

# ***Geostatistics for Environmental Scientists***

Second Edition

# **Statistics in Practice**

*Advisory Editors*

**Stephen Senn**

University of Glasgow, UK

**Marion Scott**

University of Glasgow, UK

*Founding Editor*

**Vic Barnett**

Nottingham Trent University, UK

---

*Statistics in Practice* is an important international series of texts which provide detailed coverage of statistical concepts, methods and worked case studies in specific fields of investigation and study.

With sound motivation and many worked practical examples, the books show in down-to-earth terms how to select and use an appropriate range of statistical techniques in a particular practical field within each title's special topic area.

The books provide statistical support for professionals and research workers across a range of employment fields and research environments. Subject areas covered include medicine and pharmaceutics; industry, finance and commerce; public services; the earth and environmental sciences, and so on.

The books also provide support to students studying statistical courses applied to the above areas. The demand for graduates to be equipped for the work environment has led to such courses becoming increasingly prevalent at universities and colleges.

It is our aim to present judiciously chosen and well-written workbooks to meet everyday practical needs. Feedback of views from readers will be most valuable to monitor the success of this aim.

A complete list of titles in this series appears at the end of the volume.

# **Geostatistics for Environmental Scientists**

Second Edition

**Richard Webster**  
*Rothamsted Research, UK*

**Margaret A. Oliver**  
*University of Reading, UK*



John Wiley & Sons, Ltd

Copyright © 2007

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester,  
West Sussex PO19 8SQ, England

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): [cs-books@wiley.co.uk](mailto:cs-books@wiley.co.uk)

Visit our Home Page on [www.wileyeurope.com](http://www.wileyeurope.com) or [www.wiley.com](http://www.wiley.com)

All Rights Reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except under the terms of the Copyright, Designs and Patents Act 1988 or under the terms of a licence issued by the Copyright Licensing Agency Ltd, 90 Tottenham Court Road, London W1T 4LP, UK, without the permission in writing of the Publisher. Requests to the Publisher should be addressed to the Permissions Department, John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, or emailed to [permreq@wiley.co.uk](mailto:permreq@wiley.co.uk), or faxed to (+44) 1243 770620.

This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the Publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

#### ***Other Wiley Editorial Offices***

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 42 McDougall Street, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 6045 Freemont Blvd, Mississauga, ONT, L5R 4J3

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Anniversary Logo Design: Richard J. Pacifico

#### ***British Library Cataloguing in Publication Data***

A catalogue record for this book is available from the British Library

ISBN-13: 978-0-470-02858-2 (HB)

Typeset in 10/12 photina by Thomson Digital

Printed and bound in Great Britain by TJ International, Padstow, Cornwall

This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

*Geostatistics for Environmental Scientists/2nd Edition* R. Webster and M.A. Oliver

© 2007 John Wiley & Sons, Ltd

# ***Contents***

<b>Preface</b>	<b>xi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Why geostatistics?	1
1.1.1 Generalizing	2
1.1.2 Description	5
1.1.3 Interpretation	5
1.1.4 Control	5
1.2 A little history	6
1.3 Finding your way	8
<b>2 Basic Statistics</b>	<b>11</b>
2.1 Measurement and summary	11
2.1.1 Notation	12
2.1.2 Representing variation	13
2.1.3 The centre	15
2.1.4 Dispersion	16
2.2 The normal distribution	18
2.3 Covariance and correlation	19
2.4 Transformations	20
2.4.1 Logarithmic transformation	21
2.4.2 Square root transformation	21
2.4.3 Angular transformation	22
2.4.4 Logit transformation	22
2.5 Exploratory data analysis and display	22
2.5.1 Spatial aspects	25
2.6 Sampling and estimation	26
2.6.1 Target population and units	28
2.6.2 Simple random sampling	28
2.6.3 Confidence limits	29
2.6.4 Student's <i>t</i>	30
2.6.5 The $\chi^2$ distribution	31
2.6.6 Central limit theorem	32
2.6.7 Increasing precision and efficiency	32
2.6.8 Soil classification	35

<b>3 Prediction and Interpolation</b>	<b>37</b>
3.1 Spatial interpolation	37
3.1.1 Thiessen polygons (Voronoi polygons, Dirichlet tessellation)	38
3.1.2 Triangulation	38
3.1.3 Natural neighbour interpolation	39
3.1.4 Inverse functions of distance	40
3.1.5 Trend surfaces	40
3.1.6 Splines	42
3.2 Spatial classification and predicting from soil maps	42
3.2.1 Theory	43
3.2.2 Summary	45
<b>4 Characterizing Spatial Processes: The Covariance and Variogram</b>	<b>47</b>
4.1 Introduction	47
4.2 A stochastic approach to spatial variation: the theory of regionalized variables	48
4.2.1 Random variables	48
4.2.2 Random functions	49
4.3 Spatial covariance	50
4.3.1 Stationarity	52
4.3.2 Ergodicity	53
4.4 The covariance function	53
4.5 Intrinsic variation and the variogram	54
4.5.1 Equivalence with covariance	54
4.5.2 Quasi-stationarity	55
4.6 Characteristics of the spatial correlation functions	55
4.7 Which variogram?	60
4.8 Support and Krige's relation	60
4.8.1 Regularization	63
4.9 Estimating semivariances and covariances	65
4.9.1 The variogram cloud	65
4.9.2 h-Scattergrams	66
4.9.3 Average semivariances	67
4.9.4 The experimental covariance function	73
<b>5 Modelling the Variogram</b>	<b>77</b>
5.1 Limitations on variogram functions	79
5.1.1 Mathematical constraints	79
5.1.2 Behaviour near the origin	80
5.1.3 Behaviour towards infinity	82
5.2 Authorized models	82
5.2.1 Unbounded random variation	83
5.2.2 Bounded models	84

5.3	Combining models	95
5.4	Periodicity	97
5.5	Anisotropy	99
5.6	Fitting models	101
5.6.1	What weights?	104
5.6.2	How complex?	105
<b>6</b>	<b>Reliability of the Experimental Variogram and Nested Sampling</b>	<b>109</b>
6.1	Reliability of the experimental variogram	109
6.1.1	Statistical distribution	109
6.1.2	Sample size and design	119
6.1.3	Sample spacing	126
6.2	Theory of nested sampling and analysis	127
6.2.1	Link with regionalized variable theory	128
6.2.2	Case study: Youden and Mehlich's survey	129
6.2.3	Unequal sampling	131
6.2.4	Case study: Wyre Forest survey	134
6.2.5	Summary	138
<b>7</b>	<b>Spectral Analysis</b>	<b>139</b>
7.1	Linear sequences	139
7.2	Gilgai transect	140
7.3	Power spectra	142
7.3.1	Estimating the spectrum	144
7.3.2	Smoothing characteristics of windows	148
7.3.3	Confidence	149
7.4	Spectral analysis of the Caragabal transect	150
7.4.1	Bandwidths and confidence intervals for Caragabal	150
7.5	Further reading on spectral analysis	152
<b>8</b>	<b>Local Estimation or Prediction: Kriging</b>	<b>153</b>
8.1	General characteristics of kriging	154
8.1.1	Kinds of kriging	154
8.2	Theory of ordinary kriging	155
8.3	Weights	159
8.4	Examples	160
8.4.1	Kriging at the centre of the lattice	161
8.4.2	Kriging off-centre in the lattice and at a sampling point	169
8.4.3	Kriging from irregularly spaced data	172
8.5	Neighbourhood	172
8.6	Ordinary kriging for mapping	174

8.7 Case study	175
8.7.1 Kriging with known measurement error	180
8.7.2 Summary	180
8.8 Regional estimation	181
8.9 Simple kriging	183
8.10 Lognormal kriging	185
8.11 Optimal sampling for mapping	186
8.11.1 Isotropic variation	188
8.11.2 Anisotropic variation	190
8.12 Cross-validation	191
8.12.1 Scatter and regression	193
<b>9 Kriging in the Presence of Trend and Factorial Kriging</b>	<b>195</b>
9.1 Non-stationarity in the mean	195
9.1.1 Some background	196
9.2 Application of residual maximum likelihood	200
9.2.1 Estimation of the variogram by REML	200
9.2.2 Practicalities	203
9.2.3 Kriging with external drift	203
9.3 Case study	205
9.4 Factorial kriging analysis	212
9.4.1 Nested variation	212
9.4.2 Theory	212
9.4.3 Kriging analysis	213
9.4.4 Illustration	218
<b>10 Cross-Correlation, Coregionalization and Cokriging</b>	<b>219</b>
10.1 Introduction	219
10.2 Estimating and modelling the cross-correlation	222
10.2.1 Intrinsic coregionalization	224
10.3 Example: CEDAR Farm	226
10.4 Cokriging	228
10.4.1 Is cokriging worth the trouble?	231
10.4.2 Example of benefits of cokriging	232
10.5 Principal components of coregionalization matrices	235
10.6 Pseudo-cross-variogram	241
<b>11 Disjunctive Kriging</b>	<b>243</b>
11.1 Introduction	243
11.2 The indicator approach	246
11.2.1 Indicator coding	246
11.2.2 Indicator variograms	247
11.3 Indicator kriging	249

11.4	Disjunctive kriging	251
11.4.1	Assumptions of Gaussian disjunctive kriging	251
11.4.2	Hermite polynomials	252
11.4.3	Disjunctive kriging for a Hermite polynomial	254
11.4.4	Estimation variance	256
11.4.5	Conditional probability	256
11.4.6	Change of support	257
11.5	Case study	257
11.6	Other case studies	263
11.7	Summary	266
<b>12</b>	<b>Stochastic Simulation</b>	<b>267</b>
12.1	Introduction	267
12.2	Simulation from a random process	268
12.2.1	Unconditional simulation	270
12.2.2	Conditional simulation	270
12.3	Technicalities	271
12.3.1	Lower–upper decomposition	272
12.3.2	Sequential Gaussian simulation	273
12.3.3	Simulated annealing	274
12.3.4	Simulation by turning bands	276
12.3.5	Algorithms	277
12.4	Uses of simulated fields	277
12.5	Illustration	278
<b>Appendix A</b>	<b>Aide-mémoire for Spatial Analysis</b>	<b>285</b>
A.1	Introduction	285
A.2	Notation	285
A.3	Screening	285
A.4	Histogram and summary	286
A.5	Normality and transformation	287
A.6	Spatial distribution	288
A.7	Spatial analysis: the variogram	288
A.8	Modelling the variogram	290
A.9	Spatial estimation or prediction: kriging	291
A.10	Mapping	292
<b>Appendix B</b>	<b>GenStat Instructions for Analysis</b>	<b>293</b>
B.1	Summary statistics	293
B.2	Histogram	294
B.3	Cumulative distribution	294
B.4	Posting	295
B.5	The variogram	295

B.5.1	Experimental variogram	295
B.5.2	Fitting a model	296
B.6	Kriging	297
B.7	Coregionalization	297
B.7.1	Auto- and cross-variograms	297
B.7.2	Fitting a model of coregionalization	298
B.7.3	Cokriging	298
B.8	Control	298
<b>References</b>		<b>299</b>
<b>Index</b>		<b>309</b>

# Preface

When the first edition of *Geostatistics for Environmental Scientists* was published six years ago it was an instant success. The book had a long gestation as we tested our presentation on newcomers to the subject in our taught courses and on practitioners with a modicum of experience. Responses from readers and from our students showed that they wanted to understand more, and that wish coincided with the need to produce a new revised edition. That feedback has led us to change the emphasis and content. The result is that new material comprises about 20% of the new edition, and we have revised and reorganized Chapters 4, 5 and 6.

The focus of the book remains straightforward linear geostatistics based on least-squares estimation. The theory and techniques have been around in mineral exploration and petroleum engineering for some four decades. For much of that time environmental scientists could not see the merits of the subject or appreciate how to apply it to their own problems, because of the context, the jargon and the mathematical presentation of the subject by many authors. This situation has changed dramatically in the last ten years as soil scientists, hydrologists, ecologists, geographers and environmental engineers have seen that the technology is for them if only they could know how to apply it. Here we have tried to satisfy that need.

The structure of the book follows the order in which an environmental scientist would tackle an investigation. It begins with sampling, followed by data screening, summary statistics and graphical display. It includes some of the empirical methods that have been used for mapping, and the shortcomings of these that lead to the need for a different approach. This last is based on the theory of random processes, spatial covariances, and the variogram, which is central to practical geostatistics. Practitioners will learn how to estimate the variogram, what models they may legitimately use to describe it mathematically, and how to fit them. Their attention is also drawn to some of the difficulties of variography associated with the kinds of data that they might have to analyse. There is a brief excursion into the frequency domain to show the equivalence of covariance and spectral analysis.

The book then returns to the principal reason for geostatistics, local estimation by kriging, in particular ordinary kriging. Other kinds of kriging, such as lognormal kriging, kriging in the presence of trend and factorial kriging, are described for readers to put into practice as they become more skilled. Coregionalization is introduced as a means of improving estimates of a primary variable where data on one or more other variables are to hand or can be

obtained readily. There is an introduction to non-linear methods, including disjunctive kriging for decision-making. The final chapter is on geostatistical simulation, which is widely used in the petroleum industry and in hydrology.

In environmental applications the problems are nearly always ones of estimation in two dimensions and of mapping. Rarely do they extend to three dimensions or are restricted to only one.

Geostatistics is not easy. No one coming new to the subject will read this book from cover to cover and remember everything that he or she should do. We have therefore added an aide-mémoire, which can be read and reread as often as necessary. This will remind readers of what they should do and the order in which to do it. It is followed by some simple program instructions in the GenStat language for carrying out the analyses. These, with a few other commands to provide the necessary structures to read data and to write and display output, should enable practitioners to get started, after which they can elaborate their programs as their confidence and competence grow.

We illustrate the methods with data that we have explored previously in our research. The data are of soil properties, because we are soil scientists who use geostatistics in assessing soil resources. Nevertheless, there are close analogies with other aspects of the environment at or near the land surface, which we have often had to include in our analyses and which readers will see in the text.

The data come from surveys made by us or with our collaborators. The data for Broom's Barn Farm, which we can provide for readers thanks to Dr J. D. Pidgeon, are from an original survey of the farm soon after Rothamsted bought it in 1959. Those for the Borders Region (Chapter 2) were collected by the Edinburgh School of Agriculture over some 20 years between 1960 and 1980, and are provided by Mr R. B. Speirs. The data from the Jura used to illustrate coregionalization (Chapter 10) are from a survey made by the École Polytechnique Fédérale de Lausanne in 1992 under the direction of Mr J.-P. Dubois. Chapter 7 is based on a study of gilgai terrain in eastern Australia in 1973 by one of us when working with CSIRO, and the data from CEDAR Farm used to illustrate Chapter 10 were kindly provided by Dr Z. L. Frogbrook from her original study in 1998. The data from the Yattendon Estate (Chapters 6 and 9) are from a survey by Dr Z. L. Frogbrook and one of us at the University of Reading for the Home-Grown Cereals Authority. We are grateful to the organizations and people whose data we have used. Finally, we thank our colleagues Dr R. M. Lark and Dr B. P. Marchant for their help with some of the computing.

The data from Broom's Barn Farm, CEDAR Farm, the Borders Region of Scotland and the Swiss Jura and all of the maps in colour are on the book's website at <http://www.wiley.com/go/geostatistics2e>

Finally, we thank Blackwell Publishing Ltd for allowing us to reproduce Figures 6.7, 6.9 and 6.10 from a previous paper of ours.

**Richard Webster**

**Margaret Oliver**

March 2007

# 1

## ***Introduction***

### **1.1 WHY GEOSTATISTICS?**

Imagine the situation: a farmer has asked you to survey the soil of his farm. In particular, he wants you to determine the phosphorus content; but he will not be satisfied with the mean value for each field as he would have been a few years ago. He now wants more detail so that he can add fertilizer only where the soil is deficient, not everywhere. The survey involves taking numerous samples of soil, which you must transport to the laboratory for analysis. You dry the samples, crush them, sieve them, extract the phosphorus with some reagent and finally measure it in the extracts. The entire process is both time-consuming and costly. Nevertheless, at the end you have data from all the points from which you took the soil—just what the farmer wants, you might think!

The farmer's disappointment is evident, however. 'Oh', he says, 'this information is for a set of points, but I have to farm continuous tracts of land. I really want to know how much phosphorus the soil contains everywhere. I realize that that is impossible; nevertheless, I should really like some information at places between your sampling points. What can you tell me about those, and how do your small cores of soil relate to the blocks of land over which my machinery can spread fertilizer, that is, in bands 24 m wide?'

This raises further issues that you must now think about. Can you say what values to expect at intervening places between the sample points and over blocks the width of the farmer's fertilizer spreader? And how densely should you sample for such information to be reliable? At all times you must consider the balance between the cost of providing the information and the financial gains that will accrue to the farmer by differential fertilizing. In the wider context there may be an additional gain if you can help to avoid over-fertilizing and thereby protect the environment from pollution by excess phosphorus. Your task, as a surveyor, is to be able to use sparse affordable data to estimate, or predict, the average values of phosphorus in the soil over blocks of land 24 m × 24 m or perhaps longer strips. Can you provide the farmer with spatially referenced values that he can use in his automated fertilizer spreader?

This is not fanciful. The technologically minded farmer can position his machines accurately to 2 m in the field, he can measure and record the yields of his crops continuously at harvest, he can modulate the amount of fertilizer he adds to match demand; but providing the information on the nutrient status of the soil at an affordable price remains a major challenge in modern precision farming (Lake *et al.*, 1997).

So, how can you achieve this? The answer is to use geostatistics—that is what it is for.

We can change the context to soil salinity, pollution by heavy metals, arsenic in ground water, rainfall, barometric pressure, to mention just a few of the many variables and materials that have been and are of interest to environmental scientists. What is common to them all is that the environment is continuous, but in general we can afford to measure properties at only a finite number of places. Elsewhere the best we can do is to estimate, or predict, in a spatial sense. This is the principal reason for geostatistics—it enables us to do so without bias and with minimum error. It allows us to deal with properties that vary in ways that are far from systematic and at all spatial scales.

We can take the matter a stage further. Alert farmers and land managers will pounce on the word ‘error’. ‘Your estimates are subject to error’, they will say, ‘in other words, they are more or less wrong. So there is a good chance that if we take your estimates at face value we shall fertilize or remediate where we need not, and waste money, because you have underestimated, and not fertilize or fail to remediate where we should.’ The farmer will see that he might lose yield and profit if he applies too little fertilizer because you overestimate the nutrient content of the soil; the public health authority might take too relaxed an attitude if you underestimate the true value of a pollutant. ‘What do you say to that?’, they may say.

Geostatistics again has the answer. It can never provide complete information, of course, but, given the data, it can enable you to estimate the probabilities that true values exceed specified thresholds. This means that you can assess the farmer’s risks of losing yield by doing nothing where the true values are less than the threshold or of wasting money by fertilizing where they exceed it.

Again, there are analogies in many fields. In some situations the conditional probabilities of exceeding thresholds are as important as the estimates themselves because there are matters of law involved. Examples include limits on the arsenic content of drinking water (what is the probability that a limit is exceeded at an unsampled well?) and heavy metals in soil (what is the probability that there is more cadmium in the soil than the statutory maximum?)

### **1.1.1 Generalizing**

The above is a realistic, if colourful, illustration of a quite general problem. The environment extends more or less continuously in two dimensions. Its

properties have arisen as the result of the actions and interactions of many different processes and factors. Each process might itself operate on several scales simultaneously, in a non-linear way, and with local positive feedback. The environment, which is the outcome of these processes varies from place to place with great complexity and at many spatial scales, from micrometres to hundreds of kilometres.

The major changes in the environment are obvious enough, especially when we can see them on aerial photographs and satellite imagery. Others are more subtle, and properties such as the temperature and chemical composition can rarely be seen at all, so that we must rely on measurement and the analysis of samples. By describing the variation at different spatial resolutions we can often gain insight into the processes and factors that cause or control it, and so predict in a spatial sense and manage resources.

As above, measurements are made on small volumes of material or areas a few centimetres to a few metres across, which we may regard as point samples, known technically as *supports*. In some instances we enlarge the supports by taking several small volumes of material and mixing them to produce bulked samples. In others several measurements might be made over larger areas and averaged rather than recorded as single measurements. Even so, these supports are generally very much smaller than the regions themselves and are separated from one another by distances several orders of magnitude larger than their own diameters. Nevertheless, they must represent the regions, preferably without bias.

An additional feature of the environment not mentioned so far is that at some scale the values of its properties are positively related—*autocorrelated*, to give the technical term. Places close to one another tend to have similar values, whereas ones that are farther apart differ more on average. Environmental scientists know this intuitively. Geostatistics expresses this intuitive knowledge quantitatively and then uses it for prediction. There is inevitably error in our estimates, but by quantifying the spatial autocorrelation at the scale of interest we can minimize the errors and estimate them too.

Further, as environmental protection agencies set maximum concentrations, thresholds, for noxious substances in the soil, atmosphere and water supply, we should also like to know the probabilities, given the data, that the true values exceed the thresholds at unsampled places. Farmers and graziers and their advisers are more often concerned with nutrients in the soil and the herbage it grows, and they may wish to know the probabilities of deficiency, i.e. the probabilities that true values are less than certain thresholds. With some elaboration of the basic approach geostatistics can also answer these questions.

The reader may ask in what way geostatistics differs from the classical methods that have been around since the 1930s; what is the effect of taking into account the spatial correlation? At their simplest the classical estimators, based on random sampling, are linear sums of data, all of which carry the same

weight. If there is spatial correlation, then by stratifying we can estimate more precisely or sample more efficiently or both. If the strata are of different sizes then we might vary the weights attributable to their data in proportion. The means and their variances provided by the classical methods are regional, i.e. we obtain just one mean for any region of interest, and this is not very useful if we want local estimates. We can combine classical estimation with stratification provided by a classification, such as a map of soil types, and in that way obtain an estimate for each type of class separately. Then the weights for any one estimate would be equal for all sampling points in the class in question and zero in all others. This possibility of local estimation is described in Chapter 3. In linear geostatistics the predictions are also weighted sums of the data, but with variable weights determined by the strength of the spatial correlation and the configuration of the sampling points and the place to be estimated.

Geostatistical prediction differs from classical estimation in one other important respect: it relies on spatial models, whereas classical methods do not. In the latter, survey estimates are put on a probabilistic footing by the design of the sampling into which some element of randomization is built. This ensures unbiasedness, and provides estimates of error if the choice of sampling design is suitable. It requires no assumptions about the nature of the variable itself. Geostatistics, in contrast, requires the assumption that the variable is random, that the actuality on the ground, in the sea or in the air is the outcome of one or more random processes. The models on which predictions are based are of these random processes. They are not of the data, nor even of the actuality that we could observe completely if we had infinite time and patience. Newcomers to the subject usually find this puzzling; we hope that they will no longer do so when they have read Chapter 4, which is devoted to the subject. One consequence of the assumption is that sampling design is less important than in classical survey; we should avoid bias, but otherwise even coverage and sufficient sampling points are the main considerations.

The desire to predict was evident in weather forecasting and soil survey in the early twentieth century, to mention just two branches of environmental science. However, it was in mining and petroleum engineering that such a desire was matched by the financial incentive and resources for research and development. Miners wanted to estimate the amounts of metal in ore bodies and the thicknesses of coal seams, and petroleum engineers wanted to know the positions and volumes of reservoirs. It was these needs that constituted the force originally driving geostatistics because better predictions meant larger profits and smaller risks of loss. The solutions to the problems of spatial estimation are embodied in geostatistics and they are now used widely in many branches of science with spatial information. The origins of the subject have also given it its particular flavour and some of its characteristic terms, such as 'nugget' and 'kriging'.

There are other reasons why we might want geostatistics. The main ones are description, explanation and control, and we deal with them briefly next.

### 1.1.2 Description

Data from classical surveys are typically summarized by means, medians, modes, variances, skewness, perhaps higher-order moments, and graphs of the cumulative frequency distribution and histograms and perhaps box-plots. We should summarize data from a geostatistical survey similarly. In addition, since geostatistics treats a set of spatial data as a sample from the realization of a random process, our summary must include the spatial correlation. This will usually be the experimental or sample variogram in which the variance is estimated at increasing intervals of distance and several directions. Alternatively, it may be the corresponding set of spatial covariances or autocorrelation coefficients. These terms are described later. We can display the estimated semivariances or covariances plotted against sample spacing as a graph. We may gain further insight into the nature of the variation at this stage by fitting models to reveal the principal features. A large part of this book is devoted to such description.

In addition, we must recognize that spatial positions of the sampling points matter; we should plot the sampling points on a map, sometimes known as a ‘posting’. This will show the extent to which the sample fills the region of interest, any clustering (the cause of which should be sought), and any obvious mistakes in recording the positions such as reversed coordinates.

### 1.1.3 Interpretation

Having obtained the experimental variogram and fitted a model to it, we may wish to interpret them. The shape of the points in the experimental variogram can reveal much at this stage about the way that properties change with distance, and the adequacy of sampling. Variograms computed for different directions can show whether there is anisotropy and what form it takes. The variogram and estimates provide a basis for interpreting the causes of spatial variation and for identifying some of the controlling factors and processes. For example, Chappell and Oliver (1997) distinguished different processes of soil erosion from the spatial resolutions of the same soil properties in two adjacent regions with different physiography. Burrough *et al.* (1985) detected early field drains in a field in the Netherlands, and Webster *et al.* (1994) attempted to distinguish sources of potentially toxic trace metals from their variograms in the Swiss Jura.

### 1.1.4 Control

The idea of controlling a process is often central in time-series analysis. In it there can be a feedback such that the results of the analysis are used to change

the process itself. In spatial analysis the concept of control is different. In many instances we are unlikely to be able to change the spatial characteristics of a process; they are given. But we may modify our response. Miners use the results of analysis to decide whether to send blocks of ore for processing if the estimated metal content is large enough or to waste if not. They may also use the results to plan the siting of shafts and the expansion of mines. The modern precision farmer may use estimates from a spatial analysis to control his fertilizer spreader so that it delivers just the right amount at each point in a field.

## 1.2 A LITTLE HISTORY

Although mining provided the impetus for geostatistics in the 1960s, the ideas had arisen previously in other fields, more or less in isolation. The first record appears in a paper by Mercer and Hall (1911) who had examined the variation in the yields of crops in numerous small plots at Rothamsted. They showed how the plot-to-plot variance decreased as the size of plot increased up to some limit. ‘Student’, in his appendix to the paper, was even more perceptive. He noticed that yields in adjacent plots were more similar than between others, and he proposed two sources of variation, one that was autocorrelated and the other that he thought was completely random. In total, this paper showed several fundamental features of modern geostatistics, namely spatial dependence, correlation range, the support effect, and the nugget, all of which you will find in later chapters. Mercer and Hall’s data provided numerous budding statisticians with material on which to practise, but the ideas had little impact in spatial analysis for two generations.

In 1919 R. A. Fisher began work at Rothamsted. He was concerned primarily to reveal and estimate responses of crops to agronomic practices and differences in the varieties. He recognized spatial variation in the field environment, but for the purposes of his experiments it was a nuisance. His solution to the problems it created was to design his experiments in such a way as to remove the effects of both short-range variation, by using large plots, and long-range variation, by blocking, and he developed his analysis of variance to estimate the effects. This was so successful that later agronomists came to regard spatial variation as of little consequence.

Within 10 years Fisher had revolutionized agricultural statistics to great advantage, and his book (Fisher, 1925) imparted much of his development of the subject. He might also be said to have hidden the spatial effects and therefore to have held back our appreciation of them. But two agronomists, Youden and Mehlich (1937), saw in the analysis of variance a tool for revealing and estimating spatial variation. Their contribution was to adapt Fisher’s concepts so as to analyse the spatial scale of variation, to estimate the variation from different distances, and then to plan further sampling in the light of the knowledge gained. Perhaps they did not appreciate the significance of their

research, for they published it in the house journal of their institute, where their paper lay dormant for many years. The technique had to be rediscovered not once but several times by, for example, Krumbein and Slack (1956) in geology, and Hammond *et al.* (1958) and Webster and Butler (1976) in soil science. We describe it in Chapter 6.

We next turn to Russia. In the 1930s A. N. Kolmogorov was studying turbulence in the air and the weather. He wanted to describe the variation and to predict. He recognized the complexity of the systems with which he was dealing and found a mathematical description beyond reach. Nowadays we might call it chaos (Gleick, 1988). However, he also recognized spatial correlation, and he devised his ‘structure function’ to represent it. Further, he worked out how to use the function plus data to interpolate optimally, i.e. without bias and with minimum variance (Kolmogorov, 1941); see also Gandin (1965). Unfortunately, he was unable to use the method for want of a computer in those days. We now know Kolmogorov’s structure function as the variogram and his technique for interpolation as kriging. We deal with them in Chapters 4 and 8, respectively.

The 1930s saw major advances in the theory of sampling, and most of the methods of design-based estimation that we use today were worked out then and later presented in standard texts such as Cochran’s *Sampling Techniques*, of which the third edition (Cochran, 1977) is the most recent, and that by Yates, which appeared in its fourth edition as Yates (1981). Yates’s (1948) investigation of systematic sampling introduced the semivariance into field survey. Von Neumann (1941) had by then already proposed a test for dependence in time series based on the mean squares of successive differences, which was later elaborated by Durbin and Watson (1950) to become the Durbin–Watson statistic. Neither of these leads were followed up in any concerted way for spatial analysis, however.

Matérn (1960), a Swedish forester, was also concerned with efficient sampling. He recognized the consequences of spatial correlation. He derived theoretically from random point processes several of the now familiar functions for describing spatial covariance, and he showed the effects of these on global estimates. He acknowledged that these were equivalent to Jowett’s (1955) ‘serial variation function’, which we now know as the variogram, and mentioned in passing that Langsaetter (1926) had much earlier used the same way of expressing spatial variation in Swedish forest surveys.

The 1960s bring us back to mining, and to two men in particular. D. G. Krige, an engineer in the South African goldfields, had observed that he could improve his estimates of ore grades in mining blocks if he took into account the grades in neighbouring blocks. There was an autocorrelation, and he worked out empirically how to use it to advantage. It became practice in the gold mines. At the same time G. Matheron, a mathematician in the French mining schools, had the same concern to provide the best possible estimates of mineral grades from autocorrelated sample data. He derived solutions to the problem of

estimation from the fundamental theory of random processes, which in the context he called the theory of regionalized variables. His doctoral thesis (Matheron, 1965) was a *tour de force*.

From mining, geostatistics has spread into several fields of application, first into petroleum engineering, and then into subjects as diverse as hydrogeology, meteorology, soil science, agriculture, fisheries, pollution, and environmental protection. There have been numerous developments in technique, but Matheron's thesis remains the theoretical basis of most present-day practice.

### **1.3 FINDING YOUR WAY**

We are soil scientists, and the content of our book is inevitably coloured by our experience. Nevertheless, in choosing what to include we have been strongly influenced by the questions that our students, colleagues and associates have asked us and not just those techniques that we have found useful in our own research. We assume that our readers are numerate and familiar with mathematical notation, but not that they have studied mathematics to an advanced level or have more than a rudimentary understanding of statistics.

We have structured the book largely in the sequence that a practitioner would follow in a geostatistical project. We start by assuming that the data are already available. The first task is to summarize them, and Chapter 2 defines the basic statistical quantities such as mean, variance and skewness. It describes frequency distributions, the normal distribution and transformations to stabilize the variance. It also introduces the chi-square distribution for variances. Since sampling design is less important for geostatistical prediction than it is in classical estimation, we give it less emphasis than in our earlier *Statistical Methods* (Webster and Oliver, 1990). Nevertheless, the simpler designs for sampling in a two-dimensional space are described so that the parameters of the population in that space can be estimated without bias and with known variance and confidence. The basic formulae for the estimators, their variances and confidence limits are given.

The practitioner who knows that he or she will need to compute variograms or their equivalents, fit models to them, and then use the models to kriging can go straight to Chapters 4, 5, 6 and 8. Then, depending on the circumstances, the practitioner may go on to kriging in the presence of trend and factorial kriging (Chapter 9), or to cokriging in which additional variables are brought into play (Chapter 10). Chapter 11 deals with disjunctive kriging for estimating the probabilities of exceeding thresholds.

Before that, however, newcomers to the subject are likely to have come across various methods of spatial interpolation already and to wonder whether these will serve their purpose. Chapter 3 describes briefly some of the more popular methods that have been proposed and are still used frequently for prediction, concentrating on those that can be represented as linear sums of

data. It makes plain the shortcomings of these methods. Soil scientists are generally accustomed to soil classification, and they are shown how it can be combined with classical estimation for prediction. It has the merit of being the only means of statistical prediction offered by classical theory. The chapter also draws attention to its deficiencies, namely the quality of the classification and its inability to do more than predict at points and estimate for whole classes.

The need for a different approach from those described in Chapter 3, and the logic that underpins it, are explained in Chapter 4. Next, we give a brief description of regionalized variable theory or the theory of spatial random processes upon which geostatistics is based. This is followed by descriptions of how to estimate the variogram from data. The usual computing formula for the sample variogram, usually attributed to Matheron (1965), is given and also that to estimate the covariance.

The sample variogram must then be modelled by the choice of a mathematical function that seems to have the right form and then fitting of that function to the observed values. There is probably not a more contentious topic in practical geostatistics than this. The common simple models are listed and illustrated in Chapter 5. The legitimate ones are few because a model variogram must be such that it cannot lead to negative variances. Greater complexity can be modelled by a combination of simple models. We recommend that you fit apparently plausible models by weighted least-squares approximation, graph the results, and compare them by statistical criteria.

Chapter 6 is in part new. It deals with several matters that affect the reliability of estimated variograms. It examines the effects of asymmetrically distributed data and outliers on experimental variograms and recommends ways of dealing with such situations. The robust variogram estimators of Cressie and Hawkins (1980), Dowd (1984) and Genton (1998) are compared and recommended for data with outliers. The reliability of variograms is also affected by sample size, and confidence intervals on estimates are wider than many practitioners like to think. We show that at least 100–150 sampling points are needed, distributed fairly evenly over the region of interest. The distances between sampling points are also important, and the chapter describes how to design nested surveys to discover economically the spatial scales of variation in the absence of any prior information. Residual maximum likelihood (REML) is introduced to analyse the components of variance for unbalanced designs, and we compare the results with the usual least-squares approach.

For data that appear periodic the covariance analysis may be taken a step further by computation of power spectra. This detour into the spectral domain is the topic of Chapter 7.

The reader will now be ready for geostatistical prediction, i.e. kriging. Chapter 8 gives the equations and their solutions, and guides the reader in programming them. The equations show how the semivariances from the modelled variogram are used in geostatistical estimation (kriging). This chapter

shows how the kriging weights depend on the variogram and the sampling configuration in relation to the target point or block, how in general only the nearest data carry significant weight, and the practical consequences that this has for the actual analysis.

