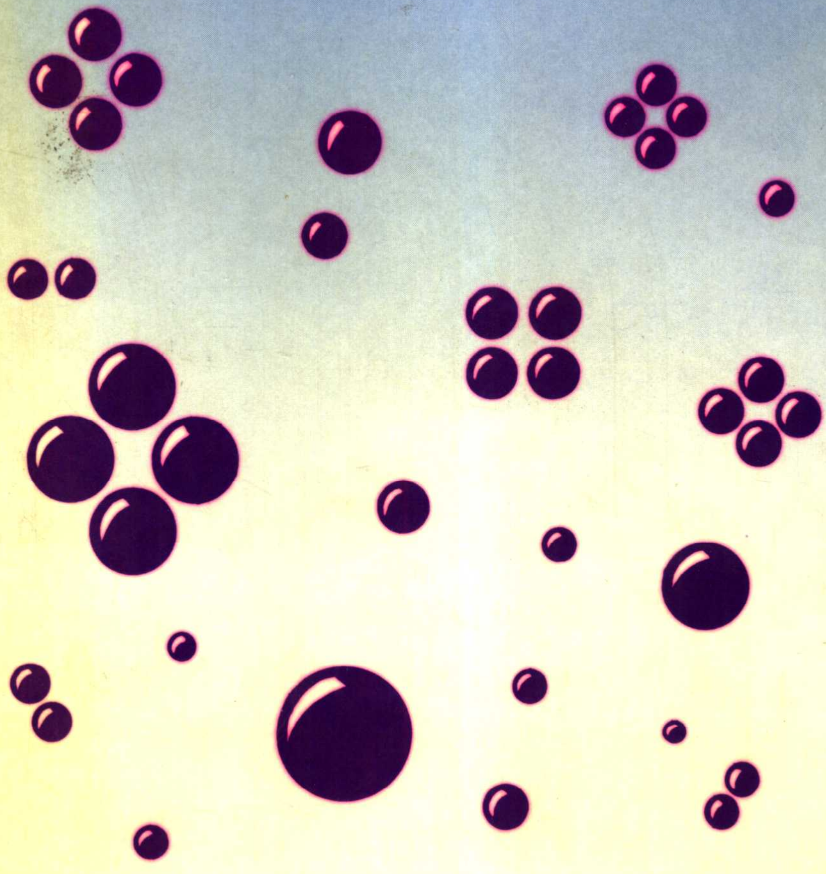
Spatial Data Analysis by Example

VOLUME 1
Point Pattern and Quantitative Data

Graham Upton and Bernard Fingleton



Spatial Data Analysis by Example

VOLUME I

Point Pattern and Quantitative Data

GRAHAM J. G. UPTON
Department of Mathematics,
University of Essex

BERNARD FINGLETON
Cambridge College of Arts and Technology

JOHN WILEY & SONS
Chichester · New York · Brisbane · Toronto · Singapore

Library of Congress Cataloging in Publication Data:

Upton, Graham J. G.
Spatial data analysis by example.

Bibliography: v. 1, p.
Includes indexes.
Contents: v. 1. Point pattern and quantitative data.
1. Spatial analysis (Statistics) I. Fingleton, B.
(Bernard), 1949- . II. Title.
QA278.2.U68 1985 519.5 84–11806
ISBN 0 471 90542 9 (v. 1)

British Library Cataloguing in Publication Data:

Upton, Graham J. G.
Spatial data analysis by example.—(Wiley series in probability and statistics)
Vol. 1: Point pattern and quantitative data
1. Spatial analysis (Statistics)
I. Title II. Fingleton, Bernard
519.4 QA278.2

ISBN 0 471 90542 9

Photosetting by Thomson Press (India) Limited, New Delhi and Printed by Page Bros. (Norwich) Ltd.

G. J. G. U.: to Sue, Robin and Christopher

B. F.: to Eileen, Paddy, Ellen, Meave and James

Preface

The idea of producing a single volume 'user's guide' to the recent developments in spatial statistics arose from a sequence of phone-calls and conversations during the summer of 1981 which culminated in the production of a list of contents mildly resembling those of this volume and its successor. However, two failings of this dialogue were our underestimation of the likely length of the resulting compendium and also of the length of time required to complete it in its entirety. We had originally envisaged that the chapters would naturally fall into four groups, which themselves were arranged in pairs, and the change to two volumes therefore seems not unreasonable. At the least we hope that our readers will find that the two volumes make a convenient pair of bookends!

