GEO BOOKS

ISSN 0306-6142
ISBN 0-86094-228-7

Published by Geo Books, Norwich
Printed by W:H. Hutchins Sons, Norwich

SPATIAL APPLICATIONS OF EXPLORATORY DATA ANALYSIS

David Sibley

CATMOG

49

ERRATA SLIP FOR CATMOG 49 BY DAVID SIBLEY

ERRATA

page 3 para 2 line 6:	'diagrammetrically' should read 'diagrammatically.'
page 13 para 4:	The fourth sentence should read: 'Proceeding to the outer eighths of the distribution, depth of eighth = (depth of fourth + 1)/2 and, generally, the depth of an outer fraction of the distribution is (previous depth + 1)/2.'
page 15 para 3 line 7:	delete 'For example,'
page 20	captions for Figures 6 and 7 are transposed.
page 24 para 2 line 5:	the formula for the tri-mean should read 'TRI = (FL + F_u + 2M)/4'.
page 25	caption for figure 9: 'Boxplot' should read 'Boxplots.'
page 25	caption for figure 10 should read 'Boxplots of marsh heights (with medians removed)'.
page 27 line 1	'different' should read 'difference.'
page 29	caption for figure 11: '132' should read '1832'.
page 34 para 3 line 4	'act' should read 'fact'.
	Please note that, in all stem-and-leaf diagrams, the stem values and the leaf values should be separated by a vertical line.

CATMOG - Concepts and Techniques in Modern Geography

CATMOG has been created to fill in a teaching need in the field of quantitative methods in undergraduate geography courses. These texts are admirable guides for teachers, yet cheap enough for student purchase as the basis of classwork. Each book is written by an author currently working with the technique or concept he describes.

1.	Introduction to Markov chain analysis	- L. Collins
2.	Distance decay in spatial interactions	- P.J. Taylor
3.	Understanding canonical correlation analysis	- D. Clark
4.	Some theoretical and applied aspects of spatial interaction shopping models	- S. Openshaw
5.	An introduction to trend surface analysis	- D. Unwin
6.	Classification in geography	- R.J. Johnston
7.	An introduction to factor analysis	- J.B. Goddard & A. Kirby
8.	Principal components analysis	- S. Daultrey
9.	Causal inferences from dichotomous variables	- N. Davidson
10.	Introduction to the use of logit models in geography	- N. Wrigley
11.	Linear programming: elementary geographical applications of the transportation problem	- A. Hay
12.	An introduction to quadrat analysis (2nd edition)	- R.W. Thomas
13.	An introduction to time-geography	- N.J. Thrift
14.	An introduction to graph theoretical methods in geography	- K.J. Tinkler
15.	Linear regression in geography	- R. Ferguson
16.	Probability surface mapping. An introduction with examples and FORTRAN programs	- N. Wrigley
17.	Sampling methods for geographical research	- C.J. Dixon & B. Leach
18.	Questionnaires and interviews in geographical research	- C.J. Dixon & B. Leach
19.	Analysis of frequency distributions	- V. Gardiner & G. Gardiner
20.	Analysis of covariance and comparison of regression lines	- J. Silk
21.	An introduction to the use of simultaneous-equation regression analysis in geography	- D. Todd
22.	Transfer function modelling: relationship between time series variables	- Pong-wai Lai
23.	Stochastic processes in one dimensional series: an introduction	- K.S. Richards
24.	Linear programming: the Simplex method with geographical applications	- James E. Killen
25.	Directional statistics	- G.L. Gaile & J.E. Burt
26.	Potential models in human geography	- D.C. Rich
27.	Causal modelling: the Simon-Blalock approach	- D.G. Pringle
28.	Statistical forecasting	- R.J. Bennett
29.	The British Census	- J.C. Dewdney
30.	The analysis of variance	- J. Silk
31.	Information statistics in geography	- R.W. Thomas
32.	Centrographic measures in geography	- A. Kellerman
33.	An introduction to dimensional analysis for geographers	- R. Haynes
34.	An introduction to 0-analysis	- J. Beaumont & A. Gatrell

(continued inside back cover)

CONCEPTS AND TECHNIQUES IN MODERN GEOGRAPHY No. 49

SPATIAL APPLICATIONS OF EXPLORATORY DATA ANALYSIS

By

David Sibley

(Hull University)

ACKNOWLEDGEMENTS

I am very grateful to two anonymous referees and to Tony Gatrell for their criticism and constructive suggestions. I have also benefited from discussions with Norman Davidson and John Pethick, who provided me with some interesting data from physical geography. Particular thanks are due to Keith Scurr for his cartographic skills, and to Janet Smith for typing the paper.

DEDICATION

In memory of Andi Bolton, who drew most of the diagrams.

Sibley, David
Spatial applications of exploratory data analysis.
---(CATMOG, ISSN 0306-6142; 49).
1. Spatial analysis (Statistics)
I. Title II. Series
519.4   QA278.2

ISBN 0-86094-228-7

No responsibility is assumed by the Publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein.

1. INTRODUCTION

Much statistical analysis is concerned with finding structure or pattern in data and in this sense it is exploratory. However, exploratory data analysis (EDA) is a term applied to a distinctive approach to data description and the measurement of association. Here, the techniques, which are very simple to use, are designed to enhance pattern recognition and to uncover data structure. The methods are particularly associated with the American statistician, John Tukey (1977) although some graphical exploratory techniques have a fairly long history, having been developed initially as quick alternatives to classical methods in the precomputer age (Quenouille, 1959). An essential distinction can be made between exploratory and confirmatory methods, that is, the conventional statistical methods based on classical theory and designed to test hypotheses. There is an exploratory equivalent of most confirmatory techniques covered in introductory texts in statistics written for geographers but an informed decision about which methods to use in dealing with a particular problem - exploratory or confirmatory, or both - requires an appreciation of the broad objectives of quantitative analysis in geography. Thus, before discussing particular exploratory methods, I will say something about the place of data analysis in geography and the characteristics of an exploratory approach which make it suitable for geographical investigations.

(i) The process of scientific inquiry

To begin, it will be useful to identify differences in approach to scientific analysis as they are projected in the conventional view and the exploratory perspective, respectively. How scientists go about explaining real-world phenomena is usually portrayed as a sequential, step-by-step process, culminating in a decision about the statistical significance of a pattern or a relationship. This sequence is often represented diagrammetrically as a series of boxes connected by arrows which indicate the correct order of events in the analysis. Most accounts suggest that we proceed from an 'a priori' model to a hypothesis; we then apply a test to the data and either confirm or verify the hypothesis, in which case we move on to theory construction, or we reject the hypothesis and think again about the 'a priori' model. The investigation is represented as a systematic and logical procedure with discrete and clearly defined stages.

A more realistic account of actual practice, however, would indicate that theoretical understanding and analysis are enmeshed in such a way that we cannot separate out the stages of an investigation in the way suggested by the conventional account. We already have a theory, however poorly articulated, when we begin the analysis and this helps us to define the problem and select the data; having used statistical methods to identify pattern and relationships in the data, we may then clarify our original theory which helps us to examine the data from a modified perspective the second time round. Thus, the process of data analysis is an essentially circular process rather than a linear sequence and our theoretical understanding increases in an incremental fashion. As Marshall (1985, p. 126) suggests, 'knowledge evolves through the interplay between theory and observation'. There may be a lot of deliberation involved, with frequent return visits to the data in order to make more sense of them in terms of our theoretical expectations.

The argument in favour of an exploratory approach to analysis is that it encourages and facilitates repeated reference to the data and a cautious, sceptical attitude to theory, although this should be characteristic of any kind of analysis. In EDA, the former is emphasised in the display of data in several alternative forms and the latter by focusing on anomalous or problematic cases. It involves immersion in the initial problem. In practice, this is rather different from the conventional ayproach to quantitative analysis, which presents two particular obstacles to understanding. First, with ritual adherence to hypothesis-testing and decision-making on the basis of probability values, there is a tendency to see a result, that is, the confirmation of a hypothesis, usually at the 95 percent level of probability, as a successful conclusion of the analysis. This is implicit in the use of terms like 'verification', 'success' and 'confirmation'. Several writers have suggested that hypothesis confirmation only discourages researchers from considering new ideas. Tukey (1969, cited by Diaconis, 1985, p. 3), for example, refers to classical statistics as 'a ritual for sanctification' and Nozick (1974, xiii) has argued that, in this kind of analysis results are effectively fiddled. '...you push and shove...until finally almost everything sits unstably more or less in there; what doesn't, gets heaved far away so that it won't be noticed'. Clearly, it is not necessarily the case that scientists using conventional statistical procedures act in this way but this kind of abuse may be encouraged by the rigid adherence to textbook procedures.

The problem of sanctification, recognized by Diaconis, does not arise in exploratory data analysis because there are no results in the sense that a relationship is deemed to have been confirmed or not confirmed. Decision-making is, of course, critical for many practical problems, for example, in medicine or engineering, but in many geographical investigations, the primary use of quantitative analysis is to identify structure, which may then be explained by other methods. We often do not need the decision-making apparatus of inferential statistics, which could be a hindrance to understanding. The fact that we are not drawn into making decisions about the significance of a relationship is a positive attribute of the exploratory approach.

The second attribute of conventional quantitative analysis which presents difficulties is that it may be projected as progressive when it is not. Wilson and Kirkby (1975, p. 4), for example, maintained that geography 'had moved beyond a period of orderly description, via a so-called quantitative revolution into statistical analysis, and into a period when mathematical analysis is becoming both commonplace and fruitful'. This suggests that 'orderly description' is a rather inferior and less worthwhile pursuit than mathematical modelling and this downgrading may account for the neglect of the descriptive devices of exploratory data analysis by quantitative geographers. It is also the case that mathematical modelling in human geography has been heavily criticized both for its normative assumptions, which tend to mask fundamental economic and social cleavages, and for the tendency to translate theory into stultifying practice (Olsson, 1978). Exploratory data analysis does not have these heavy ideological undertones and an important part of the case for exploratory methods is that they fit easily into research designs using mixed modes of analysis. Thus,

apart from the possibly complementary roles of exploratory and confirmatory analysis, EDA might be used to isolate a problem which can then be investigated, say, by questionnaire surveys or by participant observation. Advocates of EDA do not, or should not, make exclusive claims for the methods.