A new Chapter 9 pursues two themes. The first part describes kriging in the presence of trend. Means of dealing with this difficulty are becoming more accessible, although still not readily so. The means essentially involve the use of REML to estimate both the trend and the parameters of the variogram model of the residuals from the trend. This model is then used for estimation, either where there is trend in the variable of interest (universal kriging) or where the variable of interest is correlated with that in an external variable in which there is trend (kriging with external drift). These can be put into practice by the empirical best linear unbiased predictor.

Chapter 10 describes how to calculate and model the combined spatial variation in two or more variables simultaneously and to use the model to predict one of the variables from it, and others with which it is cross-correlated, by cokriging.

Chapter 11 tackles another difficult subject, namely disjunctive kriging. The aim of this method is to estimate the probabilities, given the data, that true values of a variable at unsampled places exceed specified thresholds.

Finally, a completely new Chapter 12 describes the most common methods of stochastic simulation. Simulation is widely used by some environmental scientists to examine potential scenarios of spatial variation with or without conditioning data. It is also a way of determining the likely error on predictions independently of the effects of the sampling scheme and of the variogram, both of which underpin the kriging variances.

In each chapter we have tried to provide sufficient theory to complement the mechanics of the methods. We then give the formulae, from which you should be able to program the methods (except for the variogram modelling in Chapter 5). Then we illustrate the results of applying the methods with examples from our own experience.

# 2

## ***Basic Statistics***

Before focusing on the main topic of this book, geostatistics, we want to ensure that readers have a sound understanding of the basic quantitative methods for obtaining and summarizing information on the environment. There are two aspects to consider: one is the choice of variables and how they are measured; the other, and more important, is how to sample the environment. This chapter deals with these. Chapter 3 will then consider how such records can be used for estimation, prediction and mapping in a classical framework.

The environment varies from place to place in almost every aspect. There are infinitely many places at which we might record what it is like, but practically we can measure it at only a finite number by sampling. Equally, there are many properties by which we can describe the environment, and we must choose those that are relevant. Our choice might be based on prior knowledge of the most significant descriptors or from a preliminary analysis of data to hand.

### **2.1 MEASUREMENT AND SUMMARY**

The simplest kind of environmental variable is binary, in which there are only two possible states, such as present or absent, wet or dry, calcareous or non-calcareous (rock or soil). They may be assigned the values 1 and 0, and they can be treated as quantitative or numerical data. Other features, such as classes of soil, soil wetness, stratigraphy, and ecological communities, may be recorded qualitatively. These qualitative characters can be of two types: unordered and ranked. The structure of the soil, for example, is an unordered variable and may be classified into blocky, granular, platy, etc. Soil wetness classes—dry, moist, wet—are ranked in that they can be placed in order of increasing wetness. In both cases the classes may be recorded numerically, but the records should not be treated as if they were measured in any sense. They can be converted to sets of binary variables, called ‘indicators’ in geostatistics (see Chapter 11), and can often be analysed by non-parametric statistical methods.

The most informative records are those for which the variables are measured fully quantitatively on continuous scales with equal intervals. Examples include the soil's thickness, its pH, the cadmium content of rock, and the proportion of land covered by vegetation. Some such scales have an absolute zero, whereas for others the zero is arbitrary. Temperature may be recorded in kelvin (absolute zero) or in degrees Celsius (arbitrary zero). Acidity can be measured by hydrogen ion concentration (with an absolute zero) or as its negative logarithm to base 10, pH, for which the zero is arbitrarily taken as  $-\log_{10} 1$  (in moles per litre). In most instances we need not distinguish between them. Some properties are recorded as counts, e.g. the number of roots in a given volume of soil, the pollen grains of a given species in a sample from a deposit, the number of plants of a particular type in an area. Such records can be analysed by many of the methods used for continuous variables if treated with care.

Properties measured on continuous scales are amenable to all kinds of mathematical operation and to many kinds of statistical analysis. They are the ones that we concentrate on because they are the most informative, and they provide the most precise estimates and predictions. The same statistical treatment can often be applied to binary data, though because the scale is so coarse the results may be crude and inference from them uncertain. In some instances a continuous variable is deliberately converted to binary, or to an 'indicator' variable, by cutting its scale at some specific value, as described in Chapter 11.

Sometimes, environmental variables are recorded on coarse stepped scales in the field because refined measurement is too expensive. Examples include the percentage of stones in the soil, the root density, and the soil's strength. The steps in their scales are not necessarily equal in terms of measured values, but they are chosen as the best compromise between increments of equal practical significance and those with limits that can be detected consistently. These scales need to be treated with some caution for analysis, but they can often be treated as fully quantitative.

Some variables, such as colour hue and longitude, have circular scales. They may often be treated as linear where only a small part of each scale is used. It is a different matter when a whole circle or part of it is represented. This occurs with slope aspect and with orientations of stones in till. Special methods are needed to summarize and analyse such data (see Mardia and Jupp, 2000), and we shall not consider them in this book.

### **2.1.1 Notation**

Another feature of environmental data is that they have spatial and temporal components as well as recorded values, which makes them unique or deterministic (we return to this point in Chapter 4). In representing the data we must distinguish measurement, location and time. For most classical statistical

analyses location is irrelevant, but for geostatistics the location must be specified. We shall adhere to the following notation as far as possible throughout this text. Variables are denoted by italics: an upper-case  $Z$  for random variables and lower-case  $z$  for a realization, i.e. the actuality, and also for sample values of the realization. Spatial position, which may be in one, two or three dimensions, is denoted by bold  $\mathbf{x}$ . In most instances the space is two-dimensional, and so  $\mathbf{x} = \{x_1, x_2\}$ , signifying the vector of the two spatial coordinates. Thus  $Z(\mathbf{x})$  means a random variable  $Z$  at place  $\mathbf{x}$ , and  $z(\mathbf{x})$  is the actual value of  $Z$  at  $\mathbf{x}$ . In general, we shall use bold lower-case letters for vectors and bold capitals for matrices.

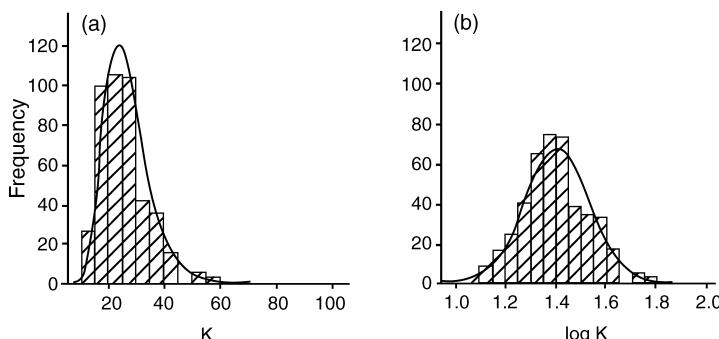
We shall use lower-case Greek letters for parameters of populations and either their Latin equivalents or place circumflexes ( $\hat{\cdot}$ ), commonly called ‘hats’ by statisticians, over the Greek for their estimates. For example, the standard deviation of a population will be denoted by  $\sigma$  and its estimate by  $s$  or  $\hat{s}$ .

### 2.1.2 Representing variation

The environment varies in almost every aspect, and our first task is to describe that variation.

#### Frequency distribution: the histogram and box-plot

Any set of measurements may be divided into several classes, and we may count the number of individuals in each class. For a variable measured on a continuous scale we divide the measured range into classes of equal width and count the number of individuals falling into each. The resulting set of frequencies constitutes the frequency distribution, and its graph (with frequency on the ordinate and the variate values on the abscissa) is the *histogram*. Figures 2.1 and 2.4 are examples. The number of classes chosen depends on the

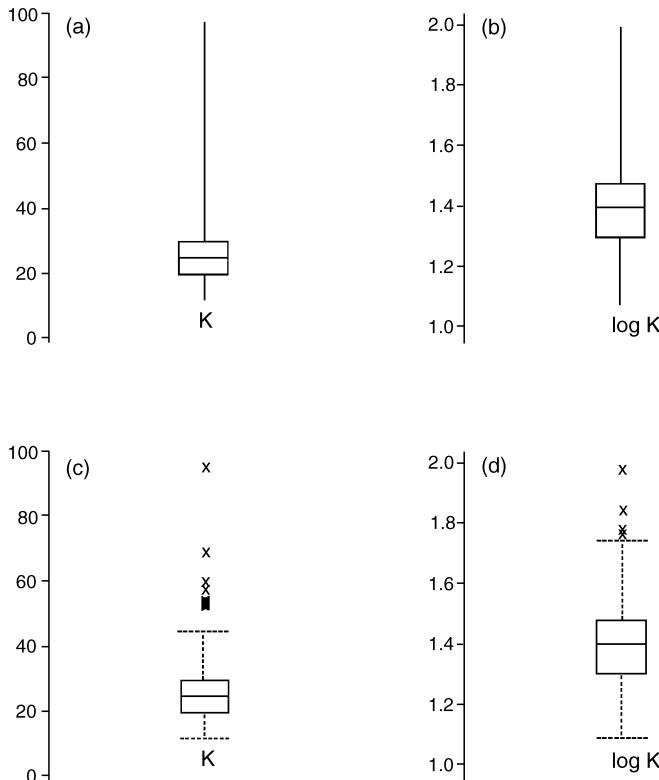


**Figure 2.1** Histograms: (a) exchangeable potassium (K) in  $\text{mg l}^{-1}$ ; (b)  $\log_{10} K$ , for the topsoil at Broom's Barn Farm. The curves are of the (lognormal) probability density.

number of individuals and the spread of values. In general, the fewer the individuals the fewer the classes needed or justified for representing them. Having equal class intervals ensures that the area under each bar is proportional to the frequency of the class. If the class intervals are not equal then the heights of the bars should be calculated so that the areas of the bars are proportional to the frequencies.

Another popular device for representing a frequency distribution is the box-plot. This is due to Tukey (1977). The plain 'box and whisker' diagram, like those in Figure 2.2, has a box enclosing the interquartile range, a line showing the median (see below), and 'whiskers' (lines) extending from the limits of the interquartile range to the extremes of the data, or to some other values such as the 90th percentiles.

Both the histogram and the box-plot enable us to picture the distribution to see how it lies about the mean or median and to identify extreme values.



**Figure 2.2** Box-plots: (a) exchangeable K; (b)  $\log_{10}K$  showing the 'box' and 'whiskers', and (c) exchangeable K and (d)  $\log_{10}K$  showing the fences at the quartiles plus and minus 1.5 times the interquartile range.

## Cumulative distribution

The cumulative distribution of a set of  $N$  observations is formed by ordering the measured values,  $z_i$ ,  $i = 1, 2, \dots, N$ , from the smallest to the largest, recording the order, say  $k$ , accumulating them, and then plotting  $k$  against  $z$ . The resulting graph represents the proportion of values less than  $z_k$  for all  $k = 1, 2, \dots, N$ . The histogram can also be converted to a cumulative frequency diagram, though such a diagram is less informative because the data are grouped.

The methods of representing frequency distribution are illustrated in Figures 2.1–2.6.

### 2.1.3 The centre

Three quantities are used to represent the ‘centre’ or ‘average’ of a set of measurements. These are the mean, the median and the mode, and we deal with them in turn.

#### Mean

If we have a set of  $N$  observations,  $z_i$ ,  $i = 1, 2, \dots, N$ , then we can compute their arithmetic average, denoted by  $\bar{z}$ , as

$$\bar{z} = \frac{1}{N} \sum_{i=1}^N z_i. \quad (2.1)$$

This, the mean, is the usual measure of central tendency.

The mean takes account of all of the observations, it can be treated algebraically, and the sample mean is an unbiased estimate of the population mean. For *capacity variables*, such as the phosphorus content in the topsoil of fields or daily rainfall at a weather station, means can be multiplied to obtain gross values for larger areas or longer periods. Similarly, the mean concentration of a pollutant metal in the soil can be multiplied by the mass of soil to obtain a total load in a field or catchment. Further, addition or physical mixing should give the same result as averaging.

*Intensity variables* are somewhat different. These are quantities such as barometric pressure and matric suction of the soil. Adding them or multiplying them does not make sense, but the average is still valuable as a measure of the centre. Physical mixing will in general not produce the arithmetic average. Some properties of the environment are not stable in the sense that bodies of material react with one another if they are mixed. For example, the average pH of a large volume of soil or lake water after mixing will not be the same as the average of the separate bodies of the soil or water that you measured previously. Chemical equilibration takes place. The same can be true for other exchangeable ions.

So again, the average of a set of measurements is unlikely to be the same as a single measurement on a mixture.

## **Median**

The median is the middle value of a set of data when the observations are ranked from smallest to largest. There are as many values less than the median as there are greater than it. If a property has been recorded on a coarse scale then the median is a rough estimate of the true centre. Its principal advantage is that it unaffected by extreme values, i.e. it is insensitive to outliers, mistaken records, faulty measurements and exceptional individuals. It is a robust summary statistic.

## **Mode**

The mode is the most typical value. It implies that the frequency distribution has a single peak. It is often difficult to determine the numerical value. If in a histogram the class interval is small then the mid-value of the most frequent class may be taken as the mode. For a symmetric distribution the mode, the mean and the median are in principle the same. For an asymmetric one

$$(\text{mode} - \text{median}) \approx 2 \times (\text{median} - \text{mean}). \quad (2.2)$$

In asymmetric distributions, e.g. Figures 2.1(a) and 2.4(a), the median and mode lie further from the longer tail of the distribution than the mean, and the median lies between the mode and the mean.

### **2.1.4 Dispersion**

There are several measures for describing the spread of a set of measurements: the range, interquartile range, mean deviation, standard deviation and its square, the variance. These last two are so much easier to treat mathematically, and so much more useful therefore, that we concentrate on them almost to the exclusion of the others.

#### **Variance and standard deviation**

The variance of a set of values, which we denote  $S^2$ , is by definition

$$S^2 = \frac{1}{N} \sum_{i=1}^N (z_i - \bar{z})^2. \quad (2.3)$$

The variance is the second moment about the mean. Like the mean, it is based on all of the observations, it can be treated algebraically, and it is little affected by sampling fluctuations. It is both additive and positive. Its analysis and use are backed by a huge body of theory. Its square root is the standard deviation,  $S$ . Below we shall replace the divisor  $N$  by  $N - 1$  so that we can use the variance of a sample to estimate  $\sigma^2$ , the population variance, without bias.

### Coefficient of variation

The standard deviation expresses dispersion in the same units as those in which the variable is measured. There are situations in which we may want to express it in relative terms, as where a property has been measured in two different regions to give two similar values of  $S$  but where the means are different. If the variances are the same we might regard the region with the smaller mean as more variable than the other in relative terms. The coefficient of variation (CV) can express this. It is usually presented as a percentage:

$$CV = 100(S/\bar{z})\%. \quad (2.4)$$

It is useful for comparing the variation of different sets of observations of the same property. It has little merit for properties with scales having arbitrary zeros and for comparing different properties except where they can be measured on the same scale.

### Skewness

The skewness measures the asymmetry of the observations. It is defined formally from the third moment about the mean:

$$m_3 = \frac{1}{N} \sum_{i=1}^N (z_i - \bar{z})^3. \quad (2.5)$$

The coefficient of skewness is then

$$g_1 = \frac{m_3}{m_2 \sqrt{m_2}} = \frac{m_3}{S^3}, \quad (2.6)$$

where  $m_2$  is the variance. Symmetric distributions have  $g_1 = 0$ . Skewness is the most common departure from normality (see below) in measured environmental data. If the data are skewed then there is some doubt as to which measure of centre to use. Comparisons between the means of different sets of observations are especially unreliable because the variances can differ substantially from one set to another.

### Kurtosis

The kurtosis expresses the peakedness of a distribution. It is obtained from the fourth moment about the mean:

$$m_4 = \frac{1}{N} \sum_{i=1}^N (z_i - \bar{z})^4. \quad (2.7)$$

The coefficient of kurtosis is given by

$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{m_4}{(S^2)^2} - 3. \quad (2.8)$$

Its significance relates mainly to the normal distribution, for which  $g_2 = 0$ . Distributions that are more peaked than normal have  $g_2 > 0$ ; flatter ones have  $g_2 < 0$ .

## 2.2 THE NORMAL DISTRIBUTION

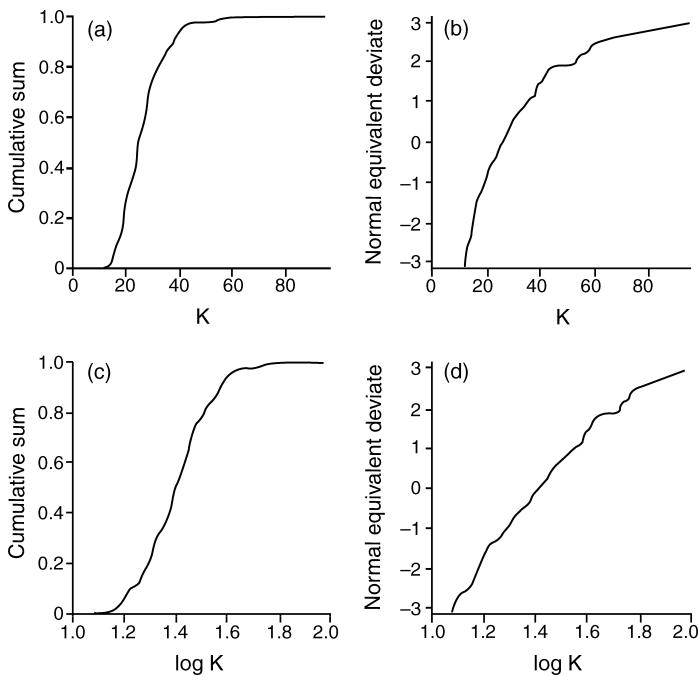
The normal distribution is central to statistical theory. It has been found to describe remarkably well the errors of observation in physics. Many environmental variables, such as of the soil, are distributed in a way that approximates the normal distribution. The form of the distribution was discovered independently by De Moivre, Laplace and Gauss, but Gauss seems generally to take the credit for it, and the distribution is often called ‘Gaussian’. It is defined for a continuous random variable  $Z$  in terms of the probability density function (pdf),  $f(z)$ , as

$$f(z) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(z-\mu)^2}{2\sigma^2}\right\}, \quad (2.9)$$

where  $\mu$  is the mean of the distribution and  $\sigma^2$  is the variance.

The shape of the normal distribution is a vertical cross-section through a bell. It is continuous and symmetrical, with its peak at the mean of the distribution. It has two points of inflexion, one on each side of the mean at a distance  $\sigma$ . The ordinate  $f(z)$  at any given value of  $z$  is the *probability density* at  $z$ . The total area under the curve is 1, the total probability of the distribution. The area under any portion of the curve, say between  $z_1$  and  $z_2$ , represents the proportion of the distribution lying in that range. For instance, slightly more than two-thirds of the distribution lies within one standard deviation of the mean, i.e. between  $\mu - \sigma$  and  $\mu + \sigma$ ; about 95% lies in the range  $\mu - 2\sigma$  to  $\mu + 2\sigma$ ; and 99.73% lies within three standard deviations of the mean.

Just as the frequency distribution can be represented as a cumulative distribution, so too can the pdf. In this representation the normal distribution



**Figure 2.3** Cumulative distribution: (a) exchangeable  $K$  in the range 0 to 1 and (b) as normal equivalent deviates, on the original scale ( $\text{mg l}^{-1}$ ); (c)  $\log_{10} K$  in the range 0 to 1 and (d) as normal equivalent deviates.

is characteristically sigmoid as in Figures 2.3(a), 2.3(c), 2.6(a) and 2.6(c). The main use of the cumulative distribution function is that the probability of a value's being less than a specified amount can be read from it. We shall return to this in Chapter 11.

In many instances distributions are far from normal, and these departures from normality give rise to unstable estimates and make inference and interpretation less certain than they might otherwise be. As above, we can be in some doubt as to which measure of centre to take if data are skewed. Perhaps more seriously, statistical comparisons between means of observations are unreliable if the variable is skewed because the variances are likely to differ substantially from one set to another.

## 2.3 COVARIANCE AND CORRELATION

When we have two variables,  $z_1$  and  $z_2$ , we may have to consider their joint dispersion. We can express this by their covariance,  $C_{1,2}$ , which for a finite set of

observations is

$$C_{1,2} = \frac{1}{N} \sum_{i=1}^N \{(z_1 - \bar{z}_1)(z_2 - \bar{z}_2)\}, \quad (2.10)$$

in which  $\bar{z}_1$  and  $\bar{z}_2$  are the means of the two variables. This expression is analogous to the variance of a finite set of observations, equation (2.3).

The covariance is affected by the scales on which the properties have been measured. This makes comparisons between different pairs of variables and sets of observations difficult unless measurements are on the same scale. Therefore, the *Pearson product-moment correlation coefficient*, or simply the correlation coefficient, is often preferred. It refers specifically to linear correlation and it is a dimensionless value.

The correlation coefficient is obtained from the covariance by

$$r = \frac{C_{1,2}}{S_1 S_2}. \quad (2.11)$$

This quantity is a measure of the relation between two variables; it can range between 1 and  $-1$ . If units with large values of one variable also have large values of the other then the two variables are positively correlated,  $r > 0$ ; if the large values of the one are matched by small values of the other then the two are negatively correlated,  $r < 0$ . If  $r = 0$  then there is no linear relation.

Just as the normal distribution is of special interest for a single variable, for two variables we are interested in a joint distribution that is bivariate normal. The joint pdf for such a distribution is given by

$$f(z) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left[ - \left\{ \frac{(z_1 - \mu_1)^2}{\sigma_1^2} - \frac{2\rho(z_1 - \mu_1)(z_2 - \mu_2)}{\sigma_1\sigma_2} + \frac{(z_2 - \mu_2)^2}{\sigma_2^2} \right\} / 2(1 - \rho^2) \right]. \quad (2.12)$$

In this equation  $\mu_1$  and  $\mu_2$  are the means of  $z_1$  and  $z_2$ ,  $\sigma_1^2$  and  $\sigma_2^2$  are the variances, and  $\rho$  is the correlation coefficient.

One can imagine the function as a bell shape standing above a plane defined by  $z_1$  and  $z_2$  with its peak above the point  $\{\mu_1, \mu_2\}$ . Any vertical cross-section through it appears as a normal curve, and any horizontal section is an ellipse—a ‘contour’ of equal probability.

## 2.4 TRANSFORMATIONS

To overcome the difficulties arising from departures from normality we can attempt to transform the measured values to a new scale on which the distribution is more nearly normal. We should then do all further analysis on

the transformed data, and if necessary transform the results to the original scale at the end. The following are some of the commonly used transformations for measured data.

### 2.4.1 Logarithmic transformation

The geometric mean of a set of data is

$$\bar{g} = \left\{ \prod_{i=1}^N z_i \right\}^{1/N}, \quad (2.13)$$

so that

$$\log \bar{g} = \frac{1}{N} \sum_{i=1}^N \log z_i, \quad (2.14)$$

in which the logarithm may be either natural ( $\ln$ ) or common ( $\log_{10}$ ). If by transforming the data  $z_i$ ,  $i = 1, 2, \dots, N$ , we obtain  $\log z$  with a normal distribution then the variable is said to be lognormally distributed. Its probability distribution is given by equation (2.9) in which  $z$  is replaced by  $\ln z$ , and  $\sigma$  and  $\mu$  are the parameters on the logarithmic scale.

It is sometimes necessary to shift the origin for the transformation to achieve the desired result. If subtracting a quantity  $a$  from  $z$  gives a close approximation to normality, so that  $z - a$  is lognormally distributed, then we have the probability density

$$f(z) = \frac{1}{\sigma(z-a)\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2} \{\ln(z-a) - \mu\}^2\right]. \quad (2.15)$$

We can write this as

$$f(z) = \frac{1}{\sigma(z-a)\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2} \left\{\ln \frac{z-a}{b}\right\}^2\right], \quad (2.16)$$

where  $b = \exp(\mu)$ . This is known as the three-parameter log-transformation; the parameters  $a$ ,  $b$  and  $\sigma$  represent the position, size and shape, respectively, of the distribution. You can read more about this distribution in Aitchison and Brown (1957).

### 2.4.2 Square root transformation

Taking logarithms will often normalize, or at least make symmetric, distributions that are strongly positively skewed, i.e. have  $g_1 > 1$ . Less pronounced

positive skewness can be removed by taking square roots:

$$r = \sqrt{z}. \quad (2.17)$$

### 2.4.3 Angular transformation

This is sometimes used for proportions in the range 0 to 1, or 0 to 100 if expressed as percentages. If  $p$  is the proportion then define

$$\phi = \sin^{-1} \sqrt{p}. \quad (2.18)$$

The desired transform is the angle whose sine is  $\sqrt{p}$ .

### 2.4.4 Logit transformation

If, as above,  $p$  is a proportion ( $0 < p < 1$ ), then its logit is

$$l = \ln \left\{ \frac{p}{1-p} \right\}. \quad (2.19)$$

Note that the limits 0 and 1 are excluded; otherwise  $l$  would either go to  $-\infty$  or  $+\infty$ . If you have proportions that include 0 or 1 then you must make some little adjustment to use the logit transformation.

In Chapter 11 we shall see a more elaborate transformation using Hermite polynomials.

## 2.5 EXPLORATORY DATA ANALYSIS AND DISPLAY

The physics of the environment might determine what transformation would be appropriate. More often than not, however, one must decide empirically by inspecting data. This is part of the preliminary exploration of the data from survey, which should always be done before more formal analysis. You should examine data by displaying them as histograms, box-plots and scatter diagrams, and compute summary statistics. You should suspect observations that are very different from their neighbours or from the general spread of values, and you should investigate abnormal values; they might be true outliers, or errors of measurement, or recording or transcription mistakes. You must then decide what to do about them.

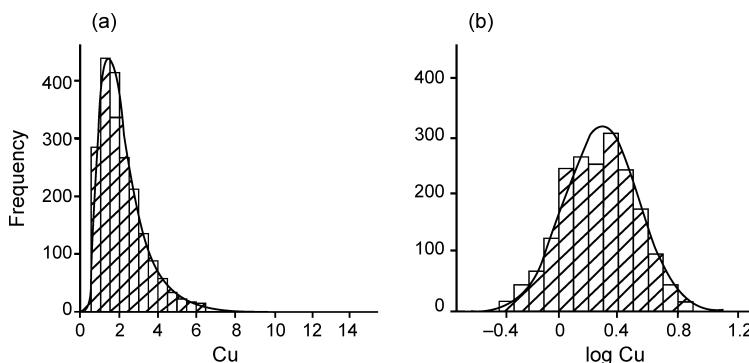
If the data are not approximately normal then you can experiment with transformation to make them so, as outlined in Section 2.4. There are formal

significance tests for normality, but these are generally not helpful, partly because they depend on the number of data and partly because they do not tell you in what way a distribution departs from normal. We illustrate this weakness below. You can try fitting theoretical distributions from the estimated parameters of the distribution to the histogram. If the histogram appears erratic then another way of examining the data for normality is to compute the cumulative distribution and plot it against the normal probability on normal probability paper. This paper has an ordinate scaled in such a way that a normal cumulative distribution appears as a straight line. Alternatively, you can compute the normal equivalent deviate for probability  $p$ ; this is the value of  $z$  to the left of which on the graph the area under the standard normal curve is  $p$ . A strong deviation from the line indicates non-normality, and you can try drawing the cumulative distributions of transformed data to see which gives a reasonable fit to the line before deciding whether to transform and, if so, in what way.

To illustrate these effects we turn to the distribution of potassium at Broom's Barn Farm. The data are from an original study by Webster and McBratney (1987). The distribution is shown as a histogram of the measured values in Figure 2.1(a). To it is fitted the curve of the lognormal distribution with parameters as given in Table 2.1. It is positively skewed. The histogram of the logarithms is shown in Figure 2.1(b). It is approximately symmetric, the normal pdf fits well, and transforming to logarithms has approximately normalized the data. Figure 2.2 shows the corresponding box-plots, as 'box and whisker' plots in which the limits of the boxes enclose the interquartile ranges and the whiskers extend to the limits of the data, Figure 2.2(a)–(b). In Figure 2.2(c)–(d) the whiskers extend only to 'fences', and any points lying beyond them are plotted individually. The upper fence is the limit of the upper quartile plus 1.5 times the interquartile range or the maximum if that is

**Table 2.1** Summary statistics for exchangeable potassium (K, mg l<sup>-1</sup>) at Broom's Barn Farm.

	K	$\log_{10} K$
Minimum	12.0	1.0792
Maximum	96.0	1.9823
Mean	26.31	1.3985
Median	25.0	1.3979
Standard deviation	9.039	0.1342
Variance	81.706	0.01800
Skewness	2.04	0.39
Kurtosis	9.51	0.57
Number of observations	434	434
$\chi^2$ for normal fit (with 18 degrees of freedom)	174.4	43.6



**Figure 2.4** Histograms: (a) extractable copper (Cu); (b)  $\log_{10}\text{Cu}$ , in the topsoil of the Borders Region. The curves are of the (lognormal) probability density.

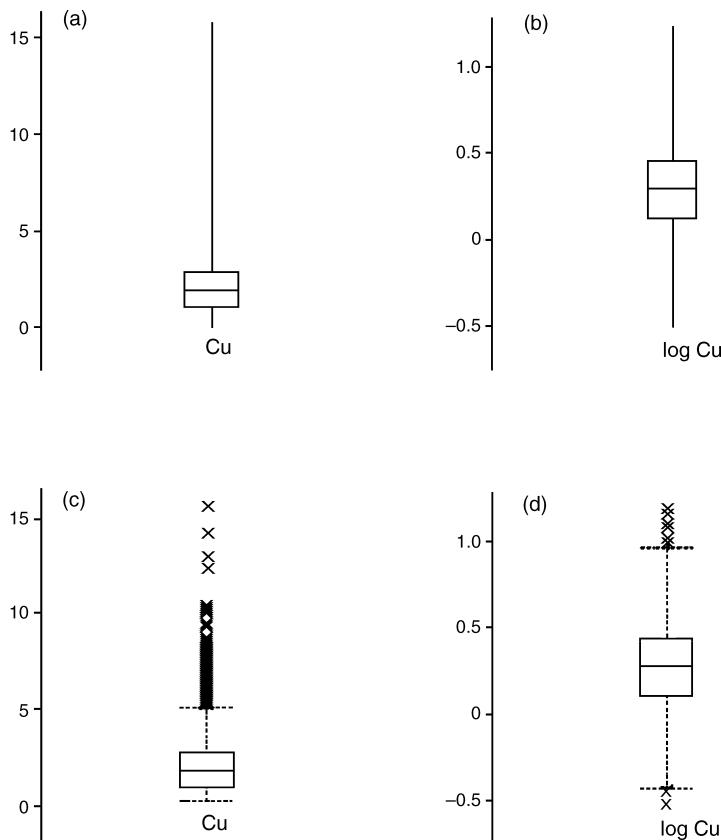
smaller; the lower fence is defined analogously. Again, skew is seen to be removed by taking logarithms. Figure 2.3(a)–(b) shows the cumulative distributions plotted on the probability scale and as normal equivalent deviates, respectively. Figure 2.3(c)–(d) shows the same graphs for  $\log_{10}K$ . These graphs are close to the normal line, and clearly transformation to logarithms yields a near-normal distribution in this instance.

Figures 2.4–2.6 show the effects of transformation to common logarithms for readily extractable copper of the topsoil in the Borders Region of Scotland (McBratney *et al.*, 1982). For these data, which are summarized in Table 2.2, taking logarithms normalizes the data very effectively.

The shortcomings of formal testing for a theoretical distribution can be seen in the  $\chi^2$  values given in Tables 2.1 and 2.2 for fitting the normal distribution. The values for the untransformed data are huge and clearly significant.

**Table 2.2** Summary statistics for extractable copper (Cu, mg kg<sup>-1</sup>) in the Borders Region.

	Cu	$\log_{10}\text{Cu}$
Minimum	0.3	-0.5214
Maximum	15.7	1.1959
Mean	2.221	0.2713
Median	1.85	0.2674
Standard deviation	1.461	0.2544
Variance	2.1346	0.064731
Skewness	2.52	0.06
Kurtosis	12.10	-0.05
Number of observations	1949	1949
$\chi^2$ for normal fit (with 18 degrees of freedom)	977.6	28.1

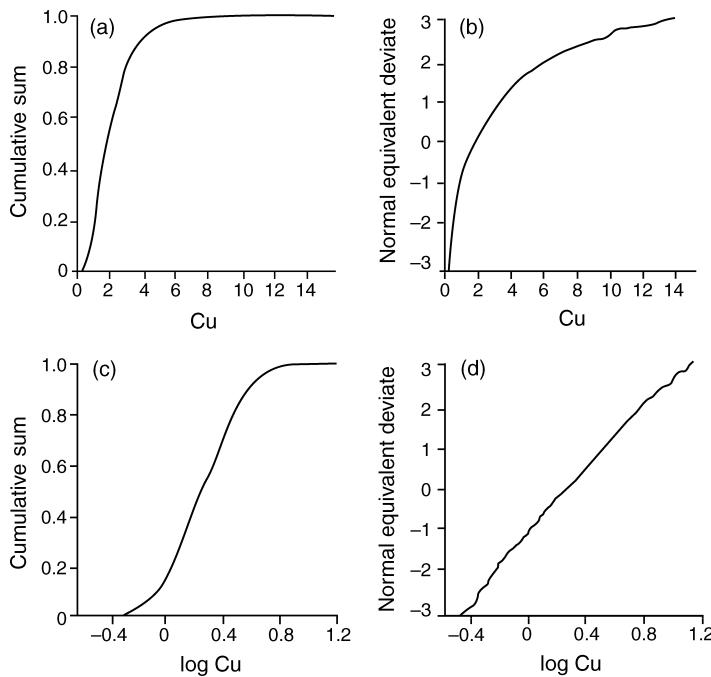


**Figure 2.5** Box-plots: (a) extractable Cu and (b)  $\log_{10}\text{Cu}$  showing the ‘box’ and ‘whiskers’; (c) extractable Cu and (d)  $\log_{10}\text{Cu}$  showing the fences at the quartiles plus and minus 1.5 times the interquartile range.

Transforming potassium to logarithms still gives a  $\chi^2$  (43.6) exceeding the 5% value ( $\chi^2_{p=0.05, f=18} = 28.87$ ), where  $p$  signifies the probability and  $f$  the degrees of freedom. Even for  $\log \text{Cu}$  the computed  $\chi^2$  (28.1) is close to the 5% value. The reason, as mentioned above, lies largely in having so many data, so that the test is very sensitive.