It is a pleasure to acknowledge the help and encouragement of the publishers in the various stages of preparation of this volume. Most especially we thank our families for their patience: the gestation period has been so long that they must often have wondered whether eventual delivery (deliverance?!) would ever come.

We are very grateful to Dr S. A. L. M. Kooijman for his assistance with the glasswort and anemone data and we gratefully acknowledge the following sources for permission to reproduce various figures and tables: The Association of American Geographers (Figure 1.30); Akademie-Verlag, Berlin (Figure 4.9); Blackwell Scientific Publications Ltd (Figures 3.14 and 4.14); *Biometrika* Trustees (Figure 1.39b, Appendices 3 and 4); Cambridge University Press (Figures 3, 1.7, 1.29 and 4.13); *Evolution* (Figures 3.6 a–d); the Geographical Association (Figures 1.34a and 1.34b); Gustav Fischer Verlag (Figure 1.11); Gordon and Breach Science Publications (Table 1.18); Harvard University Press (Figure 3.7); Institut National de la Recherche Agronomique Service des Publications, St. Cyr (Figure 2); the Institute of British Geographers (Figure 1.16); the Editors of the *Journal of the Royal Statistical Society, Series B* (Figure 4.3); dr W. Junk publishers (Figure 6); Regional Science Association (Figure 1.12).

Any typographical errors or errors of principle or practice that remain in this volume are of course our own responsibility and we would be grateful to be notified by readers if they detect such errors. We would also value any suggestions for additions or improvements to the present volume.

Contents

Preface xi

INTRODUCTION 1

Chapter 1: THE IDENTIFICATION OF PATTERN 9

1.1 Alternative Patterns 9
The random pattern. Clusters. Generalized Poisson distributions. Compound Poisson distributions. Poisson cluster processes. Inhibition models. Central place models. The county seat model.

1.2 The Use of Scattered Quadrats 26
Indices. Choice of Quadrat Size.
Example 1.1

1.3 The Use of Contiguous Quadrats 32
The use of a square grid. Randomly paired quadrats. A transect of quadrats. Refinements to the use of the transect. Tests of significance. Distinguishing between true and apparent contagion. Summary of the use of quadrats.
Examples 1.2–1.7

1.4 The Use of Sampled Point Locations 53
Distances between points and plants. Some distributional results. Testing for randomness. Recommendations.
Examples 1.8–1.10

1.5 The Use of Mapped Plant Locations 70
Mapping methods. Edge effects and interdependence. Tests based on distance measures. Refined nearest-neighbour analysis. Combined count and distance analysis. Generalized quadrat analysis. Reflexive nearest neighbours.
Examples 1.11–1.15

1.6 Circles—Not Points 94
Example 1.16

1.7 Spatial Tessellations 96
The Dirichlet tessellation. The Delaunay tessellation. The line tessellation. Curved-boundary tessellations
Example 1.17

Chapter 2: THE ESTIMATION OF SPATIAL INTENSITY **105**

2.1 Introduction 105

2.2 Quadrat Methods 105

Practical problems

2.3 The Use of Line Transects 107

The impact of detection on the estimate of intensity. Parametric descriptions of the detection function. Non-parametric procedures. The use of sighting distances.

Examples 2.1–2.2

2.4 Distance Methods 118

Point–plant methods. Combined point–plant and plant–plant calculations. Cox's procedures. Batcheler's procedure. Precision of the estimates. Recommendations.

Example 2.3

2.5 Areal Methods 133

The Dirichlet tessellation. Weighted polygons and weighted triangles.

Examples 2.4–2.5

2.6 The Angle-Count Method 144

2.7 Mosaics 146

Plant intensity. Total plant area and boundary length.

Example 2.6

Chapter 3: SPATIAL AUTOCORRELATION **151**

3.1 General Themes 151

Randomization. The general cross-product statistic. The Monte Carlo approach. The normal approximation.

Example 3.1

3.2 Join-Count Statistics 158

One-sided tests. The Poisson approximation. Multicoloured maps. The bishop's definition of contiguity. The queen's definition of contiguity. Multiple hypothesis tests.

Examples 3.2–3.6

3.3 Moran's *I* 170

The distribution of *I* under randomization. Moran's *I* with ranked data.

Examples 3.7–3.8

3.4 Generalized Proximity Values 176

Alternative models of proximity. Searching for *W* matrices. Aspatial *W* matrices.

