pollen abundances from a site in Papua and New Guinea (Walker and Wilson, 1978); soil and surface properties from a gilgai area in New South Wales (Webster, 1977) and hillslope angles from North Yorkshire (Cox, 1979b). In each case outliers and long tails are evident (and outliers were omitted rather cavalierly from their analyses by Cliff and Ord and by Walker and Wilson). The results proved of considerable interest: they cast doubt on previous interpretations in the first two cases; the methods provided a simple alternative to spectral analysis in identifying periodicities in the third; and some light was thrown on the difficult issue of scale variations in hillslope profile morphometry in the last. And this is what we should expect from exploratory methods: not only answers to some old questions, but also some interesting and provocative new questions to be considered in further work. Indeed nonlinear smoothers should be more widely used as exploratory methods in future, it being understood that the more popular linear methods will continue to be employed where shown to be appropriate.

Inference

Most of the literature on statistical methods in geography appears to be based, at least implicitly, on the view that descriptive statistics are rather obvious, if not trivial, while inferential statistics are both more challenging and more valuable intellectually. 'Confirmatory' approaches have generally been promoted at the expense of 'exploratory' approaches, in Tukey's terms. There are many reasons for this, some more respectable than others. Extended accounts of confirmatory methods in textbooks may be necessary if only because these more difficult ideas need to be explained at some length. It is more disturbing that both teaching and research in statistical geography have been strongly influenced by the recipes given in cook-books intended for other disciplines (e.g. biology, psychology, sociology) and the appropriateness of these recipes has generally received rather limited attention. A review of the basic ideas of statistical inference and of experience in quantitative geography leads us to the view that statistical inference has been oversold in geography, because it is often inappropriate or irrelevant for geographical problems (cf. Cox and Anderson, 1978, 1980). It follows that exploratory analyses deserve a greater proportion of teaching and research efforts in geographical data analysis.

It is necessary, therefore, to review the grounds for believing statistical inference to be oversold in geography. Since brevity may impart an air of dogmatism, let it be stressed that many complex issues are involved here which do remain controversial and unresolved.

In the first place, many test procedures assume normality (Gaussianity) of distributions; this may not be met by geographical data. On the other hand, recourse may be had to transformations, distribution-free procedures or outlier rejection in attempts to circumvent this difficulty. Secondly, many test procedures assume mutual independence of data or of stochastic disturbances, whereas it is natural to expect autocorrelation to be present in geographical data. Thirdly, not all geographical data sets may be usefully regarded as representative samples from a larger population. On the other hand, these two related problems may be attacked by invoking stochastic process theory or randomization procedures (but *not* distribution-free procedures). In short, since the assumptions behind simple inferential procedures are often not met in geographical problems, more complex procedures must be invoked, and these in turn may present difficulties (for further discussion cf. Gould, 1970; Cox and Hinkley, 1974; Box, 1976; Haggett *et al.*, 1977; Box *et al.*, 1978; Silk, 1979).

There are also grounds for doubting the relevance of the ideas of statistical inference in geographical data analysis. Although there are several schools of statistical inference, most geographers adhere to the Neyman-Pearson school which regards inference as essentially a matter of deciding between rival hypotheses. For instance, Johnston (1978, p. 14) cites as an example that 'our research hypothesis may be that south-facing slopes in Derbyshire have a more rapid growth rate for grass in April than do north-facing slopes, so that the null hypothesis is of no difference between the two types of slope'. The important question is whether it is really fruitful to regard data analysis in this way, reducing matters to a simple qualitative dichotomy (*either* some difference *or* no difference) with everything else set on one side. If one really were interested in the contrast between different aspects, the key question is estimating the magnitude of the difference, not establishing whether a difference exists. In any case a clear contrast between different situations is probably less likely in observational than in experimental studies. In general, testing hypotheses with the aim of producing firm decisions may be less valuable than attempts to summarize the quantitative evidence available and efforts to be open to the indications provided by the data (for further discussion cf. Edwards, 1972; Cox and Hinkley, 1974; Mosteller and Tukey, 1977; Cox and Anderson, 1978).