### 2.5.1 Spatial aspects

For spatial data the spatial coordinates must also be checked. The positions of the sampling points can be plotted on a map, referred to in Chapter 1 as a ‘posting’ of the data. Do all the points lie within the region surveyed? If not, why? Sampling points for a soil survey falling in the sea are obviously wrong,



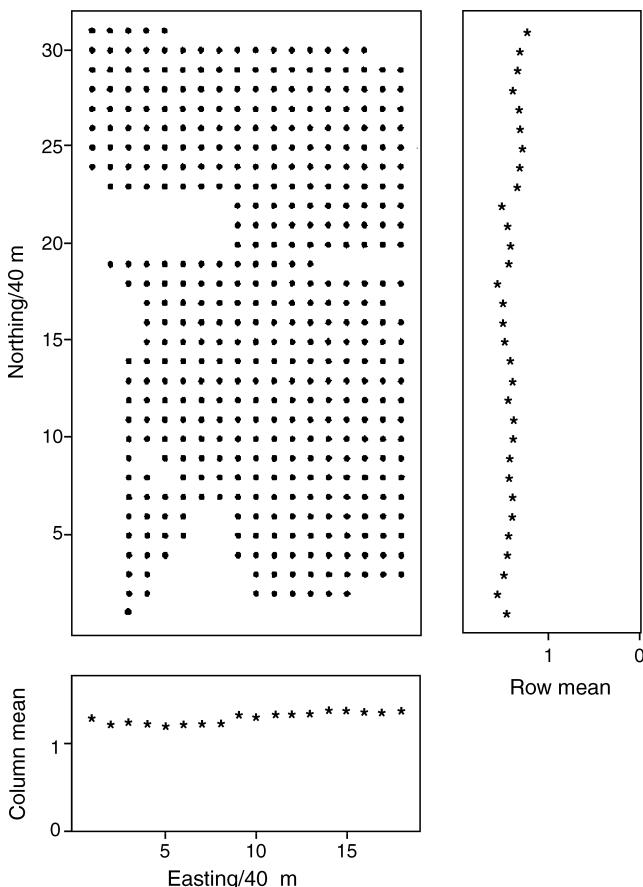
**Figure 2.6** Cumulative distributions: (a) extractable Cu in the range 0 to 1 and (b) as normal equivalent deviates, on the original scale ( $\text{mg kg}^{-1}$ ); (c)  $\log_{10}\text{Cu}$  in the range 0 to 1 and (d) as normal equivalent deviates.

but those on land just outside the region might be valid. Frequently the cause is a reversal of the coordinates, however.

The data should also be examined for trend, which might be evident as a gross regional change in the values, which is also smooth and predictable. If you have sampled on a grid then arrange the data in a two-way table, compute the means and medians of both rows and columns, and plot them. The results will show if the data embody trend, at least in the directions of the axes of the coordinate system, by a progressive increase or decrease in the row or column means. Figure 2.7 shows the distribution of the sampling points for Broom's Barn Farm. The graphs of the row and column means are on the right-hand side and at the bottom, respectively. These graphs show small fluctuations about the row and column means, but no evidence of trend.

## 2.6 SAMPLING AND ESTIMATION

We have made the point above that we can rarely have complete information about the environment. Soil, for example, forms a continuous mantle on the



**Figure 2.7** Posting of data for Broom's Barn Farm with the row and column means plotted on the right-hand side and at the bottom, respectively.

land except where it is broken by water or rock. Measurements, in contrast, are made on small cores or on bulked samples from small plots or fields. Similarly, rainfall is recorded in small gauges separated from one another by large distances. Data in this sense are fragmentary; they constitute a sample from whatever region is of interest, and from them we can try to describe the region in terms of mean values and variation.

The principal advances in sampling theory, sometimes known as classical theory, were made in the 1930s. The aim was to estimate means, and to a lesser extent higher-order moments, especially variances. It was not concerned to express spatial variation, which has become the province of geostatistics. Nevertheless, many of the ideas and formulae for geostatistics derive from the classical theory, and we therefore devote a short section to them. For fuller

treatment you should consult one of the standard texts such as Cochran (1977) and Yates (1981).

Better still in the context of environmental survey is the new book by de Gruijter *et al.* (2006) which deals with spatial sampling for both design-based estimation (the classical situation) and model-based prediction of geostatistics.

### **2.6.1 Target population and units**

The first step in sampling theory is to define a *target population*. This population comprises a set of *units*. In environmental survey a population is almost always circumscribed by the boundary of a physical region, and the units are all the places within it at which one might measure its properties. Measurements must be made on bodies of material with finite size, and so there is a finite number of non-overlapping units in the population. The units are usually so small in relation to the whole region that the population is effectively infinite. Millions of rain gauges 30 cm across could fit into a region of several hundred square kilometres without overlapping. The same is true of boreholes and soil profile pits. Even if the units were fields, there would be thousands of them. Nevertheless, they are all large enough to encompass variation, and in any one survey they should be of the same size. In fact, they should all have the same size, shape and orientation, known as the *support* of the sample.

The population is sampled by taking a subset of its units on a defined support. In classical theory this subset must be chosen with some element of randomization to ensure that the estimates from it are unbiased and to provide a probabilistic basis for inference. Perhaps paradoxically, the units must be selected according to a design to achieve this, and the technique is often called ‘design-based estimation’ in consequence.

### **2.6.2 Simple random sampling**

This is the simplest form of design. Every unit in the sample is chosen without regard to any other, and all units have the same chance of selection.

#### **Estimates from a simple random sample**

If there are  $N$  units in the sample then its *mean*,  $\bar{z}$ , estimates the mean of the parent population,  $\mu$ , by

$$\hat{\mu} = \bar{z} = \frac{1}{N} \sum_{i=1}^N z_i. \quad (2.20)$$

The *variance* of the population is the expected mean squared difference between  $\mu$  and  $z$ , i.e. it is the mean of  $(z - \mu)^2$ , denoted by  $\sigma^2$ . It is estimated by

$$\hat{\sigma}^2 = s^2 = \frac{1}{N-1} \sum_{i=1}^N (z_i - \bar{z})^2. \quad (2.21)$$

The divisor is  $N - 1$ , not  $N$ , and this difference between the formula for the estimated variance of a population and the variance of a finite set, equation (2.3), arises because we do not know the true mean, but have only an estimate of it from the data. The *standard deviation* of the sample,  $s$ , computed using equation (2.21), estimates  $\sigma$ . In like manner we estimate the population covariance between two variables by replacing the divisor  $N$  in equation (2.10) by  $N - 1$ .

### Estimation variance and standard error

All estimates are subject to error: sample information is never complete, and we want a measure of the uncertainty. This is usually expressed by the estimation variance of a mean:

$$s^2(\bar{z}) = \hat{\sigma}^2(\bar{z}) = s^2/N. \quad (2.22)$$

It estimates the variance we should expect if we were to sample repeatedly and compute the average squared difference between the mean  $\mu$  and the sample mean,  $\bar{z}$ :

$$\begin{aligned} E[s^2(\bar{z})] &= E[(\bar{z} - \mu)^2] \\ &= \sigma^2/N. \end{aligned} \quad (2.23)$$

Its square root is the standard error,  $s(\bar{z})$ . The equation introduces the symbol  $E$  to signify the expected value of something.

Naturally,  $s^2(\bar{z})$  should be as small as possible. Evidently we can decrease  $s^2(\bar{z})$ , and improve our estimates, by increasing  $N$ , the size of the sample. Unless we can measure every unit in a population, however, we cannot eliminate the error. Further, simply increasing  $N$  confers less and less benefit for the effort involved, and beyond about 25 the gain in precision is disappointing.

### 2.6.3 Confidence limits

Having obtained an estimate and its variance we may wish to know within what interval it lies for any degree of confidence. If the variable has a normal distribution and the sample is reasonably large then the confidence limits for the mean are readily obtained as follows.

**Table 2.3** Typical confidences and their associated standard normal deviates,  $y$ .

Confidence (%)	68	75	80	90	95	99
$y$	1.0	1.15	1.28	1.64	1.96	2.58

We consider a *standard normal deviate*, i.e. a normally distributed variable,  $y$ , with a mean of 0 and variance of 1, sometimes written  $\mathcal{N}(0, 1)$ . Then for any  $\mu$  and  $\sigma$ ,

$$y = \frac{z - \mu}{\sigma}. \quad (2.24)$$

Confidence limits on a mean are given by

$$\bar{z} - ys/\sqrt{N} \quad \text{and} \quad \bar{z} + ys/\sqrt{N}. \quad (2.25)$$

These are the lower and upper limits on  $\mu$ , given a sample mean  $\bar{z}$  and standard deviation  $s$  that estimates  $\sigma^2$  precisely, corresponding to some chosen probability or level of confidence. Values of standard normal deviates and their cumulative probabilities are published, and we list the values for a few typical confidences at which people might wish to work and the associated values of  $y$  in Table 2.3. The first entry is usually too liberal, and we include it only to show that approximately 68% of a normally distributed population lies within the range  $-\sigma$  to  $+\sigma$ .

## 2.6.4 Student's $t$

With small samples  $s^2$  is a poor estimate of  $\sigma^2$ , and in these circumstances one should replace  $y$  in expressions (2.25) by Student's  $t$ , which is defined by

$$t = \frac{\bar{z} - \mu}{s/\sqrt{N}}. \quad (2.26)$$

The true mean,  $\mu$ , is unknown of course, but  $t$  has been worked out and tabulated for  $N$  up to 120. So one chooses the confidence level, and then finds from the published table the value of  $t$  corresponding to  $N - 1$  *degrees of freedom*. The confidence limits of the mean are then

$$\bar{z} - ts/\sqrt{N} \quad \text{and} \quad \bar{z} + ts/\sqrt{N}. \quad (2.27)$$

As  $N$  increases so  $t$  approaches  $y$ , and for  $N \geq 60$  the differences are trivially small. So we need use  $t$  only when  $N < 60$ .

### 2.6.5 The $\chi^2$ distribution

Let  $y_1, y_2, \dots, y_m$  be  $m$  values drawn from a standard normal distribution. Their sum of squares is

$$\chi^2 = \sum_{i=1}^m y_i^2. \quad (2.28)$$

This quantity has the distribution

$$f(x) = \{2^{f/2} \Gamma(f/2)\}^{-1} x^{(f/2)-1} \exp(-x/2) \quad \text{for } x \geq 0, \quad (2.29)$$

where  $f$  is the number of degrees of freedom, equal to  $N - 1$  in our case, and  $\Gamma$  is the gamma function defined for any  $k > 0$  by

$$\Gamma(k) = \int_0^\infty x^{k-1} \exp(-x) dx.$$

Values of  $\chi^2$  have been worked out and tabulated, and can be found in any good book of statistical tables, such as that by Fisher and Yates (1963). They are also available in many statistical packages on computers.

The variance estimated from a sample is, from equation (2.21),

$$s^2 = \frac{1}{N-1} \sum_{i=1}^N (z_i - \mu)^2. \quad (2.30)$$

Dividing through by  $\sigma^2$  gives

$$\frac{s^2}{\sigma^2} = \frac{1}{N-1} \sum_{i=1}^N \frac{(z_i - \mu)^2}{\sigma^2}, \quad (2.31)$$

and so

$$s^2/\sigma^2 = \chi^2/(N-1) \quad \text{and} \quad \chi^2 = (N-1)s^2/\sigma^2$$

with  $N - 1$  degrees of freedom, provided the original population was normally distributed.

Rearranging the last expression gives the following limits for a variance:

$$\frac{(N-1)s^2}{\chi_{p_1}^2} \leq \sigma^2 \leq \frac{(N-1)s^2}{\chi_{p_2}^2}, \quad (2.32)$$

where  $p_1$  and  $p_2$  are the probabilities and for which we can obtain values of  $\chi^2$  from the published tables.

### **2.6.6 Central limit theorem**

In the foregoing discussion of confidence limits (Section 2.6.3) we have restricted the formulae to those for the normal distribution, the properties of which are so well established. It lends weight to our argument for transforming variables to normal if that is possible. However, even if a variable is not normally distributed it is often still possible to use the tabulated values and formulae when working with grouped data. As it happens, the distributions of sample means tend to be more nearly normal than those of the original populations. Further, the bigger is a sample the closer is the distribution of the sample mean to normality. This is the central limit theorem. It means that we can use a large body of theory when studying samples from the real world.

We might, of course, have to work with raw data that cannot readily be transformed to normal, and in these circumstances we should see whether the data follow some other known distribution. If they do then the same line of reasoning can be used to arrive at confidence limits for the parameters.

### **2.6.7 Increasing precision and efficiency**

The confidence limits on means computed from simple random samples can be alarmingly wide, and the sizes of sample needed to obtain satisfactory precision can also be alarmingly large. One reason when sampling space with a simple random design is that it is *inefficient*. Its cover is uneven; there are usually parts of the region that are sparsely sampled while elsewhere there are clusters of sampling points. If a variable  $z$  is spatially *autocorrelated*, which is likely at some scale, then clustered points duplicate information. Large gaps between sampling points mean that information that could have been obtained is lacking. Consequently, more points are needed to achieve a given precision, as measured by  $s^2(\bar{z})$ , than if the points are spread more evenly. There are several better designs for areas, and we consider the two most common ones, *stratified random* and *systematic*.

#### **Stratified sampling**

In stratified designs the region of interest,  $R$ , is divided into small subdivisions (*strata*). These are typically small squares, but they may be other shapes, of equal area. At least two sampling points are chosen randomly within each stratum. For this scheme the largest possible gap is then less than four strata.

The variance within a stratum  $k$  is estimated from  $n_k$  data in it by

$$s_k^2 = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (z_{ik} - \bar{z}_k)^2, \quad (2.33)$$

in which  $z_{ik}$  are the measured values and  $\bar{z}_k$  is their mean. If there are  $K$  strata then by averaging their variances we can obtain the estimated variance for the region:

$$s^2(\bar{z}, \text{stratified}) = \frac{1}{K^2} \sum_{k=1}^K \frac{s_k^2}{n_k}. \quad (2.34)$$

Its square root is the standard error.

The quantity  $(1/K) \sum_{k=1}^K s_k^2$  is the pooled within-stratum variance, denoted by  $s_W^2$ . If there is any spatial dependence then it will be less than  $s^2$ , and so the variance and standard error of a stratified sample will be less than that of a simple random sample for the same effort, the same size of sample.

The ratio  $s^2(\bar{z})/s^2(\bar{z}, \text{stratified})$  is the *relative precision* of stratification.

If we were happy with the precision achieved by simple random sampling then we could get the same precision by stratification with a smaller sample. Stratified sampling is more *efficient* by the factor

$$N_{\text{random}}/N_{\text{stratified}}.$$

## Systematic sampling

Systematic sampling provides the most even cover. In one dimension the sampling points are placed at equal intervals along a line, a transect. In two dimensions the points may be placed at the intersections of an equilateral triangular grid for maximum precision or efficiency. With this configuration the maximum distance between any unsampled point and the nearest point on the sampling grid is the least. However, rectangular grids are more practical, and the loss of precision compared with triangular ones is usually so small that they are preferred.

The main disadvantage of systematic sampling is that classical theory provides no means of determining the variance or standard error without bias from the sample because once one sampling point has been chosen (and the orientation in two dimensions) there is no randomization. An approximation may be obtained by dividing the region into strata and computing the pooled within-stratum variance as if sampling were random within the strata. The result will almost certainly be an overestimate, and conservative therefore. A closer approximation, and one that will almost certainly be close enough, can usually be obtained by Yates's method of *balanced differences* (Yates, 1981).

Estimates of error by balanced differences are computed as follows. Consider first regular sampling on a transect, i.e. in one dimension. The transect is viewed through a small window containing, say,  $m$  sampling points with values  $z_1, z_2, \dots, z_m$ . We then compute for the window the differences:

$$d_m = \frac{1}{2}z_1 - z_2 + z_3 - z_4 + \cdots + \frac{1}{2}z_m. \quad (2.35)$$

A value of  $m = 9$  is convenient. We then move the window along the transect in steps and compute  $d_m$  at each new position. If the transect is short then the positions should overlap; if not, a satisfactory procedure is to choose the first sampling point in a new position as the last one in the previous position. In this way every sampling point contributes, and with equation (2.35) all contribute equally. Then the variance for the transect mean is the sum

$$s^2(\text{balanced differences}) = \frac{1}{J(m - 2 + 0.5)} \sum_{j=1}^J d_{mj}^2, \quad (2.36)$$

where  $J$  is the number of steps or positions of the window, and the quantity  $m - 2 + 0.5$  is the sum of the squares of the coefficients in equation (2.35).

For a two-dimensional grid the procedure is analogous. One chooses a square window. For illustration let it be of side 4. The coefficients can be assigned as follows:

$$\begin{array}{cccc} -0.25 & +0.5 & -0.5 & +0.25 \\ +0.5 & -1.0 & +1.0 & -0.5 \\ -0.5 & +1.0 & -1.0 & +0.5 \\ +0.25 & -0.5 & +0.5 & -0.25 \end{array}$$

The variance is calculated as in equation (2.36), now with the divisor  $J \times 6.25$ , the value 6.25 being the sum of the squares of the coefficients above. Again, the positions of the window may overlap, but usually it is sufficient to arrange them so that only the sides are in common, and with this arrangement and the coefficients listed all points count and carry equal weight.

What these schemes do in both one and two dimensions, and in three if the scheme is extended, is to filter out long-range fluctuation, just as stratification does.

Where there is trend across the sampled region or periodicity, as, for example, in an orchard or as a result of land drainage, systematic sampling can give biased estimates of means. Such bias can be avoided by randomizing systematically within the grid. The result is *unaligned sampling* (see Webster and Oliver, 1990). It gives almost even cover. The disadvantage is the same as that of strict grid sampling in that the error cannot be estimated very accurately. The best procedure again is to stratify the region and compute the pooled within-stratum

variance. Empirical studies have shown some big gains in precision and efficiency from both systematic and unaligned sampling (again, see Webster and Oliver, 1990, for an example).

### 2.6.8 Soil classification

Another way of stratifying a region to improve the precision of estimates is to divide it on the basis of certain attributes. This practice is widespread in land resource surveys, and it was the norm in soil survey. Soil surveyors stratify, i.e. classify, regions on the appearance of the soil in profile and on related features in the landscape.

#### Regional mean

If the classification is good then the within-class variance of a stratum, i.e. the pooled within-stratum variance, is smaller than the total variance. Classification should therefore improve the precision or efficiency in estimating the regional mean.

The classes of soil are rarely equal in area, and so the formula, equation (2.34), must be adjusted accordingly. We define a weight,  $w_k$ , for the  $k$ th stratum or class in proportion to the area it covers:

$$w_k = \frac{\text{area of stratum } k}{\text{total area}}.$$

The mean,  $\mu$ , for the whole area is then estimated by the weighted average:

$$\bar{z} = \sum_{k=1}^K w_k \bar{z}_k, \quad (2.37)$$

where  $\bar{z}_k$  is the estimated mean of the  $k$ th stratum. The estimation variance is

$$s^2(\bar{z}, \text{stratified}) = \left\{ \sum_{k=1}^K \frac{w_k^2 s_k^2}{n_k} \right\}. \quad (2.38)$$

The average within-class variance and other diagnostics of a classification can be estimated from data by *analysis of variance*, which is both elegant and powerful. It can also serve for prediction, and we therefore defer its treatment to the next chapter.

# 3

## ***Prediction and Interpolation***

### **3.1 SPATIAL INTERPOLATION**

As mentioned in Chapter 2, measurements of the environment, of soil, weather, rock and water, are made on small bodies of material (supports) separated from one another by relatively large distances. They constitute a sample from a continuum that cannot be recorded everywhere. Yet the people who make the measurements or their clients would almost always like to know what the values are in the intervening space; they want to predict in a spatial sense from their more or less sparse data. For example, meteorologists want to predict rainfall from their rain gauges, hydrologists want to predict flow properties in rock from their measurements in boreholes, mining engineers want to estimate ore grades from diamond drill cores, and pedologists and agronomists want to estimate concentrations of elements in the soil from auger samples. Further, they usually want to map the spatial distributions of these variables. Their desires are almost as old the subjects themselves, and there have been many attempts to satisfy them quantitatively. They constitute the principal force driving geostatistics to meet practical needs; first in ore evaluation because of the huge costs of mining and metal extraction, but now in other branches of environmental science such as those we have listed.

Most attempts at spatial prediction have been mathematical, based on geometry and some appreciation of the physical nature of the phenomena. Most take account of only systematic or deterministic variation, but not of any error. In these respects, as we shall see, they fall short of what is needed practically. In some ways geostatistical prediction, kriging, is the logical conclusion of these attempts in that it builds on them and overcomes their weaknesses.

Nearly all the methods of prediction, including the simpler forms of kriging, can be seen as weighted averages of data. Thus we have the general prediction formula

$$z^*(\mathbf{x}_0) = \sum_{i=1}^N \lambda_i z(\mathbf{x}_i), \quad (3.1)$$

where  $\mathbf{x}_0$  is a target point for which we want a value; the  $z(\mathbf{x}_i)$ ,  $i = 1, 2, \dots, N$ , at places  $\mathbf{x}_i$  are the measured data; and  $\lambda_i$  are the weights assigned to them. For now we shall denote the prediction by  $z^*(\mathbf{x}_0)$ . First we examine how the weights are assigned for some of the common methods, and we leave kriging until Chapter 8 after we have dealt with its underlying theory.

### **3.1.1 Thiessen polygons (Voronoi polygons, Dirichlet tessellation)**

This method is one of the earliest and simplest. The region sampled,  $R$ , is divided by perpendicular bisectors between the  $N$  sampling points into polygons or tiles,  $V_i$ ,  $i = 1, 2, \dots, N$ , such that in each polygon all points are nearer to its enclosed sampling point  $\mathbf{x}_i$  than to any other sampling point. The prediction at each point in  $V_i$  is the measured value at  $\mathbf{x}_i$ , i.e.  $z^*(\mathbf{x}_0) = z(\mathbf{x}_i)$ . The weights are

$$\lambda_i = \begin{cases} 1 & \text{if } \mathbf{x}_i \in V_i, \\ 0 & \text{otherwise.} \end{cases} \quad (3.2)$$

The shortcomings of the method are evident; each prediction is based on just one measurement, there is no estimate of the error, and information from neighbouring points is ignored. When used for mapping the result is crude; the interpolated surface consists of a series of steps.

### **3.1.2 Triangulation**

Another early group of interpolators comprises those deriving from triangulation. The sampling points are linked to their neighbours by straight lines to create triangles that do not contain any of the points. The measured values are envisaged as standing above the basal plane at a height proportional to those values so that the whole set of data forms a polyhedron consisting of more or less tilted triangular plates. The aim is to determine the height of the plate at  $\mathbf{x}_0$  from the apices of the triangle by linear interpolation.

This can be represented as a weighted average with weights determined as follows. We denote the coordinates of the three apices by  $\{x_{11}, x_{12}\}$ ,  $\{x_{21}, x_{22}\}$  and  $\{x_{31}, x_{32}\}$  and those of the target point by  $\{x_{01}, x_{02}\}$ . Then the weights are given by

$$\lambda_1 = \frac{(x_{01} - x_{31})(x_{22} - x_{32}) - (x_{02} - x_{32})(x_{21} - x_{31})}{(x_{11} - x_{31})(x_{22} - x_{32}) - (x_{12} - x_{32})(x_{21} - x_{31})}. \quad (3.3)$$

An analogous equation for  $\lambda_2$  is obtained by the exchange of  $x_{11}$  for  $x_{21}$ ,  $x_{12}$  for  $x_{22}$ ,  $x_{12}$  for  $x_{21}$  and  $x_{21}$  for  $x_{11}$ . A similar set of exchanges will give the value  $\lambda_3$ . All other weights are 0.

The technique is simple and local. The disadvantages are that, although it is somewhat better than the Thiessen method, each prediction still depends on only three data; it makes no use of data further away, and there is again no measure of error. Unlike the Thiessen method, the resulting surface is continuous, but it has abrupt changes in gradient at the margins of the triangles. If the principal aim is to predict rather than to make a map with smooth isolines then the discontinuities in the derivative are immaterial. Another difficulty is that there is no obvious triangulation that is better than any other; even for a rectangular grid there are two options.

### 3.1.3 Natural neighbour interpolation

Sibson (1981) combined the best features of the two methods above in what he called ‘natural neighbour interpolation’. The first step is a triangulation of the data by Delauney’s method in which the apices of the triangles are those sampling points in adjacent Dirichlet tiles. This triangulation is unique except where the data are on a regular rectangular grid. To determine the value at any other point,  $\mathbf{x}_0$ , that point is inserted into the tessellation, and its neighbours, the set  $T$  (the points within its bounding Dirichlet tiles), are used for the interpolation. Sibson called these points ‘natural neighbours’.

For each neighbour the area,  $A$ , of the portion of its original Dirichlet tile that became incorporated in the tile of the new point is calculated. These areas, when scaled to sum to 1, become the weights. We can represent this by the general formula:

$$\lambda_i = \frac{A_i}{\sum_{k=1}^N A_k} \quad \text{for all } i = 1, 2, \dots, N. \quad (3.4)$$

This means that if a point  $\mathbf{x}_i$  is a natural neighbour, i.e.  $\mathbf{x}_i \in T$ , then  $A_i$  has a value and the point carries a positive weight. If  $\mathbf{x}_i$  is not a natural neighbour then it has no area in common with the target and its weight,  $\lambda_i$ , is zero.

This interpolator is continuous and smooth except at the data points where its derivative is discontinuous. Sibson called it the *natural neighbour*  $C^0$  *interpolant*.

He did not like abrupt change in the surface at the data points, and so he elaborated the method by calculating the gradients of the statistical surface at these from their natural neighbours. These gradients were then combined with the weighted measurements to provide the height at the new point. The result is a smooth, once differentiable surface. Like the simple polyhedral interpolator, it returns the actual values at the measured points, i.e. it is an exact interpolator. Sibson showed that it reproduces continuous mathematical functions faithfully. However, both we and Laslett *et al.* (1987) have found that it produces unacceptable results where data are noisy. At local maxima and minima in such data it generates ‘Prussian helmets’, which Sibson wished to avoid.

### 3.1.4 Inverse functions of distance

Somewhat more elaborate than triangulation, and much more popular, are the methods based on inverse functions of distance in which the weights are defined by

$$\lambda_i = 1/|\mathbf{x}_i - \mathbf{x}_0|^\beta \quad \text{with } \beta > 0, \quad (3.5)$$

and again scaled so that they sum to 1. The result is that data points near to the target point carry larger weight than those further away. The most popular choice of  $\beta$  is 2 so that the data are inversely weighted as the square of distance. As with triangulation, if  $\mathbf{x}_0$  coincides with any  $\mathbf{x}_i$  then  $\lambda_i$  becomes infinite, the other weights are immaterial, and  $z(\mathbf{x}_0)$  takes the value  $z(\mathbf{x}_i)$ . Interpolation is exact. An attractive feature of weighting by inverse squared distance is that the relative weights diminish rapidly as the distance increases, and so the interpolation is sensibly local. Further, because the weights never become zero there are no discontinuities. Its disadvantages are that the choice of the weighting function is arbitrary, and there is no measure of error. Further, it takes no account of the configuration of the sampling. So where data are clustered two or more may be at approximately the same distance and direction from  $\mathbf{x}_0$ , and each point will carry the same weight as an isolated point a similar distance away but in a different direction. This is clearly undesirable, and some implementations for mapping have elaborated the scheme to overcome this—see, for example, Shepard's (1968) solution in the once popular SYMAP program. The interpolated surface will have a gradient of zero at the data points, and maxima and minima can occur only there.

### 3.1.5 Trend surfaces

A method that became popular among earth scientists, especially petroleum geologists, when they first had access to computers was trend surface analysis. This is a form of multiple regression in which the predictors are the spatial coordinates. For example,

$$z(x_1, x_2) = f(x_1, x_2) + \varepsilon, \quad (3.6)$$

where  $z(x_1, x_2)$  is the predicted value at  $\{x_1, x_2\}$  and  $f$  denotes a function of the spatial coordinates there. The model contains an error term,  $\varepsilon$ , and in regression this is assumed to be independently and identically distributed with mean 0 and variance  $\sigma_\varepsilon^2$ . Plausible functions, usually simple polynomials such as planes, quadratics or cubics, are fitted by least squares to the spatial coordinates, and the resulting regression equation is used for the prediction. Thus for a plane the regression equation would be

$$z = b_0 + b_1 x_1 + b_2 x_2, \quad (3.7)$$

and for a quadratic surface

$$z = b_0 + b_1x_1 + b_2x_2 + b_3x_1^2 + b_4x_2^2 + b_5x_1x_2. \quad (3.8)$$

The predictor can be expressed as a weighted average of the data used to obtain the trend surface as follows. We represent the spatial coordinates and their powers by a matrix  $\mathbf{X}$  with  $N$  rows for the  $N$  sampling points and as many columns as coefficients  $b$  to be estimated. For a first-order surface we can write the spatial coordinates as the matrix

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{12} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{N1} & x_{N2} \end{bmatrix},$$

in which the first column is a dummy variate of 1s, and the recorded values of  $z$  at those places as the vector

$$\mathbf{z} = \begin{bmatrix} z(\mathbf{x}_1) \\ z(\mathbf{x}_2) \\ \vdots \\ z(\mathbf{x}_N) \end{bmatrix}.$$

The coefficients  $b$  are obtained from the matrix multiplication

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{z}, \quad (3.9)$$

and the predictions are then given by

$$z_0^* = \mathbf{x}_0 \mathbf{b}, \quad (3.10)$$

in which  $\mathbf{x}_0$  is the row vector  $[1 \ x_{01} \ x_{02}]$ . Thus the weights are given by equation (3.9). For a more complex surface the matrix  $\mathbf{X}$  is simply extended by adding columns for the additional powers of  $x_1$  and  $x_2$ .

Initially trend surfaces seemed attractive, but enthusiasm soon turned to disappointment. In most instances spatial variation is so complex that a polynomial of very high order is needed to describe it, and the resulting matrix equations are usually unstable. The residuals from the trend are autocorrelated, and so one of the assumptions of regression is violated. As a consequence the errors calculated by the usual formula, such as equation (6.49) in Webster and Oliver (1990), are incorrect. The random component is often large and masks the deterministic trend, and fitting in one part of a region affects the predictions

everywhere. Thus, in a region containing both mountains and plain the prediction of topographic height on the plain will be determined by the much larger fluctuations in the mountains. Trend surfaces are not sufficiently local, and they do not return the values at data points.

Nevertheless, simple regression surfaces can represent long-range trend in some instances. The technique has its merits therefore in revealing long-range structures and filtering them to leave variation of shorter range that can be analysed by other techniques—see Moffat *et al.* (1986) for an example. We return to this matter in Chapter 9.

### **3.1.6 Splines**

A spline function also consists of polynomials, but each polynomial of degree  $p$  is local rather than global. The polynomials describe pieces of a line or surface, and they are fitted together so that they join smoothly, and their  $p - 1$  derivatives are continuous. The places at which the pieces join are known as ‘knots’, and the choice of knots confers an arbitrariness on the technique. Splines can be constrained to pass through the data. Alternatively, by choosing knots away from the data points they can be fitted by least squares or some other method to produce smoothing splines. Typically the splines are of degree 3; these are cubic splines.

## **3.2 SPATIAL CLASSIFICATION AND PREDICTING FROM SOIL MAPS**

We conclude this chapter with a look at prediction using spatial classification. Surveyors in most branches of environmental science divide the regions that they study into classes by boundaries, and they characterize each class so derived from sample data. The maps, choropleth maps to give them their technical name, are commonplace. The soil map showing a patchwork of colour is a familiar one, as are similar maps in geology and ecology. The intention, usually implied rather than expressed, is that the characteristic information for any one class, any one colour on the map, may be used to predict conditions elsewhere in the same class. Rarely, however, is this put in statistical terms, yet it is the only sound application of classical statistics to spatial prediction. Indeed, it combines classical survey in such fields as geology and soil science with classical statistics. Several scientists have analysed soil maps in this framework—for example, Kantey and Williams (1962), Morse and Thornburn (1961), and Webster and Beckett (1970)—but as far as we know, the subject is not covered in any textbook. We therefore describe it in some detail, drawing on the thorough analysis by Leenhardt *et al.* (1994).

### 3.2.1 Theory

We start with a region of interest,  $R$ , that has been classified into  $K$  classes, separated by boundaries. Every point in  $R$ , i.e. every unit of the population, belongs to one and only one class.

For any class  $k$  in  $R$  we can express the value of any point  $\mathbf{x}_i$ , i.e. any unit  $i$ , selected at random as

$$Z_{ik} = \mu + \alpha_k + \varepsilon_{ik}, \quad (3.11)$$

where  $Z_{ik}$  is the value of  $z$  at  $\mathbf{x}_i$  in class  $k$ ,  $\mu$  is the general mean of  $z$ ,  $\alpha_k$  is the difference between  $\mu$  and the mean of the class  $k$ , and  $\varepsilon_{ik}$  is a random component with mean zero and variance  $\sigma_k^2$ , the variance within class  $k$ .

The mean of class  $k$ ,  $\mu_k = \mu + \alpha_k$ , is estimated from  $n_k$  observations by

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} z_i, \quad (3.12)$$

with variance

$$\sigma^2(\hat{\mu}_k) = \sigma_k^2/n_k. \quad (3.13)$$

In the absence of other information  $\mu_k$  is also the best predictor of  $z$  at any  $\mathbf{x}_i, i \in k$ , and in keeping with our general formula for linear predictors, equation (3.1), we can represent it as

$$z^*(\mathbf{x}_0) = \hat{\mu}_k = \sum_{i=1}^N \lambda_i z(\mathbf{x}_i), \quad (3.14)$$

in which

$$\lambda_i = \begin{cases} 1/n_k & \text{for } \mathbf{x}_i \in k, \\ 0 & \text{otherwise.} \end{cases} \quad (3.15)$$

Its prediction variance is the expected mean squared difference (MSE) between the true value and the predicted one:

$$\text{MSE}_k = E_i[\{Z_{ik} - \mu_k\}^2] = \sigma_k^2. \quad (3.16)$$

In practice we never know  $\mu_k$ ; we only ever have an estimate,  $\hat{\mu}_k$ . So its variance,  $\text{var}[\hat{\mu}_k]$ , is an additional source of error in our prediction. Further, there is the possibility that our estimate of  $\mu_k$  is biased. So a term representing the bias should be added, and the full squared prediction error becomes

$$\text{MSE}_k = \sigma_k^2 + \text{var}[\hat{\mu}_k] + \text{bias}^2[\hat{\mu}_k]. \quad (3.17)$$

If we sample randomly and estimate  $\mu_k$  by the arithmetic average of  $n_k$ , the observed values in  $k$ , then there is no bias, and  $\text{var}[\hat{\mu}_k]$  equals  $\sigma_k^2/n_k$ . Equation (3.17) then becomes

$$\text{MSE}_k = \sigma_k^2 + \sigma_k^2/n_k = \sigma_k^2(1 + 1/n_k). \quad (3.18)$$

An immediate practical problem is to estimate  $\sigma_k^2$ . Confidence limits on variances are typically wide (Chapter 2) for small samples, and for a map with many classes surveyors rarely have the resources to record at more than a few sites within each class. Consequently equation (3.18) is likely to lead to crude estimates of the MSE. To solve this problem we therefore make a further assumption, namely that the variance within classes is the same for all. In conventional soil mapping, for example, surveyors try to maintain the same categoric level for all the classes in any one survey, say, all soil series or all soil families. The intention, expressed or implied, is that classes are equally variable. In these circumstances  $\sigma_k^2$  in the above equations may be replaced by  $\sigma_W^2$ , the average or pooled within-class variance. Our task now is to estimate it, and this is best done by an analysis of variance (see Chapter 2).