Examples 3.9–3.11

3.5 Correlograms 186

Spatial correlograms. Definition of distance classes. Rook's

definition S statistics for the correlogram on a regular lattice. Spatial correlograms for join-count statistics. Correlograms, the bishop's definition, and multicoloured maps. The S statistics for the correlogram using the queen's definition. Correlograms with Moran's I. General distance classes. The modified autocorrelation coefficient. The partial correlogram. The spatio-temporal correlogram. Partial spatio-temporal autocorrelation.
Examples 3.12–3.22

3.6 Related Approaches 204
Space–time interaction tests. Relative conditional intensity. Ordinal autocorrelation measures. Alternative sampling models. Difference maps. The choice of a test.
Example 3.23

Chapter 4: INTER-TYPE RELATIONS **214**

4.1 Bivariate Point Patterns 214
Processes displaying attraction. Processes displaying repulsion. Summary

4.2 Analysis Using Quadrats 223
Presence/absence cross-classifications. Inter-quadrat correlations.
Examples 4.1–4.4

4.3 Analysis Using Transects 237
The runs test. Runs of healthy and diseased plants.
Example 4.5

4.4 Sophisticated Distance Methods 242
The nearest-neighbour table. Point–plant and plant–plant distributions. Paired point–plant distances. Toroidal shifts. Combined count and distance analysis.
Examples 4.6–4.10

4.5 Spatial Rank Correlation 258
Example 4.11

Chapter 5: REGRESSION AND AUTOREGRESSION **264**

5.1 Regression, Structural, and Sampling Models 264

5.2 Ordinary Least Squares 265
The Gauss–Markov theorem. Inference under Gauss–Markov conditions. Variance ratio tests. Stepwise procedures.
Examples 5.1–5.2

5.3 Generalised Least Squares 277
Weighted least squares. Logit regression. Intermediate weighted least squares. Generalized least squares with spatially autocor-

related errors. Spatial differencing.
Example 5.1 (*cont.*)

5.4 Maximum Likelihood 283
Maximum likelihood with spatially autocorrelated errors. Computational problems. Inference and maximum likelihood. Likelihood-ratio tests. The unrestricted autocorrelated errors model. More complex autocorrelated errors models. Poisson responses. Poisson responses with logarithmic link.
Examples 5.3–5.7

5.5 Non-Linear Regression 312
The family of power transformations. Parameter bias in log-linear regression. Trend-surface analysis.
Examples 5.8–5.9

5.6 The Analysis of Residuals 329
Abnormal residuals. Spatially autocorrelated residuals. Moran's I for residuals. The I_k distribution under the null hypothesis. A randomization approach to autocorrelated residuals. Residual correlograms. Alternative residual correlograms. Error covariance from correlograms. Autoregressive model residuals and residual maps.
Examples 5.10–5.18

5.7 Autoregressive Response Models 349
Pure autoregression. Conditional autoregression. Moving averages. Boundary effects.
Examples 5.19–5.20

5.8 Alternative Approaches 366
Prediction methods related to Kriging. Space–time models.

5.9 A Recapitulation of Spatial Regression Models 371

REFERENCES 374

APPENDICES 395
A: The normal distribution function. B: Percentage points of the χ^2 distribution. C: Percentage points for the F distribution. D: Percentage points for the Eberhardt statistic

AUTHOR INDEX 400

SUBJECT INDEX 406

EXAMPLE INDEX 410

Introduction

This is the first volume of a two-volume work that deals with the analysis of spatial data. Spatial data comes in many varieties and it is not easy to arrive at a system of classification that is simultaneously exclusive, exhaustive, imaginative, and satisfying. We have identified four major clusters of techniques which appear, to us, to include the majority of the diverse examples of spatial data that we have encountered. We treat point pattern and quantitative data in this first volume and categorical and directional data in the subsequent volume. We do not claim that these volumes entirely account for spatial data in all its manifestations, nor that we have exhausted all aspects of the four clusters of techniques. We can, however, claim with confidence that what we have covered has certainly exhausted us!

Before giving a more detailed introduction to the subject-matter of the five chapters of this volume, we present an overview derived from the nearly 500 references included at the end of the book. We find lists fascinating and therefore present Table 1, which shows, in descending order, the five most commonly cited journals in our bibliography. Table 1 is somewhat misleading, however, since the apparent emphasis on biometrics is a side-effect of there being fewer journals in this field.