(ii) Statistical properties of exploratory methods

What is it, then, that makes EDA an attractive alternative to conventional methods of data analysis? The essential characteristic of exploratory methods is that they are resistant. This is a very similar idea to robustness which, according to Kendall and Buckland (1971, p.131) applies to a statistical procedure which is not very sensitive to departure from the assumptions on which it depends. Likewise, resistance means that summary statistics are not unduly affected by extreme or anomalous values. However, while all exploratory techniques are resistant, not all confirmatory methods are robust. In the case of descriptive statistics used in conventional analysis, which are based on Gaussian or normally distributed probabilities, the arithmetic mean, variance and standard deviation are very much influenced by unusually high or low values. In EDA, much use is made of the median, for example, as a resistant summary statistic by itself, or to derive other statistics. As the middle value in a sample, it has much greater resistance than the arithmetic mean which may well be useless as a measure of central tendency. Resistance is also important in describing relationships between variables. A least-squares regression, for example, will be unduly affected by an extreme data point; a resistant regression line fitted from median values will not.

If exploratory techniques are resistant, we can attach more significance to extreme values or residuals. As Besag (1981) has suggested, residuals are important in classical analysis but they are of paramount importance in exploratory work. From a philosophical point of view, we could argue that an emphasis on residuals is essential in developing theoretical ideas if we are to use empirical analysis in critical tests of existing theory. If we accept only statistically significant fits as 'results', the tendency will be for existing theory to remain unchallenged. Thus, theoretical understanding will not progress. As Diaconis (1985, p. 1) suggests 'there is a disinclination to revise belief after further observation'. More generally, Rock (1979, p. 64) has argued 'the very word 'knowledge' is mischievous because it suggests a finality or consummation. It should not be a noun but a verb. So defined, knowledge becomes knowing...it constitutes and is constituted by its objects as it unfolds'. The most interesting aspect of a problem may be not, for example, that a regression line adequately summarizes a relationship but that a few values clearly do not fit. It is these residuals that may provide clues to other operating factors not hypothesized initially and the difference between the 'a priori' model and the actual observations is more important than the fit in understanding what is going on. Thus, data exploration along these lines creates more problems that solutions. The inspection of residuals for structure or pattern, which is central to exploratory analysis, helps the analysis to unfold.

A further important feature of EDA, which relates particularly to the question of scientific practice discussed above, is the emphasis on the display of numerical data and the use of graphical summaries. Exploratory methods encourage repeated reference to the original data and to patterns in the data, which makes it easy to identify interesting cases as the analysis proceeds. In conventional analysis, the data are necessarily summarized to the point where individual cases are lost from view. One obvious but often neglected point in support of graphical statistical methods is that the assessment of the outcomes of tests depends very much on visual perception, even when much of the analysis is numerical, as, for example, in regression analysis. We need graphics but graphics which clarify variability and structure in data and do not disguise or mystify.

In the examples of exploratory methods discussed below, the spatial dimension of the problem is emphasized because it is often important in geography and it can be used to provide an additional perspective on the data. An exploratory analysis in a spatial dimension usefully extends analyses of frequency distributions. Having isolated exceptional or anomalous cases on a frequency curve, mapping these cases will enable us to assess their importance as elements of a spatial distribution. In particular, the contiguity of apparently exceptional values and the rest of the data will suggest whether or not these cases are different and deserving particular scrutiny. In this sense, mapping data, particularly residuals, is an obvious complement to Tukey's methods of data display and analysis and this makes EDA particularly appropriate for geography. There should be continuous movement between the individual data point, the frequency distribution and spatial representations of components of the frequency distribution. What Balchin (1970) termed 'graphicacy' and numeracy are essentially fused. (Figure 1).

Combining these three elements in the analysis can be seen as a means of getting to grips with a problem, possibly as a prelude to a confirmatory analysis although it may be decided that exploratory methods alone are sufficient. In the following account, there is no attempt to provide a comprehensive guide to exploratory techniques. Rather, I will focus on methods which have immediate spatial applications, namely, methods for describing and analysing single variables and a resistant regression procedure. For a fairly comprehensive introduction to the subject, particularly for social science applications, Erickson and Nosanchuk (1979) Understanding data is highly recommended. Recent developments in EDA are discussed in two companion volumes, edited by Hoaglin, Mosteller and Tukey, Understanding robust and exploratory data analysis (Hoaglin, 1983) and Exploring data tables, trends and shapes (1985). These two books discuss applications in both physical and social science and they are intelligible to anyone with only a little knowledge of mathematics.

II THE STEM-AND-LEAF DISPLAY

(i) The basic display

The first problem in'analysing a data set is to characterize the sample or batch of numbers as a whole. Before discussing the method, it is necessary to introduce a number of alternative terms to those customarily employed in descriptive statistics. Employing more jargon may seem perverse but the object of this is to provide a more graphic and intelligible terminology than that used in classical statistics. It also serves to set exploratory methods apart from classical methods.

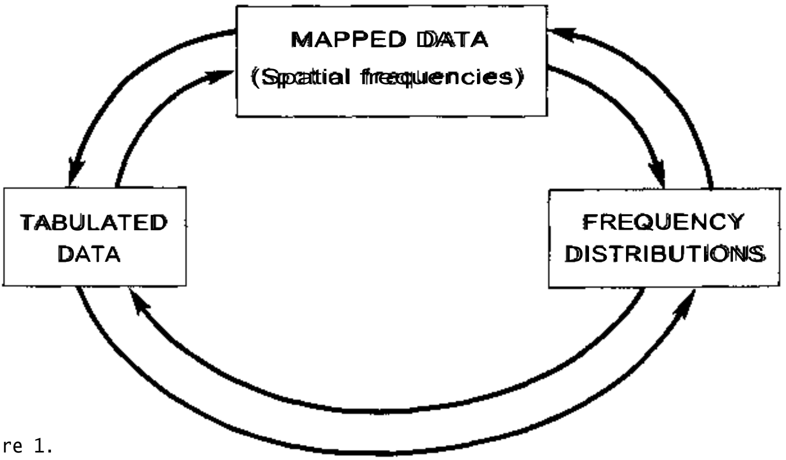

Figure 1.

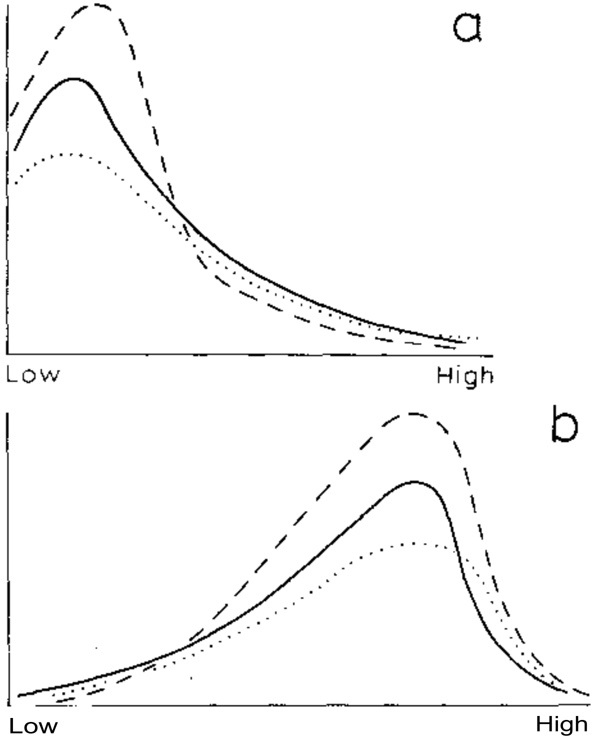

Figure 2. Asymmetric frequency distributions a) with upward straggle b) with downward straggle, and with varying levels of peakedness.

Thus, when describing the shape of a frequency distribution, we refer to symmetr , and <u>peakedness.</u> A curve which is asymmetric but with a sing e peak will either straggle upwards towards the high values (positive skewness), which mean that there is a relatively high frequency of low values and the frequencies are attenuated towards the high end, or straggle downwards towards the low-end (negative skewness). Peakedness, as the term suggests, describes the concentration of values in particular frequencies (Figure 2). Apart from the general shape, we note gaps in the frequency curve and the presence of any values which are detached from the main body of the data. The point of this first inspection is that it may provide clues to the processes which have contributed to variability within the batch.

In order to describe the general shape characteristics of a distribution, some kind of graphical representation is necessary, the histogram being the most familiar device used for this purpose. The grouping of values into class intervals involved in constructing the histogram, however, means losing sight of the numerical values in the first stage of analysis which, for the ruminant quantifier, is undesirable. The *stem-and-leaf* display is a better alternative because the data values comprise the building blocks used in constructing the frequency distribution. It is a histogram where the columns, or rows, are strings of numbers. This makes it possible to identify individual cases in the context of the frequency curve. This is particularly important in the identification of unusual or exceptional individuals, which will be located in the tails of the distribution or will be detached from the continuous part of the curve.

Table 1: Cholera deaths in London, 1849 (per 10,000 population)

Kensington	24	City of London	38
Chelsea	46	Shoreditch	76
St. George, Hanover Sq.	18	Bethnal Green	90
Westminster	68	Whitechapel	64
St. Martin in the Fields	37	St. George-in-the-East	36
St. James, Westminster	16	Stepney	38
Marylebone	17	Poplar	71
Hampstead	8	St. Saviour, Southwark	153
St. Pancras	22	St. Olave, Southwark	181
Islington	22	Bermondsey	161
Hackney	25	St. George, Southwark	164
St. Giles	53	Newington	144
Strand	35	Lambeth	120
Holborn	35	Wandsworth	100
Clerkenwell	19	Camberwell	97
St. Luke	34	Rotherhithe	205
East London	45	Greenwich	75
West London	96	Lewisham	30

The procedure for constructing the basic stem-and-leaf display can be demonstrated using the data in Table 1, showing cholera deaths per 10,000 population for London districts in 1849 (British Parliamentary Papers, 1854-1896). Of particular concern in this case was the threat of infection to the bourgeoisie - this prompted parliamentary attention to the question - so one point of geographical interest would be the extent to which the disease spread beyond the working-class districts of the metropolis. A more general problem in the case of an epidemic disease will be in the relationship of extreme values to the rest of the data because this will provide some clues about the diffusion process. Here, we are concerned essentially with the rate and extent of spread from one or more sources.