The total variance of  $Z$  in the region, designated by  $\sigma_T^2$ , can be written as

$$\sigma_T^2 = \sigma_W^2 + \sigma_B^2, \quad (3.19)$$

where  $\sigma_B^2$  is the between-class variance. These quantities immediately lead to expressions of the efficacy of a classification at partitioning the variance of  $Z$  by, for example, the intraclass correlation:

$$\rho_i = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2} = 1 - (\sigma_W^2/\sigma_T^2). \quad (3.20)$$

Evidently, the larger is  $\sigma_B^2$  and the smaller is  $\sigma_W^2$ , and hence the larger is  $\rho_i$ , the better we should regard the classification.

### **Prediction using a random sample**

If we sample a region by selecting points at random and with numbers proportional to the areas covered by the classes then  $\sigma_W^2$  and  $\sigma_B^2$  are estimated by  $s_W^2$  and  $s_B^2$ , respectively, without bias in a one-way analysis of variance. This leads equally to an unbiased estimate of  $\rho_i$  by  $r_i$  (Webster and Beckett, 1968). Alternatively, we may take  $s_T^2$  as an estimate of  $\sigma_T^2$  and compute  $R_i^2 = 1 - (s_W^2/s_T^2)$ , which is the proportion of variance in the data explained by the classification and analogous to  $R^2$  in regression analysis.

We can now insert the pooled within-class variance,  $\sigma_W^2$ , into equation (3.18) in place of  $\sigma_k^2$  to obtain

$$\text{MSE}_k = \sigma_W^2(1 + 1/n_k). \quad (3.21)$$

Further, we can compute an average prediction error,  $\overline{\text{MSE}}$ , for the region by

$$\overline{\text{MSE}} = \sum_{k=1}^K A_k \sigma_W^2(1 + 1/n_k), \quad (3.22)$$

where  $A_k$  is the proportion of the area of class  $k$  in  $R$ .

### Prediction from a purposively chosen sample

Consider now predicting  $Z$  from a purposively chosen representative profile,  $p$ , in class  $j$  with value  $z_{pj}$ . The latter replaces  $\hat{\mu}_j$  as an estimate of  $Z_{ij}$ . It is fixed, however, so  $\text{var}[z_{pj}] = 0$ , and the difference  $d_j = z_{pj} - \mu_j$  is the bias of equation (3.17). The prediction variance is

$$\text{MSE}_{pj} = \sigma_j^2 + d_j^2. \quad (3.23)$$

Under the assumption of a common within-class variance, we obtain the expected mean squared error of prediction from class representatives for the whole region by

$$\overline{\text{MSE}}_p = E_j[\text{MSE}_{pj}] = \sigma_W^2 + \sum_{j=1}^J A_j d_j^2. \quad (3.24)$$

The minimum value of  $\overline{\text{MSE}}_p$  is  $\sigma_W^2$ , which is reached when the  $z_{pj} = \mu_j$  for all  $j$ .

#### 3.2.2 Summary

Whether we predict  $Z$  using the means of random samples or from purposively chosen representatives,  $\sigma_W^2$  sets a lower limit to the mean squared error of prediction. In the former case we can approach this minimum by increasing the size of sample; in the latter by selecting the representatives to match the mean values as closely as possible. If we want to improve prediction further using the conventional approach, we must diminish the within-class variance by refining the classification. This might be done by increasing the scale so that boundaries can be delineated more accurately and intricately, or by subdividing the soil more finely, i.e. by increasing the number of classes. In practice the second is

likely to demand the first: there is no point in creating classes that cannot be displayed at the chosen scale. Alternatively, we might devise a special classification for each property we wish to predict.

The effectiveness of the conventional procedure for soil survey depends both on the quality of the classification and its mapping and on the ability of the surveyor to select representative soil profiles in the field, where the values of the properties of interest approximate the class means. In particular,  $\overline{\text{MSE}}_p$  should be less than  $2\sigma_W^2$ , otherwise the selection is worthless, and  $2\sigma_W^2$  should be less than  $\sigma_T^2 + \sigma_T^2/N$ , where  $N$  is the total size of sample, otherwise classification confers no benefit.

For the whole procedure to be successful we want

$$\sigma_W^2 < \overline{\text{MSE}}_p < 2\sigma_W^2 < \sigma_T^2 + \sigma_T^2/N. \quad (3.25)$$

To complete the picture we have to estimate  $\overline{\text{MSE}}_p$ . Let us assume that we have for each class  $j$ ,  $j = 1, 2, \dots, J$ , one representative with value  $z_{pj}$  and  $V(j)$  validation points chosen probabilistically with values  $Z_{vj}$ ,  $v = 1, 2, \dots, V(j)$ , and that there are  $N_v$  validation points in all. Then

$$\hat{d}_p = \frac{1}{N_v} \sum_{j=1}^J \sum_{v=1}^{V(j)} (Z_{vj} - z_{pj}) \quad (3.26)$$

and

$$\widehat{\text{MSE}}_p = \frac{1}{N_v} \sum_{j=1}^J \sum_{v=1}^{V(j)} (Z_{vj} - z_{pj})^2. \quad (3.27)$$

# 4

## *Characterizing Spatial Processes: The Covariance and Variogram*

### **4.1 INTRODUCTION**

The previous chapter describes several of the common methods of spatial interpolation. Some of them are crude, so that maps made using them display the spatial variation poorly. The interpolators also fail to provide any estimates of the error, which are desirable for prediction. The conventional approach to spatial prediction in soil science combines classical estimation with spatial classification and thereby overcomes some of these weaknesses. It is the only method described in Chapter 3 that gives sound estimates of error. However, it requires replicated sampling for each class to provide individual estimates for that class and some degree of randomization of the sample. The sampling effort can be large, but even with such effort the predicted values at all points within a given class are simply the mean of that class for the property of interest. The precision of prediction is limited by the goodness of the classification; variation within classes is ignored, and local variation is not resolved.

Mathematical functions of the spatial coordinates seemed at one time to have promise. They could be defined fully, and they could therefore be used repeatably. Most were also intuitively reasonable. Some, such as the inverse functions of distance and triangulation, however, were also quite arbitrary, taking no account of more general knowledge of the variation in the region. Trend surface analysis, the only function described in Chapter 3 that does recognize the generality, has other defects.

The methods are deterministic, and to that extent they accord with our understanding that the variation in the environment has physical causes, i.e. is physically determined. However, the environment and its component attributes, such as the soil, result from many physical and biological processes

that interact, some in highly non-linear and chaotic ways. The outcome is so complex that the variation appears to be random. This complexity, together with our current, far from complete, understanding of the processes, means that mathematical functions are not adequate to describe any but the simplest components.

A fully deterministic solution to our problems seems out of reach at present. To make progress we must look at spatial variation differently. Recapitulating, we have two needs: to describe quantitatively how soil varies spatially, and to predict its values at places where we have not sampled. In addition we want estimates of the errors on these predictions so that we can judge what confidence to place in them; estimates of errors are lacking in the classical methods of interpolation. We need a model for prediction, and since there is no deterministic one the solution seems to lie in a probabilistic or stochastic approach.

## **4.2 A STOCHASTIC APPROACH TO SPATIAL VARIATION: THE THEORY OF REGIONALIZED VARIABLES**

### **4.2.1 Random variables**

The fact that spatial variation appears to be random suggests a way forward. Consider throwing a die; on any one throw we obtain a number, for instance, a 6. This is the outcome of throwing the die once, of drawing one value from a distribution that consists of the set  $\{1, 2, 3, 4, 5, 6\}$  with equal probability. One can argue that the result is physically determined in that it depends on the position of the die in the cup and of the cup itself at the start, the forces imparted to it by the thrower, and the nature of the surface on which it lands (Matheron, 1989). Nevertheless, these are so imperfectly known and so far beyond our control that we regard the process as random and as unbiased. Similarly, since the factors that determine the values of environmental variables are numerous, largely unknown in detail, and interact with a complexity that we cannot disentangle, we can regard their outcomes as random.

If we adopt a stochastic view then at each point in space there is not just one value for a property but a whole set of values. We regard the observed value there as one drawn at random according to some law, from some probability distribution. This means that at each point in space there is variation, a concept that has no place in classical estimation. Thus, at a point  $\mathbf{x}$  a property,  $Z(\mathbf{x})$ , is treated as a random variable with a mean,  $\mu$ , a variance,  $\sigma^2$ , and higher-order moments, and a cumulative distribution function (cdf). It has a full probability distribution, and it is from this that the actual value is drawn. If we know approximately what that distribution might be we can estimate values at unrecorded places from data in the neighbourhood and put errors on our estimates.

Most environmental variables, such as the soil's pH and potassium concentration, are continuous. For these a value  $z(\mathbf{x})$  can be thought of as one of an infinite number of possible values, with a cdf that is the probability that  $Z$  takes any value less than or equal to a particular value  $z_c$ :

$$F\{Z(\mathbf{x}; z)\} = \text{Prob}[Z(\mathbf{x}) \leq z_c] \quad \text{for all } z. \quad (4.1)$$

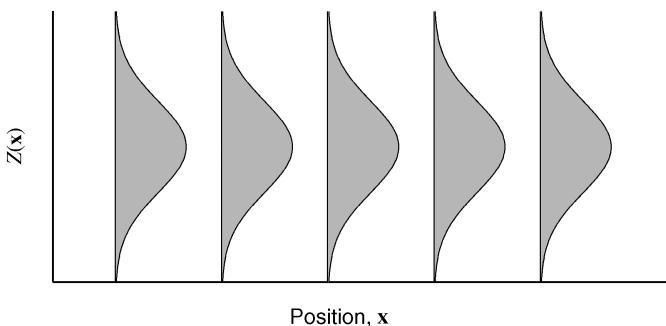
The probability  $F\{Z(\mathbf{x}; z)\}$  takes values between 0 and 1. Its derivative is the probability density function, the pdf:

$$f\{Z(\mathbf{x})\} = \frac{dF\{Z(\mathbf{x}; z)\}}{dz}, \quad (4.2)$$

which we described in Chapter 2. The distribution may be bounded, as in the case of a proportion or percentage, but the most useful assumption is that it is not, so that  $-\infty \leq Z(\mathbf{x}) \leq +\infty$ .

#### 4.2.2 Random functions

The description above for an individual point  $\mathbf{x}$  applies to the infinitely many points in the space; at each point  $\mathbf{x}_i, i = 1, 2, \dots, Z(\mathbf{x}_i)$  has its own distribution and cdf. The range of possible values constitutes an *ensemble*, and one member of the ensemble is the realization. The idea is illustrated in Figure 4.1 in which the curves are imagined to protrude vertically out of the plane of the page. The set of random variables,  $Z(\mathbf{x}_1), Z(\mathbf{x}_2), \dots$ , constitute a *random function*, a *random process*, or a *stochastic process*. The set of actual values of  $Z$  that comprise the realization of the random function is known as a *regionalized variable*. Just as in Chapter 2 we regarded a region as made up of a population of units, so we can think of a random function  $Z(\mathbf{x})$  as a superpopulation, with an infinite number of units in space and an infinite number of values of  $Z$  at each point in the space. It is doubly infinite.



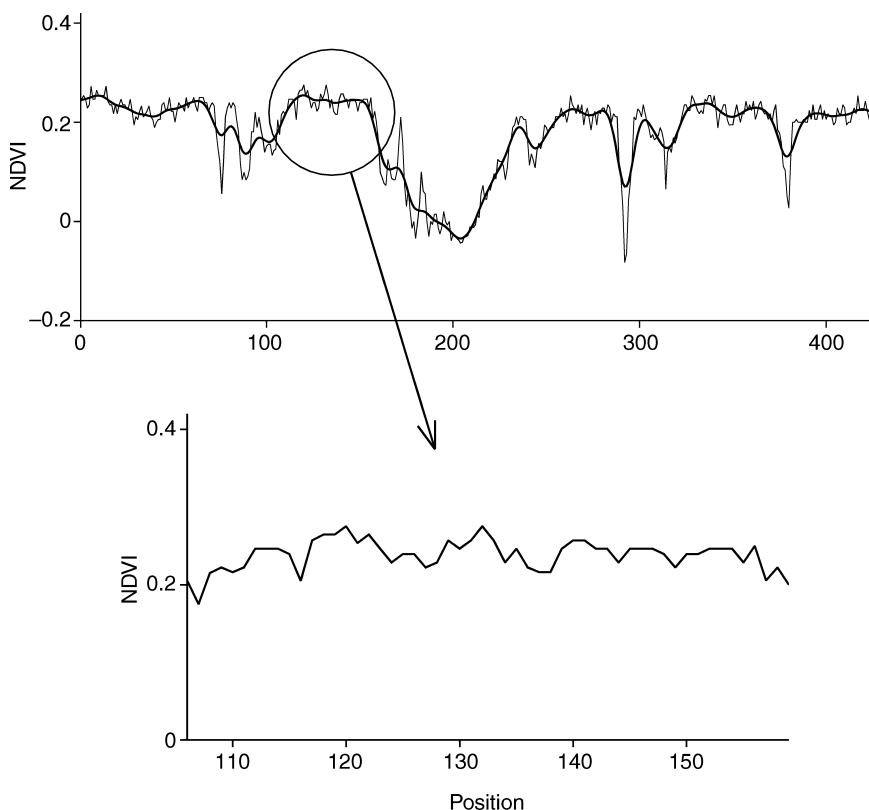
**Figure 4.1** The normal distributions of the random variables at five sites.

### 4.3 SPATIAL COVARIANCE

To define the variation we need to describe the ensemble simply. For the possible outcomes of throwing the die it is easy because they are independent. The values of regionalized variables, on the other hand, tend to be related. In general, values at two places near to one another are similar, whereas those at more widely separated places are less so. This can be seen in Figure 4.2, which represents pixel values for the normalized difference vegetation index (NDVI), where

$$\text{NDVI} = (\text{infrared} - \text{red}) / (\text{infrared} + \text{red}), \quad (4.3)$$

along one row of a SPOT (Système Probatoire de l'Observation de la Terre) image (from Oliver *et al.*, 2000). The fine line joining the pixel values in the



**Figure 4.2** Transect across a SPOT image for normalized difference vegetation index. The fine line in the upper graph joins the data, and the bold line is a smoothing spline fitted through them. The lower graph is an enlarged version of the section in the circle.

upper graph illustrates the locally erratic nature of the variation. Wherever we look we see some fluctuation, but in most short sections of the transect the values are similar. Over longer distances, however, the values vary more substantially, with some sections having small values on average and others where they are large. This becomes clear when the locally erratic variation has been filtered by a smoothing spline, the bold line in the graph. Where the property is continuous, as in this example and as is the case for most properties of the environment, its values must be related at some scale. This is illustrated further in the lower graph where a small section of the transect is magnified and the pixel values are plotted in more detail. What appeared to be entirely erratic in the upper graph can be seen at the larger resolution as structured in the sense that neighbouring values are similar to one another on average. We want to describe these relations, and we do so using the concept of covariance.

We are likely to be familiar with using the covariance to determine the relation between two variables for paired observations. For  $n$  pairs of observations,  $z_{i,1}, z_{i,2}, i = 1, 2, \dots, n$ , of two variables,  $z_1$  and  $z_2$ , the covariance is given by

$$\hat{C}(z_1, z_2) = \frac{1}{n} \sum_{i=1}^n \{z_{i,1} - \bar{z}_1\} \{z_{i,2} - \bar{z}_2\}, \quad (4.4)$$

where  $\bar{z}_1$  and  $\bar{z}_2$  are the means of  $z_1$  and  $z_2$ , respectively. If the units,  $i = 1, 2, \dots, n$ , on which the observations were made were drawn at random then  $\hat{C}(z_1, z_2)$  estimates the population covariance without bias.

We can extend this definition for relating two random variables. The concept and its mathematical expression were developed originally for analysing time series during the 1920s and 1930s, and they have been much used for processing signals and for forecasting. They are now described in many textbooks, of which we can recommend Jenkins and Watts (1968) and Priestley (1981). They have their analogies in space, and Yaglom (1987) presents them in this context as underpinning spatial prediction.

In our new spatial setting  $z_1$  and  $z_2$  become  $Z(\mathbf{x}_1)$  and  $Z(\mathbf{x}_2)$ , i.e. they are the sets of values of the same property,  $Z$ , at the two places  $\mathbf{x}_1$  and  $\mathbf{x}_2$ , and we have switched the notation to capital  $Z$  to signify that they are random variables. Their covariance is

$$C(\mathbf{x}_1, \mathbf{x}_2) = E[\{Z(\mathbf{x}_1) - \mu(\mathbf{x}_1)\}\{Z(\mathbf{x}_2) - \mu(\mathbf{x}_2)\}], \quad (4.5)$$

where  $\mu(\mathbf{x}_1)$  and  $\mu(\mathbf{x}_2)$  are the means of  $Z$  at  $\mathbf{x}_1$  and  $\mathbf{x}_2$ . The equation is analogous to equation (4.4). Unfortunately, however, its solution is unavailable because we have only the one realization of  $Z$  at each point; we cannot know the means. Thus we seem to have reached an impasse, and we can progress only by making further assumptions of *stationarity* which allow us to treat the values at different places as though they are different realizations of the property.

### 4.3.1 Stationarity

By stationarity we mean that the distribution of the random process has certain attributes that are the same everywhere. Starting with the first moment, we assume that the mean,  $\mu = E[Z(\mathbf{x})]$ , about which individual realizations fluctuate, is constant for all  $\mathbf{x}$ . This enables us to replace  $\mu(\mathbf{x}_1)$  and  $\mu(\mathbf{x}_2)$  by the single value  $\mu$ , which we can estimate by repetitive sampling.

We next consider the second moments. Equation (4.5) as written is restricted to the two particular points  $\mathbf{x}_1$  and  $\mathbf{x}_2$ , which is not very useful. We want to generalize it so that it describes the process, and to do so we must make further assumptions. The first concerns what happens when  $\mathbf{x}_1$  and  $\mathbf{x}_2$  coincide. Equation (4.5) then defines the variance,  $\sigma^2 = E[\{Z(\mathbf{x}) - \mu\}^2]$ , sometimes called the *a priori* variance of the process. We assume this to be finite and, like the mean, to be the same everywhere. Second, when  $\mathbf{x}_1$  and  $\mathbf{x}_2$  do not coincide their covariance depends on their separation and not on their absolute positions: this applies to any pair of points  $\mathbf{x}_i, \mathbf{x}_j$  separated by the vector  $\mathbf{h} = \mathbf{x}_i - \mathbf{x}_j$ , so that we have

$$C(\mathbf{x}_i, \mathbf{x}_j) = E[\{Z(\mathbf{x}_i) - \mu\}\{Z(\mathbf{x}_j) - \mu\}], \quad (4.6)$$

which is constant for any given  $\mathbf{h}$ . This constancy of the mean, variance and covariances that depend only on separation and not on absolute positions, i.e. constancy of the first and second moments of the ensemble or process, constitutes *second-order stationarity* or *weak stationarity*. Note that the moments are of the imaginary random process of which we have the one realization and that we can never know their values exactly. We can estimate them, and a general formula for doing so is given below in equation (4.43).

Just as each random function has its cdf, each pair of random functions  $Z(\mathbf{x}_i)$  and  $Z(\mathbf{x}_j)$  will have a joint cdf:

$$F\{Z(\mathbf{x}_i, \mathbf{x}_j; z)\} = \text{Prob}[Z(\mathbf{x}_i) \leq z, Z(\mathbf{x}_j) \leq z] \quad \text{for all } z, \quad (4.7)$$

and a corresponding pdf, the derivative of equation (4.7). Chapter 2 gives the formula for the pdf of a bivariate normal distribution. As an example, if we have a set of points regularly spaced along a line at positions  $x_1, x_2, \dots, x_N$  then we expect the joint cdf  $F\{Z(x_1, x_2; z)\}$  to be the same as  $F\{Z(x_2, x_3; z)\}$ , as  $\dots$ , and as  $F\{Z(x_{N-1}, x_N; z)\}$ . Further, it enables us to obtain a picture of the joint distribution of pairs of points one interval apart by sampling at these positions and plotting their values on a scatter diagram as a representation of the pdf. This is described in greater detail below and illustrated in Figure 4.10. There will be  $N - 1$  pairs one interval apart,  $N - 2$  pairs two intervals apart, and  $N - h$  pairs  $h$  intervals apart. This is illustrated in Figure 4.10(a). In two and

three dimensions the separation is a vector with both distance and direction, which we denote by  $\mathbf{h}$ , and is known as the *lag*.

The joint cdf will have higher-order moments. If these also depend on the separation only then the process is said to be strictly or fully stationary. It is not always wise to assume such strong stationarity, but in practice it might not matter. If the distribution is normal (Gaussian) then the moments of order 3 and more are known constants, and we need not concern ourselves with them. This is another motivation for transforming non-normal data to normality if possible. Therefore, we can usually limit ourselves to nothing more demanding than second-order stationarity and concentrate on the covariance.

### 4.3.2 Ergodicity

Ergodicity is closely related to stationarity. A process is said to be ergodic when the moments of the single observable realization in space approach those of the ensemble as the regional bounds expand towards infinity. It is of mainly theoretical interest rather than of practical value because the regions we study are finite, and we never know the ensemble averages. Nevertheless, we sometimes have to distinguish, especially when choosing estimators.

## 4.4 THE COVARIANCE FUNCTION

We can rewrite equation (4.6) as

$$\begin{aligned}\text{cov}[Z(\mathbf{x}), Z(\mathbf{x} + \mathbf{h})] &= E[\{Z(\mathbf{x}) - \mu\}\{Z(\mathbf{x} + \mathbf{h}) - \mu\}] \\ &= E[\{Z(\mathbf{x})\}\{Z(\mathbf{x} + \mathbf{h})\} - \mu^2] \\ &= C(\mathbf{h}).\end{aligned}\tag{4.8}$$

In words, the covariance is a function of the lag,  $\mathbf{h}$ , and the lag only. It is the *autocovariance function*—*auto* because it represents the covariance of  $Z$  with itself. Unless there is any ambiguity, we shall refer to it simply as the covariance function. It describes the dependence between values of  $Z(\mathbf{x})$  with changing lag. If  $Z(\mathbf{x})$  has a multivariate normal distribution for all positions then the mean and the covariance function completely characterize the process because all of the higher-order moments are constant.

The autocovariance depends on the scale on which  $Z$  is measured, and it is often more convenient and easier to appreciate if we make it dimensionless by converting it to the *autocorrelation*:

$$\rho(\mathbf{h}) = C(\mathbf{h})/C(\mathbf{0}),\tag{4.9}$$

where  $C(\mathbf{0})$  is the covariance at lag  $\mathbf{0}$ , i.e.  $\sigma^2$ .

## 4.5 INTRINSIC VARIATION AND THE VARIOGRAM

We can represent a stationary random process by the model

$$Z(\mathbf{x}) = \mu + \varepsilon(\mathbf{x}). \quad (4.10)$$

This states simply that the value of  $Z$  at  $\mathbf{x}$  is the mean of the process plus a random component drawn from a distribution with mean zero and covariance function

$$C(\mathbf{h}) = E[\varepsilon(\mathbf{x})\varepsilon(\mathbf{x} + \mathbf{h})]. \quad (4.11)$$

Quite the most serious worry and widespread departure from weak stationarity is that the mean appears to change across a region and the variance to increase without bound as the area of interest increases. In these circumstances the covariance cannot be defined. We cannot insert a value for  $\mu$  in equation (4.8), for example.

Matheron (1965) recognized the problem that this created, and his solution was a major contribution to practical geostatistics. He took the view that, whereas in general the mean might not be constant, it would be so for small  $|\mathbf{h}|$  at least, so that the expected differences would be zero:

$$E[Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})] = 0. \quad (4.12)$$

Further, he replaced the covariances by the variances of differences as measures of spatial relation, which, like the covariance, depended on the lag and not on absolute position. This led to

$$\begin{aligned} \text{var}[Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})] &= E[\{Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})\}^2] \\ &= 2\gamma(\mathbf{h}). \end{aligned} \quad (4.13)$$

Equations (4.12) and (4.13) constitute Matheron's *intrinsic hypothesis*. This step released practitioners from the constraints of second-order stationarity where the assumptions either did not hold or were doubtful. It opened up a wider field of application. The quantity  $\gamma(\mathbf{h})$  is known as the *semivariance* at lag  $\mathbf{h}$ . The 'semi' evidently refers to the fact that it is half of a variance; it is half the variance of a difference in this instance. It is, nevertheless, the variance per point when the points are considered in pairs, and it had been recognized by Yates (1948). As a function of  $\mathbf{h}$  it is the *semivariogram*, now usually termed simply the *variogram*.

### 4.5.1 Equivalence with covariance

For second-order stationary processes the variogram and the covariance are equivalent, and from their definitions in equations (4.8) and (4.13) we have

$$\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h}). \quad (4.14)$$

Thus, a graph of the variogram is simply a mirror image of the covariance function about a line or plane parallel to the abscissa. We can also relate the semivariance to the autocorrelation coefficient by combining equations (4.14) and (4.9):

$$\gamma(\mathbf{h}) = \sigma^2 \{1 - \rho(\mathbf{h})\}. \quad (4.15)$$

If a process is intrinsic only there is no equivalence because the covariance function does not exist. The variogram is still valid, nevertheless, and it is its validity in the wider range of circumstances that has made it so much more useful than the covariance. As a consequence its has become the cornerstone of practical geostatistics. For this reason we look at its properties in detail both in the remainder of this chapter and in the following two.

#### 4.5.2 Quasi-stationarity

In practice it often happens that the variogram is of interest only very locally—we shall see this later when we deal with kriging. In these circumstances we can restrict the mean,  $\mu$ , to that in small neighbourhoods,  $V$ , so that equation (4.10) becomes

$$Z(\mathbf{x}) = \mu_V + \varepsilon(\mathbf{x}). \quad (4.16)$$

Provided  $\mathbf{h}$  remains within the bounds of  $V$  the variogram is unaffected.

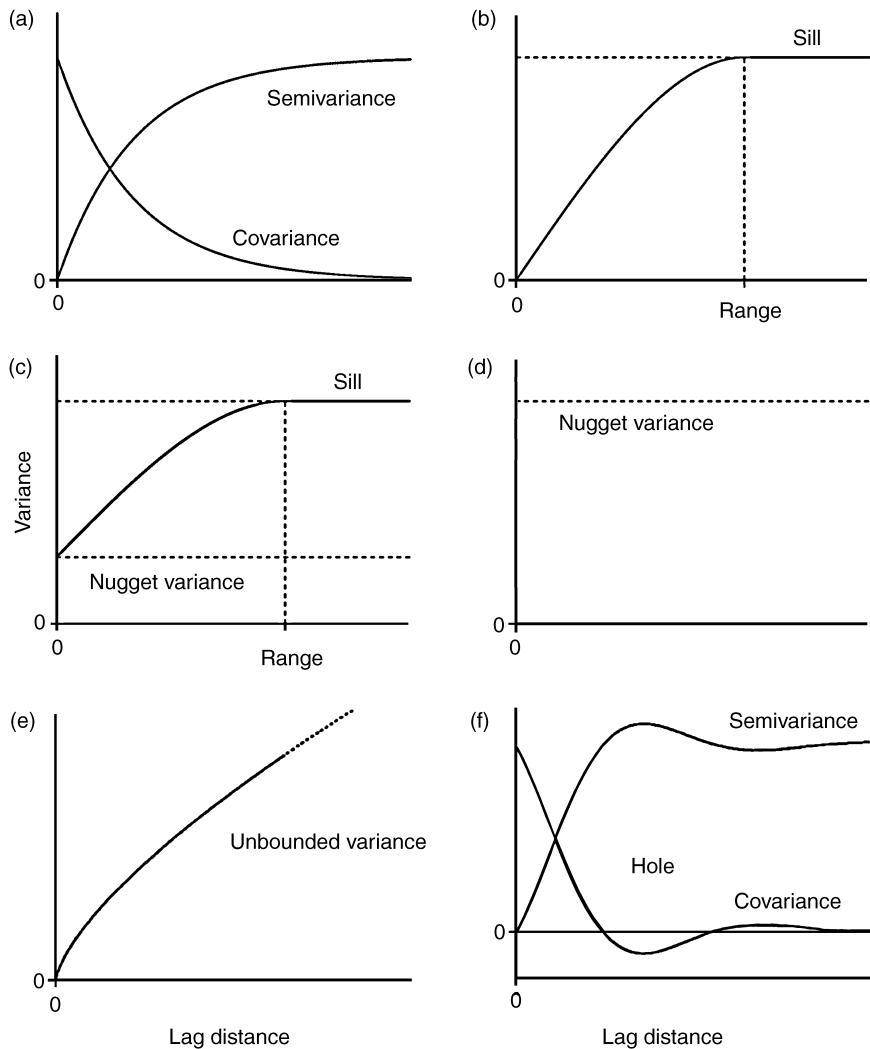
### 4.6 CHARACTERISTICS OF THE SPATIAL CORRELATION FUNCTIONS

We now consider the more important characteristics of the covariance and autocorrelation functions and the variogram. Figure 4.3 illustrates some of these.

*Autocorrelation.* Like the ordinary product–moment correlation coefficient, the autocorrelation function varies between 1 and  $-1$ . From equation (4.9), its value at lag  $\mathbf{0}$  is 1.

*Symmetry.* Because of our assumption of stationarity,

$$\begin{aligned} C(\mathbf{h}) &= E[\{Z(\mathbf{x}) - \mu\}\{Z(\mathbf{x} + \mathbf{h}) - \mu\}] \\ &= E[Z(\mathbf{x} - \mathbf{h})Z(\mathbf{x}) - \mu^2] \\ &= E[Z(\mathbf{x})Z(\mathbf{x} - \mathbf{h}) - \mu^2] \\ &= C(-\mathbf{h}). \end{aligned} \quad (4.17)$$



**Figure 4.3** Theoretical functions for spatial correlation: (a) typical variogram and equivalent covariance function; (b) bounded variogram showing the sill and range; (c) bounded variogram with a nugget variance; (d) pure nugget variogram; (e) unbounded variogram; (f) variogram and covariance function illustrating the hole effect.

In words, the autocovariance is symmetric in space. The same is true of the variogram; i.e.  $\gamma(\mathbf{h}) = \gamma(-\mathbf{h})$  for all  $\mathbf{h}$ . So all three functions are even. This means that we need consider only the positive lags, and indeed this is the convention. In the graphs of the functions, such as those in Figure 4.3, we show only the right-hand halves of the functions.

*Positive semidefiniteness.* The covariance matrix for any number of points is positive semidefinite. That is to say that for a matrix of order  $n$  its determinant

$$\begin{vmatrix} C(\mathbf{x}_1, \mathbf{x}_1) & C(\mathbf{x}_1, \mathbf{x}_2) & \cdots & C(\mathbf{x}_1, \mathbf{x}_n) \\ C(\mathbf{x}_2, \mathbf{x}_1) & C(\mathbf{x}_2, \mathbf{x}_2) & \cdots & C(\mathbf{x}_2, \mathbf{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ C(\mathbf{x}_n, \mathbf{x}_1) & C(\mathbf{x}_n, \mathbf{x}_2) & \cdots & C(\mathbf{x}_n, \mathbf{x}_n) \end{vmatrix}$$

and all its principal minors are positive or zero. This is necessary because the variance of any linear sum of the random variables,

$$Y(\mathbf{x}) = \lambda_1 Z(\mathbf{x}_1) + \lambda_2 Z(\mathbf{x}_2) + \cdots + \lambda_n Z(\mathbf{x}_n), \quad (4.18)$$

must be positive or zero; a variance cannot be negative. The covariance and autocorrelation functions are positive semidefinite. In like manner, the variogram must be negative semidefinite. We shall develop this in Chapter 5 where we shall see that this limits the choice of legitimate mathematical functions to describe the covariance function.

*Continuity.* As mentioned above, most environmental variables are continuous; the stochastic processes that we believe to represent them are continuous, and so also are the autocovariance functions and variograms of a continuous lag. Crucially,  $C(\mathbf{h})$  and  $\gamma(\mathbf{h})$  are continuous at  $\mathbf{h} = \mathbf{0}$ , and if that is so they must be continuous everywhere. So  $C(\mathbf{h})$  declines from some positive value,  $C(\mathbf{0}) = \sigma^2$ , at  $\mathbf{0}$  to smaller values at longer lag distances; see Figure 4.3(a). Its mirror image, the variogram, increases from 0 at  $\mathbf{h} = \mathbf{0}$ , i.e. it must pass through the origin if the process is continuous; see Figure 4.3(a)–(b).

If this were not so then we should have a continuous sequence of positions in space, the values at which are not related. It seems impossible, yet in practice data often suggest that a spatial process is discontinuous. It manifests itself most evidently in the sample variogram; the calculated values appear to approach some positive value on the ordinate as the lag distance approaches 0, whereas, at  $\mathbf{h} = \mathbf{0}$ ,  $\gamma(\mathbf{0})$  must be 0; Figure 4.3(c). This discrepancy is known as the *nugget variance*. The term arose in gold mining from the notion that gold nuggets occur quite independently of one another at random; they are sparse and certainly not continuous at the working scale. They have a variance that jumps from 0 at lag zero to positive immediately away from the origin, and we can recognize this by defining

$$\gamma(\mathbf{h}) = \sigma^2 \{1 - \delta(\mathbf{h})\}, \quad (4.19)$$

where  $\delta(\mathbf{h})$  is the Kronecker delta function taking the values 1 when  $\mathbf{h} = \mathbf{0}$  and 0 otherwise.

The data themselves differ from their neighbours in irregular steps, large or small, rather than in smooth progression. It seems as though they derive from two or more components, one uncorrelated superimposed on another that is correlated. In other words, we seem to have one source of variation in which contiguous positions in space do take values of  $Z$  that are totally unrelated.