Figure 1 attempts to give a truer indication of the emphasis of the book. When we attempt to classify the references it quickly becomes apparent that there are roughly equal numbers falling in the four categories displayed in the diagram. Thus, roughly half the references have appeared in primarily theoretical statistical or biometrical journals and half in the more subject-oriented journals.

Figure 1 deliberately situates the two applied fields side by side because we wish to emphasize the similarities of their statistical problems. We greatly hope

***Table 1** The five most commonly cited journals (in descending order)*

1. *Biometrics*
2. *Biometrika*
3. *Geographical Analysis*
4. *Journal of the Royal Statistical Society, Series B*
5. *Journal of Ecology*

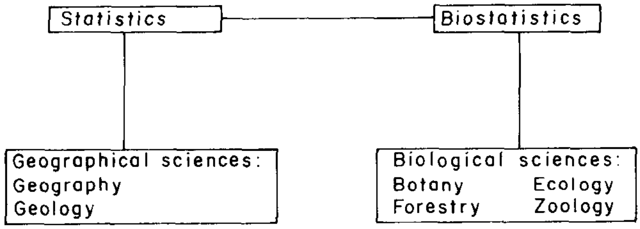


Figure 1. The contents of the book

that a side-effect of this work on spatial data analysis will be an increased cross-fertilization between these two subject-areas. The breadth of potential application of the methods that we describe is easily seen by glancing at the example index, which includes subjects as diverse as South-East-Asian birds, Southampton junior schools, desert shrubs, and forest trees! The common feature with all these examples is that individual observations can be located, more or less accurately, on the surface of the earth, so that a set of such observations is therefore capable of being displayed in the form of a map. This spatial attribute is, in many cases, a highly revealing piece of information concerning the causal processes and mechanisms operating amongst the data.

So that the reader may decide at the outset whether this work is likely to be suitable for his or her purposes, it seems sensible to set out here our guiding philosophy in the construction of this work. Throughout our writing we have had three precepts in mind:

(i) the practitioner seeking to *use* a statistical procedure does not want his or her understanding of that procedure obscured by immense numbers of theorems, lemmas, and other mathematical excrescences, but

(ii) it is manifestly unwise for the practitioner to use a statistical 'black-box' without having, at the least, a reasonable idea of its rationale, and

(iii) some readers will require more detail than we have provided and will welcome a reference to a source paper that provides that detail.

Our aim has therefore been to explain, clearly but succinctly, the ideas and principles underlying each statistical procedure without getting lost in too many mathematical formulae. Undoubtedly the idea of how many is 'too many' will differ from reader to reader and we shall not please everybody all of the time. However, we have *tried* to avoid proving results that could reasonably be stated and to avoid stating results if they were not helpful.

The degree of statistical and mathematical background that is required varies somewhat from chapter to chapter. As a rule, we presume that the reader is familiar with (or has met in the past) basic statistical ideas such as probability, probability distributions, variance, expectation, estimation, tests of significance, and the like. The chapter on regression (Chapter 5) anticipates rather more than this because of the impossibility of including a general introduction to the subject along with everything else. In that chapter we have assumed that the reader can stand (and understand) apparently complicated matrix expressions and is familiar with the general (non-spatial) methodology. Notice that in the second sentence of this paragraph we used the word 'ideas' rather than the word 'formulae'—the latter are supplied as required. Thus, in Chapter 1 we provide the formula for the Poisson distribution, though we certainly anticipate that most readers will have met that formula before.

The difference between point pattern and quantitative data is illustrated by the contrast between Figure 2 and Figure 3. Figure 2 is a map of 71 Swedish Pine saplings in a 10 m × 10 m square (see Strand, 1972, and Ripley, 1981). Since the saplings are presumably all roughly the same size, and if we regard their diameter as being negligible by comparison with the inter-sapling distances, then size variations can be ignored and the matter of interest may be regarded as being simply their existence at the points indicated on the map. This particular point pattern shows signs of being regularly spaced, and of not being representative of the sort of pattern to be expected if the saplings grew at random places within the plot. The various traditional methods for detecting departures from randomness (mostly based on quadrat counts) together with the very recently developed methods based on inter-item distances form the subject material for Chapter 1.