It must be admitted that there is something rather attractive about statistical inference. A moderate investment of intellectual effort yields a fair grasp of what is going on, yet the ideas are

sufficiently abstruse to impart a feeling of sophistication, and (best of all) attention to the rules of the game provides simple definite answers: results are or are not significant at some conventional level, something easily recorded. Exploratory methods are much less satisfactory, for most are so simple that they appear suspiciously trivial, and (worst of all) practitioners are thrust into a messy and chaotic world where 'problems may often not have neat answers or a single correct solution' (Mosteller and Tukey, 1977, p. xi). Irony aside, it is to be hoped that quantitative geography in the 1980s will be less afflicted than in the past by a craving for the semblance of elegance, exactness and rigour exuded by inferential ideas, and that geographers will show more willingness to engage in uninhibited exploration of their data, guided but not dominated by the procedures devised by statisticians.

Acknowledgement

We are very grateful to Ian Evans for his comments on a draft of this article.

References

Anderson, E.W., and Cox, N.J. (1978) 'A comparison of different instruments for measuring soil creep', *Catena*, 5, 81-93.

Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., and Tukey, J.W. (1972) *Robust estimates of location. Survey and advances*, Princeton University Press, NJ.

Anscombe, F.J. (1973) 'Graphs in statistical analysis', *American Statistician*, 27, 17-21.

Balchin, W.G.V. (1976) 'Graphicacy', *American Cartographer*, 3, 33-8.

Barnard, K.C. (1978) 'The residential geography of the elderly: a multiple-scale approach', unpublished PhD thesis, University of Southampton.

Barnett, V.D. (1978) 'The study of outliers: purpose and model', *Applied Statistics*, 27, 242-50.

Beaton, A.E., and Tukey, J.W. (1974) 'The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data', *Technometrics*, 16, 147-85.

Besag, J.E., and McNeil, D.R. (1976) 'On the use of exploratory data analysis in human geography' [abstract], *Advances in Applied Probability*, 8, 652.

Box, G.E.P. (1976) 'Science and statistics', *Journal of the American Statistical Association*, 71, 791-9.

Box, G.E.P., Hunter, W.G., and Hunter, J.S. (1978) *Statistics for experimenters*, Wiley: New York.

Chatterjee, S., and Price, B. (1977) *Regression analysis by example*, Wiley: New York.

Claerbout, J.F. (1976) *Fundamentals of geophysical data processing*, McGraw-Hill: New York.

Claerbout, J.F., and Muir, F. (1973) 'Robust modeling with erratic data', *Geophysics*, 38, 826-44.

Cleveland, W.S., and Guarino, R. (1976) 'Some robust statistical procedures and their application to air pollution data', *Technometrics*, 18, 401-9.

Cliff, A.D., and Ord, J.K. (1973) *Spatial autocorrelation*, Pion: London.

Cox, N.J. (1978) 'Exploratory data analysis for geographers', *Journal of Geography in Higher Education*, 2 (2), 51-4.

Cox, N.J. (1979a) 'Nonlinear smoothing in one and two dimensions', Paper presented to Institute of British Geographers Annual Conference, Manchester.

Cox, N.J. (1979b) 'Models and methods in hillslope profile morphometry', unpublished PhD thesis, University of Durham.

Cox, N.J., and Anderson, E.W. (1978) 'Teaching geographical data analysis: problems and possible solutions', *Journal of Geography in Higher Education*, 2 (2), 29-37.

Cox, N.J., and Anderson, E.W. (1980) 'In defence of exploratory data analysis', *Journal of Geography in Higher Education*, 4 (1), 85-9.

Cox, D.R., and Hinkley, D.V. (1974) *Theoretical statistics*, Chapman & Hall: London.

Crewe, I., and Payne, C. (1971) 'Analysing the census data' in D. Butler and M. Pinto-Duschinsky (eds), *The British general election of 1970*, 416-36, Macmillan: London.

Crewe, I., and Payne, C. (1976) 'Another game with nature: an ecological regression model of the British two-party vote ratio in 1970', *British Journal of Political Science*, 6, 43-81.

Crowe, P.R. (1933) 'The analysis of rainfall probability', *Scottish Geographical Magazine*, 49, 73-91.

Edwards, A.W.F. (1972) *Likelihood*, Cambridge University Press.