Engineers recognize this uncorrelated variation as ‘white noise’. They usually express it by its covariance function:

$$C(\mathbf{h}) = \sigma^2 \delta(\mathbf{h}), \quad (4.20)$$

where now  $\delta(\mathbf{h})$  is the Dirac function taking the values 0 when  $|\mathbf{h}| \neq 0$  and infinity when  $|\mathbf{h}| = 0$ . Thus for white noise  $C(\mathbf{h}) = 0$  for all  $|\mathbf{h}| > 0$  and  $C(\mathbf{0}) = \infty$ . The representation might seem bizarre, but it is the only way that we can describe white noise using covariances. Its equivalent is a ‘pure nugget’ variogram; Figure 4.3(d).

For properties that vary continuously in space, such as the soil’s pH, the concentrations of trace metals, air temperature and rainfall, the apparent nugget variance comprises measurement error plus variation that occurs over distances less than the shortest sampling interval. The latter is usually dominant.

*Monotonic increasing.* The variograms in Figure 4.3(b)–(c) are monotonically increasing functions, i.e. the variance increases with increasing lag distance. The small values of  $\gamma(\mathbf{h})$  at short  $|\mathbf{h}|$  show that the  $Z(\mathbf{x})$  are similar, and that as  $|\mathbf{h}|$  increases  $Z(\mathbf{x})$  and  $Z(\mathbf{x} + \mathbf{h})$  become increasingly dissimilar on average. Looked at from the point of view of correlation,  $\rho(\mathbf{h})$  increases as the lag distance shortens, and the process is therefore said to be *autocorrelated* or *spatially dependent*.

*Sill and range.* The variograms of second-order stationary processes reach upper bounds at which they remain after their initial increases, as in Figure 4.3(b)–(c). The maximum is known as the *sill* variance; it is the *a priori* variance,  $\sigma^2$ , of the process.

A variogram may reach its sill at a finite lag distance, in which case it has a *range*, also known as the *correlation range* since this is the range at which the autocorrelation becomes 0; Figure 4.3(c). This separation marks the limit of spatial dependence. Places further apart than this are spatially independent. Some variograms approach their sills asymptotically, and so they have no strict ranges. For practical purposes their effective ranges are usually taken as the lag distances at which they reach 0.95 of their sills.

*Unbounded variogram.* If, as in Figure 4.3(e), the variogram increases indefinitely with increasing lag distance then the process is not second-order stationary. It might be intrinsic, but the covariance does not exist.

*Hole effect.* In some instances the variogram decreases from its maximum to a local minimum and then increases again, as in Figure 4.3(f). This maximum is

equivalent to a minimum in the covariance function, which appears as a ‘hole’. This form arises from fairly regular repetition in the process. A variogram that continues to fluctuate with a wave-like form with increasing lag distance signifies greater regularity.

*Anisotropy.* Spatial variation is not necessarily the same in all directions. If the process is anisotropic then so is the variogram, as is the covariance function if it exists. Anisotropy may take several forms. The initial gradient may vary. If the variogram has a sill then variation in the gradient will lead to variation in the range, or effective range. If the variation with changing direction is such that a simple transformation of the spatial coordinates will remove it then we have *geometric anisotropy* (see Chapter 5).

A region may contain preferentially oriented zones with different mean values. In these circumstances the variance encountered changes with change in direction so that the sill fluctuates. This is called *zonal anisotropy*.

*Trend.* In some instances the experimental variograms (see below for their definition) follow smooth curves that approach the origin with decreasing gradient: the curves have concave upwards forms. This shape can arise from local *trend* or *drift*, i.e. smooth change in the underlying variable. The dashed line in Figure 5.3 is an example. In other instances the experimental estimates increase sharply after having appeared to reach sills, as in Figure 9.2(a); this is often a sign of long-range trend in the variation superimposed on relatively short-range random variation. In both circumstances the expected value,  $E[Z(\mathbf{x})]$ , is not constant, even within small neighbourhoods, but is a function of position. We have then to elaborate our model for spatial variation to

$$Z(\mathbf{x}) = u(\mathbf{x}) + \varepsilon(\mathbf{x}). \quad (4.21)$$

The quantity  $u(\mathbf{x})$  is the local trend, and it replaces the means in equations (4.10) and (4.16). The assumption of second-order stationarity does not hold, nor does the intrinsic hypothesis. The experimental semivariances calculated from the raw data no longer estimate the expected squared differences between the residuals at two places. The residuals are given by

$$\varepsilon(\mathbf{x}) = Z(\mathbf{x}) - u(\mathbf{x}). \quad (4.22)$$

They constitute the random process with its associated variogram,

$$\gamma(\mathbf{h}) = \frac{1}{2} E \left[ \{ \varepsilon(\mathbf{x}) - \varepsilon(\mathbf{x} + \mathbf{h}) \}^2 \right]. \quad (4.23)$$

A more general description of non-stationarity is as an *intrinsic random function of order k* (IRFk):

$$Z(\mathbf{x}) = Z_k(\mathbf{x}) + u(\mathbf{x}). \quad (4.24)$$

We describe some ways of dealing with the difficulties of non-stationarity in Chapter 9.

## 4.7 WHICH VARIOGRAM?

The variogram (and covariance function) as treated above is a function of an underlying stochastic process. We may regard it as the *theoretical variogram*. It may be thought of as the average of the variograms from all possible realizations of the process. Following Matheron (1965), we need to distinguish it from two others, namely the regional and the experimental.

The *regional variogram* is the variogram of the particular realization in a finite region,  $R$ . It is the one that you might compute if you had complete information of the region, as, for example, from the simulated fields in Figures 5.5, 5.6, 5.8 and 5.11, and from many digital images (see Muñoz-Pardo, 1987). The regional variogram does not necessarily represent the whole ensemble. A process that is second-order stationary might appear unbounded in a small region, especially if the distance across the region is smaller than the correlation range. The regional variogram is called the non-ergodic variogram by some workers, e.g. Brus and de Gruijter (1994), for this reason. It is more or less accessible, depending on the effort we are prepared to devote to sampling the realization, and this leads us to the third variogram, below.

The *experimental variogram* is computed from data,  $z(\mathbf{x}_i), i = 1, 2, \dots$ , which constitute a sample from the region. It is also called the *sample variogram*. We describe it in Section 4.9. It necessarily applies to an actual realization, and it estimates the regional variogram for that realization. It is usually the only variogram that we know, and any inference from it requires modelling, as described in Chapter 5.

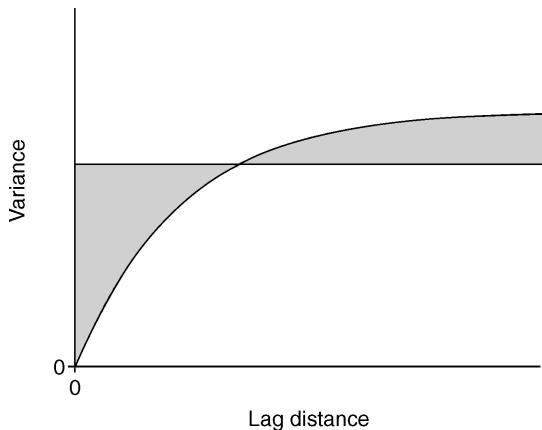
## 4.8 SUPPORT AND KRIGE'S RELATION

Spatial dependence within a finite region has both theoretical and practical consequences, which we now explore.

The variance of  $Z(\mathbf{x})$  within a region  $R$  of area  $|R|$  is the double integral of the variogram:

$$\sigma_R^2 = \bar{\gamma}(R, R) = \frac{1}{|R|^2} \int_R \int_R \gamma(\mathbf{x} - \mathbf{x}') d\mathbf{x} d\mathbf{x}', \quad (4.25)$$

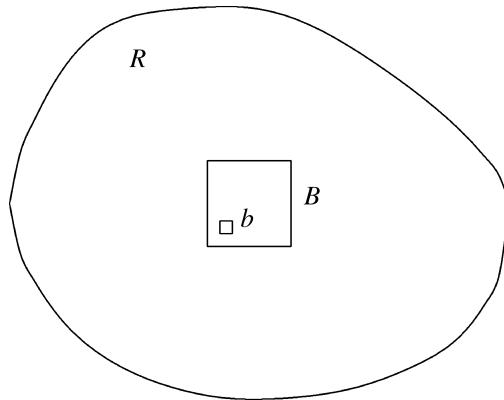
where  $\mathbf{x}$  and  $\mathbf{x}'$  sweep independently over  $R$ . In geostatistics this variance is called the *dispersion variance* of  $Z(\mathbf{x})$  in  $R$ . Unless the variogram is all nugget the dispersion variance for a finite  $R$  is less than the *a priori* variance of the process,



**Figure 4.4** Relation between the variogram and the dispersion variance in a finite region,  $R$ .

if it is second-order stationarity. Figure 4.4 shows the relation between the two for a one-dimensional process; in it the shaded areas are equal. Evidently, as  $R$  is made smaller  $\sigma_R^2$  diminishes, until in the limit we are left with a point, at which  $\sigma_R^2$  disappears.

The region  $R$  (see Figure 4.5) limits the extent of a realization. At the small end of our spatial scale we encounter another limit. Measurements must be made on finite volumes, whether of samples taken into the laboratory or the surroundings of instruments placed in the field. The volume, with its particular size, shape and orientation, is known as the *support* of the sample. The supports have finite cross-sectional areas in  $\mathbb{R}^2$ , and they are effectively small but finite regions, each with its own dispersion variance. If we denote them by  $b$ , each



**Figure 4.5** Krige's relation for a region,  $R$ , a block,  $B$ , and a small support,  $b$ .

with area  $|b|$  (see Figure 4.5), then their dispersion variances are given by the analogue of equation (4.25):

$$\sigma_b^2 = \bar{\gamma}(b, b) = \frac{1}{|b|^2} \int_b \int_b \gamma(\mathbf{x} - \mathbf{x}') d\mathbf{x} d\mathbf{x}'. \quad (4.26)$$

One practical consequence of this is that the support of the sample sets a minimum to the resolution of the spatial variation that can be detected and measured by that sample: engineers will understand this as ‘band-limited’ measurement.

In many applications we are interested in blocks,  $B$ , of intermediate size,  $|B|$  (see Figure 4.5). They may be mining blocks, plots in an experiment, or fields on a farm, as examples. They too will have dispersion variances,  $\sigma_B^2$ , defined in a way analogous to  $\sigma_R^2$  and  $\sigma_b^2$ , and with intermediate values. We now relate the three.

Consider first the supports  $b$ . Though small, they have finite size, and so in a finite region they are finite in number if they do not overlap. If there are  $n_R^b$  of them with values  $z_i^b, i = 1, 2, \dots, n_R^b$ , then their variance in  $R$  is

$$s^2(b \in R) = \frac{1}{n_R^b} \sum_{i=1}^{n_R^b} \{\bar{z}_R - z_i^b\}^2, \quad (4.27)$$

where  $\bar{z}_R$  is the mean of the  $z_i^b$ . In like manner their variance in a block  $B$  with mean  $\bar{z}_B$  is

$$s^2(b \in B) = \frac{1}{n_B^b} \sum_{i=1}^{n_B^b} \{\bar{z}_B - z_i^b\}^2, \quad (4.28)$$

which can be averaged over all  $B \in R$  to give  $\bar{s}^2(b \in B)$ . Finally, we consider the blocks,  $B$ , themselves. Their variance in  $R$  is

$$s^2(B \in R) = \frac{1}{n_R^B} \sum_{j=1}^{n_R^B} \{\bar{z}_R - \bar{z}_j^B\}^2, \quad (4.29)$$

where  $\bar{z}_j^B$  is the mean of  $Z$  in the  $j$ th block.

For any finite region that is divided in the above way into blocks, which in turn are further subdivided, whether into small supports or smaller blocks, the dispersion variance is partitioned quite simply as

$$s^2(b \in R) = \bar{s}^2(b \in B) + s^2(B \in R). \quad (4.30)$$

In words, the dispersion variance of  $Z$  of supports  $b$  in region  $R$  is the sum of the variance of the supports with blocks  $B$  plus the variance of the blocks within  $R$ . This is *Krige’s relation*. It is strictly analogous to the partition of the total

variance into within and between classes in the simple one-way analysis of variance.

The expectations of the dispersion variances are all readily obtained from the variogram by

$$\begin{aligned}\sigma^2(b \in R) &= \bar{\gamma}(R, R) - \bar{\gamma}(b, b), \\ \sigma^2(B \in R) &= \bar{\gamma}(R, R) - \bar{\gamma}(B, B), \\ \sigma^2(b \in B) &= \bar{\gamma}(B, B) - \bar{\gamma}(b, b),\end{aligned}\quad (4.31)$$

and so Krige's relation applies to them equally.

#### 4.8.1 Regularization

Another consequence of the finite sample support is that the variogram in practice is a function of the support. The larger the support is the more variation each measurement encompasses, and the less there is in the intervening space to appear in the variogram. This inevitably diminishes the sill or gradient and tends to make the variogram concave upwards near to the origin. It is a physical regularization, the statistical aspects of which we describe below. Results should always refer specifically to the particular support, which should therefore remain the same throughout any one investigation.

The variogram on one support can be related, at least theoretically, to that on another. The semivariance for two supports  $b(\mathbf{x})$  and  $b(\mathbf{x} + \mathbf{h})$ , the centroids of which are  $\mathbf{h}$  apart, is

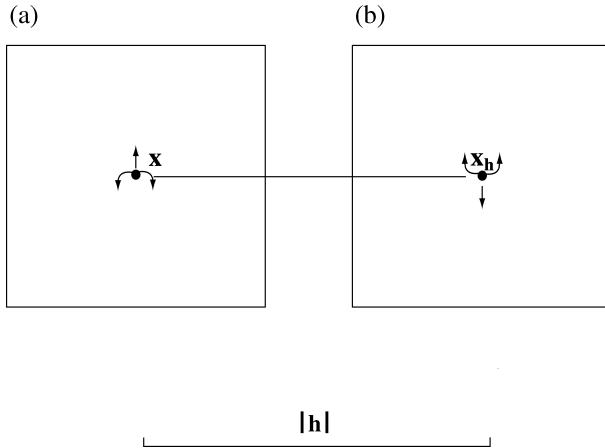
$$\gamma_b(\mathbf{h}) = E[\{Z_b(\mathbf{x}) - Z_b(\mathbf{x} + \mathbf{h})\}^2], \quad (4.32)$$

where  $Z_b(\mathbf{x})$  and  $Z_b(\mathbf{x} + \mathbf{h})$  are the integrals of  $Z(\mathbf{x})$  over the supports  $b$ . This is composed of two parts, the average squared difference between the points in one support and those in the other less the dispersion variance within supports. The first is given by

$$\bar{\gamma}(b, b_{\mathbf{h}}) = \frac{1}{|b|^2} \int_b \int_b \gamma(\mathbf{x} - \mathbf{x}_{\mathbf{h}}) d\mathbf{x} d\mathbf{x}_{\mathbf{h}}, \quad (4.33)$$

where  $\mathbf{x}$  sweeps one support and  $\mathbf{x}_{\mathbf{h}}$  sweeps the other independently, as in Figure 4.6. The second is the integral of the variogram within the support  $b$ :

$$\bar{\gamma}(b, b) = \frac{1}{|b|^2} \int_b \int_b \gamma(\mathbf{x} - \mathbf{x}') d\mathbf{x} d\mathbf{x}', \quad (4.34)$$



**Figure 4.6** The block-to-block integration of the variogram.

where  $\mathbf{x}$  and  $\mathbf{x}'$  sweep  $b$  independently, illustrated in Chapter 8 (Figure 8.1). The variogram on the new supports thus becomes

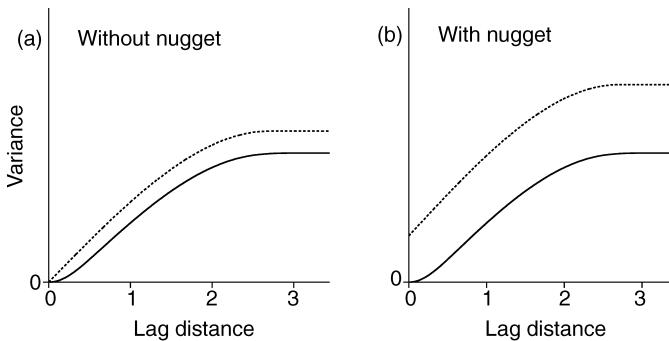
$$\gamma_b(\mathbf{h}) = \bar{\gamma}(b, b_{\mathbf{h}}) - \bar{\gamma}(b, b). \quad (4.35)$$

If  $|\mathbf{h}|$  is large relative to the distances across the support then  $\bar{\gamma}(b, b_{\mathbf{h}})$  is approximately the punctual semivariance at lag  $\mathbf{h}$ , and

$$\gamma_b(\mathbf{h}) \approx \gamma(\mathbf{h}) - \bar{\gamma}(b, b). \quad (4.36)$$

So when  $|\mathbf{h}| \gg \sqrt{\text{area of } b}$  the regularized variogram is derived from the punctual one simply by subtraction of the dispersion variance of the support.

This procedure in which the variogram for one support is obtained from that of a smaller support is known as *regularization*. Figure 4.7 shows what can happen to the variogram. In this figure two punctual variograms appear as the dashed lines, (a) without a nugget component and (b) with one. The solid lines are variograms derived by regularization with blocks of size  $0.5 \times 0.5$  units. Notice the sills are diminished, the nugget variance disappears and the approach of the variogram at the origin is somewhat concave upwards. It is especially important when bulking samples, for two reasons. The first is that the supports can be large. Second, if the variogram is known for very small supports on which the variable has been measured then that for samples bulked over larger areas, the regularized variogram, can be determined from it and surveys be planned with greater efficiency.



**Figure 4.7** Regularization of punctual variograms (dashed) to ones with block supports of  $0.5 \times 0.5$  (solid lines): (a) without a nugget component; (b) with a nugget component.

## 4.9 ESTIMATING SEMIVARIANCES AND COVARIANCES

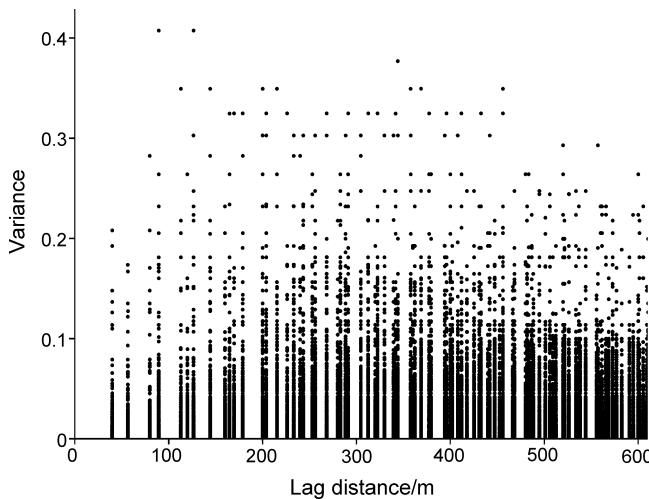
As mentioned above, the variogram is the cornerstone of geostatistics, and it is therefore vital to estimate, interpret and model it correctly. This section concerns its estimation using the usual computing equation, Matheron's method-of-moments estimator, and how it applies to various kinds of sampling. We also describe how to determine the possible presence of anisotropy and non-stationarity in the process of interest.

### 4.9.1 The variogram cloud

For any set of data we can compute the variances for every pair of points,  $\mathbf{x}_i$  and  $\mathbf{x}_j$ , as

$$\gamma(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{2} \{z(\mathbf{x}_i) - z(\mathbf{x}_j)\}^2. \quad (4.37)$$

These values can then be plotted against the lag distance as a scatter diagram, called the 'variogram cloud' by Chauvet (1982). Figure 4.8 shows the variogram cloud for  $\log_{10}K$  at Broom's Barn to a lag of 600 m. It contains all of the information on the spatial relations in the data to that lag. In principle we could fit a model to it to represent the regional variogram, but in practice it is almost impossible to judge from it if there is any spatial correlation present, what form it might have, and how we could model it. A more practicable approach is to average the variances for each of a few lags and then examine the result. Nevertheless, the variogram cloud shows the spread of values at each lag, and it might enable us to detect outliers or anomalies. The tighter this distribution is the stronger is the spatial continuity in the data.



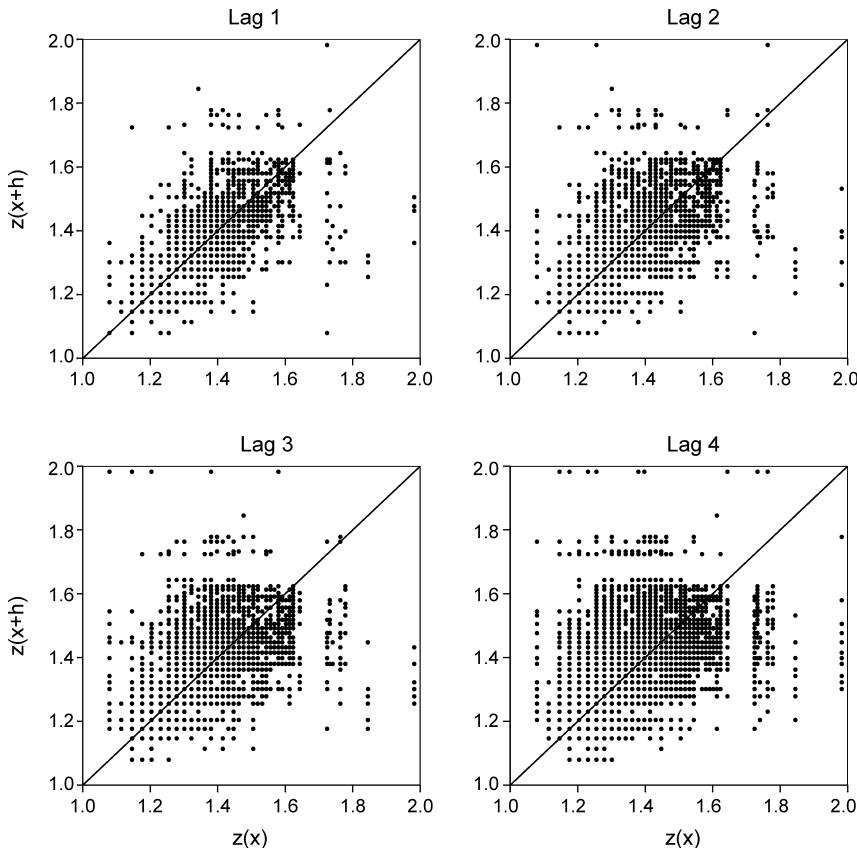
**Figure 4.8** The variogram cloud of  $\log_{10}K$  at Broom's Barn Farm.

#### 4.9.2 h-Scattergrams

The **h**-scattergram of  $z(\mathbf{x})$  plotted against  $z(\mathbf{x} + \mathbf{h})$  for each lag interval shows the joint distribution of pairs of points that interval apart as mentioned above (section 4.3.1), which represents the pdf. The closer the points lie to the diagonal line with gradient 1, the stronger is the correlation,  $\hat{\rho}(\mathbf{h})$ , and the smaller is the semivariance,  $\hat{\gamma}(\mathbf{h})$ . Figure 4.9 shows the **h**-scattergrams for four lag intervals, 40 m, 80 m, 120 m and 160 m, computed from the data on  $\log_{10}K$  on Broom's Barn Farm. The autocorrelation coefficients and semivariances listed in Table 4.1 describe quantitatively what happens as the lag interval increases; the correlations between pairs of points decrease and the semivariances increase.

**Table 4.1** Autocorrelation coefficients and semivariances for  $\log_{10}K$  at Broom's Barn Farm computed for lag distances 40 m (lag 1), 80 m (lag 2), 120 m (lag 3) and 1600 m (lag 4).

Lag distance/m	Autocorrelation coefficient	Semivariance
40	0.590	0.00726
80	0.470	0.00942
120	0.399	0.01065
160	0.311	0.01228



**Figure 4.9** The  $\mathbf{h}$ -scattergrams for four lags computed from the data of  $\log_{10} K$  at Broom's Barn Farm.

### 4.9.3 Average semivariances

If we recall the definition of the semivariance from equation (4.13) as

$$\gamma(\mathbf{h}) = \frac{1}{2} E[\{Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})\}^2] \quad (4.38)$$

then its estimator is

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2} \text{mean} [\{z(\mathbf{x}) - z(\mathbf{x} + \mathbf{h})\}^2], \quad (4.39)$$

where the  $z(\mathbf{x})$  and  $z(\mathbf{x} + \mathbf{h})$  represent actual values of  $Z$  at places separated by  $\mathbf{h}$ . For a set of data  $z(\mathbf{x}_i), i = 1, 2, \dots$ , we can compute

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2m(\mathbf{h})} \sum_{i=1}^{m(\mathbf{h})} \{z(\mathbf{x}_i) - z(\mathbf{x}_i + \mathbf{h})\}^2, \quad (4.40)$$

where  $m(\mathbf{h})$  is the number of pairs of data points separated by the particular lag vector  $\mathbf{h}$ . By changing  $\mathbf{h}$  we obtain an ordered set of semivariances, which constitute the *experimental variogram* or *sample variogram*. Equation (4.40) is the usual formula for computing semivariances; it is commonly known as Matheron's method-of-moments estimator. The way that it is implemented as an algorithm depends on the configuration of the data, and we consider the possibilities below.

### **Regular sampling in one dimension**

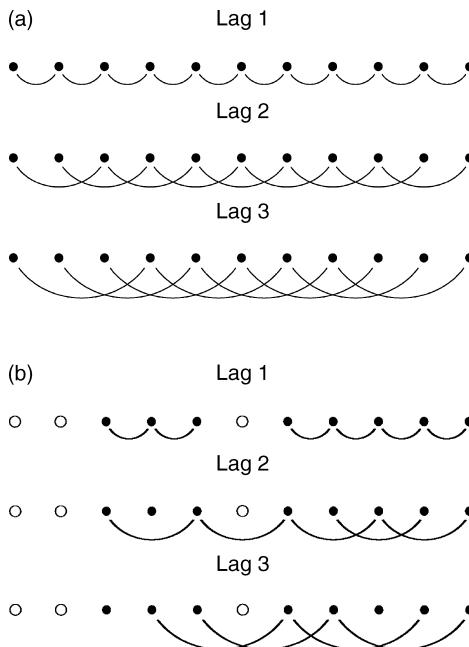
For regular sampling in one dimension along transects and down boreholes we can denote the data by  $z_i = z(\mathbf{x}_i), i = 1, 2, \dots, N$ . The lag becomes a scalar,  $h = |\mathbf{h}|$ , for which  $\hat{\gamma}$  can be computed only at integral multiples of the sampling interval. The semivariance is then computed as

$$\hat{\gamma}(h) = \frac{1}{2(N-h)} \sum_{i=1}^{N-h} \{z_i - z_{i+h}\}^2. \quad (4.41)$$

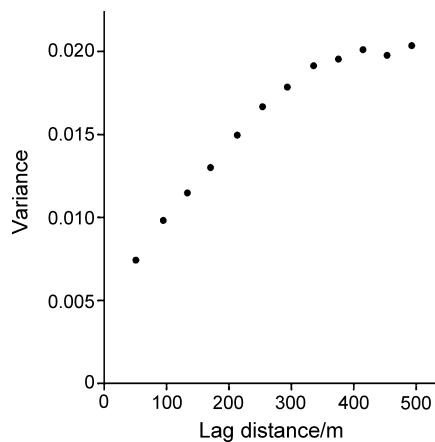
Figure 4.10(a) shows the situation. First, the squared differences between neighbouring pairs of values,  $z_1$  and  $z_2$ ,  $z_2$  and  $z_3$ , and so on, i.e. for  $h = 1$ , are determined for each position and averaged. All of the observations at lag interval  $h$  are used twice except for those at the ends of the transect, and so there are  $N - 1$  comparisons. If there are missing values at some locations, as in Figure 4.10(b), then there will be fewer comparisons, and the divisor is diminished accordingly. By increasing  $h$  to 2 the comparisons are then  $z_1$  with  $z_3$ ,  $z_2$  with  $z_4$ , etc., and we can repeat the procedure for  $h = 3, 4, \dots$ . The result is a set of semivariances  $\hat{\gamma}(1), \hat{\gamma}(2), \hat{\gamma}(3), \dots$  that is ordered as a function of  $h$ . It is a one-dimensional experimental variogram, and we can plot  $\hat{\gamma}(h)$  against  $h$  as in Figure 4.11.

### **Irregular sampling in one dimension**

If data are irregularly scattered then the average semivariance for any particular lag can be derived only by grouping the individual lag distances between pairs of points into 'bins' as in a histogram. Otherwise we have individual semivariances as in the variogram cloud. Typically the averaging is done by choosing a set of lags,  $h_j, j = 1, 2, \dots$ , at arbitrary constant increments  $d$ , and then associating with each  $h_j$  a bin of width  $d$ , bounded by  $h_j - d/2$  and  $h_j + d/2$ . Each pair of points separated by  $h_j \pm d/2$  is used to estimate  $\gamma(h_j)$ . In this way each comparison contributes to one and only one



**Figure 4.10** Comparisons for computing a variogram from regular sampling on a transect: (a) with a complete set of data, indicated with •; (b) with missing values, indicated by ○.



**Figure 4.11** Sample variogram of  $\log_{10}K$  at Broom's Barn, obtained by discretization of the lags into bins as in Figure 4.13.

estimate. Sometimes there are more comparisons at the shorter lags, especially where sampling has been nested (see Chapter 6), and then it can be advantageous to increase the increments, and with them  $d$ , as  $h$  increases.

The lag increments can affect the resulting variogram, and so  $d$  should be chosen with care. If the increment is small then there might be too few comparisons at each lag, leading to semivariances that are estimated crudely and an experimental variogram that appears erratic. If, on the other hand,  $d$  is large then there are likely to be few estimates and detail is lost by unnecessary smoothing. The best compromise will depend on the number of data, the evenness of the sampling and the form of the underlying variogram. A useful starting point is to use the average separation between nearest neighbours as the interval.

### **Sampling on transects to represent variation in two dimensions**

Sometimes investigators sample regularly along transects to explore variation in two dimensions and, in particular, to identify and estimate anisotropy, i.e. directional differences. The computational procedure is the same as for the regular one-dimensional sampling, and equation (4.41) produces a separate set of estimates for each transect. These need to be seen together as a whole and not as separate variograms. The variogram in two dimensions is itself two-dimensional, and the ordered sets of semivariances computed from transects are effectively samples of sections through the two-dimensional function. To identify and estimate anisotropy, transects must be aligned in at least three directions. If the directional variogram appears to have markedly different gradients or ranges in the different directions then it is likely that the underlying variation is anisotropic, and it should be modelled accordingly (see Chapter 5). If the variation seems isotropic, i.e. if there are no directional differences, then the separate estimates can be averaged over all directions to give the isotropic variogram where the vector  $\mathbf{h}$  can be replaced by the scalar  $h = |\mathbf{h}|$ .

### **Regular sampling in two dimensions**

For data recorded at regular intervals on a rectangular grid the above formula (4.41), for one dimension, is readily extended. If the grid has  $m$  rows and  $n$  columns then we compute

$$\begin{aligned}\hat{\gamma}(p, q) &= \frac{1}{2(m-p)(n-q)} \sum_{i=1}^{m-p} \sum_{j=1}^{n-q} \{z(i, j) - z(i+p, j+q)\}^2, \\ \hat{\gamma}(p, -q) &= \frac{1}{2(m-p)(n-q)} \sum_{i=1}^{m-p} \sum_{j=q+1}^{n-q} \{z(i, j) - z(i+p, j-q)\}^2,\end{aligned}\quad (4.42)$$

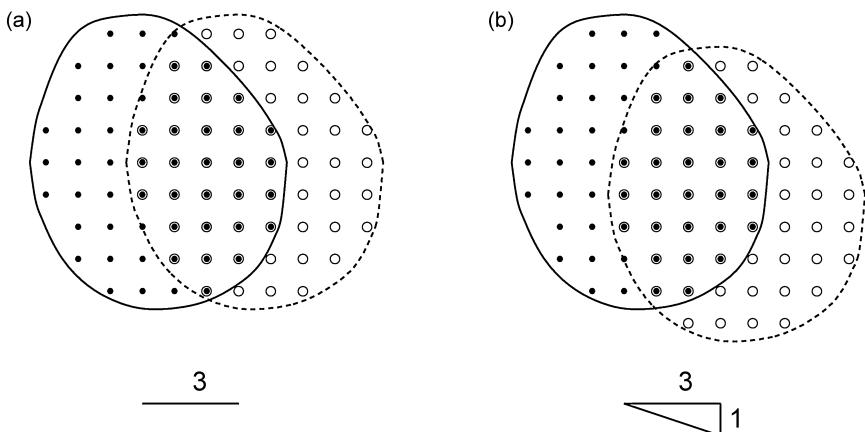
where  $p$  and  $q$  are the lags along the rows and down the columns of the grid, respectively. In general, the lag increment is simply the grid interval. These equations enable half the variogram to be computed for lags from  $-q$  to  $q$  and from 0 to  $p$ . The variogram is symmetrical about its centre, and the full set of semivariances is obtained by computing

$$\begin{aligned}\hat{\gamma}(-p, q) &= \hat{\gamma}(p, -q), \\ \hat{\gamma}(-p, -q) &= \hat{\gamma}(p, q).\end{aligned}$$

The procedure can be envisaged as moving the grid over itself to the right and up or down to new positions, as in Figure 4.12, and making the comparisons between the values at the points that coincide. In Figure 4.12(a) the grid has been moved to the right by three units, i.e.  $p = 3$  and  $q = 0$ , as represented by the horizontal line. In Figure 4.12(b) the grid has been moved down one unit in addition, so that now  $q = -1$ ; the horizontal and vertical shifts are shown in the triangle, with its hypotenuse showing the resultant.

Figure 4.12 also shows that as  $p$  and  $q$  are increased so the number of coincident points diminishes rapidly from the original 55. As a consequence the semivariances become less and less well estimated, a matter to which we return in Chapter 6.