To build up the display, each number in the table is partitioned into the leading digit(s) and the trailing digit. The leading digits form the stem (equivalent to the class interval in the histogram) and the trailing digits the *leaves*. Noting the range of values in this case, a suitable stem interval would be 10. Thus, the first entry in the table of cholera deaths, Kensington, appears as:

stem	leaf
2	4

and all other values in the interval 20 to 29 have their trailing digits entered against the 2 stem. In the display, the stems appear in ascending or descending order, as follows:

0
1
2

etc.

Then, the leaves are assigned to their stems. This can be done in two steps, first, by writing down all the trailing digits without regard for order, which creates a *dirty* stem-and-leaf, and then ranking the leaves on each line to give a *clean* stem-and-leaf. Note that, in the process of constructing the display, all the values in the original table are put in rank order so that prior ranking of the data is unnecessary. The two displays are shown below for the cholera data.

dirty		clean	
0	8	0	8
1	8679	1	6789
2	4225	2	2245
3	75548680	3	04556788
4	65	4	56
5	3	5	3
6	84	6	48
7	615	7	156
8		8	
9	607	9	067
10	0	10	0
11		11	
12	0	12	0
13		13	
14	4	14	4
15	3	15	3
16	14	16	14
17		17	
18	1	18	1
20	5	(205)	

One additional feature of the display should be noted. There is a single detached value which is bracketed in order to indicate that it may warrant special attention. It is an essential feature of the exploratory approach that we focus on data values which may be exceptional because they could provide clues to unsuspected processes. It is also evident that, without any analysis, a few summary values and significant points in the distribution can either be read-off or worked out with a little mental arithmetic, for example, the median (45.5) and the modal class (30 to 39). Clearly, it is just as easy to characterize the shape of the frequency distribution as it would be using a histogram. This completes the first stage of the analysis but there are other ways of compiling the frequency distribution which may be more appropriate when the data are in a different form.

(ii) Alternative formats for stem-and-leaf displays

For some batches, using an interval of one leading digit per stem may give too many leaves per line - that is, the display looks too crowded and it is difficult to read. If we think of the problem of displaying the data as one of scaling the stem- and-leaf according to the number of ways we can factor 10, that is, 1x10, 2x5 and 5x2 (Emerson and Hoaglin, 1983, p. 18), there are then three ways of dividing up the data values to give different allocations of leaves per stem.

The 10x1 case is described above. For a 2x5 arrangement, the stem is split, with leaves from 0 to 4 in a* row and those from 5 to 9 in a . row. Suppose we had a sample of house prices in £000's covering a small range, say, from £25,000 to £30,000 the display would be like this:

```
26*     3
26.     7889
27*     01144
27.     558889
28*     0014

        etc.
```

More ingenuity is required to divide the stem into five classes. This can be done by assigning zeros and one's to a* row, two's and three's to a t row, four's and five's to an f row, sixes and sevens to an s row and eight's and nine's to a . row, as in the following plot of beta radio-activity values for 46 states of the United States (Mason, 1970):

```
1*
t     33
f     5
s     667777
1.    99
2*    0000000111
t     2333
f     555
s     66
2.    8899
3*    000
t     33
f     55
s
3.    8
4*    0
t     2

(0.53)
(0.55)
```

(range = 0.13 - 0.55 picocuries per m^3)

If, alternatively, a display appears too straggly, it is possible to make it more compact by having more than one stem per line, separating the respective leaves by a colon. Thus, we could represent the frequencies of cholera deaths more compactly in the following form:

```
0,1    8:6789
2,3    2245:04556788
4,5    56:3
6,7    48:156

       etc.
```

This format is messy, however, and difficult to read. Although it may convey the overall shape of the distribution better than the 10x1 arrange-ment, it should be avoided if possible. In fact, there is often no single best design for the stem-and-leaf diagram. There is a rule of thumb for the maximum number of lines, L$_{ax}$ (10 x to n), which will indicate if splitting or doubling-up stem values is advisable. However, it is probably best to experiment with alternatives to see what information different versions suggest.

Two further variants of the display may be useful. If we want to compare two closely related samples, one possibility is to put them back-to-back with a common stem (as in an age-sex pyramid). For example, comparing gross population densities for London boroughs in 1961 and 1981, we can see similarities in the shape of the distributions but a change in the spread of values:

1961		1981
9	1	9
11	2	019
99766543	3	23455799
841	4	0345
420	5	1478
73	6	77
3	7	3349
64	8	8
6	9	023
94	10	7
3	11	6
65	12	
72	13	

If we want to complement the display of numerical values with geographically specific information, names can be substituted for numbers. This may be useful in analysing a number of batches of data pertaining to areas, where we do not want to lose sight of individual cases while examining overall variability. In the following example, the 1981 population densities for London boroughs are shown separately for inner and outer boroughs, using abbreviated names. The difference in this case does suggest that the 'inner' and 'outer' dichotomy has some demographic significance:

```
          Inner  Low          Outer

                              BROM
                              HAV, HILL, RICH
                              ENF, BARN, NOUNS, BEX, KING
                              CROY, HARR, SUTT
                              RED, MERT, BARK, GREEN
                              EAL, WAL, BRENT
                        NEW
                  LEW, HARI
   T. HAM, WAND, STHK, CAMD
                       WEST
           LAMB, HAM, HACK
                        ISL
                        KEN
                           high
```

A third possibility, not illustrated, would be to put numerical values on one side of a back-to-back diagram and the corresponding names on the other.

The purpose of these diagrams is to give a clear indication of the position of an individual case within the overall frequency distribution, particularly with a view to noting anomalies which may be pursued in greater detail. In a particular analysis, not all the formats illustrated here will be appropriate but it would be a good idea to combine alternatives in an experimental way in order to extract the maximum amount of information from the data. More emphasis is put on this stage of the analysis in exploratory work than in confirmatory statistics precisely because it may yield important clues to problem formulation. Thus, experimenting with displays is to be encouraged.

III. LETTER VALUES AND THE ANALYSIS OF SEVERAL BATCHES

This display of data in a stem-and-leaf diagram is the first step in a search for structure but the displays can also be used in conjunction with summary statistics to provide more specific information on the characteristics of a frequency distribution. These summary statistics are called letter values. The properties of letter values are described here before returning to stem-and-leaf displays to consider their utility in analysing spatial data.

Letter values are 'a collection of observations drawn systematically from the batch, more densely from the tails than from the middle' (Hoaglin, 1983, p. 33). Their most important characteristics, by comparison with conventional summary statistics, such as the sample arithmetic mean and variance, is that they are resistant measures and are, therefore, more reliable indicators of variability and central tendency than their confirmatory equivalents.

(i) Defining letter values

Any letter value is described in terms of its depth which is its upward rank (from the lowest to the highest value) or its downward rank, whichever is smallest. Starting with the median, we can now define a number of letter values which locate increasingly outlying fractions of the frequency distribution.

Thus, the median or middle value has a depth of (n+1)/2 (if n is even, we interpolate between the two middle values). Moving out from the middle of the distribution, the depth of the fourth (also known as the hinge, or more familiarly, as the quartile) is: ¯ depth of median + 1)/2. (The term 'fourth' is preferred to the alternatives because we can then refer to other partitions in the same terms, that is, as fractions). Proceeding to the outer eighths of the distribution (depth of fourth + 1)/2 and, generally, the depth of an outer fraction of the distribution is: (previous depth) +1/2.

In practice, it is unlikely that the outer 1/32 or 1/16th would provide additional useful information unless the sample size is very large but it will often be the case that we want to examine progressively smaller portions of the tail of a distribution or, alternatively, move in from the outer areas towards the middle in order to identify any interesting features of a variable. In order to identify the depth of any value in a distribution quickly, the depths can be entered as a separate column in a stem-and-leaf display. This is simply the accumulated frequency of values from the top and bottom of the distribution. Thus, for the population densities of London boroughs, 1981, the depths column appears as follows:

Depth		
1	1	9
4	2	019
12	3	23455799
16	4	0345
16	5	1478
12	6	77
10	7	3349
6	8	8
5	9	023
2	10	7
1	11	6

In this case, the median falls between 45 and 51. If the median value had been one of the leaves or came between two leaves on a line, that line would be omitted from the accumulated frequencies and the entry against it in the depths column would be the number of values on that line (shown in brackets in order to indicate on which line the median is located). An example is given shortly.

(ii) Letter value displays

The values indicating the various fractions of the frequency distribution, together with the extremes, can be entered in a diagram called a letter value display which provides a concise summary of batch characteristics. The minimal diagram is a 5-number summary:

	name of sample	
M	median	
F	lower fourth	upper fourth
1	lower extreme	upper extreme

where the tags or labels at the side of the diagram indicate sample size (n), the depth of the median (M), the depth of the fourths (F) and the depth of the extreme values (which is of course, 1). This summary is particularly useful if we have to compare a number of batches (Emerson and Strenio, 1983, pp. 68-69). To give an example, in the case of the 1981 population densities of the inner and outer London boroughs, the 5-number summaries give a neat and concise indication of the salient differences between the two zones:

#13	inner London	
M 7	79	
F 4	73	72
I	8	116

#19	outer London	
M10	37	
F 5.5	32.5	43.5
1	19.0	57.0

Similarly, this format could be used for a 7-number summary:

#n	name of sample	
M	M	
F	F_L	F_U
E	E_L	E_U
1	lowest	highest

where F_L, F_U and E_L, E_U indicate lower and upper fourths and lower and upper eights, respectively. Additional information, which can be added to the display, includes the fourth-spread and any outside values, which are discussed below.

(iii) Measures derived from letter values

In addition to such overall summaries, it is desirable to have other measures of spread and concentration based on letter values. The most useful resistant measures of the bunching of sample values are based on the fourths. The first is the fourth-spread or mid-spread, dF, which is simply FU-FL (more familiar as the inter-quartile range). The second is the tri-mean, TRI for short, which is defined as (F_L + F_U + 2M)/4. Although this uses more data points than other measures of central tendency, it is still resistant.