Ehrenberg, A.S.C. (1975) *Data reduction*, Wiley: London.

Ehrenberg, A.S.C. (1979a) [Review of Tukey, 1977], *Applied Statistics*, 28, 79-83.

Ehrenberg, A.S.C. (1979b) 'A note of dissent on data analysis', *Journal of Geography in Higher Education*, 3 (2), 113-16.

Erickson, B.H., and Nosanchuk, T.A. (1977) *Understanding data*, McGraw-Hill Ryerson: Toronto.

Everitt, B.S. (1978) *Graphical techniques for multivariate data*, Heinemann: London.

Gardiner, V. (1973) 'Univariate distributional characteristics of some morphometric variables', *Geografiska Annaler*, 54A, 147-53.

Gnanadesikan, R. (1977) *Methods for statistical*

data analysis of multivariate observations, Wiley: New York.
Gould, P.R. (1970) 'Is *Statistix inferens* the geographical name for a wild goose?', *Economic Geography*, 46, (supplement) 439-48.
Gregory, S. (1978) *Statistical methods and the geographer*, Longman: London.
Haggett, P., Cliff, A.D., and Frey, A. (1977) *Locational analysis in human geography*, Arnold: London.
Hoaglin, D.C. (1977) 'Mathematical software and exploratory data analysis' in J.R. Rice (ed.) *Mathematical software III*, 139-59, Academic Press: New York.
Hoyle, M.H. (1973) 'Transformations - an introduction and a bibliography', *International Statistical Review*, 41, 203-23.
Johnston, R.J. (1978) *Multivariate statistical analysis in geography*, Longman: London.
Jones, K. (1980) 'Geographical variation in mortality: an exploratory analysis', unpublished PhD thesis, University of Southampton.
Larsen, W.A., and McCleary, S.J. (1972) 'The use of partial residual plots in regression analysis', *Technometrics*, 14, 781-90.
McGill, R., Tukey, J.W., and Larsen, W.A. (1978) 'Variations of box plots', *American Statistician*, 32, 12-16.
McNeil, D.R. (1977) *Interactive data analysis*, Wiley: New York.
Mather, P.M. (1976) *Computational methods of multivariate analysis in physical geography*, Wiley: London.
Mosteller, F., and Tukey, J.W. (1977) *Data analysis and regression*, Addison-Wesley: Reading, Mass.
Norcliffe, G.B. (1977) *Inferential statistics for geographers*, Hutchinson: London.
Silk, J.A. (1979) *Statistical concepts in geography*, Allen & Unwin: London.
Taylor, P.J. (1977) *Quantitative methods in geography*, Houghton Mifflin: Boston, Mass.
Tufte, E.R. (1974) *Data analysis for politics and policy*, Prentice-Hall: Englewood Cliffs, NJ.
Tukey, J.W. (1970) 'Some further inputs' in D.F. Merriam (ed.), *Geostatistics: a colloquium*, 163-74, Plenum: New York.
Tukey, J.W. (1972) 'Some graphic and semigraphic displays' in T.A. Bancroft and S.A. Brown (eds), *Statistical papers in honor of George W. Snedecor*, 293-316, Iowa State University Press: Ames, Iowa.
Tukey, J.W. (1977) *Exploratory data analysis*, Addison-Wesley: Reading, Mass.
Tukey, J.W., and Wilk, M.B. (1966) 'Data analysis and statistics: an expository overview', *American Federation of Information Processing Societies Conference Proceedings*, 29, 695-709.
Velleman, P.F. (1977) 'Robust nonlinear data smoothers: definitions and recommendations', *Proceedings, National Academy of Sciences*, 74, 434-6.
Wainer, H. (1976) 'Robust statistics: a survey and some prescriptions', *Journal of Educational Statistics*, 1, 285-312.
Wainer, H. (1977) [Review of Tukey, 1977] *Psychometrika*, 42, 635-8.
Walker, D., and Wilson, S.R. (1978) 'A statistical alternative to the zoning of pollen diagrams', *Journal of Biogeography*, 5, 1-21.
Webster, R. (1977) 'Spectral analysis of gilgai soil', *Australian Journal of Soil Research*, 15, 191-204.