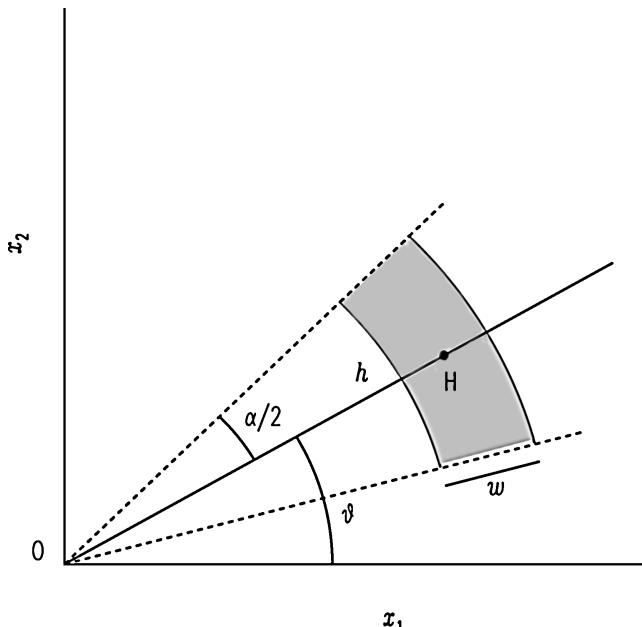
Where data are missing, the quantities  $(m - p)(n - q)$  in the denominators of equation (4.42) must be replaced by the actual numbers of comparisons.



**Figure 4.12** Computing a two-dimensional variogram from a regular grid of data by sliding the grid over itself: (a) by three units to the right; (b) by one unit down in addition, with resultant lag given by the hypotenuse of the triangle.

### Irregular sampling in two dimensions

Survey data in two dimensions are often unevenly distributed. Each pair of observations is separated by a potentially unique lag in both distance and direction. To obtain averages containing directional information we must group the separations by direction as well as by distance. Figure 4.13 shows the geometry of the grouping. We choose a lag interval, the multiples of which will form a regular progression of nominal lag distances as in the one-dimensional case. We then choose a range in distance,  $w$  in Figure 4.13, usually equal to the lag interval. The nominal lag distance is represented by the line OH of length  $h$ . We also choose a set of directions, one of which is shown as  $\vartheta$  in Figure 4.13, and a range in direction,  $\alpha$ , such that  $\alpha = \pi/n$ , where  $n$  is the number of directions, and  $\vartheta$  progresses in steps of  $\alpha$  from 0 to  $\pi/(n - 1)$ . For example, if we choose four directions ( $n = 4$ ) then a sensible progression for  $\vartheta$  would be  $0, \pi/4, \pi/2, 3\pi/4$ , i.e.  $0, 45, 90, 135$  degrees, with  $\alpha = \pi/4$  ( $45^\circ$ ). This ensures complete coverage and no overlap between the different directions. For six directions  $\alpha$  would be  $30^\circ$ . Then for a point  $\mathbf{x}_i$  at O with a second point  $\mathbf{x}_i + \mathbf{h}$  within the stippled zone  $\{z(\mathbf{x}_i) - z(\mathbf{x}_i + \mathbf{h})\}^2$  contributes to  $\hat{\gamma}(\mathbf{h}) = \hat{\gamma}(h, \vartheta)$ . When all comparisons have been made the experimental variogram will consist of the set of averages for the nominal lags in both distance and direction. We can extend this further by



**Figure 4.13** The geometry for discretizing the lag into bins by distance and direction in two dimensions.

computing the average experimental variogram over all directions (omnidirectional) by setting  $\alpha = \pi$  ( $180^\circ$ ). Appendix B gives the GenStat instructions for computing directional and omnidirectional variograms.

### Exploring and displaying anisotropy

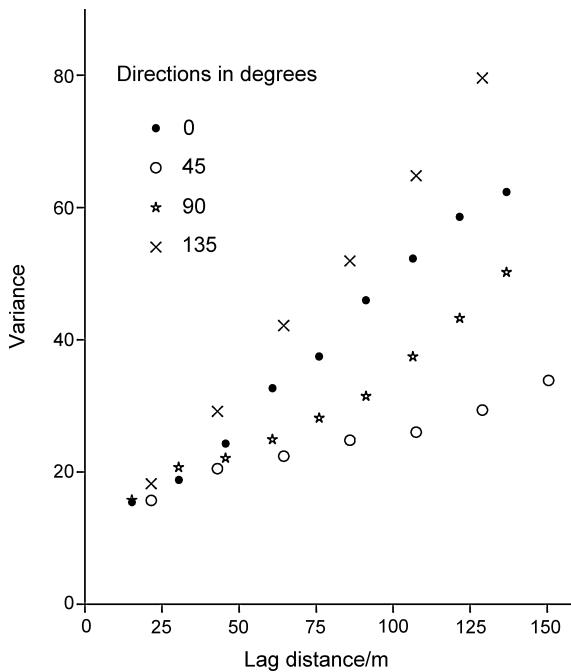
So far we have concentrated on explaining the computation in one and two dimensions, but there is also the matter of representing the results of the two spatial dimensions on a plane, and of exploring differences in the variation in two dimensions.

Where data are on a rectangular grid we can plot the semivariances along the rows and columns and those on the principal diagonals separately, bearing in mind that the lag intervals will not be the same in all four directions. No directional information is lost, and the results can then be examined for directional differences. Where data are irregularly scattered and we have to group the angular separations then we inevitably lose some of the directional information. The wider is  $\alpha$  the more information we lose, until when  $\alpha = \pi$  ( $180^\circ$ ) all is lost. Choosing  $\alpha$  is therefore a compromise between a stable estimate based on many comparisons over a wide angle that will underestimate variance in the direction of the maximum and overestimate that in the direction of the minimum, and one that is subject to large error but which gets closer to the true values in the directions of maximum and minimum. At the outset a reasonable rule of thumb is to let  $\alpha = \pi/4$ . If this appears to reveal anisotropy then try reducing  $\alpha$  until the resulting variogram becomes too erratic. The larger is  $\alpha$ , the more the anisotropy ratio will be underestimated when models are fitted (see Chapter 6). If the variation is isotropic the vector  $\mathbf{h}$  can be replaced by the scalar  $h = |\mathbf{h}|$  in distance only, and the general computing formula, equation (4.40), can be used. In this case we set  $\alpha = \pi$  to compute the omnidirectional variogram.

Whereas it is easy to draw and comprehend a graph of the experimental variogram for either one-dimensional data or one averaged over all directions in two dimensions, it is much less so for the two-dimensional experimental variogram. One simple way is to plot the values with a unique symbol for each direction on the same pair of axes (Figure 4.14). Alternatively, some kind of statistical surface can be fitted to the two-dimensional variogram to represent it as an isarithmic chart or perspective diagram (Figures 4.15 and 4.16). When the variogram has been modelled, this surface can be that of the model. The ideal solution would be to draw it as a stereogram.

#### 4.9.4 The experimental covariance function

All of the above considerations also apply to the estimation of spatial covariances, and the equations are analogous. Remember, however, that the



**Figure 4.14** A two-dimensional variogram with a distinct symbol for each of four directions.

covariance requires stationarity of the mean and the variance of the underlying process. The general computing formula for the experimental covariance at lag  $\mathbf{h}$ , the analogue of equation (4.40), is

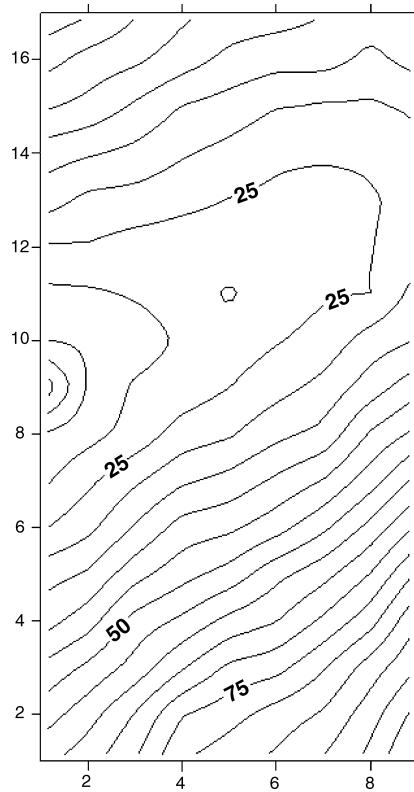
$$\hat{C}(\mathbf{h}) = \frac{1}{m(\mathbf{h})} \sum_{i=1}^{m(\mathbf{h})} \{z(\mathbf{x}_i)z(\mathbf{x}_i + \mathbf{h})\} - \bar{z}^2, \quad (4.43)$$

where  $\bar{z}$  is the mean of all the data. The analogous correlation function, the sample correlogram, is readily derived from  $\hat{C}(\mathbf{h})$  by

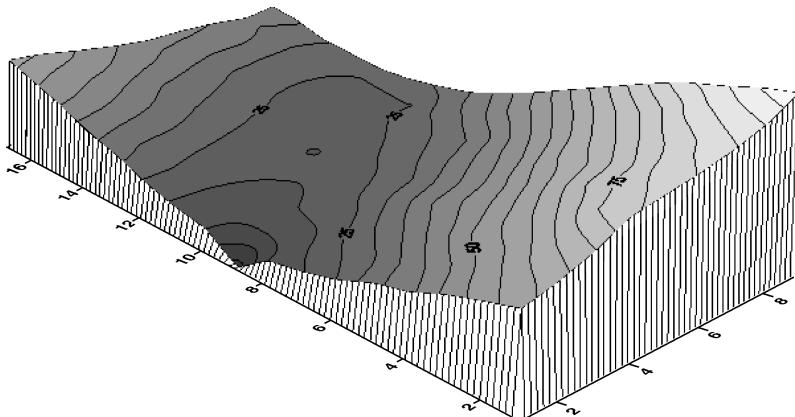
$$\hat{\rho}(\mathbf{h}) = \frac{\hat{C}(\mathbf{h})}{s^2}, \quad (4.44)$$

where  $s^2$  is the variance of the data.

If  $Z(\mathbf{x})$  is second-order stationary then  $\hat{C}(\mathbf{h}) \approx \hat{C}(\mathbf{0}) - \hat{\gamma}(\mathbf{h})$  for all  $\mathbf{h}$ . If there is trend in the variation then  $\hat{C}(\mathbf{0}) - \hat{\gamma}(\mathbf{h})$  will tend to be larger than  $\hat{C}(\mathbf{h})$  computed by equation (4.43). This tendency can be counteracted by replacing the regional mean  $\bar{z}$  by two distinct means, one the mean of the



**Figure 4.15** An isarithmic chart of a two-dimensional variogram. The origin is in the middle of the left-hand side.



**Figure 4.16** A perspectice diagram of a two-dimensional variogram. The origin is in the middle at the left front.

$z(\mathbf{x}_i), i = 1, 2, \dots$ , say  $\bar{z}_1$ , and the other the mean of the  $z(\mathbf{x}_i + \mathbf{h})$ ,  $\bar{z}_2$ , and computing

$$\hat{C}(\mathbf{h}) = \frac{1}{m(\mathbf{h})} \sum_{i=1}^{m(\mathbf{h})} \{z(\mathbf{x}_i) - \bar{z}_1\} \{z(\mathbf{x}_i + \mathbf{h}) - \bar{z}_2\}. \quad (4.45)$$

This measure of the covariance corresponds with that often used in statistics. We no longer assume implicitly that  $\bar{z}_1$  is the same as  $\bar{z}_2$ . Several spatial analysts, e.g. Deutsch and Journel (1992) and Isaaks and Srivastava (1989), use this formula as a matter of course. They call the quantities  $\bar{z}_1$  and  $\bar{z}_2$  the means of the ‘heads’ and of the ‘tails’, respectively.

Similarly the autocorrelation coefficients can be estimated by

$$\hat{\rho}(\mathbf{h}) = \frac{\hat{C}(\mathbf{h})}{s_1 s_2}, \quad (4.46)$$

where  $s_1$  and  $s_2$  are the standard deviations of the heads and tails. These formulae are used by time-series analysts, but Yule and Kendall (1950) warn against using equation (4.46) if you have rather few data.

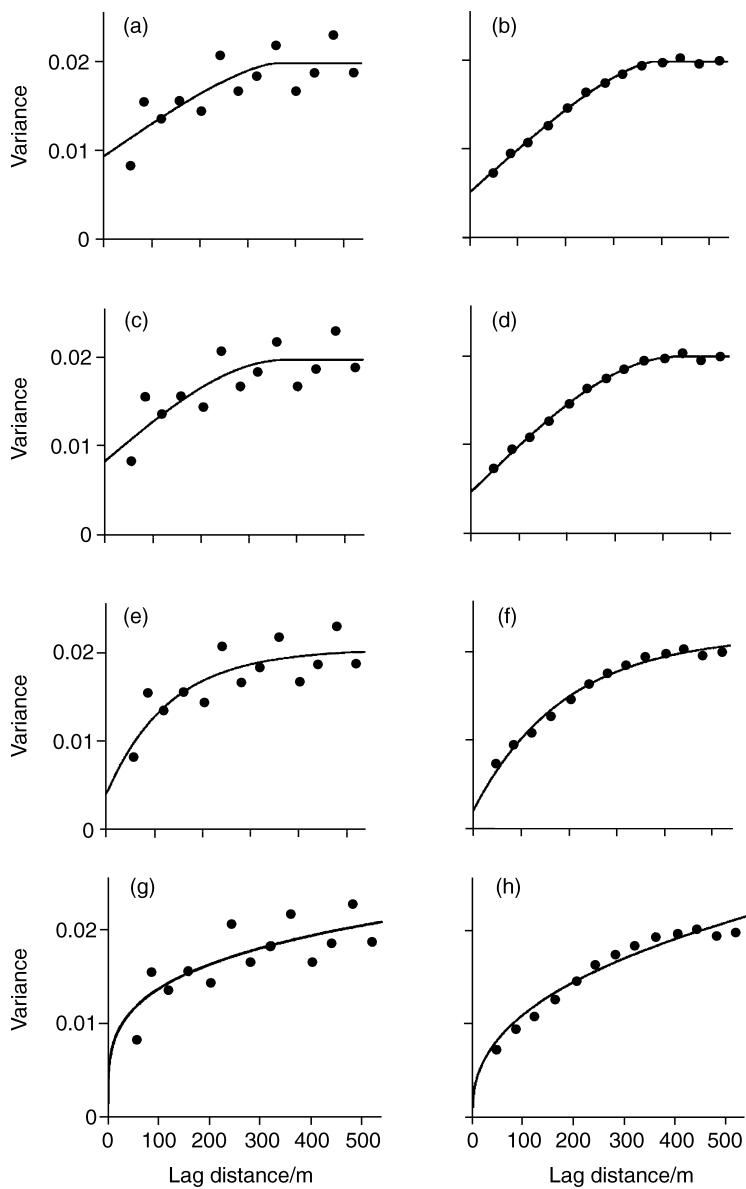
# **Modelling the Variogram**

In Chapter 4 we saw that when we compute an empirical variogram we obtain an ordered set of values, the experimental or sample variogram, consisting of  $\hat{\gamma}(\mathbf{h}_1), \hat{\gamma}(\mathbf{h}_2), \dots$ , at particular lags,  $\mathbf{h}_1, \mathbf{h}_2, \dots$ . This variogram summarizes the spatial relations in the data. We usually want more than that, however; we want a variogram to describe the variance of the region. Each calculated semivariance for a particular lag is only an estimate of a mean semivariance for that lag. As such it is subject to error.

This error, which arises largely from sampling fluctuation, can give the experimental variogram a more or less erratic appearance, as depicted by the plotted points in the graphs in the left-hand column of Figure 5.1. We computed this experimental variogram from 87 values of  $\log_{10}K$  from the Broom's Barn data by taking every fifth value from the file. In the right-hand column is the experimental variogram computed from all 434 data; the points lie on a relatively smooth curve. Evidently the sampling fluctuation is more pronounced where the data points are further apart and there are fewer of them. We explore the effects of the number of data points on the reliability of the variogram in more detail in Chapter 6, and we explain the smooth curves drawn through the experimental values later in this chapter.

The true variogram representing the regional variation is continuous, and it is this variogram that we should really like to know. We can use our observed values as approximations to the function by imagining a curve passing through them, such as the ones we have drawn in Figure 5.1. In two dimensions we have to imagine a surface, for the variogram of a two-dimensional field is itself two-dimensional. How closely should we attempt to follow the experimental variogram? Answering this is difficult because we do not know how much of the observed fluctuation is due to error and how much is structural.

The solution usually taken is that of Occam's razor; namely, fit the simplest function that makes sense, subject to certain mathematical constraints which are considered below. We ignore the point-to-point fluctuation and concentrate on the general trends.



**Figure 5.1** Experimental variograms plotted, as points, of  $\log_{10}K$  at Broom's Barn computed from 87 data in the left-hand column and from all 434 data in the right-hand column. The solid lines from top to bottom are: the circular, spherical, exponential and power models fitted to them.

Another reason for fitting a continuous function is to describe the spatial variation so that we can estimate or predict values at unsampled places and in larger blocks of land optimally by kriging (see Chapter 8). This requires semivariances at lags for which we have no direct comparisons, and we must be able to calculate these from such a function. The function must therefore be mathematically defined for all real  $\mathbf{h}$ .

There are a few principal features that a function must be able to represent. These include:

- (1) a monotonic increase with increasing lag distance from the ordinate of appropriate shape;
- (2) a constant maximum or asymptote, or ‘sill’;
- (3) a positive intercept on the ordinate, or ‘nugget’;
- (4) periodic fluctuation, or a ‘hole’;
- (5) anisotropy.

## 5.1 LIMITATIONS ON VARIOGRAM FUNCTIONS

### 5.1.1 Mathematical constraints

Not any close-fitting function will serve. The model we choose must describe random variation, and the function must be such that it will not give rise to ‘negative variances’ of combinations of random variables. This is explained below.

Let  $z(\mathbf{x}_i), i = 1, 2, \dots, n$ , be a realization of the random variable  $Z(\mathbf{x})$  with covariance function  $C(\mathbf{h})$  and variogram  $\gamma(\mathbf{h})$ . Now consider the linear sum

$$y = \sum_{i=1}^n \lambda_i z(\mathbf{x}_i),$$

where the  $\lambda_i$  are any arbitrary weights.

The variable  $Y$  from which  $y$  derives is itself random with variance

$$\text{var}[Y] = \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j C(\mathbf{x}_i - \mathbf{x}_j), \quad (5.1)$$

where  $C(\mathbf{x}_i - \mathbf{x}_j)$  is the covariance of  $Z$  between  $\mathbf{x}_i$  and  $\mathbf{x}_j$ . The variance of  $Y$  may be positive or zero; but it may not be negative. The right-hand side of equation (5.1) must ensure this. The covariance function,  $C(\mathbf{h})$ , must be *positive semidefinite*. Equation (5.1) can be written

$$\text{var}[Y] = \boldsymbol{\lambda}^T \mathbf{C} \boldsymbol{\lambda} \geq 0, \quad (5.2)$$

where  $\boldsymbol{\lambda}$  is the vector of weights and  $\mathbf{C}$  is the matrix of covariances. If the latter is positive semidefinite then so is the covariance function. In fact, since we are dealing with ‘variables’, the variance cannot be zero, and so  $C(\mathbf{h})$  must be positive definite.

If the covariance does not exist, because the variable is intrinsic only and not second-order stationary, then we rewrite equation (5.1) as

$$\text{var}[Y] = C(\mathbf{0}) \sum_{i=1}^n \lambda_i \sum_{j=1}^n \lambda_j - \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \gamma(\mathbf{x}_i - \mathbf{x}_j), \quad (5.3)$$

where  $\gamma(\mathbf{x}_i - \mathbf{x}_j)$  is the semivariance of  $Z$  between  $\mathbf{x}_i$  and  $\mathbf{x}_j$ . The first term on the right-hand side of equation (5.3) contains  $C(\mathbf{0})$ , the covariance at lag  $\mathbf{0}$ , which we do not know, but we can eliminate it by making the weights sum to 0 without loss of generality. Then

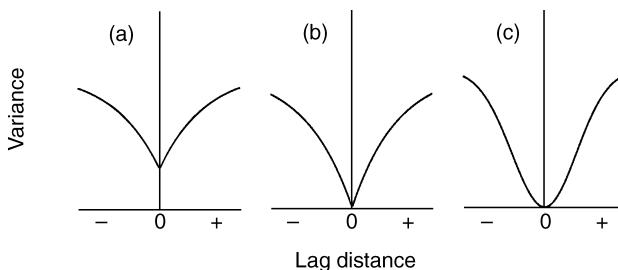
$$\text{var}[Y] = - \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \gamma(\mathbf{x}_i - \mathbf{x}_j). \quad (5.4)$$

This may not be negative either; but notice the minus sign. So, the variogram must be *conditional negative semidefinite* (CNSD), the condition being that the weights in equation (5.4) sum to zero.

Only functions that ensure non-zero variances may be used for variograms. They are called *authorized* models or functions in much of the literature.

### 5.1.2 Behaviour near the origin

The way in which the variogram approaches the origin is determined by the continuity (or lack of continuity) of the variable,  $Z(\mathbf{x})$ , itself. We may distinguish the following features, which are illustrated in Figure 5.2. As mentioned in Chapter 4, the variogram is symmetric about the origin, and for reasons that will become evident both halves appear in this figure.



**Figure 5.2** Behaviour of the variogram near to the origin: (a) positive intercept (nugget); (b) linear approach, not differentiable; (c) continuous and differentiable.

*Positive intercept.* The semivariance at  $|\mathbf{h}| = 0$  is by definition 0. It often happens, however, that any line or surface projected through the experimental values to the ordinate intersects it at some positive value, as in Figure 5.2(a). This implies a discontinuity in  $Z(\mathbf{x})$ . The feature appeared often in gold mining, and the mining engineers attributed it to the spatially independent occurrence of gold nuggets in ore bodies. They called the phenomenon the ‘nugget effect’ and the intercept ‘nugget variance’. It is easy to imagine discontinuities arising from the dispersal of small nuggets of gold in a large body of rock. The same might be true for certain features of the soil, such as stones and concretions among the fine earth. Discontinuities in the soil’s physical and chemical properties are harder to imagine, and any apparent nugget variance usually arises from errors of measurement and spatial variation within the shortest sampling interval.

*Linear approach.* The variogram may approach the origin approximately linearly with decreasing lag distance:

$$\gamma(\mathbf{h}) \approx b|\mathbf{h}| \quad \text{as } |\mathbf{h}| \rightarrow 0, \quad (5.5)$$

where  $b$  is the gradient. The variogram passes through the origin, as in Figure 5.2(b), unlike in Figure 5.2(a), but its first derivative is discontinuous there: its gradient changes abruptly from negative to positive. Nevertheless, it signifies continuity in  $Z(\mathbf{x})$  itself, and because

$$\lim E[\{Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})\}^2] = 0 \quad \text{as } |\mathbf{h}| \rightarrow 0, \quad (5.6)$$

$Z(\mathbf{x})$  is often said to be ‘mean-square’ continuous. It is not differentiable, however, nor is the process it describes because it is random (see Chapter 4).

*Parabolic approach.* Figure 5.2(c) illustrates the situation in which a variogram is parabolic at the origin; it passes smoothly through the origin with a gradient of 0 there, so that

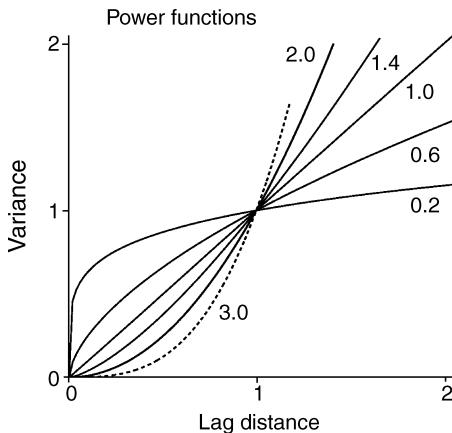
$$\gamma(\mathbf{h}) = b|\mathbf{h}|^2 \quad \text{as } |\mathbf{h}| \rightarrow 0. \quad (5.7)$$

The variogram is twice differentiable at the origin, and  $Z(\mathbf{x})$  is itself differentiable: it varies smoothly, and it is no longer random. The exponent 2 represents a strict limit to power functions for describing random processes.

A raw variogram that appears parabolic at the origin suggests that there is local *trend*, i.e. short-range deterministic variation. This feature is described in Chapter 4, and we deal with it again in Chapter 9. The expectation of  $Z(\mathbf{x})$  is not stationary but depends on position  $\mathbf{x}$ , thus:

$$E[Z(\mathbf{x})] = u(\mathbf{x}), \quad (5.8)$$

from equation (4.21).



**Figure 5.3** Graphs of the power function,  $\gamma(h) = wh^\alpha$ , with  $\alpha = 0.2, 0.6, 1.0, 1.4$ , and  $2.0$  (the limiting value for  $\alpha$ ), and with  $w$  set to  $1$ , shown by solid lines. The dashed line represents  $\gamma(h) = h^3$  and is not an authorized function for a variogram.

### 5.1.3 Behaviour towards infinity

The way that a variogram behaves with increasing lag distance is constrained by

$$\lim_{|\mathbf{h}| \rightarrow \infty} \frac{\gamma(\mathbf{h})}{|\mathbf{h}|^2} = 0 \quad \text{as } |\mathbf{h}| \rightarrow \infty. \quad (5.9)$$

The variogram must increase less than the square of the lag distance as the latter approaches infinity; if it does not then the process is not entirely random. The limit is shown in Figure 5.3, in which the parameter  $\alpha$  in the power function is set to 2. Any function that increases more, such as that shown by the dashed line with  $\alpha = 3$ , is not CNSD and so is not compatible with the intrinsic hypothesis.

A variogram that increases faster than  $|\mathbf{h}|^2$  suggests that there is long-range trend, again deterministic in the statistical sense (see Chapters 4 and 6). As above, the expectation of  $Z(\mathbf{x})$  is not stationary but depends on position  $\mathbf{x}$ ; see equation (5.8).

## 5.2 AUTHORIZED MODELS

There are two main families of simple functions that encompass the features listed above and that are CNSD. One represents unbounded variation, the other bounded. We deal with them in turn in their isotropic form, so that the lag vector  $|\mathbf{h}|$  becomes a scalar measure in distance only,  $h$ . All of the ones that we describe are used in practice.

### 5.2.1 Unbounded random variation

The idea of unbounded, i.e. infinite, variance may seem strange. After all, we live on a finite earth, and there must be some limit to the amount of variation in the soil. Yet the evidence from surveys of small parts of the planet suggests that if we were to increase the region surveyed we should encounter ever more variation; our extrapolation of the experimental variogram is one that continues to increase.

The simplest models for unbounded variation are the power functions:

$$\gamma(h) = wh^\alpha \text{ for } 0 < \alpha < 2, \quad (5.10)$$

where  $w$  describes the intensity of variation and  $\alpha$  describes the curvature. If  $\alpha = 1$  then the variogram is linear, and  $w$  is simply the gradient. If  $\alpha < 1$  then the variogram is convex upwards. If  $\alpha > 1$  then the variogram is concave upwards. The limits 0 and 2 are excluded. If  $\alpha = 0$  then we are left with a constant variance for all  $h > 0$ ; if  $\alpha = 2$  then the function is parabolic with gradient 0 at the origin and represents differentiable variation in the underlying process, which is not random, as mentioned above.

Figure 5.3 shows examples with several values of  $\alpha$ , including the upper bound,  $\alpha = 2$ ; at the lower limit  $\alpha = 0$  would represent white noise, and hence discontinuous variation. Nevertheless, some experimental variograms seem flat, and we return to this matter below.

One way of looking at these unbounded functions is to consider Brownian motion in one dimension. Suppose a particle moves in this dimension with a velocity or momentum at position  $\mathbf{x} + \mathbf{h}$  that depends on its velocity or momentum at a close previous position  $\mathbf{x}$ . It can be represented by the equation

$$Z(\mathbf{x} + \mathbf{h}) = \beta Z(\mathbf{x}) + \varepsilon, \quad (5.11)$$

where  $\varepsilon$  is an independent Gaussian random deviate and  $\beta$  is a parameter. At its simplest  $\beta = 1$ , and its variogram is then

$$2\gamma(\mathbf{h}) = E[\{Z(\mathbf{x} + \mathbf{h}) - Z(\mathbf{x})\}^2] = |\mathbf{h}|^k. \quad (5.12)$$

If the exponent  $k$  in equation (5.12) is 1 then we obtain the linear model, with  $\gamma(|\mathbf{h}|) \rightarrow \infty$  as  $|\mathbf{h}| \rightarrow \infty$ . This is also known as a *random walk* model.

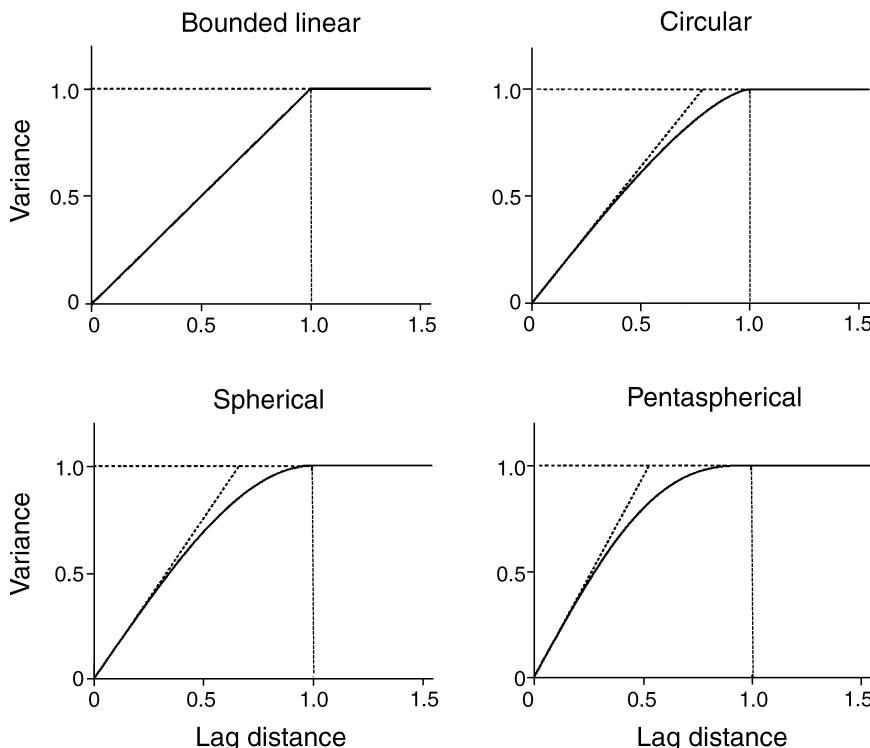
In ordinary Brownian motion the  $\varepsilon$ s are independent of one another. If, however, the  $\varepsilon$ s in equation (5.11) are spatially correlated then a trace is generated that is smoother than that of pure Brownian motion. The exponent,  $k$ , now exceeds 1, and the curve is concave upwards. If, on the other hand, the  $\varepsilon$ s are negatively correlated then a trace is generated that is rougher, or ‘noisier’, than that of pure Brownian motion. The exponent  $k$  in equation (5.12) is now less than 1, and the curve is convex upwards.

If the  $\varepsilon$ s are perfectly correlated then  $k = 2$  and the trace is completely smooth, i.e. there is no longer any randomness. As  $k \rightarrow 0$ , the noise increases until in the limit we have white noise, or pure nugget, as described in Chapter 4.

Priestley (1981) gives a much more comprehensive account of these random processes. Chapter 3 of that book is especially relevant, and we must leave the reader to pursue the matter there.

### 5.2.2 Bounded models

In our experience bounded variation is more common than unbounded variation, and the variograms have more varied shapes. In most of these models the variance has a maximum, which is the *a priori* variance of the process, known in geostatistics as the *sill* variance. The variogram may reach its sill at a finite lag distance, the *range*. Alternatively, the variogram may approach its sill asymptotically. In some models the semivariance reaches a maximum, only to decrease again and perhaps fluctuate about its *a priori* variance. These variograms represent second-order stationary processes and so have equivalent covariance functions. They are illustrated in Figures 5.4 and 5.5.



**Figure 5.4** Bounded models with fixed ranges: (a) bounded linear; (b) circular; (c) spherical; (d) pentaspherical.

## Bounded linear model

The simplest function for describing bounded variation consists of two straight lines, as in Figure 5.4(a). The first increases and the other has a constant variance:

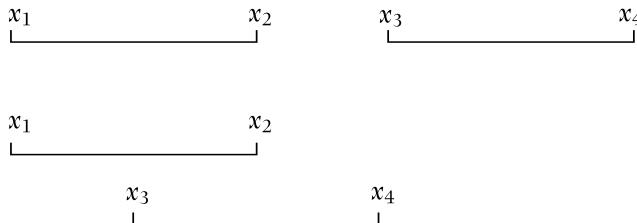
$$\gamma(h) = \begin{cases} c\left(\frac{h}{a}\right) & \text{for } h \leq a \\ c & \text{for } h > a, \end{cases} \quad (5.13)$$

where  $c$  is the sill variance and  $a$  is the range. Evidently its slope at the origin is  $c/a$ . It is CNSD in one dimension ( $\mathbb{R}^1$ ) only; it may not be used to describe variation in two and three dimensions.

We can derive the variogram for the bounded linear model heuristically as follows. We start with a stationary ‘white noise’ process,  $Y(x)$ , in one dimension, i.e. a random process with random variables at all positions along a line but in which there is no spatial dependence or autocorrelation. It has a mean  $\mu$  and variance  $\sigma_Y^2$ . Suppose that we pass the process through a simple linear filter of finite length  $a$  to obtain

$$Z(x) - \mu = \frac{1}{a} \int_x^{x+a} Y(v) dv. \quad (5.14)$$

Thus, we average  $Y(x)$  within the interval  $a$  to obtain the corresponding  $Z(x)$ . Consider now the variable  $Z(x)$  derived from two segments of the process  $Y(x)$ , one from  $x_1$  to  $x_2$  and the other from  $x_3$  to  $x_4$ . They may overlap or not, as below.



Evidently, if the two segments do not overlap, as in the upper example, then we should expect their means in  $Z(x)$  to be independent. But if they do overlap, as in the lower example, then they will share some of the original white noise series; their means will not be independent, and we should expect some autocorrelation. In general, the closer is  $x_1$  to  $x_3$  (and  $x_2$  to  $x_4$ ) and the longer is  $a$ , the stronger should be the correlation. In fact when  $x_1$  coincides with  $x_3$  (and  $x_2$  with  $x_4$ ) we should have perfect correlation. The only question is what form the correlation takes as  $x_3$  approaches  $x_1$ .