The fourth-spread has particular application in the detection of exceptional values. In locating such values, we need to find a distance from that part of the distribution contained within dF to the tails of the distribution, beyond which we can claim that any data points are outside values. This term can be used interchangeably with 'outliers' because both suggest that these extreme values may have the same underlying behaviour as the rest of the distribution. As Kendall and Buckland (1971, p. 109) put it: in a sample of n observations, it is possible for a limited number to be so far separated in value from the remainder that they give rise to the question whether they are not from different populations or that the sampling technique is at fault'. In an exploratory analysis, however, we are concerned only with isolating possible anomalies, not with confirming their difference from the remainder of the observations. Generally, we remain cautious about identifying exceptional cases. From inspection of the array in a stem-and-leaf display, observations may appear to be outside values because they are detached from the frequency curve. However, arbitrary cut-offs can be calculated to provide a consistent basis for the isolation of outside values. Conventionally, these are F_L - 1.5 d_F and F_U + 1.5 d_F. Because most distributions that we encounter in geography straggle upwards, high outside values are more likely than low outside values. In fact, the lower outside cut-off is usually beyond the range of the data. It is important to bear in mind that the outside cut-off does not unambiguously partition the data. It is preferable to think of it as a fuzzy boundary around which observations are most likely to exhibit different behaviour to the main part of the distribution. Thus, we should inspect values which lie just inside the cut-offs - adjacent values - as well as outside values. In fact, having used letter values to define outer fractions of a frequency distribution, it is desirable to work in from the extremes as far as the mid-spread looking for further elements of pattern in the distribution of values.

There are direct spatial applications of this approach to the description of a variable. Mapping different fractions of a distribution may indicate areas which are anomalous or, alternatively, form part of a larger pattern and this will suggest directions for subsequent investigations. For example, if we look again at the beta radio-activity values for the coterminous states of the USA in the following stem-and-leaf display, there are two detached values which clearly warrant a closer look. For example, these two states, Nevada and Arizona, are, in fact, beyond the outside cut-off. If we map these states; then map the upper eighth (minus the outside values), then the upper-fourth (minus the upper-eighth plus outside values), we can build up a picture of the concentration of radio-activity (Figure 3). Adding the upper fourths and upper eighths to the map puts the outside values into context and provides additional information which helps in deciding whether or not the outside values are exceptional or simply 'highs' in a larger pattern.

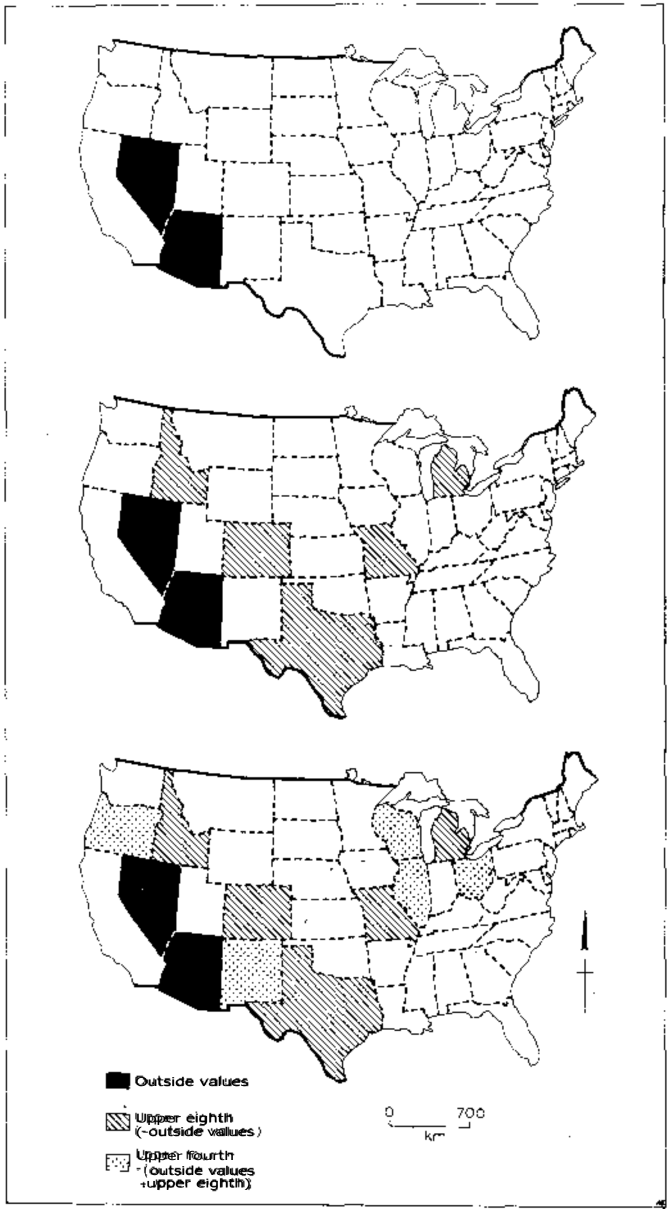
Outside values
Upper eighth (-outside values)
Upper fourth -(outside values +upper eighth)
0
700
km

Figure 3. Levels of atmospheric radioactivity in the United States, 1966 (in picocuries per m^3).

Depths

```
  2      t    33
  3      f    5
  9      s    667777
 11     1.    99
 21     2*    0000000111
(4)      t    2333
 21      f    555
 18      s    66
 16     2.    8899
 12     3*    000
  9      t    33
  7      f    55
         s
              8
  5     3.
  4     4*    0
  3      t    2

  2       (0.53)
  1       (0.55)
```

While the state is not a particularly useful spatial unit for analysis, it is evident that there are some contiguous elements in the pattern which need further investigation, particularly in relation to weapons-testing sites, nuclear power generation, natural background radiation and the atmospheric diffusion of radio-activity. Apart from noting that examination of the spatial dimension of the problem in an incremental fashion may discourage the analyst from rushing to conclusions, it is worth stressing that we can easily identify the original data values while summarizing the characteristics of the frequency distribution and the spatial distribution, which is an important advantage of using a stem-and-leaf display in conjunction with letter values in a mapping exercise. They provide complementary perspectives on a problem.

IV. BOX PLOTS

While letter value displays are concise and can be used to give a fairly full specification of a frequency distribution, it may be the case that we need a more immediate impression of spread and asymmetry than can be indicated by a set of numbers, particularly when there is a large number of batches to compare. For this purpose, the *boxplot* (alternatively known as the box-and-whisker plot) is a useful graphical device, providing a diagrammatic summary of the information in a basic letter-value display.

(i) Using a boxplot

If we have a letter-value display of the summary values, it is an easy step to portray this information in graphical form as a boxplot. For example, the radiation data for the United States can be summarized as:

n=46	picocuries per m³		
M 23.5	0.23		
F 12	0.20	0.30	outside cut-offs:
1	0.13	0.55	0.04, 0.46

(outliers: 0.53, 0.55)

and the corresponding boxplot is shown in Figure 4.

The vertical line inside the box locates the median. Spread is indicated by the length of the box, defined by the position of the fourths, and the position of the median line in the box conveys something of the skewness or straggle of the distribution. Thus, in Figure 4, the closeness of the median line to the left side of the box indicates upward straggle. The line extending from the box to the shorter vertical line gives the position of the upper and lower outside cut-offs in relation to the main body of the data and the dots locate the outside values. It is, therefore, easy to read off the essential features of a frequency distribution from the boxplot. Apart from providing an economical summary, it is also resistant because it is based on resistant statistics. By comparison, using the sample mean and standard deviation to indicate the same characteristics of a distribution would be misleading in the presence of a single 'wild' value.

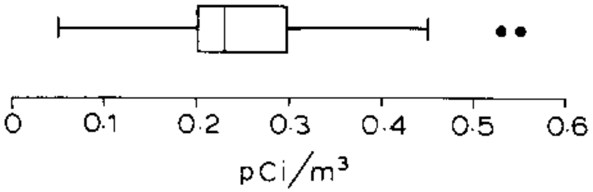

Figure 4. A boxplot

The benefits of the boxplot become clearer when it is used to compare a number of related samples. The population density data for London boroughs might be appropriate for this purpose because the stem-and-leaf displays shown earlier suggest significant changes over time and between inner and outer boroughs which it would be helpful to examine simultaneously. The plots of densities in Figure 5, showing spatial and temporal variability, suggest a clear convergence of inner and outer boroughs although the lack of overlap is maintained until 1981. This convergence will have resulted primarily from change in inner London and we might reasonably suggest that housing renewal and outmigration have caused a reduction in densities in inner London, bringing it closer to the densities of the suburban boroughs. Although the information is presented in a highly condensed form in the diagrams, the essential summary characteristics of the samples are retained. Information loss is slight and we would be able to refer to the individual data in the stem-and-leaf displays which were used to calculate the summary statistics.

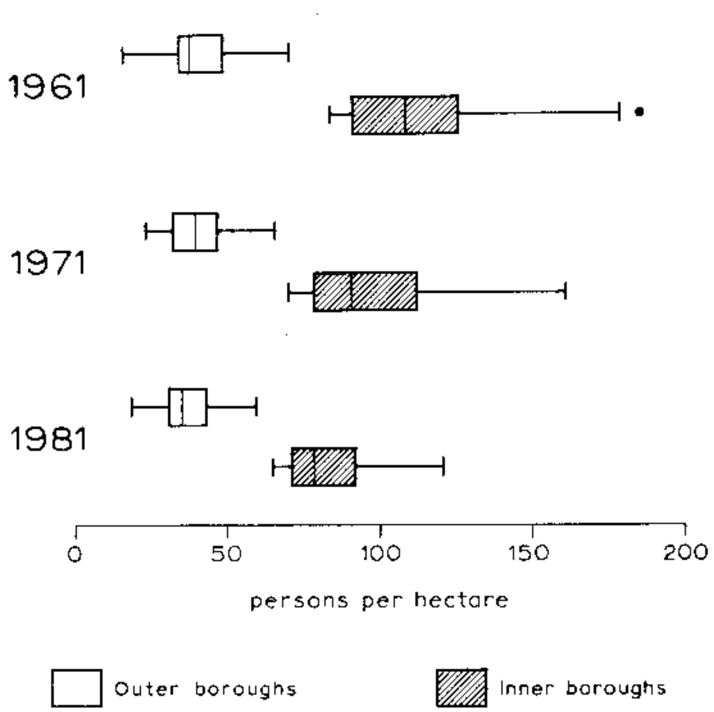

Figure 5. Boxplots of population densities for inner- and outer-London boroughs

(ii) Removing levels and spread

While it is possible to recognize patterns of variability by looking at boxplots, as in the previous example, it is occasionally useful to focus on one feature of a frequency distribution at a time when comparing a number of batches. It could be the case that differences between the original boxplots are not obvious or that there is a large number of comparisons to make and the information contained in the plots is difficult to digest.