Figure 3 is an example of what we call quantitative data. Ostensibly it is merely

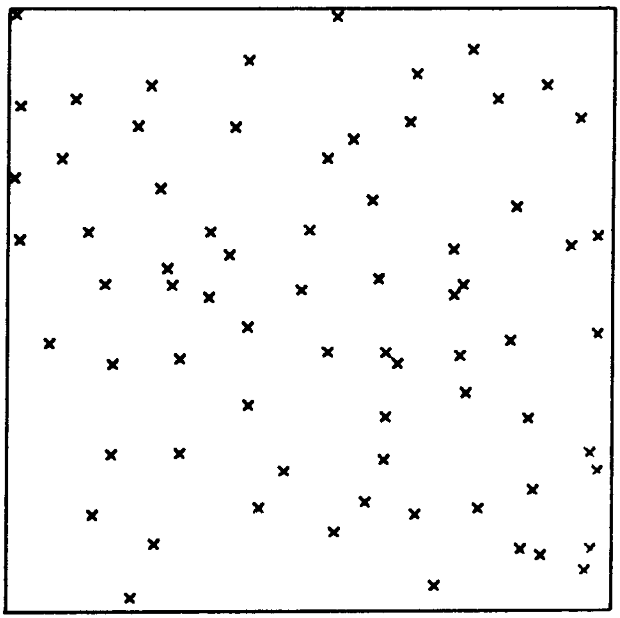


Figure 2. The distribution of 71 Swedish Pine saplings. (Redrawn from Strand, 1972. Reproduced by permission of INRA)

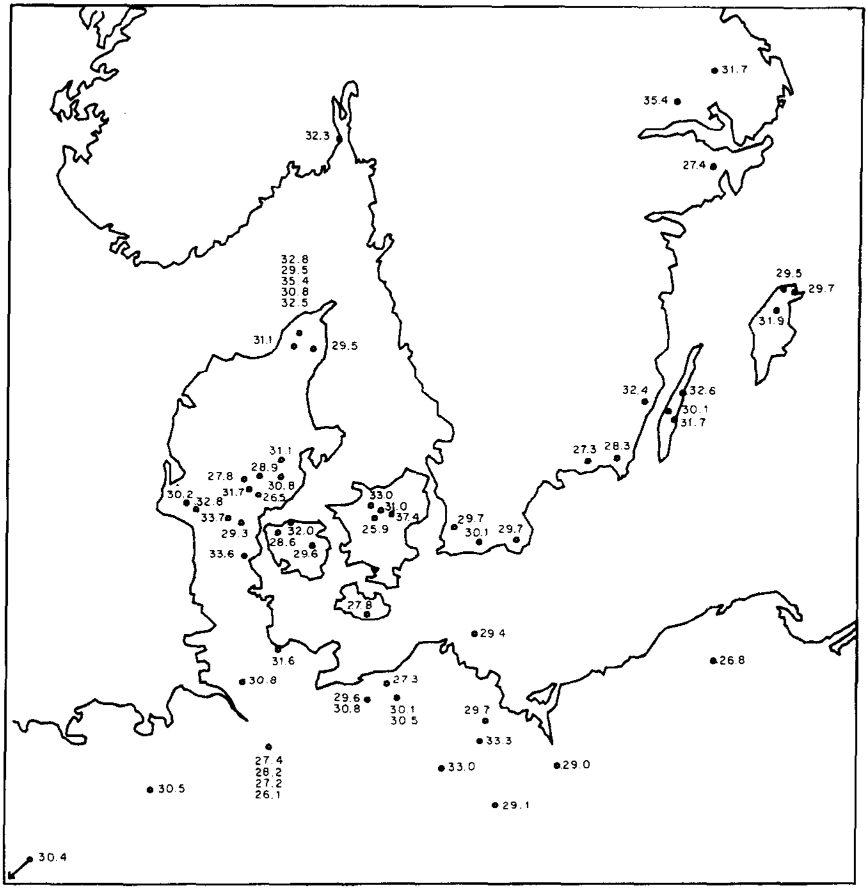


Figure 3. The distribution of the length/breadth index for the Bagterp spearheads. (Redrawn from Hodder and Orton, 1976. Reproduced by permission of Cambridge University Press)

another point pattern, but in this case the points possess values and, while the location of the points is not uninteresting, our principal concern is with the possibility of spatial variation in the values. The data displayed in Figure 3 relate to finds of so-called Bagterp spearheads which emanate from the early Bronze Age. A possible method of analysing the degree of localization of the manufacture of these spearheads (which have particular easily recognized characteristics) is based on the variation in the shape of the spearheads. A simple measure of shape is the ratio of the length of each spearhead to its breadth, and these are the values that are recorded in Figure 3, which is taken from Hodder and Orton (1976, p. 175) who plot data collated by Jacob-Friesen (1967).