To answer this we consider the discrete analogue of equation (5.14):

$$\begin{aligned} Z(x+d) - \mu &= \lambda_0 Y(x+d) + \lambda_1 Y(x+d+1) + \lambda_2 Y(x+d+2) \\ &\quad + \cdots + \lambda_{a-1} Y(x+d+a-1), \end{aligned} \quad (5.15)$$

where the  $\lambda_0, \lambda_1, \dots, \lambda_{a-1}$  are weights, here all equal to  $1/a$ , and  $d = 1/2a$  is half the distance between two successive points in the sequence. All more distant members, say  $Y(x+d+a-1+b)$ , of the series carry zero weight. Suppose that  $Y(x)$  is a white noise process; then  $Z(x)$  is a moving average process of order  $a-1$ . Further, if the variance of  $Y(x)$  is  $\sigma_Y^2$  then that of  $Z(x)$  is

$$\begin{aligned} \sigma_Z^2 &= \lambda_0^2 \sigma_Y^2 + \lambda_1^2 \sigma_Y^2 + \lambda_2^2 \sigma_Y^2 + \cdots + \lambda_{a-1}^2 \sigma_Y^2 \\ &= \sigma_Y^2 \sum_{i=0}^{a-1} \lambda_i \lambda_i \\ &= \sigma_Y^2 / a, \end{aligned} \quad (5.16)$$

which is familiar as the variance of a mean. It is also the covariance at lag 0,  $C(0)$ . We now want the covariances for the larger lags. These are obtained simply by extension from the above equation:

$$C(h) = \sigma_Y^2 \sum_{i=0}^{a-1-h} \lambda_i \lambda_{i+h} = \sigma_Y^2 \frac{a-h}{a^2}. \quad (5.17)$$

The covariances are in order, for  $h = 0, 1, 2, \dots, a-1, a$ ,

$$\frac{a-0}{a^2} \sigma_Y^2, \frac{a-1}{a^2} \sigma_Y^2, \frac{a-2}{a^2} \sigma_Y^2, \dots, \frac{a-a+1}{a^2} \sigma_Y^2, \frac{a-a}{a^2} \sigma_Y^2.$$

Dividing through by the  $C(0)$  we obtain the autocorrelations,  $\rho(h)$ , as

$$1, (a-1)/a, (a-2)/a, \dots, (a-a+1)/a, 0.$$

In words, the covariance and autocorrelation functions decay linearly with increasing  $h$  until  $h=a$ , at which point it is 0. Then the autocorrelation coefficient at any  $h$  is simply equal to the proportion of the filter that overlaps when the filter is translated by  $h$ . The variogram is obtained simply from relation (4.14) by

$$\begin{aligned} \gamma(h) &= C(0) - C(h) \\ &= \sigma_Y^2 \frac{a-h}{a^2} = \frac{\sigma^2}{a} \left( \frac{h}{a} \right) = c \left( \frac{h}{a} \right), \end{aligned} \quad (5.18)$$

since  $c = \sigma_Y^2/a = C(0)$ .

### Circular model

The formula for the circular variogram is

$$\gamma(h) = \begin{cases} c \left\{ 1 - \frac{2}{\pi} \cos^{-1} \left( \frac{h}{a} \right) + \frac{2h}{\pi a} \sqrt{1 - \frac{h^2}{a^2}} \right\} & \text{for } h \leq a, \\ c & \text{for } h > a. \end{cases} \quad (5.19)$$

The parameters  $c$  and  $a$  are again the sill and range. The function curves tightly as it approaches the range (see Figure 5.4(b)) and its gradient at the origin is  $4c/\pi a$ . It is CNSD in  $\mathbb{R}^1$  and  $\mathbb{R}^2$ , but not in  $\mathbb{R}^3$ .

This model can be derived in a way analogous to that of the bounded linear model from the area of intersection,  $A$ , of two discs of diameter  $a$ , the centres of which are separated by distance  $h$ . Matérn (1960) did this by considering the densities with which points are distributed at random by a Poisson process in two overlapping circles. This area is

$$A = \begin{cases} \frac{1}{2} a^2 \cos^{-1} \left( \frac{h}{a} \right) - \frac{h}{2\pi} \sqrt{a^2 - h^2} & \text{for } h \leq a, \\ 0 & \text{for } h > a. \end{cases} \quad (5.20)$$

If we express this as a fraction of the area,  $\pi a^2/4$ , of one of the circles, in the same way as we expressed the fraction of the linear filter that overlapped along the line above, then we obtain the autocorrelation for the separation:

$$\rho(h) = \frac{2}{\pi} \left\{ \cos^{-1} \left( \frac{h}{a} \right) - \frac{h}{a} \sqrt{1 - \frac{h^2}{a^2}} \right\} \quad \text{for } h \leq a. \quad (5.21)$$

Then from relation (4.14) the variogram, equation (5.19) above, follows.

### Spherical model

By a similar line of reasoning we can derive the three-dimensional analogue of the circular model to obtain the spherical correlation function and variogram. The volume of intersection of two spheres of diameter  $a$  with their centres  $h$  apart is

$$V = \begin{cases} \frac{\pi}{4} c \left( \frac{2}{3} a^3 - a^2 h + \frac{1}{3} h^3 \right) & \text{for } h \leq a, \\ 0 & \text{otherwise.} \end{cases} \quad (5.22)$$

The volume of a sphere is  $\frac{1}{6}\pi a^3$ , and so dividing by it gives the autocorrelation

$$\rho(h) = \begin{cases} 1 - \frac{3h}{2a} + \frac{1}{2} \left( \frac{h}{a} \right)^3 & \text{for } h \leq a, \\ 0 & \text{for } h > a, \end{cases} \quad (5.23)$$

and the variogram is

$$\gamma(h) = \begin{cases} c \left\{ \frac{3h}{2a} - \frac{1}{2} \left( \frac{h}{a} \right)^3 \right\} & \text{for } h \leq a, \\ c & \text{for } h > a. \end{cases} \quad (5.24)$$

The spherical model seems the obvious one to describe variation in three-dimensional bodies of rock, and it has proved well suited to them. It would seem less obviously suited for describing the variation in one and two dimensions, which is usually what is needed in soil and land resource survey. Yet it nearly always fits experimental results from soil sampling better than the one- and two-dimensional analogues. The function curves more gradually than they do Figure 5.4(c), and the reason is probably that there are additional sources of variation at other scales that it can represent. Its gradient at the origin is  $3c/2a$ . It is CNSD in  $\mathbb{R}^2$  and  $\mathbb{R}^1$  as well as in  $\mathbb{R}^3$ .

The spherical function is one of the most frequently used models in geostatistics, in one, two and three dimensions. It represents transition features that have a common extent and which appear as patches, some with large values and others with small ones. The average diameter of the patches is represented by the range of the model. One can see this interpretation by simulating a large field of values using the function as the generator. Figures 5.5 and 5.6(a) are examples in which values have been simulated on a  $256 \times 256$  square grid with unit interval. The model had a sill variance,  $c = 1.0$ , and ranges of  $a = 15, 25$  and  $50$  units in Figures 5.5(a), 5.5(b) and 5.6(a), respectively. The maps show that the extents of the patches with large and small values increase as the range increases. The patches have a fairly regular form.

### Pentaspherical model

Following Matérn (1960), McBratney and Webster (1986) extended the line of reasoning to obtain the five-dimensional analogue of the above, the pentaspherical function:

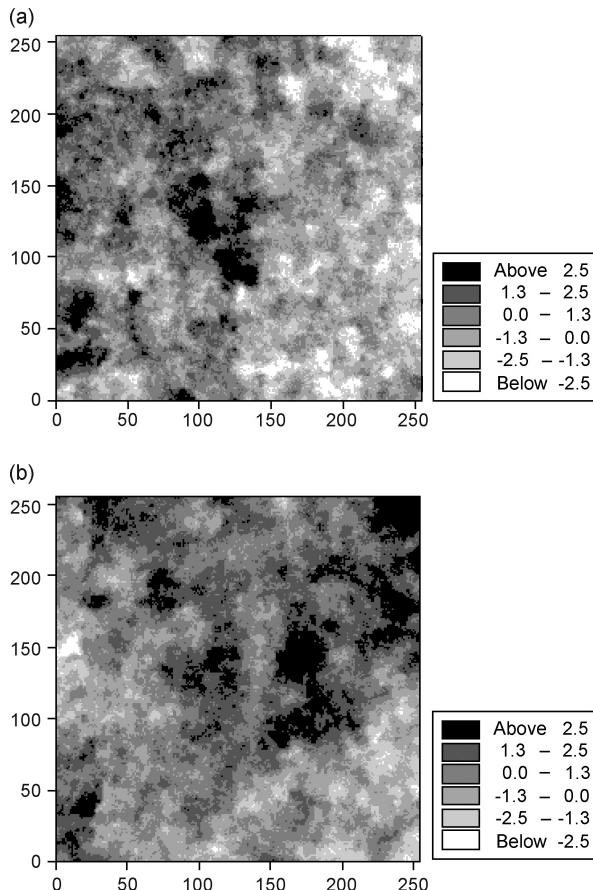
$$\gamma(h) = \begin{cases} c \left\{ \frac{15h}{8a} - \frac{5}{4} \left( \frac{h}{a} \right)^3 + \frac{3}{8} \left( \frac{h}{a} \right)^5 \right\} & \text{for } h \leq a, \\ c & \text{for } h > a. \end{cases} \quad (5.25)$$

It is useful in that its curve is somewhat more gradual than that of the spherical model Figure 5.4(d). Its gradient at the origin is  $15c/8a$ . Again it is CNSD in  $\mathbb{R}^1$ ,  $\mathbb{R}^2$  and  $\mathbb{R}^3$ .

### Exponential model

A function that is also much used in geostatistics is the negative exponential:

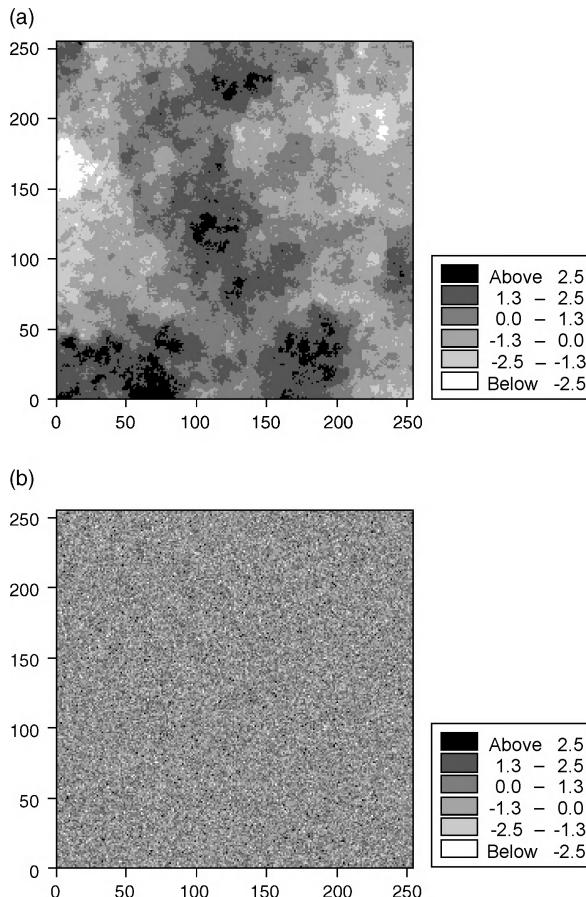
$$\gamma(h) = c \left\{ 1 - \exp \left( -\frac{h}{r} \right) \right\}, \quad (5.26)$$



**Figure 5.5** Simulated fields of values from spherical functions, equation (5.24), with distance parameters (a)  $a = 15$ , (b)  $a = 25$ .

with sill  $c$ , and a distance parameter,  $r$ , that defines the spatial extent of the model. The function approaches its sill asymptotically, and so it does not have a finite range. Nevertheless, for practical purposes it is convenient to assign it an effective range, and this is usually taken as the distance at which  $\gamma$  equals 95% of the sill variance, approximately  $3r$ . Its slope at the origin is  $c/r$ . Figure 5.7(a) shows it.

The function has an important place in statistical theory. It represents the essence of randomness in space. It is the variogram of first-order autoregressive and Markov processes. Its equivalent autocorrelation function has been the basis of several theoretical studies of the efficiency of sampling designs by, for example, Cochran (1946), Yates (1948), Quenouille (1949) and Matérn (1960). We should expect variograms of this form where differences in soil

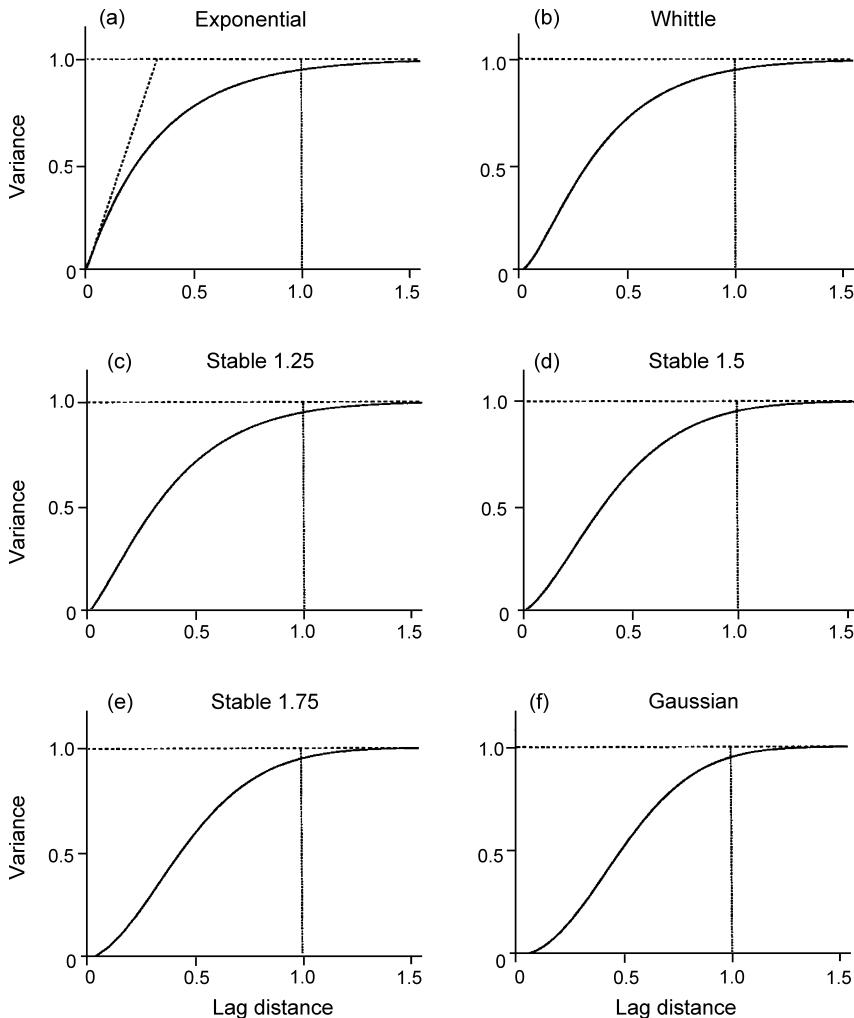


**Figure 5.6** Simulated fields of values using: (a) a spherical function, equation (5.24), with distance parameter  $a = 50$ ; (b) a pure nugget variogram, equation (5.33).

type are the main contributors to soil variation and where the boundaries between types occur at random as a Poisson process. Burgess and Webster (1984) found this to be the situation in many instances. If the intensity of the process is  $\eta$  then the mean distance between boundaries is  $\bar{d} = 1/\eta$  and the variogram is

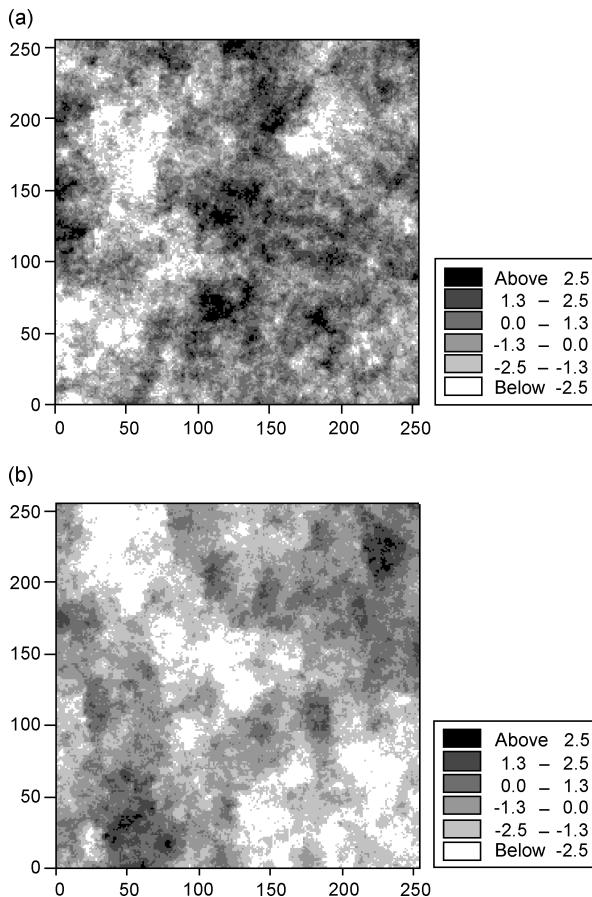
$$\begin{aligned}\gamma(h) &= c\{1 - \exp(-h/\bar{d})\} \\ &= c\{1 - \exp(-\eta h)\}. \end{aligned} \tag{5.27}$$

Put another way, this is the variogram of a transition process in which the structures have random extents.



**Figure 5.7** Models with asymptotic bounds. All are scaled so that the effective range where the function reaches 0.95 of its sill is approximately 1, marked by the vertical lines on the graphs. (a)  $\alpha = 1$  (exponential),  $r = 0.333$ ; (b) Whittle,  $r = 0.25$ ; (c)  $\alpha = 1.25$  (stable),  $r = 0.416$ ; (d)  $\alpha = 1.5$  (stable),  $r = 0.478$ ; (e)  $\alpha = 1.75$  (stable),  $r = 0.533$ ; (f)  $\alpha = 2$  (Gaussian),  $r = 1/\sqrt{3}$ .

Simulated fields obtained from an exponential function with an asymptote approaching 1.0 and distance parameters,  $r$ , of 5 and 16 are shown in Figure 5.8(a) and 5.8(b), respectively. The patches of large and small values in the two fields are similarly irregular, but the average sizes of the patches show the different spatial scales of the generator.



**Figure 5.8** Simulated fields of values from exponential functions (equation (5.26)), with distance parameters (a)  $r = 5$ , (b)  $r = 16$ .

### Whittle's elementary correlation

Whittle (1954) showed that a simple stochastic diffusion process also has an exponential variogram in one and three dimensions. In  $\mathbb{R}^2$ , however, the process leads to Whittle's *elementary correlation*, given by

$$\gamma(h) = c \left\{ 1 - \frac{h}{r} K_1 \left( \frac{h}{r} \right) \right\}. \quad (5.28)$$

The parameter  $c$  is the sill, as before, the *a priori* variance of the process,  $r$  is a distance parameter, and  $K_1$  is the modified Bessel function of the second kind. Like the exponential function, Whittle's function approaches its sill

asymptotically and so has no definite range. Its effective range may be chosen as for the exponential function where the semivariance reaches 95% of the sill, and this is at approximately  $4r$ . The function approaches the origin with a decreasing gradient, however, and appears slightly sigmoid when plotted, Figure 5.7(b).

### Gaussian model

Another function with reverse curvature near the origin recurs again and again in geostatistical texts and software packages. It is the so-called Gaussian model, Figure 5.7(f), with equation

$$\gamma(h) = c \left\{ 1 - \exp \left( -\frac{h^2}{r^2} \right) \right\}. \quad (5.29)$$

Once more,  $c$  is the sill and  $r$  is a distance parameter. The function approaches its sill asymptotically, and it can be regarded as having an effective range of approximately  $\sqrt{3}r$  where it reaches 95% of its sill variance.

A serious disadvantage of the model is that it approaches the origin with zero gradient, which we saw above as the limit for random variation and at which the underlying variation becomes continuous and twice differentiable. This can lead to unstable kriging equations, which we present in Chapter 8, and bizarre effects when used for estimation—see Wackernagel (2003) for examples.

In general we deprecate this model. If a variogram appears somewhat sigmoid then we recommend the theoretically attractive Whittle function. Alternatively, if the reverse curvature is stronger you may replace the exponent 2 in equation (5.29) by an additional parameter, say  $\alpha$ , with a value less than 2:

$$\gamma(h) = c \left\{ 1 - \exp \left( -\frac{h^\alpha}{r^\alpha} \right) \right\}. \quad (5.30)$$

Wackernagel (2003) calls these ‘stable models’. Some examples of them are shown in Figure 5.7(c)–(e) with various values of  $\alpha$ , and we have used the model with  $\alpha = 1.965$  to describe topographic variation (Webster and Oliver, 2006).

### Cubic model

Another bounded model with reverse curvature near the origin is the cubic function. Its formula is

$$\gamma(h) = \begin{cases} c \left\{ 7 \left( \frac{h}{a} \right)^2 - 8.75 \left( \frac{h}{a} \right)^3 + 3.5 \left( \frac{h}{a} \right)^5 - 0.75 \left( \frac{h}{a} \right)^7 \right\} & \text{for } h \leq a, \\ c & \text{for } h > a. \end{cases} \quad (5.31)$$

The parameter  $a$  is a finite range which is approached much more gradually than in the spherical and pentaspherical models.

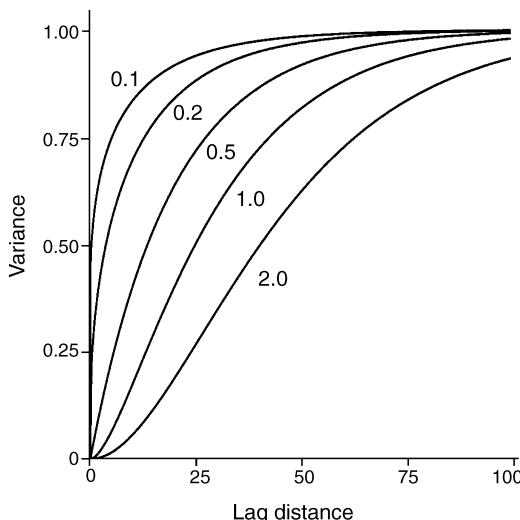
There are other simple models used in particular disciplines because of their theoretical attractions. Examples include the prismato-gravimetric and prismato-magnetic functions developed in geophysics to model gravimetric and magnetic anomalies (see Armstrong, 1998). If you work in such a special field then you should ask whether there are preferred functions for the particular applications.

### Matérn function

The Matérn function is a generalization of several of the functions mentioned above and so appears attractive for this reason. Its formula is

$$\gamma(h) = c \left\{ 1 - \frac{1}{2^{v-1}\Gamma(v)} \left( \frac{h}{r} \right)^v K_v \left( \frac{h}{r} \right) \right\}. \quad (5.32)$$

As in the exponential, Whittle and Gaussian models the function has a distance parameter  $r$ , and  $c$  is the sill. It also has a smoothness parameter,  $v$ , analogous to  $\alpha$  in the stable models, equation (5.30), though whereas  $\alpha$  is limited to between 0 and 2,  $v$  can vary in the range 0 (very rough) to infinity (very smooth). It includes the special cases of exponential when  $v = 0.5$  and Whittle's function when  $v = 1$ . Figure 5.9 shows variograms for several values of  $v$ .



**Figure 5.9** The Matérn function (5.32) with *a priori* variance  $c = 1$  and distance parameter  $r = 20$  and five values of the smoothness parameter  $v$ , giving the five curves. The curve with  $v = 0.5$  is the exponential and that with  $v = 1$  is Whittle's function. After Minasny and McBratney (2005).

Unfortunately, when Minasny and McBratney (2005) examined its potential for describing soil properties they had difficulty fitting it to experimental variograms. They found that  $v$  was poorly estimated by the usual method of weighted least squares (see below).

### Pure nugget

Although the limiting value 0 of the exponent of equation (5.10) for the power function was excluded because it would give a constant variance, we do need some way of expressing such a constant because that is what appears in practice. We do so by defining a ‘pure nugget’ variogram as follows:

$$\gamma(h) = c_0 \{1 - \delta(h)\}, \quad (5.33)$$

where  $c_0$  is the variance of the process, and  $\delta(h)$  is the Kronecker  $\delta$  which takes the value 1 when  $h = 0$  and is zero otherwise. If the variable is continuous, as almost all properties of the soil and natural environment are, then a variogram that appears as pure nugget has almost certainly failed to detect the spatially correlated variation because the sampling interval was greater than the scale of spatial variation.

Since the nugget variance is constant for all  $\mathbf{h}$ ,  $|\mathbf{h}| > 0$ , it is usually denoted simply by the variance  $c_0$ . Figure 5.6(b) shows the simulated field from a pure nugget variogram. There is no detectable pattern in the variation as there is in Figures 5.5, 5.6(a) and 5.8.

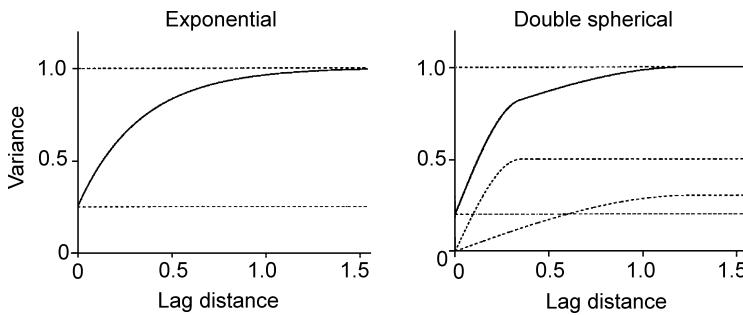
## 5.3 COMBINING MODELS

As is apparent in Figures 5.3, 5.4 and 5.7, all the above functions have simple shapes. In many instances, however, especially where we have many data, variograms appear more complex, and we may therefore seek more complex functions to describe them. The best way to do this is to combine two or more simple models. Any combination of CNSD functions is itself CNSD. Do not look for complex mathematical solutions the properties of which are unknown.

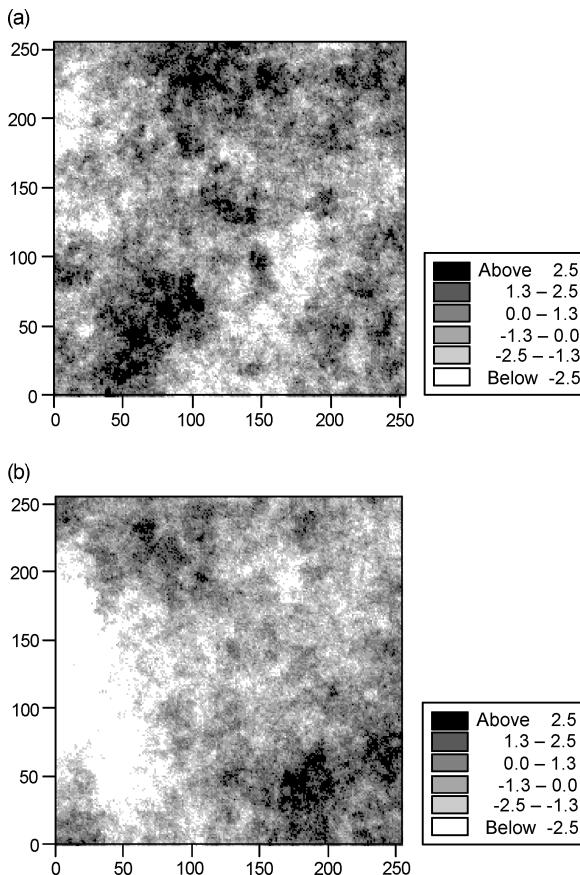
The most common requirement is for a model that has a nugget component in addition to an increasing, or structured, portion. So, for example, the equation for an exponential variogram with a nugget may be written as

$$\gamma(h) = c_0 + c \left\{ 1 - \exp\left(-\frac{h}{r}\right) \right\}, \quad (5.34)$$

and an example is shown in Figure 5.10(a). Figure 5.11 shows the simulated fields for an exponential variogram with parameters  $c_0 = 0.333$ ,  $c = 0.667$  and distance parameters,  $r$ , of 5 and 16 as before. The speckled appearance within the patches is the result of the nugget variance.



**Figure 5.10.** Combined (nested) models: (a) single exponential with sill 0.75 plus a nugget variance of 0.25; (b) double spherical with ranges 0.35 and 1.25 and corresponding sills 0.3 and 0.5 plus a nugget variance of 0.2 with the components shown separately.



**Figure 5.11** Simulated fields of values from exponential functions with nugget variance one-third of the total variance, equation (5.34): with distance parameters (a)  $r = 5$ ; (b)  $r = 16$ .

Spatial dependence may occur at two distinct scales, and these may be represented in the variogram as two spatial components. The nested spherical, or double spherical, function is the one that has been used most often in these circumstances. Its equation is

$$\gamma(h) = \begin{cases} c_1 \left\{ \frac{3h}{2a_1} - \frac{1}{2} \left( \frac{h}{a_1} \right)^3 \right\} + c_2 \left\{ \frac{3h}{2a_2} - \frac{1}{2} \left( \frac{h}{a_2} \right)^3 \right\} & \text{for } 0 < h \leq a_1, \\ c_1 + c_2 \left\{ \frac{3h}{2a_2} - \frac{1}{2} \left( \frac{h}{a_2} \right)^3 \right\} & \text{for } a_1 < h \leq a_2, \\ c_1 + c_2 & \text{for } h > a_2, \end{cases} \quad (5.35)$$

where  $c_1$  and  $a_1$  are the sill and range of the short-range component of the variation, and  $c_2$  and  $a_2$  are the sill and range of the long-range component. If it appears to need a nugget then that can be added as a third component, and Figure 5.10(b) shows this combination.

## 5.4 PERIODICITY

A variogram may seem to fluctuate more or less periodically, rather than increase monotonically, and we might try to describe it with a periodic function. The simplest such function is a sine wave, as shown in Figure 5.12(a), with equation

$$\gamma(h) = W \left\{ 1 - \cos \left( \frac{2\pi h}{\omega} \right) \right\}, \quad (5.36)$$

where  $W$  and  $\omega$  are the amplitude and length of the wave, respectively.

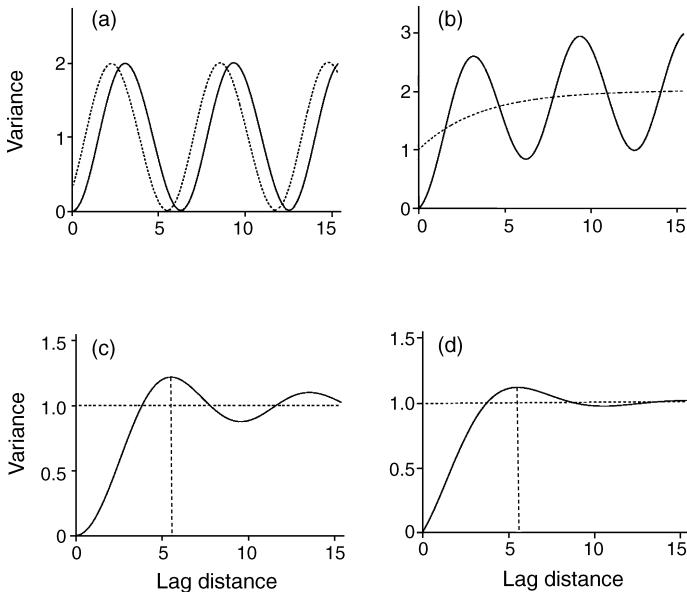
The gradient at the origin is 0, which, as mentioned above, is undesirable. Usually, however, we find that the periodicity is superimposed on some other source of variation and that the combined model increases from the origin more steeply. Figure 5.12(b) shows an example of it superimposed on an exponential function. An example from actual soil survey is illustrated in Chapter 7.

We might be tempted to move the curve along the abscissa to fit the experimental values so that it increases more nearly linearly from lag 0. We have drawn such a function as the dashed line in Figure 5.12(a). In other words, we have introduced a phase shift,  $\phi$ . If we designate the angle  $2\pi h/\omega$  as  $\theta$  for simplicity then the equation becomes

$$\gamma(h) = W \{ 1 - \cos(\theta - \phi) \}. \quad (5.37)$$

Unfortunately, the resulting function is not guaranteed to be CNSD, and so the temptation should be resisted.

Equation (5.36) is valid for one dimension only; it is not CNSD in  $\mathbb{R}^2$  and  $\mathbb{R}^3$ . In two and three dimensions the fluctuation must damp, i.e. become less



**Figure 5.12** Periodic and hole effect models: (a) simple sine wave of length  $2\pi$  (solid) and with phase shift (dashed line); (b) sine wave superimposed on an exponential function; (c) damped sine wave of period  $2.5\pi$  (with maximum marked at  $h \approx 5.6$ ); (d) model with Bessel function  $J_0$ , equation (5.40), with distance parameter set to 10 (giving maximum at  $h \approx 5.6$ ).

pronounced with increasing lag distance. Damping can be achieved by division of the sine or cosine function by a function of the lag distance. Choosing again the simplest function, we can write

$$\gamma(h) = W \left( 1 - \frac{1}{\theta} \sin \theta \right), \quad (5.38)$$

which increases from zero at  $h = 0$ , and this appears in Figure 5.12(c).

This model is valid in one, two and three dimensions. Journel and Huijbregts (1978) show that it has a relative amplitude, which they define as

$$\beta = \frac{1}{c} \{ \max[\gamma(h)] - c \}, \quad (5.39)$$

where  $\max[\gamma(h)]$  is the maximum semivariance of the function, and  $c$  is the *a priori* variance, the horizontal line in Figure 5.12(c). In equation (5.38)  $\beta$  is approximately 0.217, and it occurs where  $2.5\pi h/\omega \approx 5.6$ . It is the maximum for a periodic model in  $\mathbb{R}^3$ .

Usually in such cases the damping is such that only the first undulation is substantial. The corresponding covariance function, see Figure 4.3(f), appears to have a single depression in it: it is said to exhibit a *hole effect*.

Another model that might describe a less pronounced hole effect satisfactorily embodies the Bessel function  $J_0$ :

$$\gamma(h) = c \left\{ 1 - \exp\left(-\frac{h}{r}\right) J_0\left(\frac{2\pi h}{\omega_J}\right) \right\}, \quad (5.40)$$

where  $J_0$  is the Bessel function of the first kind, and  $\omega_J$  is a distance parameter corresponding roughly to wavelength. The maximum is only 0.118 times the variance of the process. Figure 5.12(d) shows an example in which  $\omega_J$  has been set to 10 to give a maximum at the same lag distance, 5.6, as in the truly periodic models.