A systematic approach to this problem is to set aside summary values in order to observe other effects. As a first step, for each sample in turn, the median can be subtracted from the original sample values so that differences between sample medians are removed as a source of variation. This is called removing a level and the effect is simply to centre all sample plots on zero, which makes it easier to observe differences in spread. The second step is to divide the original values in each sample, with the median subtracted, by the fourth-spread. This removes the effect of spread and the data are now in standard form. Thus, the standard score = (observation-level)/spread which is the exploratory equivalent of the z-score for values in a normal distribution, that is, z = (x-x)/standard deviation. Diagrammatically, the change from plots of raw data to standarized plots is shown in Figure 6.

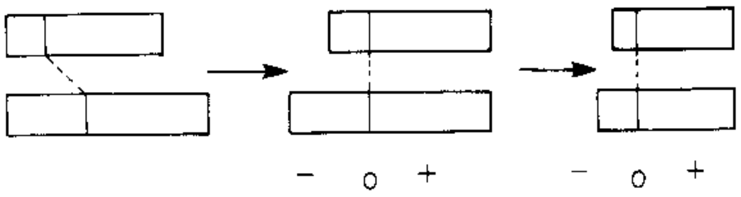

Figure 6. Boxplots of city-size distributions, with medians removed

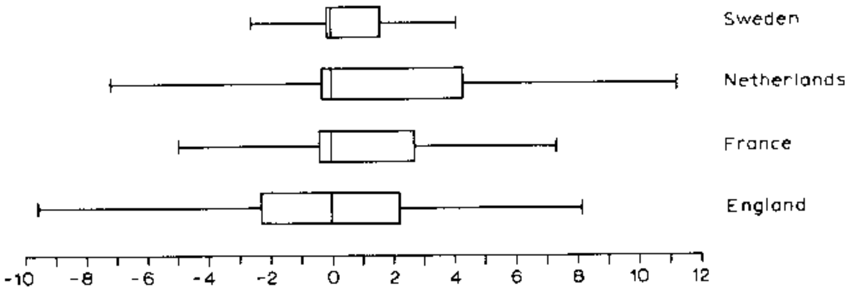

Figure 7. Standardisation of boxplots

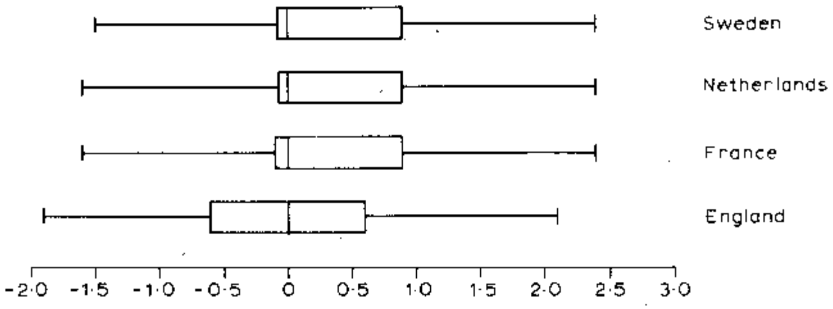

Figure 8. Standardized boxplots of city-size distributions

Emerson and Strenio (1983, pp. 65-70), in discussing applications of boxplots, give the example of city-size distributions, where the boxplot can be used as an alternative to rank-size curves to describe regularities in city-size frequencies. It is a method which might be preferred to rank-size models, when the latter are used as descriptive tools. Boxplots are more simple to construct and they are solely descriptive, unlike rank-size models, which have had an ambiguous role, as both descriptive devices and as process models, with the processes ill-defined. The plots for sixteen countries are rather difficult to read, however, and comparisons could be easier to make if level and spread were removed. The method is illustrated here with a selection of European states taken from Emerson and Strenio's data set (from the World Almanac, 1967). First, with the median removed (Figure 7), variation in the spread of city sizes becomes the most obvious source of difference, with a marked contrast between Sweden and England, for example. This diagram also suggests variation in the symmetry of the distributions but this shows up more clearly in the standardized plots (Figure 8); where England is clearly different from the rest in the sense that city sizes (excluding the largest city) are symmetrically distributed around the median whereas the other states all exhibit upward straggle (note that, in the case of Sweden, France, and England, the largest cities are outside values which cannot be shown at this scale). There are serious problems of comparability in using national census data on city sizes which suggest that it would not be worthwhile pursuing this example very far, but it does illustrate the possibilities, particularly where we want to compare a large number of batches (see Erickson and Nosanchuk, 1979, chapter 4, for other examples).

(iii) A summary example

The value of using a number of displays in combination can be demonstrated in an investigation of a problem in coastal geomorphology where theoretical knowledge is presently rather meagre and the discovery of patterns in the data may suggest some directions for further research. The problem concerns the development of salt marshes. From a general knowledge of processes, it could be argued, first, that the height of the marsh increases with age and, secondly, that the processes of sediment deposition lead to increasing uniformity of height; an irregular surface in the early stage of development gives way to a smoother, level surface. The data which are used to examine these propositions, and to examine other unanticipated problems, comprise samples of marsh levels (surface heights) in metres, collected for three marshes on Scolt Head Island, Norfolk. These are, in order of increasing age, Anchor, Missel, and Hut. Anchor Marsh is c. 20 years old, Missel Marsh c. 80 years, and Hut Marsh c. 120 years.

Before any analysis, it was clear from inspection of the data that Hut Marsh divided into two zones, referred to here as Hut 1 and Hut 2, with the former markedly higher than the latter. Thus in the first look at the distribution of values in stem-and-leaf displays we have four samples:

Hut 1

```
4.6 9
4.7 0333455677888888899
4.8 0001122222233344555557778888999
4.9 000001111222222223333333334444455556666777777788888899
5.0 011111112233334444444555666667889
5.1 012334455569
5.2 01246
5.3 39
5.4 348
5.5 5
5.6 4
5.7 6889
5.8 9
5.9
6.0 278
6.1  4
6.2
6.3
6.4 5
6.5 5
6.6 26
6.7
6.8
6.9
7.0
7.1  5
.7.2
7.3 5
7.4  1
7.5 5
7.6 5
7 .7 44
7.8
7.9 3
8.0
8.1
8.2
8.3
8.4
8.5 4
8.6
8.7 16
8.8
8.9 8

 (9.21)
 (9.71)
(10.21)
```

Missel Marsh

```
1.5 6
1.6
1.7 35
1.8 259
1.9 05679
2.0 38
2.1 0125567788
2.2 1223335566667777888999
2.3 0011111111122222223333444445555555666666666666677777777888888888899999999
2.4 000000000111111222222334444455555567 8999
2.5 000011122222355556688999
2.6 000112344445556677 78889
2.7 0011344466778888899
2.8 1122455666778889
2.9 03366
3.0  1
3.1  4
3.2 68
3.3 9
3.4 56
3.5 4
3.6 0
```

Hut 2

```
1.7     1
1.8
1.9
2.0
2.1     0788
2.2
2.3     579
2.4     023689
2.5     000224555556777788889
2.6     0002222222333444555555666666677788889999
2.7     111111222333333344444444444444666666667777788888899
2.8     000123334455556677 7889
2.9     02223446668
3.0     02344456679
3.1     0124556666899
3.2     0001234468
3.3     599
3.4     1
3.5     18
3.6     33
3.7     33
3.8     147
3.9     04
4.0     034
4.1     2
```

Anchor Marsh

```
1.7     9
1.8
1.9     7
2.0     1
2.1     445799
2.2     0112344446677
2.3     1223456
2.4     0245568
2.5     01123478999
2.6     2345599
2.7     12223357999
2.8     0034
2.9     0447
3.0
3.1     0
```

The absence of a single modal frequency on Anchor suggests that the initial idea about the morphology of a recent marsh was correct. In contrast, both Missel and the two Hut samples are highly peaked and the distribution of heights appears to be in a lower frequency range on Missel than the latter which is consistent with the argument that height increases, but variability decreases, with age. Hut 1 is anomalous, however. It is much higher than the other marshes and the height distribution shows much more upward straggle than the other samples. Thus, we must have reservations about the comparability of the batches. Does Hut 1 reflect the operation of different processes to those forming the other marshes?

In order to put this information into a more readily digestible form, it is helpful to summarize the key parameters in letter value displays. These indicate that, overall, there is no clear pattern:

n=184	Hut 1	
M 92.5	4.97	
F 47	4.89	5.14
E 24	4.81	5.96
1	4.69	10.21

n=188	Hut 2	
M 94.5	2.74	
F 48	2.64	3.04
E 24.5	2.55	3.54
1	1.71	4.12

n=252	Missel	
M 126.5	2.41	
F 64	2.26	2.81
E 32.5	2.13	2.92
1	1.56	3.60

n=34	Anchor	
M 17.5	2.51	
F 9	2.26	2.72
E 5	2.19	2.80
1	1.79	3.10

While the individual letter values do not indicate a general increase in elevation from the youngest to the oldest marsh, the tri-mean might be a useful summary because it incorporates a-high proportion of the data, while excluding the extremes. These means are: Anchor, 2.50; Missel, 2.47; Hut 2,2.79; and Hut 1, 4.99, where TRI=(F + F + 2.median)/4. Again, the results are not entirely consistent with the initial hypothesis.

We can now try boxplots, which may give a clearer indication of the spread of values. The plots based on raw data (Figure 9) only underline the difference between Hut 1 and the others. In particular, the large number of upper outside values is curious. The other three batches are similar but, of these, Hut 2 is the only one with outliers. Possibly, the presence of outliers is connected with variation in marsh building material, sand giving higher elevations than mud. If we subtract the median in order to get a better view of spread (Figure 10), there are some suggestive patterns. In particular, the reduction in spread with age (with the exception of Missel) would be consistent with the idea of a levelling of the surface with increased elevation but more samples would be needed to establish this with conviction. A standardized plot was also tried but this provided no useful additional information.