Notice that the point locations in Figure 3 do not comprise the complete set of locations for these Bagterp spearheads, but rather reflect the locations of organized archaeological activity. Thus, the recorded locations are indicative but not definitive. By contrast, Figure 2 displays *all* the saplings in the region illustrated.

The arrangement of the book is that Chapters 1, 2, and 4 are principally concerned with point pattern data, while Chapters 3 and 5 are principally concerned with quantitative data. However, this division is not precise and there are many occasions on which quantitative methods are used to analyse point pattern data. An example is illustrated by Figure 4, which displays the numbers of bitterbrush (*Purshia tridentata*) plants in a 700 ft × 500 ft rectangular region to the north-east of Florence, Montana (Lyon, 1968). Here, the original point locations have been summarized as a series of quadrat counts. We shall see in Chapter 1 that the most sensitive insights into the pattern of objects are those obtained using functions of the distance between objects and their nearest neighbours, whereas in Chapter 2 we shall see that quadrat counts of this type provide the simplest means of obtaining a reliable estimate of the total number of plants in the larger region of interest.

As already intimated, in Chapter 1 we seek to identify the characteristics of a pattern, with the hope that these characteristics will reveal something helpful concerning the origins of that pattern, whereas in Chapter 2 the emphasis is on estimating the numbers present in a region by means of a detailed examination of a small part of that region. At first sight one might imagine that this latter task scarcely deserved a chapter to itself! However, the difficulties of the task are made apparent by considering some of the possible contexts:

(i) animals that run away from an approaching observer;

(ii) trees or bog plants in inhospitable terrain;

(iii) trees that have contracted a disease but may not yet show any outward signs of that disease.

While remote sensing, using satellite photographs or some other device, can ameliorate the problems caused by these types of data, such an approach will be expensive and may also be impracticable.

In Chapter 3 we turn to quantitative data, such as that exemplified in Figure 3. This chapter is concerned with the measurement of spatial autocorrelation: that

154	74	67	168	172	136	128
79	89	69	104	126	102	73
114	120	143	117	41	160	139
117	53	84	184	74	155	172
188	115	62	38	83	108	145

Figure 4. Numbers of bitterbrush (*Purshia tridentata*) plants in 35 100 ft × 100 ft quadrats (Lyon, 1968)

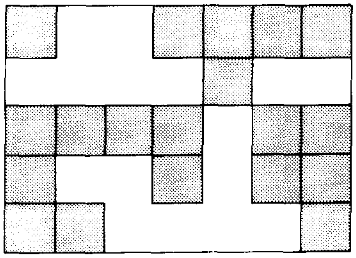


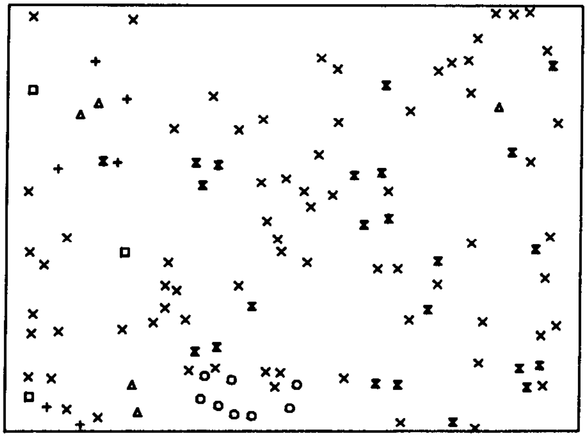
Figure 5. Black/white representation of bitterbrush counts

is, the pattern or organization in the values at fixed localities. We seek to discover whether these values are distributed haphazardly over the locations or whether there is some spatial order to their arrangement, so that, for example, all the high values are found to be grouped close together. The actual metric used may be the continuous metric of the length/breadth ratios of the Bagterp spearheads (this is the case with our subsequent analysis in Example 3.7 of variations of the wing length of a species of bird in the Far East) or it may be a much 'clumsier' metric such as 'lowest, lower, low, medium, high, higher, highest'. In the limit, the latter metric can be reduced to a two-case 'higher-than-average, lower-than-average' metric which we could refer to as a 'black and white' pattern.