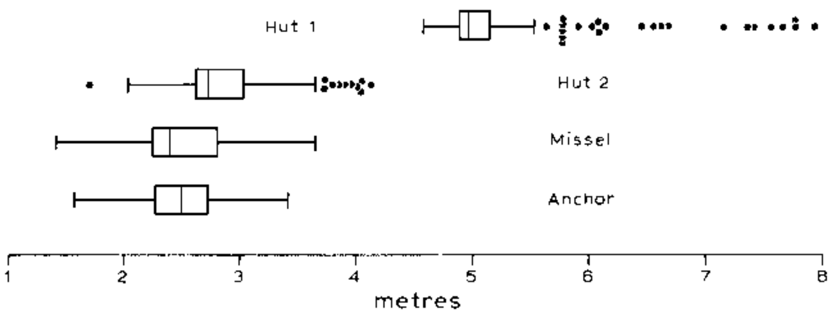

Figure 9. Boxplot of marsh heights (in raw form)

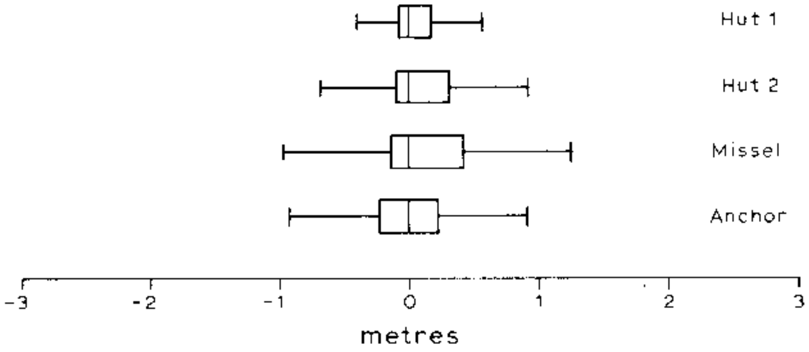

Figure 10. Boxplot of marsh heights (standardised)

Thus, from this inspection, we could at least decide to consider Hut 1 apart from the other sample areas. This is, in fact, the inner zone of Hut Marsh which has been subject to considerable human interference, with the deposition of sand and construction of brushwood fences to encourage accretion. The processes operating here are distinctly different from those affecting the other marshes. In addition, the sample area of Hut 1 may include the boundary between marsh and dune coast, which makes its inclusion in the analysis more inappropriate. The exploratory investigation also suggests that more data collection and analysis would be required to test the development model adequately. For example, an increase in the number of sample points on Anchor Marsh would be desirable if valid comparisons were to be made between this and other areas. Thus, the exploratory approach does appear to have served a purpose. It has picked up features of the data which more highly aggregated methods of data description would have missed and it has provided some pointers for subsequent investigations.

Thus, we can conclude that, in trying to make sense of a data set, it is advisable to use a number of exploratory methods because they make use of different amounts of information and may pick up different characteristics of a frequency distribution, or have varying emphasis. This apart, sifting and organizing the data repeatedly helps the investigator to gain a perspective on the problem and to shape questions about patterns that emerge from the analysis. For measured or counted data, stem-and-leaf displays, letter value displays and boxplots could be used routinely in preliminary analysis. When describing data collected by areas, maps of letter-values constitute a useful complement to the methods for describing frequency distributions, particularly where the objective is to identify problem areas or regions. This theme is pursued in the next section, which deals with an exploratory analysis of association between spatially-distributed data.

V. RESISTANT LINES AND RESIDUALS

In the linear regression model, we are trying to predict values of y in terms of x, assuming that there is a systematic linear relationship between the two variables, plus a random component. Thus, in the equation $y = a + bx + e$, the principal interest is in the line determined by the intercept, a, and the slope coefficient, b, with randomness in the residuals, e, being one test of an adequate fit. As I suggested in the introduction, however, in an exploratory analysis, structure in the residuals may be more interesting than the slope of the line because residual patterns may show up processes which have not been predicted by the equation. Because the regression line is resistant, it is possible to have both pattern in the residuals and a reasonable estimate of the linear relationship (although there may be other relationships not accounted for by the linear term). In the more commonly used least-squares fitting procedure, by contrast, first, correlated residuals violate the assumption of the model that the error term, e, is a random component and, secondly, a single exceptional value can influence the slope of the line to the extent that it becomes a fairly meaningless summary of the overall relationship between the variables. Silk (1979, pp. 242-3), for example, discusses a case where change in service employment in the Reading urban area was regressed on the distribution of service employment at the beginning of the period; apparently on the grounds that growth in employment would be positively related to the preexisting pattern. However, the least-squares regression included one exceptional data point (the central area of Reading) which reflected very rapid development of offices in this locality during the period being considered. Because of the undue influence of this data point on the parameters of the equation, the regression line is of dubious value as a measure of the overall relationship between the variables. As Emerson and Hoaglin (1983, p. 129) put it rather dramatically, 'A wild data point can easily seize control of the fitted line and cause it to give a totally misleading summary of the relationship between y and x.'

There are several ways in which this problem can be handled. With least-squares regression, one approach is to calculate the regression line using all the data and, then, to recalculate the equation having removed

aberrant data points. The different between the two lines indicates the influence of the exceptional values on the overall pattern. Alternatively, observations with large residuals can be weighted to reduce the influence of these points on the slope of the line. Both these methods could be termed robust. The exploratory approach is more straightforward, however. It involves fitting one line which is derived from the medians of the observations, divided into three more-or-less equal sized groups. Because the median is a resistant measure, the regression line is also resistant and exceptional values do not present a problem. At the end of this chapter, the difference between a least-squares regression and an exploratory, resistant regression, where exceptional values are present, is illustrated, using Silk's Reading service employment example.

As with other exploratory techniques, the resistant regression procedure is a means of probing the data to uncover structure. We start with an approximation of the linear relationship:

$$y = a + bx + e$$

which can be used to calculate a set of residuals from the predicted values, where

$$r_i = y_i - (a + bx_i)$$

As a second step, an iterative procedure to improve the fit may be desirable (Emerson and Hoaglin, 1983, p. 134), but in many practical situations in geography it may not be possible to fit a line that is anything more than a rough approximation, in which case the iteration will be irrelevant. If the data used in the regression are for areal units, for example, the fit will be influenced by the level of aggregation of the data which, in many cases, we can do nothing about so fine adjustment of the slope parameter will only give an illusion of increased accuracy. Thus, a single fit, followed by residual analysis, may be justifiable in some geographical analyses. This kind of problem is discussed here, following an account of the basic fitting procedure for a resistant line based on medians. For a full account of iterative procedures, with worked examples, see Emerson and Hoaglin (1983, pp. 129-163).

(i) Fitting a line

The resistant method involves estimating three summary points derived from the medians of the x, y data divided into three groups. Before Tukey (1977), similar methods had been suggested by Bartlett (1949) and Quenouille (1959).

First, the x values are ranked so that $x_1 < x_2 < \ldots x_n$ i.e. the order that occurs from left to right on a regression curve. The paired values, (x,y) are divided into three groups, left, middle and right, as nearly equal in size as possible. If there are 'odd' sized groups the allocation of values to left, middle and right groups in order to achieve a balanced distribution is:

	Size of group		
Group	n = 3k	n = 3k + 1	n = 3k + 2
Left	k	k	k + 1
Middle	k	k + 1	k
Right	k	k	k + 1

In the event of tied ranks, however, all x's having the same value are allocated to the same group, making the method rather unsatisfacotry if there is a large number of ties. There is no guide as to how many ties are necessary before an analysis is invalidated, however.

The next step is to calculate summary points which are given by the coordinates of the median x and median y within each group. Labelling the coordinates Left, Middle and Right, we have (x_L, y_L), (x_M, y_M) and (x_R, y_R). These coordinates are then used to estimate the slope and intercept. The slope is given by:

$$b = \frac{y_R - y_L}{x_R - x_L}$$

and the slope parameter is used to find the intercept, as follows:

$$a = 1/3\ \{(y_L - bx_L) + (y_M - bx_M) + y_R - bx_R)\}$$

If the x, y scatterplot suggested that a linear equation was appropriate, this procedure should give a reasonable estimate of the relationship.

The final, and arguably, most important step in the analysis is to calculate the residuals, $r_i = y_i - (a + bx_i)$, and to inspect them for pattern. The most effective way to approach the latter is to construct a stem-and-leaf display of residual values and to identify outliers and letter values, which will help to locate those residuals which depart substantially from the estimated relationship. In residual analysis, transformations may be advocated but there is a good case for using unmodified or raw residuals rather than reaching for a transformation. As Goodall (1983, pp. 220-221) has pointed out they are in the same scale as the original y values so we can apply subject matter judgement to them'. With spatial data, the residuals can be mapped and patterns of contiguity will provide some indication of structure.

(ii) Cholera in London

The use of a resistant linear regression is demonstrated in an analysis of the 19th century cholera epidemics in London. Specifically, the problem is to identify the spatial correspohdence between the 1849 and 1832 epidemics for thirty districts for which data are available in both years. We could argue that the pattern of mortality in the two epidemics would be broadly similar because of the relationship of the disease to environmental conditions, assuming relative stability of areas of poor environment in London during this period. From inspection of the data (Table 2) and a scatterplot of the values (Figure 11), it is clear that a least-squares regression would be inappropriate because there are a few very extreme values, notably St. Botolph in 1832 and Rotherhithe in 1849, which would have an excessive influence on the parameter estimates.

Table 2: Cholera deaths in London, 1832 and 1849

	1832(x)	1849(y)		1832(x)	1849(y)
Kensington	17	24	Bethnal Green	25	90
Hanover Square	13	18	Stepney	45	38
Chelsea	25	46	Poplar	57	71
St. James	38	68	St. George, E.	32	36
Marylebone	18	17	Hackney	2	35
St. Pancras	11	22	St. George	74	164
Islington	10	22	St. Olave	145	181
Clerkenwell	14	19	St. Saviour	131	153
Strand	32	35	Lambeth	38	120
Holborn	12	35	Wandsworth	14	100
St. Giles	53	53	Newington	45	162
City	29	38	Camberwell	37	97
Shoreditch	8	76	Bermondsey	71	161
Whitechapel	92	64	Rotherhithe	14	205
St. Boltoph	289	45	Deptford	28	65

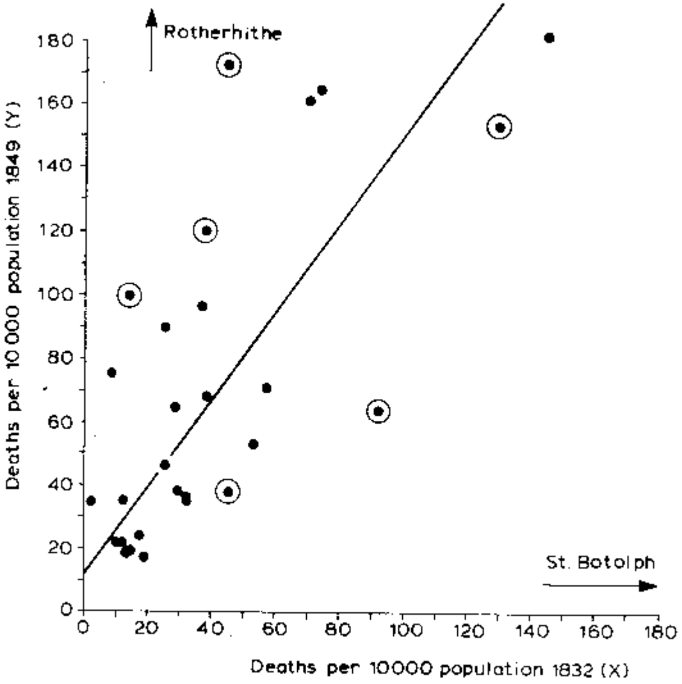

Figure 11. Resistant regression of 1849 against 132 cholera deaths in London (outside values are circled).

Calculating a resistant line

There are 30 observations, giving three groups of 10 for which the medians are:

$$x_L, y_L = 12.5,\ 28.5 \qquad x_M, y_M = 30.5,\ 55.5 \qquad x_R, y_R = 72.5,\ 112.0$$

To calculate the slope parameter:

$$b = \frac{y_R - y_L}{x_R - x_L} = \frac{112.0 - 28.5}{72.5 - 12.5} = \frac{83.5}{60.0} = 1.39$$

The intercept is given by:

$$a = 1/3\ \{(y_L - bx_L) + (y_M - bx_M) + (y_R - bx_R)\}$$

$$= 1/3\ (11.125 + 13.11 + 11.225)$$

$$= 11.82$$

Thus, the equation for the resistant regression line is y = 11.82 + 1.39x, which, judging by the scatter of residuals around the line (Figure 11), gives a reasonable fit. The slope parameter does not suggest very much apart from an overall increase in rates. A residual plot in the form of a stem-and-leaf display, however, suggests some interesting features in the relationship:

```
(-369)
 -7 6
 -6
 -5
 -4 1
 -3 226
 -2 001
 -1 224
 -0 145
  0 37
  1 4
  2 0
  3 457
  4 39
  5 035
  6 9
  7
  8 8

 (174)
```

There are two predictable outside values - St. Botolph which had the highest mortality in 1832 but was closer to the median in 1849 and Rotherhithe, which had the highest mortality in 1849 but was low in 1832. We did not need a regression analysis to identify these extreme cases but when they are mapped, together with the outer eighths and fourths (figures 12 and 13) more interesting patterns emerge. The contiguity of the positive residuals, locating markedly higher death rates in 1849 than 1832, suggests

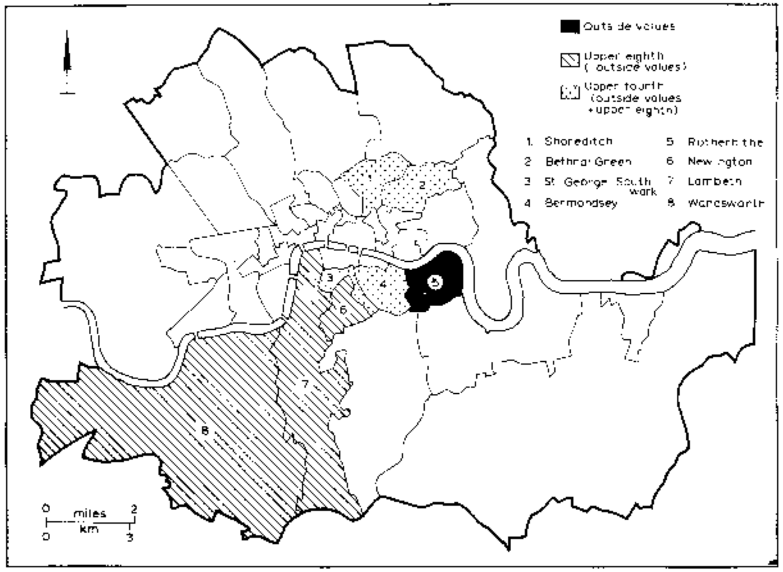

Figure 12. Cholera deaths in London, 1832 and 1849: positive residuals.

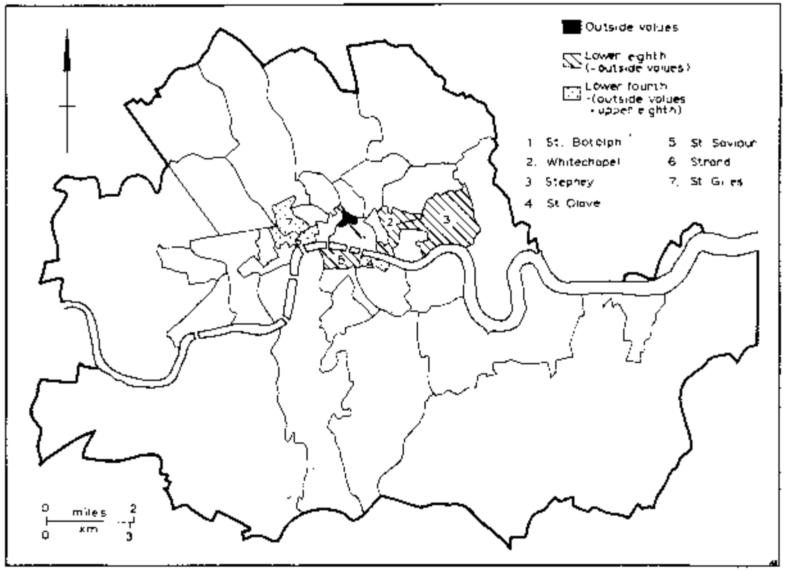

Figure 13. Cholera deaths in London, 1832 and 1849: negative residuals.

a clear pattern of diffusion in south London. The negative residuals present a more complex picture. An imaginative interpretation of the results would be possible but, really, no consistent pattern emerges and the analysis would have to move in a different direction in order to produce meaningful information. A next step might be to consult the archives and the epidemiological literature in order to identify the particular conditions which were conducive to the spread of cholera.

In a case like this, where we should be conscious of imperfections in the data, including incomplete reporting and wrong diagnosis of illness, plus the problem posed by fixed areal units which are highly variable in size and shape, a method of analysis which is unencumbered by statistical requirements derived from probability theory seems a sensible one to use. In a similarly problematic case involving air pollution data of doubtful reliability for cities in the United States, Diaconis (1985, p. 13) maintains that exploratory analyses of association have not only proved more appropriate than confirmatory methods because of measurement errors and high levels of temporal and spatial variability in the data but have been more successful in identifying pollution mechanisms. In another example, Emerson and Hoaglin (1985, pp. 241-275) demonstrate how the resistant procedure can be used in a multivariate analysis, involving mental hospital admissions, where data are similarly problematic.

(iii) Comparing a resistant and least-squares regression.

Fitting a resistant line to the service employment data used by Silk (1979) to illustrate the application of least-squares regression, nicely illustrates the affect of an extreme value on the slope parameter using the latter procedure. The data are listed below, where x is service employment 1961, and y is change in service employment 1961-66:

x	y	x	y
200	0	1009	112
227	0	1056	266
234	9	1672	310
369	11	1880	984
399	6	2167	401
427	48	2462	456
550	18	2493	699
570	0	2827	535
707	141	3091	409
797	124	4409	816
834	129	29288	5187
903	21		

The parameters for the resistant line are:

$$b = \frac{617.0 - 7.5}{2660.0 - 384.0} = \frac{609.5}{2276.0} = 0.268$$

and

$$a = 1/3\{(-95.12) + (-113.0) + (-95.88)\} = -101.432$$

Thus,

$$y = -101.432 + 0.268x$$

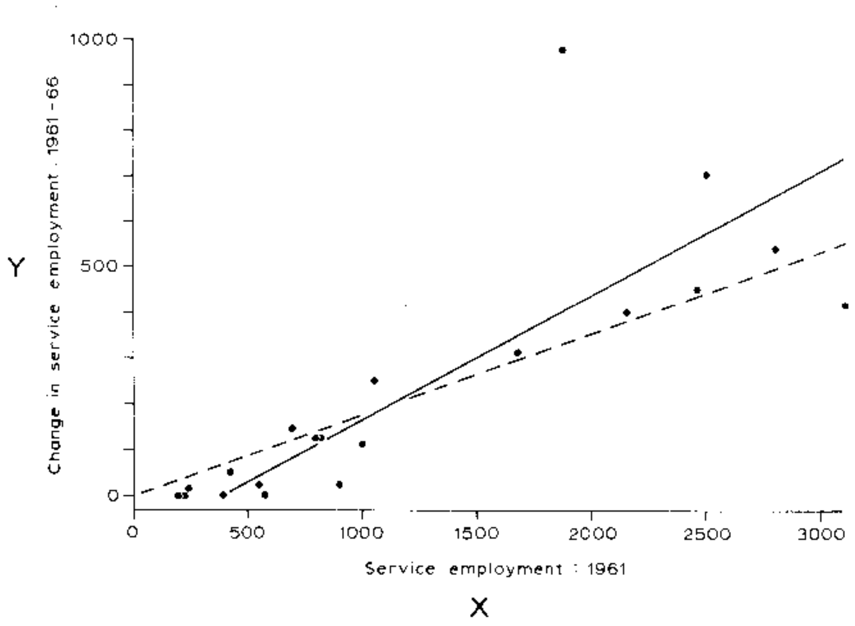

Figure 14. Service employment in Reading, 1961 and 1966: least-squares (broken line) and resistant fits. (maximum xy value not shown).

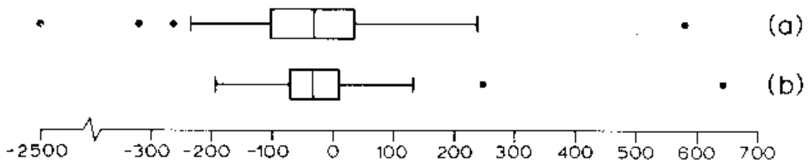

Figure 15. Service employment in Reading: boxplots of residuals from resistant fit (a) and least-squares fit (b).