Figure 5 illustrates how the detailed data illustrated in Figure 4 can be summarized using the black/white metric. It incidentally also illustrates that what was basically point pattern data can also be analysed in a relevant manner by methods recorded in Chapters 3 and 5. A major class of methods for measuring spatial autocorrelation hinge upon the numbers of occasions on which 'black' cells border on other 'black' cells. Figure 5 certainly provides us with a picture that suggests that there is spatial organization present in the location of the bitterbrush plants—there are few isolated cells.

Chapters 4 and 5 are concerned with bivariate and multivariate data, and therefore themes such as mutual attraction and repulsion, or the correlation between spatial distributions, come into prominence in place of the 'self-oriented' autocorrelation, inhibition, and clustering mechanisms which feature as the explanatory processes of the univariate spatial distributions of Chapters 1–3. Some of the procedures discussed in these chapters are still in the process of being developed and have methodological problems that are, as yet, unresolved. We have attempted to give the reader guidance and to provide some perspective on the available alternatives, including some illustration of the pitfalls that may arise from the uncritical application of non-spatial methods for spatial data.

Figure 6, taken from Erschbamer *et al.* (1983), is typical of the data discussed in Chapter 4. In the figure are recorded the point positions of several different types of plant which compete with one another for the available water supply in a



Key: x *Opuntia humifusa*; ɪ *Sempervivum tectorum*;
+ *Sedum reflexum*; △ *Cheilanthes maranthae*;
o *Sempervivum arachnoideum*; □ *Asplenium septentrionale*;

Figure 6. The distribution of six plant species in an Alpine valley. (Redrawn from Erschbamer *et al.*, 1983. Reproduced by permission of JUNK)

Central alpine valley in the South Tyrol. The ecologist is interested in the factors that determine their positions. Viewed as a point pattern we may use the methods described in Chapter 4 to determine the extent to which plants of different types are isolated from one another. Of course part of the isolation will be due not to physical repulsion on the part of the plants but rather to variations in the nature of the terrain. Erschbamer *et al.* note that there are several rocky outcrops within the study area, while the soil depth varies down to about 90 cm and also varies in its water content. Using the regression methods of Chapter 5 we could analyse this data from a rather different standpoint, comparing the spatial density of a particular species of plant with the (literally) underlying physical conditions: soil depth and available water.

An important consideration when applying regression procedures to spatial data is the presence of self-correlation (autoregression). As an instance, in Example 5.20 we examine a model that attempts to explain the variation in the number of different plant species among Californian localities, $\{Y_i\}$, using as explanatory factors the maximum altitude in the locality, the latitude of the locality, the area of the locality, and also an autoregressive component whereby Y_i is related to Y_j $(i \neq j)$, the number of species in the 'neighbouring' locality j. The methods for dealing with autoregressive spatial interaction have been the subject of some debate (cf. Ord, 1980 and Griffith, 1980, 1981).

At this point it is timely to make clear that the intention of our analyses is primarily to demonstrate the methodology and we make no attempt simultaneously to clinch some argument concerning the controlling mechanisms of Californian flora, or whatever. For this reason our analyses are presented 'warts and all' in order to give a true representation of our actual experiences with the methods and models that we are illustrating. Having said this, we do nevertheless hope that the actual results that we obtain will have intrinsic interest for their own sake.

Many of the procedures that we discuss require the use of a computer. After considerable hesitation we decided that it was not feasible to append the various computer programs that we wrote in order to achieve our results. Apart from a reluctance to reveal just how clumsy these programs were (!), there was the added consideration that some were written in FORTRAN, some in BASIC, some used the GENSTAT package, others the TILE4 package, and still others used subroutines from the NAG library. We concluded that not all our readers would know, or have access to, all these acronyms! Instead we have attempted to reduce the complexity of formulae wherever possible and to concentrate our efforts on describing the successful simpler techniques.

Obviously this book was not written in a void, and it is apposite to draw up a short list of major sources from which our ideas and inspiration spring. We have already presented, in Table 1, a list of the five most cited journals, though there are a further dozen that figure prominently in the list of references. To these journals must be added a quartet of books that can be warmly recommended if additional background reading is required. These are:

Cliff and Ord, *Spatial Processes* (1981), Haggett, Cliff, and Frey, *Locational Analysis in Human Geography* (1977), Pielou, *Mathematical Ecology* (1977), and Ripley, *Spatial Statistics* (1981). Further details are given in the list of references.