In Figure 14, the broken line is the least-squares regression line and it is evident that it has been flattened by the extreme x value (29,288) for Reading central area) when compared to the steeper resistant line which has a slightly more symmetrical distribution of points around the line. Inevitably, therefore, the two methods give different patterns of residuals, and as the boxplots of residuals in Figure 15 indicate, a different pattern of outlying values. While the resistant method only approximates the relationship, it is arguable that it gives a more reliable estimate than the non-robust least-squares regression.

Although the resistant regression procedure is advocated here as an alternative to conventional methods where wild values are present, the primary emphasis has been on residuals because the smooth or predictable component of the data - that aspect of the problem expressed in the regression equation - may well be unremarkable, often confirming existing, rather weak theoretical expectations. Residual analysis helps to move the study on because it is the identification of differences and anomalies that points to unexamined problems. While the analysis of residuals from least-squares regression is fraught with difficulties, particularly where the investigation has a spatial dimension, an exploratory approach to residuals, using a combination of stem-and-leaf displays, boxplots and residual maps, appears to be statistically sound and potentially fruitful.

IV. CONCLUSION

Since Gregory (1963, p. 18) published a quartile dispersion diagram of rainfall variability, geographers have shown little interest in exploratory approaches to data analysis. Seduction by inferential statistics and then by systems analysis left more basic descriptive approaches out in the cold. Many quantitative human geographers in particular pursued the goal identified by Warntz (1973, p. 89), according to whom '...a heightening of the level of abstraction is the significant thing'. Some erstwhile quantifiers then rejected quantitative data analysis completely although at the time (the early 1970s) they were probably unaware of Tukey's work on exploratory data analysis.

Generally, the use of exploratory data analysis might be seen as a return to methodological simplicity, to methods which are more in tune with the objectives of geography as an empirical science than the classical statistics and mathematics which have characterized quantitative analysis in the subject. This does not mean that analysis is necessarily less sophisticated or profound but it might be more realistic. In human geography, specifically, it would seem possible to incorporate EDA in critical analysis, combining modes in the way suggested by Castells in The City and the Grassroots (1983), for example, where quantitative ordering of data and qualitative interpretation are integrated. EDA could be used quite effectively in empirical research in human geography set within a broader framework of social and political theory. Data analysis and investigations informed by political and social theory are not incompatible so long as we remain conscious of the fact that the production of data is part of the problem. We need to be aware of the intentions of those who produced the data.

Two features of EDA make it suitable for this kind of critical enquiry. The first concerns the question of fallibility. Practitioners of quantitative social science have been criticized for making claims of total calculability and for assuming a separation of act and value (Young, 1979, pp. 63-74). This follows from an attachment to statistically significant results which then assume an objective or reified existence. As Cochran (1972, cited by Good, 1983) suggests, however, we can only claim to be groping towards the truth' - an admission which exploratory analysts may be more inclined to accept than those wedded to a model-based paradigm and using classical statistical procedures. The former group may become more aware of the complexities in data in the process of working through a problem. In this regard, the implicit difference in the objectives of EDA and inferential statistics has been expressed quite neatly by Good (1983). He argues that, in probabilistic terms, EDA is concerned with partially ordered subjective probabilities rather than 'sharp' probabilities, where a clear boundary is drawn between what is probable and what is improbable. In describing variability in a sample using letter values, for example, we make a subjective ordering of probabilities from the extremes to the middle but do not rely on numerical cut-off points for the confirmation of hypotheses. Rather, the process is one of 'successive deepening', that is, removing successive layers of the data and suppressing different amounts of information at a time, in order to gain an understanding of data structure.

Implicit in much of the foregoing discussion of methods have been claims for the utility of EDA in assessing the validity of established theory. In regard to conventional scientific practice, Kuhn (1962) and Feyerabend (1975) have argued that science is a highly conservative enterprise in the sense that theories tend to be confirmed rather than rejected, particularly if they support the dominant ideology. Paradoxically, however, modellers may have a cavalier attitude to data, ignoring irregularities when they are impressed by the virtues of a model. As Kennedy (1979, p. 558) put it, in her parody of systems modelling '...let the mathematical modelling of the naughty world continue apace but let us not confuse those models with reality.' Because the real world is in the forefront of an exploratory analysis, or, rather, that part of it which can be represented legitimately in quantitative form, this confusion should be less likely to occur. For example, we have argued that, in regression analysis where most attention is directed to anomalies or secondary patterns, the analyst should not be seduced by the fitted line. Divorced from theory, EDA is useless, but as a component of amixed-mode analysis, it could add strength to critical analysis by ordering reality in suggestive ways.

In a narrower geographical sense, exploratory methods might help in the resurrection of cartography, as has been attempted, without notable success, in the use of inferential statistics to analyse spatial data. The use of letter values to describe spatially distributed data is an instance of successive deepening which may have value in cartographic analysis and the same could be said for exploratory approaches to residuals where secondary and subsequent patterns are inspected for contiguity. The simplicity of this approach, rather than providing an excuse for rejection (because, it might be argued, deep insights will only be gained by the use of sophisticated methods) should be appreciated as a virtue. Thus, exploratory data analysis, cartography and other graphics could be combined in a modest attempt to uncover spatial structure and the routine use of these methods would strengthen the claim that the study of geography is concerned both with numeracy and graphicacy.

BIBLIOGRAPHY

Bartlett, M.S. (1949), Fitting a straight line when both variables are subject to error. *Biometrics,* 5, 207-212.

Besag, J. (1981), On resistant techniques and statistical analysis. *Biometrika,* 68(2), 463-469.

British Parliamentary Papers, Sessions. (1854-1896), London.

Castells, M. (1983), *The city and the grassroots,* (Edward Arnold, London).

:ox, N.J. and K. Jones (1981), Exploratory data analysis. In: *Quantitative geography: a British view,* (eds) N. Wrigley and R.J. Bennett, (Routledge and Kegan Paul, London), 135-143

Cochran, W.G. (1972), Observational studies. In: *Statistical papers in honour of George W. Snedecor,* (eds) T.A. Bancroft and S.A. Brown, (Iowa State University Press, Ames, Iowa), 70-90.

Diaconis, P. (1985), Theories of data analysis: from magical thinking through classical statistics. In: *Exploring data tables, trends and shapes,* (eds) D.C. Hoaglin, F. Mosteller and J.W. Tukey, (Wiley, New York), 1-36.

Emerson, J.D. and D.C. Hoaglin (1983), Stem-and-leaf displays. In: *Understanding robust and exploratory data analysis,* (eds) D.C. Hoaglin, F. Mosteller and J.W. Tukey, (Wiley, New York), 7-32.

Emerson, J.D. and D.C. Hoaglin (1983), Resistant lines for y versus x. In: *Understanding robust and exploratory data analysis,* (eds) D.C. Hoaglin, F. Mosteller and J.W. Tukey, (Wiley, New York), 129-163.

Emerson, J.D. and D.C. Hoaglin (1985), Resistant multiple regression one variable at a time. In: *Exploring data tables, trends and shapes,* (eds) D.C. Hoaglin, F. Mosteller and J.W. Tukey, (Wiley, New York), 241-275.

Emerson, J.D. and J. Strenio (1983), Boxplots and batch comparisons. In: *Understanding robust and exploratory data analysis,* (eds) D.C. Hoaglin, F. Mosteller and J.W. Tukey, (Wiley, New York), 58-96.

Erickson, B.H. and T.A. Nosanchuck (1979), *Understanding data.* (Open University Press, Milton Keynes).

Feyerabend, P. (1975), *Against method.* (New Left Books, London).

Good, I.J. (1983), The philosophy of exploratory data analysis. *Philosophy of Science,* 50, 283-295.

Goodall, C. (1983), Examining residuals. In: *Understanding robust and exploratory data analysis,* (eds) D.C. Hoaglin, F. Mosteller and J.W. Tukey, (Wiley, New York),. 211-243.

Gregory, S. (1963), *Statistical methods and the geographer.* (Longmans, London).

Hartwig, F. and B.E. Dearing (1979), *Exploratory data analysis.* (Sage, Beverley Hills).

Hoaglin, D.C. (1983), Letter values: a set of selected order statistics. In: *Understanding robust and exploratory data analysis,* (eds) D.C. Hoaglin, F. Mosteller and J.W. Tukey, (Wiley, New York), 33-55.

Kendall, M.G. and W,R. Buckland (1971), *A dictionary of statistical terms,* (Oliver and Boyd, Edinburgh).

Kennedy, B.A. (1979), A naughty world. *Transactions, Institute of British Geographers,* NS, 4(4), 550-558.

Kuhn, T.S. (1963), *The structure of scientific revolutions.* (Chicago University Press, Chicago).

Marshall, J.U. (1985), Geography as a scientific enterprise. In: *The future of geography,* (ed) R.J. Johnston, (Methuen, London), 113-128,

Mason, P.F. (1970), Spatial variability of atmospheric radioactivity in the United States. *Proceedings, Association of American Geographers,* 92-97.

Nozick, R. (1974), *Anarchy,* state *and utopia.* (Basil Blackwell, Oxford).

Olsson, G. (1978), On the mythology of the negative exponential or on power as a game of ontological transformations. *Geografiska Annaler,* 60B, 116-123.

Quenouille, M.H. (1959), *Rapid statistical calculations.* (Griffin, London).

Rock, P. (1979), The sociology of crime, symbolic interactionism and some problematic qualities of radical criminology. In: *Deviant interpretations,* (eds) D. Downes and P. Rock, (Martin Robertson, Oxford), 52-84.

Sibley, D. (1984), A robust analysis of a minority census: the distribution of Travelling people in England. *Environment and Planning A,* 16, 1279-1288.

Silk, J. (1979), *Statistical concepts in geography.* (George Allen and Unwin, London).

Slater, P.B. (1974), Exploratory analysis of trip distribution data. *Journal of Regional Science,* 14(3), 377-388.

Tukey, J.W. (1969), Analyzing data: sanctification or detective work. *American Psychologist,* 24, 83-91.

Tukey, J.W. (1977), *Exploratory data analysis.* (Addison-Wesley, Reading, Mass.).

Warntz, W. (1973), New geography as general spatial systems theory-old social physics writ large? In: *Directions in geography,* (ed) R.J. Chorley, (Methuen, London).

Wilson, A.G. and M.J. Kirkby (1975), *Mathematics for geographers and planners.* (Clarendon Press, Oxford).

Young, R.M. (1979), Why are figures so significant? The role and critique of quantification. In: *Demystifying social statistics,* (eds) J. Irvine, I. Miles and J. Evans, (Pluto Press, London), 63-74.