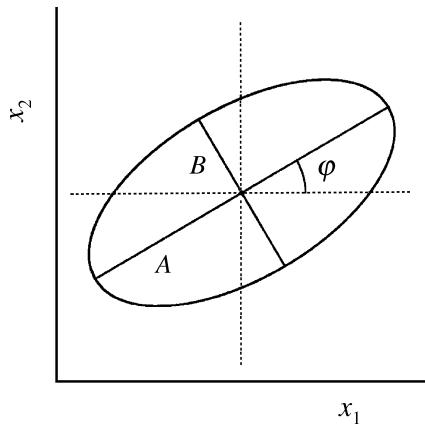
Practitioners should treat wavy experimental variograms with caution. The experimental values are themselves correlated, the more so as the correlation in the original data strengthens. One consequence of this is that any underlying wave-like fluctuation tends to be exaggerated in the estimates: the wave does not damp as much as you might expect, even with moderately long runs (200–300) of data. Before trying to fit a periodic function to such a set of points, the user should ask what evidence there is of periodicity in the phenomenon being investigated. If there is none and the apparent periodicity or hole is weak then do not try to force a periodic model on the variogram. This is a specific case of the more general advice that any variogram model should accord with what you know of the underlying variable, such as the soil, geology, landscape, or sources of pollution that you are studying.

## 5.5 ANISOTROPY

Variation can itself vary with direction. If it can be made to seem isotropic by transformation of the horizontal scales then it is called *geometric* or *affine* anisotropy. Such anisotropy can be taken into account by a simple linear transformation of the rectangular coordinates. It is perhaps best envisaged for a process with a spherical variogram in which the range, instead of being a constant, describes an ellipse in the plane of the lag. This is shown in Figure 5.13, where  $A$  is the maximum diameter of the ellipse, i.e. the range in the direction of greatest continuity (least change with separating distance), and  $B$  is the minimum diameter, perpendicular to the first, and is the range in the direction of least continuity (greatest change with separating distance). The angle  $\varphi$  is the direction in which the continuity is greatest. The equation for transformation is then

$$\Omega(\vartheta) = \{A^2 \cos^2(\vartheta - \varphi) + B^2 \sin^2(\vartheta - \varphi)\}^{1/2}, \quad (5.41)$$

where  $\Omega$  defines the anisotropy, and  $\vartheta$  is the direction of the lag.

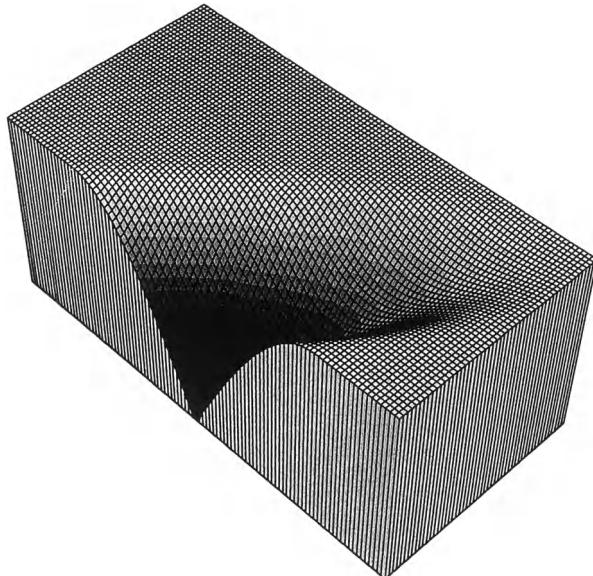


**Figure 5.13** A representation of geometric anisotropy in which the ellipse describes the range of a spherical variogram in two dimensions. The diameter  $A$  is the maximum range of the model,  $B$  is the minimum range, and  $\varphi$  is the direction of the maximum range.

If we insert  $\Omega(\vartheta)$  into the spherical function then we have

$$\gamma(h, \vartheta) = \begin{cases} c \left\{ \frac{3|\mathbf{h}|}{2\Omega(\vartheta)} - \frac{1}{2} \left( \frac{|\mathbf{h}|}{\Omega(\vartheta)} \right)^3 \right\} & \text{for } 0 < |\mathbf{h}| \leq \Omega(\vartheta), \\ c & \text{for } |\mathbf{h}| > \Omega(\vartheta). \end{cases} \quad (5.42)$$

Figure 5.14 shows an example of its surface in two dimensions as a perspective diagram.



**Figure 5.14** Perspective diagram of an anisotropic spherical variogram.

We can derive from equation (5.41) an anisotropy ratio,  $R = A/B$ , and it may be convenient to rewrite equation (5.41) with this replacing  $B$ :

$$\Omega(\vartheta) = \left\{ A^2 \cos^2(\vartheta - \varphi) + \left(\frac{A}{R}\right)^2 \sin^2(\vartheta - \varphi) \right\}^{1/2}. \quad (5.43)$$

This transformation in effect allows the variation to be represented in an isotropic form. It is as if the soil were on a rubber sheet and stretched in the direction parallel to  $B$  until  $B = A$  and the ellipse becomes a circle.

The function  $\Omega(\vartheta)$  can be applied to the power function for unbounded variation:

$$\begin{aligned} \gamma(h, \vartheta) &= [A^2 \cos^2(\vartheta - \varphi) + B^2 \sin^2(\vartheta - \varphi)]^{1/2} |\mathbf{h}|^\alpha \\ &= [\Omega(\vartheta) |\mathbf{h}|]^\alpha. \end{aligned} \quad (5.44)$$

Here the roles of  $A$  and  $B$  are inverted; they are now gradients with  $A$ , the larger, being the gradient in the direction of greatest rate of change and  $B$ , the smaller, being the gradient in the direction of the smallest rate.

## 5.6 FITTING MODELS

The models described above are those that are commonly used for variograms in resource survey. All are theoretically based. Our task now is to fit them to the experimental or sample values. One might have thought that after so many years of geostatistical development and practice—more than 40 since Matheron (1965) published his seminal thesis and more than 30 since the first textbooks appeared—that the task would be straightforward with standard algorithms and well-tried software. If so one would be wrong. Choosing models and fitting them to data remain among the most controversial topics in geostatistics.

There are still practitioners who fit models by eye and who defend their practice with vigour. They may justify their attitude on the grounds that when kriging the resulting estimates are much the same for all reasonable models of the variogram—so why worry about refinement? There are others who fit models numerically and automatically using ‘black-box’ software, often without any choice, judgement or control. This too can have unfortunate consequences. However, there is controversy among those who fit models mathematically about which methods to use and by what criteria they should judge success.

Fitting models is difficult for several reasons, including the following:

- (i) the accuracy of the observed semivariances is not constant.
- (ii) the variation may be anisotropic.

- (iii) the experimental variogram may contain much point-to-point fluctuation.
- (iv) most models are non-linear in one or more parameters.

Items (i)–(iii) make fitting by eye unreliable. The first two impair one's intuition, firstly because the brain cannot judge the weights to attribute to the semivariances, and secondly because one cannot see the variogram in three dimensions without constructing a stereogram or physical model, and for three-dimensional variation one needs a fourth dimension. Scatter, item (iii), usually means that any one of several models might be drawn through the values. It can also lead to unstable mathematical solutions, and it exacerbates the consequences of item (iv) because the non-linear parameters must be found by iteration. Further, at the end one should be able to put standard errors on the estimates of the parameters.

We also warn against a practice, still common, of choosing the dispersion variance in a finite region to estimate the sill of a bounded model for the regional variogram. For such a region the sill is always greater than the dispersion variance. Their relation is shown in Figure 4.4. The curve is the variogram of a second-order stationary process in one dimension of finite length, as on a transect. The variogram is extended to the limit of the transect, and in these circumstances the two shaded portions of the graph should be equal. Clearly the sill, the *a priori* variance of the process, must exceed the dispersion variance, which is estimated by the variance of the data.

We recommend a procedure that embodies both visual inspection and statistical fitting, as follows. First plot the experimental variogram. Then choose, from the models listed above, one or more with approximately the right shape and with sufficient detail to honour the principal trends in the experimental values that you wish to represent. Then fit each model in turn by weighted least squares, i.e. by minimizing the sums of squares, suitably weighted (see below), between the experimental and fitted values. Finally, inspect the result graphically by plotting the fitted model on the same pair of axes as the experimental variogram. Does the fitted function look reasonable? If all the plausible models seem to fit well you might choose from among them the one with smallest residual sum of squares or smallest mean square.

The experimental isotropic variogram on the left-hand side of Figure 5.1 was computed from a fairly small subset of the Broom's Barn data of 87 sites. It shows how much point-to-point fluctuation can occur with rather few data (see Chapter 6), emphasizing the point in item (iii) above. We fitted circular, spherical, exponential and power functions to these experimental values, and they appear in that order as the solid lines in the figure. No one model evidently fits better than any other, and this impression is supported by the small differences between the mean squared residuals (MSR) in Table 5.1. The experimental variogram computed from the full data for Broom's Barn of 434 sites appears on the right-hand side of the figure with the same set of functions fitted. The form of this sequence is simple; it increases smoothly in a gentle

**Table 5.1** Models fitted to the variogram of  $\log_{10}K$  at Broom's Barn Farm for the full data (434 sites) and a subset of these (87 sites), their parameter values, and the mean squared residual (MSR). The symbols are as defined in the text.

Model	$c_0$	$c$	$a/m$	$r/m$	w	$\alpha$	MSR
Circular (434)	0.00512	0.01462	386.6				0.000172
Circular (87)	0.00925	0.01043	362.0				0.000777
Spherical (434)	0.00466	0.01515	432.0				0.000155
Spherical (87)	0.00824	0.01136	376.4				0.000774
Pentaspherical (434)	0.00421	0.01570	514.1				0.000248
Pentaspherical (87)	0.00757	0.01203	434.0				0.000775
Exponential (434)	0.00196	0.01973		251.0			0.001054
Exponential (87)	0.00405	0.01618		130.8			0.000776
Power function (434)	0				0.00173	0.400	0.003295
Power function (87)	0				0.00431	0.251	0.000828

curve from near the origin and seems to flatten near the maximum lag to which it has been computed. There are much larger differences among the mean squared residuals in this case because the smooth form of the experimental variogram enables a much more accurate fit of the model. The spherical function clearly fits best according to its MSR. The MSRs for the full set of data are substantially smaller than are those for the subset, except for the power function. There are also considerable differences among the model parameters for the two variograms; in particular the nugget variances are larger for the variogram of the subset and the distance parameters are all smaller. These differences in the estimates of the parameters are important because they carry through to affect the accuracy of kriged predictions (see Chapter 8).

Never accept a fit without inspecting it afterwards; it might be poor because

- (i) you chose an unsuitable model in the first place;
- (ii) you gave poor estimates of the parameters at the start of the iteration;
- (iii) there was lot of scatter in the experimental variogram; or
- (iv) the computer program was faulty.

Further, bear in mind the advice above, namely, that the model should accord with what you know of the region.

Fitting models in this way is a form of non-linear regression, and you might think of writing your own program to do it. We recommend that unless you are proficient in numerical analysis you do not. There are now several well-tried programs written by professionals that fit models by weighted least squares. These include GenStat (Payne, 2006) in which the standard models listed above are already programmed and which is what we use, and SAS (SAS Institute, 1999). The last uses the Levenberg–Marquardt method, which has almost become a standard for non-linear model fitting (Marquardt, 1963). We give an example of a GenStat program for fitting non-linear models in Appendix B. If you do not have access to any of these programs then you might take the code for the Marquardt algorithm in Fortran from Press *et al.* (1992). Ratkowsky (1983) also tackles the subject in a clear and practical way, and the book includes a suite of subroutines for modelling.

We can call the above approach ‘fit statistically, view afterwards’. Another approach is the reverse: ‘fit visually, statistics afterwards’. Pannatier (1995) takes this route with his program Variowin, which is interactive in a Windows environment. In Variowin you form the experimental variogram from sample data and you display it on the computer’s screen. You select a plausible model from those embodied in the program—there are few—and give starting values for its parameters from which the machine draws a graph. The program simultaneously computes a goodness-of-fit criterion, which is a standardized residual sum of squares. You then adjust the values of the parameters to try to improve the fit visually, and as you do so the program redraws the model in real time and recomputes the goodness-of-fit criterion. It also compares the criterion with the best it has found to date and stores the criterion’s value and the associated values of the parameters if the new fit is better. You terminate the fitting when you are satisfied with the approximation or no further improvement seems possible. In our experience it works well, though never better than GenStat (Webster and Oliver, 1997).

### **5.6.1 What weights?**

We mentioned above that the experimental semivariances,  $\hat{\gamma}(\mathbf{h})$ , vary in their reliability, partly because they are based on varying numbers of paired comparisons,  $m(\mathbf{h})$  in equation (4.40), and partly because the confidence in an estimate of variance decreases as the variance increases. In general, therefore, assigning equal weight to all  $\hat{\gamma}(\mathbf{h})$  is unsatisfactory, especially if the  $m(\mathbf{h})$  vary widely with changing  $\mathbf{h}$ . We can take the latter into account simply by weighting in proportion  $m$ . The inverse relation between the reliability of an estimate of variance and the variance itself led Cressie (1985) to propose a more elaborate weight at a lag  $\mathbf{h}_j$  in the form

$$m(\mathbf{h}_j)/\gamma^{*2}(\mathbf{h}_j),$$

where  $\gamma^*(\mathbf{h}_j)$  is the value of semivariance predicted by the model. McBratney and Webster (1986) refined this further as

$$m(\mathbf{h}_j)\hat{\gamma}(\mathbf{h}_j)/\gamma^*(\mathbf{h}_j),$$

where  $\hat{\gamma}(\mathbf{h}_j)$  is the observed value of the semivariance at  $\mathbf{h}_j$ . Both of the last two schemes tend to give more weight at the shorter lags than does weighting on the numbers of pairs alone, and so the fitting is closer there. This is usually desirable for kriging (see Chapter 8), though it might be less desirable if the aim is to estimate the spatial scale of variation.

The process of fitting must iterate even where all the parameters are linear because the weights in the two schemes depend on the values expected from the model. Our experience is that in most instances there is little change after the first iteration, which is therefore enough.

### 5.6.2 How complex?

Let us return to the question we posed in the beginning of the chapter: how closely should the model follow the fluctuation in the experimental variogram? The best simple model, with few parameters, might fit the experimental variogram poorly, especially if there is much point-to-point scatter. We might seek a more complex model, therefore, bearing in mind that it is almost always possible to improve the fit in the least-squares sense by increasing the numbers of parameters, say  $p$ . We could continue to increase  $p$  until the model fitted perfectly, but clearly that is not a sensible answer. We must compromise between parsimony (few parameters) and close fit (more parameters), and one way of achieving that is to use Akaike's (1973) information criterion (AIC):

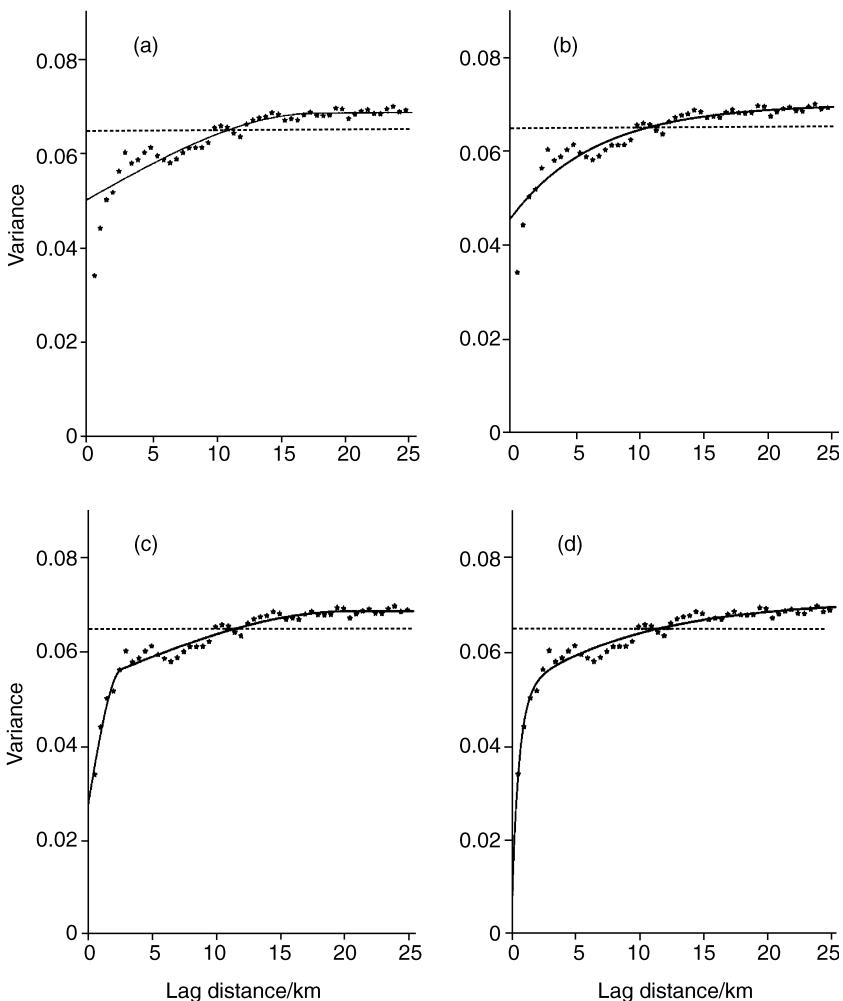
$$\text{AIC} = -2 \ln(\text{maximized likelihood}) + 2(\text{number of parameters}).$$

The AIC is estimated by

$$\widehat{\text{AIC}} = \left\{ n \ln\left(\frac{2\pi}{n}\right) + n + 2 \right\} + n \ln R + 2p, \quad (5.45)$$

where  $n$  is the number of points on the variogram,  $p$  is the number of parameters in the model, and  $R$  is the mean of the squared residuals between the experimental values and the fitted model. We may then choose the model for which  $\widehat{\text{AIC}}$  is least. The quantity in braces is constant for any one experimental variogram, so we need compute only

$$\hat{A} = n \ln R + 2p. \quad (5.46)$$



**Figure 5.15** Experimental variogram of  $\log_{10}\text{Cu}$  in the Borders Region of Scotland with fitted models: (a) single spherical; (b) exponential; (c) double (nested) spherical; (d) double exponential.

Least-squares fitting minimizes  $R$ . If, however,  $R$  is further diminished only by increasing  $p$  ( $n$  is constant) we might discover that  $\hat{A}$  is increased; we should have paid an unacceptable penalty for the greater complexity to achieve the closer fit.

We illustrate the application of the AIC to modelling the variogram of available copper in the topsoil of the Borders Region of Scotland. The data are from an original study by McBratney *et al.* (1982). There were some 2000 values from the eastern portion of the Region. They were transformed to their

**Table 5.2** Models fitted to the variogram of  $\log_{10}\text{Cu}$  in the Borders Region, estimates of their parameters, the mean squared residual (MSR), and the variable part of the Akaike information criterion ( $\hat{A}$ ). The symbols are as defined in the text.

Model	Sills			Distance parameters/km				MSR	$\hat{A}$
	$c_0$	$c_1$	$c_2$	$a_1$	$a_2$	$r_1$	$r_2$		
Spherical	0.05027	0.01805		18.0				0.06822	-128.3
Exponential	0.04549	0.02403				6.65		0.06046	-134.3
Double spherical	0.02767	0.02585	0.01505	2.7	20.5			0.02994	-165.4
Double exponential	0.00567	0.04566	0.01975			0.59	9.99	0.03616	-155.7

common logarithms to stabilize their variances, and an isotropic experimental variogram was computed. It appears as the plotted points in Figure 5.15. Any smooth curve through the points will have an intercept, so we include a nugget variance in the model. By fitting single spherical and exponential functions, with weights proportional to the numbers of pairs, we obtain the curves of best fit shown in Figure 5.15(a) and (b), respectively. Clearly both fit poorly near the ordinate. The values of the parameters, the residual sum of squares and  $\hat{A}$  are listed in Table 5.2, from which it is evident that the exponential function is the better. If we add another spherical or exponential component we obtain the more detailed curves in Figure 5.15(c) and (d), respectively. Now the double spherical is evidently the best, with the smallest mean squared residual. It also has the smallest  $\hat{A}$ , and so in this instance we choose this more elaborate model.

This solution is valid for weighted least-squares fitting provided the weights remain constant, as when they are simply set in proportion to the numbers of paired comparisons.

Webster and McBratney (1989) discuss the AIC in some detail, show its equivalence to an  $F$  test for nested models, and suggest other possible criteria.

Another method for judging the goodness of a model is *cross-validation*. This involves comparing kriged estimates and their variances, and we defer it until we have described kriging.

# 6

## ***Reliability of the Experimental Variogram and Nested Sampling***

We mentioned in Chapters 4 and 5 the importance of estimating the variogram accurately and of modelling it correctly. This chapter deals with factors that affect the reliability of the experimental variogram, in particular the statistical distribution of the data, the effect of sample size on the confidence we can have in the variogram, and the importance of a suitable separating interval between sampling points. In addition to our aim to estimate the regional variogram reliably, we show how to use limited resources wisely to determine suitable and affordable sampling intervals.

### **6.1 RELIABILITY OF THE EXPERIMENTAL VARIOGRAM**

Apart from the matter of anisotropy, equation (4.40) provides asymptotically unbiased estimates of  $\gamma(\mathbf{h})$  for  $Z$  in the region of interest,  $R$ . However, the experimental variogram obtained will fluctuate more or less, and so will its reliability. We now examine factors that affect these.

#### **6.1.1 Statistical distribution**

We made the point in Chapter 2 that sample variances of strongly asymmetric or skewed (typically  $g_1 \leq 1$  or  $g_1 \geq 1$ ) variables are unstable. The estimates obtained with the usual method-of-moments formula for the variogram, equation (4.40), are variances and so are sensitive to departures from a normal distribution. If the distribution of the variable is skewed then the confidence limits on the variogram are wider than they would be otherwise and as a result

the semivariances are less reliable. Skewness can result from a long upper or lower tail in the underlying process or from the presence of a secondary process that contaminates the primary process—values from the latter may appear as outliers. Kerry and Oliver (2007a, 2007b) have studied the effects of asymmetry in the underlying process and outliers on the variogram using simulated fields. We summarize their results below.

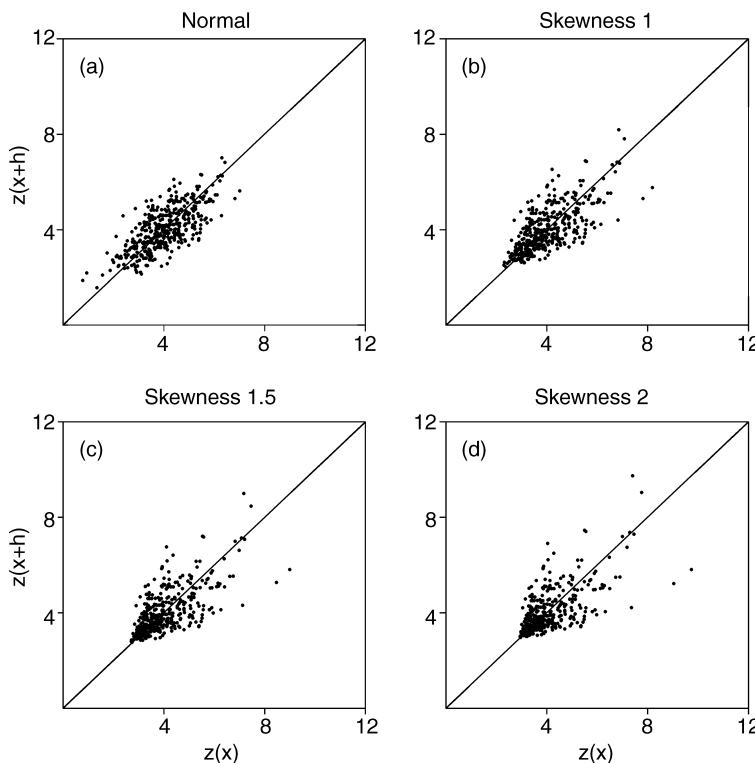
Methods of estimating variograms reliably from skewed data have been sought, and it is clear that the cause of asymmetry affects what one should do. If the skewness coefficient exceeds the bounds given above then the histogram or box-plot should be examined to reveal the detail of the asymmetry. In addition to these usual graphical methods, you can identify exceptional contributions to the semivariances by drawing an **h**-scattergram for a given lag, **h**. As described in Chapter 4, an **h**-scattergram is a graph in which the  $z(\mathbf{x})$  are plotted against the  $z(\mathbf{x} + \mathbf{h})$  with which they are compared in computing  $\hat{\gamma}(\mathbf{h})$ . In general, the plotted points appear as more or less inflated clusters, as in the usual kind of scatter graph.

### **Underlying asymmetry or skewness**

Where asymmetry arises from a long tail, especially a long upper tail, in the distribution ‘standard best practice’ has been to transform the data, as described in Chapter 2. The variogram is then computed on the transformed data. Transformation is not essential, however; the variogram computed from the original data and predictions using it are unbiased, though they are not necessarily the most precise. Perhaps more surprising is that the characteristic form of the variogram may be changed little by transformation. So, you should examine the experimental variograms of both raw and transformed data before deciding which to work with.

Kerry and Oliver (2007a) explored the effects of varying skewness and sample size, and of different transformations on random fields created by simulated annealing (see Chapter 12 for a description of the method). They simulated values on a square 5-m grid of 1600 points from a spherical function (equation (5.24)), with a range,  $a$ , of 75 m, a total sill variance,  $c_0 + c$ , of 1, and nugget:sill ratios of 0, 0.25, 0.5, 0.75 and 1. They simulated similar fields of 400 points and 100 points with grid intervals of 10 m and 20 m, respectively. Values in the fields were raised to a power  $\alpha$  to create a long upper tail in the distribution. Five values of  $\alpha$  were used to give skewness coefficients,  $g_1$ , of 0.5, 1.0, 1.5, 2.0 and 5.0. Here we illustrate what can happen with their results for  $a = 75$  m,  $c_0 = 0$  and  $c = 1$ .

Figure 6.1 shows the **h**-scattergrams at lag 10 m (lag 1) from four fields simulated on a 10 m grid. Each field has a unique coefficient of skewness,  $g_1 = 0, 1.0, 1.5$  and  $2.0$ , caused by underlying asymmetry. The scatter of points for the normal distribution is clustered fairly tightly along the diagonal line in Figure 6.1(a). As the coefficient of skewness increases, the scatter becomes more



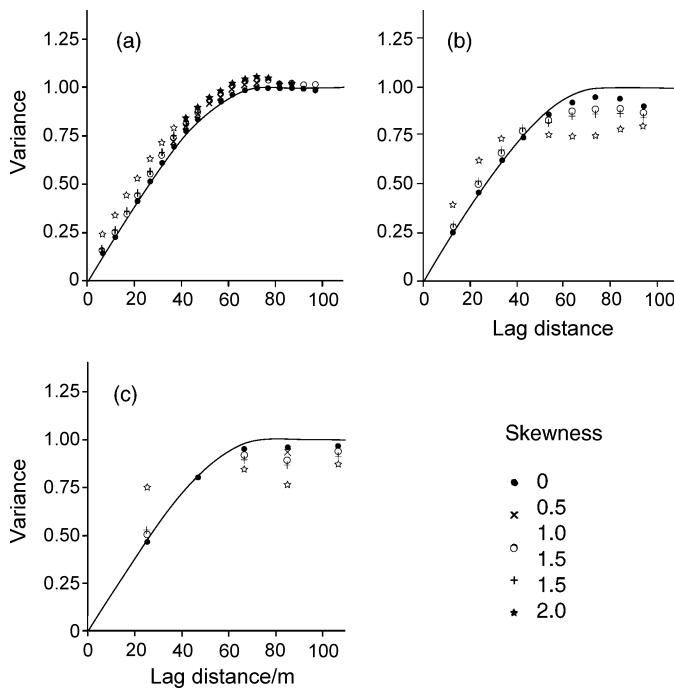
**Figure 6.1** The  $\mathbf{h}$ -scattergrams for simulated fields of 400 points with underlying asymmetry resulting in coefficients of skewness of (a) 0, (b) 1.0, (c) 1.5, (d) 2.0.

dispersed for the larger values in the tail of the distribution reflecting the positive skew, Figure 6.1(b)–(d). Table 6.1 lists the values of  $\hat{\rho}(\mathbf{h})$  and  $\hat{\gamma}(\mathbf{h})$  for the comparisons from these fields. The correlation coefficients decrease somewhat with increasing skewness, and the semivariances increase correspondingly. The effects of underlying asymmetry at the first lag interval are evident, but they are not remarkable.

**Table 6.1** Autocorrelation coefficients and semivariances for lag distance 10 m (lag 1) computed from data simulated on a 10-m grid with four degrees of underlying skewness.

Skewness coefficient	Autocorrelation coefficient	Semivariance
0	0.7188	0.2700
1.0	0.6990	0.2863
1.5	0.6860	0.2984
2.0	0.6739	0.3093

Omnidirectional variograms computed from the simulated fields by the method of moments are displayed in Figure 6.2. Figure 6.2(a) shows that as asymmetry increases the change in shape of the variogram is small for the field on a 5 m grid. This was true even for  $g_1 = 5.0$  (not shown). For the sample size of 400 (10 m grid) the change in the shapes of the variograms is not large, Figure 6.2(b), except for  $g_1 = 2.0$ . For the smaller coefficients the semivariances are close to the generating function for the first five lags, but beyond these they depart progressively from the sill of the generating model. For the sample size of 100 (20-m grid), shown in Figure 6.2(c), the semivariances at the first two lags are similar to the generating model for  $g_1 = 0.5, 1.0, 1.5$  and  $2.0$ , but beyond this they depart progressively more from the sill variance of 1. The variogram computed from data with  $g_1 = 5.0$  appeared as pure nugget. Evidently the effect of asymmetry decreases as the sample size increases; it is greatest for the sample of 100 points and least for that with 1600.



**Figure 6.2** Experimental variograms (shown by point symbols) computed from simulated fields of various sizes: (a) 1600 points (5-m grid), (b) 400 points (10-m grid), (c) 100 points (20-m grid), with skewness caused by underlying asymmetry. The solid lines represent the spherical generating function.

Kerry and Oliver (2007a) found that transformation conferred little advantage for large sets of data. Therefore, you should assess the desirability of any transformation by comparing the variograms of raw and transformed data visually before deciding whether to transform. In general, if there are no marked differences between the shapes of the experimental variograms then work on the raw data. This advice applies in particular if your aim is prediction, for then the predicted values will be on the original scale of measurement, which is what most practitioners want, and no back-transformation is needed (see Chapter 8).

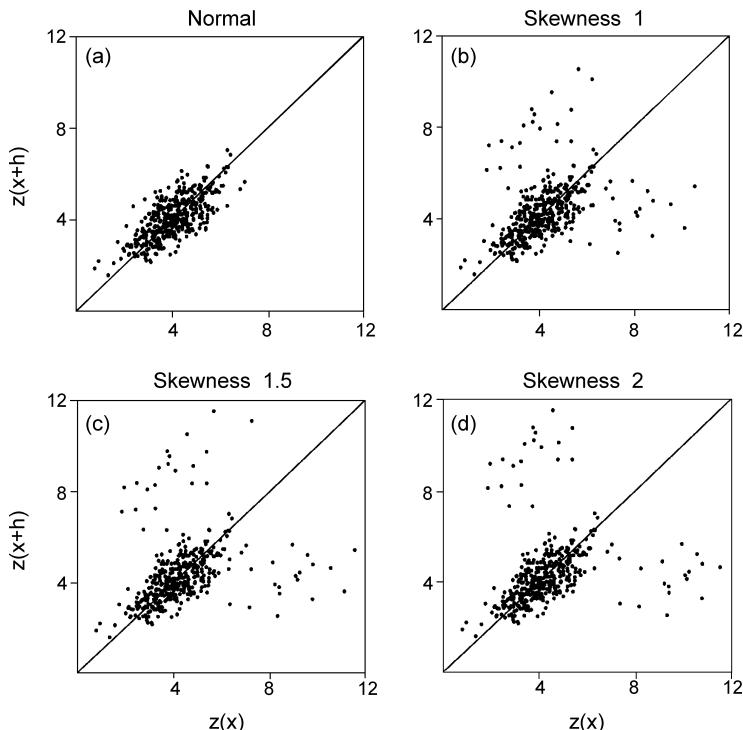
## Outliers

The variogram is sensitive to outliers and to extreme values in general. Histograms and box-plots, as described in Chapter 2, will usually reveal outliers in the marginal distributions if they are present. All outliers must be regarded with suspicion and investigated. Erroneous values should be corrected or excised, and values that remain suspect are best removed, too. If by removing one or a very few values you can reduce the skewness then it is reasonable to do so to avoid the need for transformation or the use of robust variogram estimators. For contaminated sites it is the exceptionally large values that are often of interest. In this situation the variogram can be computed without the outliers to ensure its stability, and then the values can be reinstated for estimation and other analyses. Some practitioners remove the 98th or even the 95th percentiles. This is too prescriptive in our view, and only values that are obvious outliers should be removed.

It will be evident from equation (4.40) that each observed  $z(\mathbf{x})$  can contribute to several estimates of  $\gamma(\mathbf{h})$ . So one exceptionally large or small  $z(\mathbf{x})$  will tend to swell  $\hat{\gamma}(\mathbf{h})$  wherever it is compared with other values. The result is to inflate the average. The effect is not uniform, however. If an extreme is near the edge of the region it will contribute to fewer comparisons than if it is near the centre. The end point on a regular transect, for example, contributes to the average just once for each lag, whereas points near the middle contribute many times. If data are unevenly scattered then the relative contributions of extreme values are even less predictable. The result is that the experimental variogram is not inflated equally over its range, and this can add to its erratic appearance.

Kerry and Oliver (2007b) examined the effect of outliers on the variogram for skewness coefficients 0, 0.5, 1.0, 1.5, 2.0 and 3.0, and for randomly located and grouped outliers. They created normally distributed data by simulated annealing as above for the same sizes of field. These primary fields were then contaminated by randomly located and spatially aggregated outliers from a secondary process.

Figure 6.3 shows four  $\mathbf{h}$ -scattergrams at lag 10 m (lag 1), from four fields simulated on a 10-m grid with randomly located outliers to give coefficients of skewness  $g_1 = 0, 1.0, 1.5$  and  $2.0$ . The scatter of points becomes more



**Figure 6.3** The  $h$ -scattergrams for a simulated primary Gaussian field of 400 points contaminated by outliers to give coefficients of skewness (a) 0, (b) 1.0, (c) 1.5, (d) 2.0.

pronounced as the skewness increases from a normal distribution. For a coefficient of skewness of 1.0, Figure 6.3(b), there is already a wide scatter of points around a central core that represents the primary Gaussian population. For  $g_1 = 2.0$  there are now two distinct groups of points, separated from the main group, that reflect the contaminating population. Table 6.2 supports these

**Table 6.2.** Autocorrelation coefficients and semivariances for lag distance 10 m (lag 1) computed from data simulated on a 10-m grid with skewness caused by outliers.

Skewness coefficient	Autocorrelation coefficient	Semivariance
0	0.7188	0.270
1.0	0.3938	1.013
1.5	0.3122	1.429
2.0	0.2476	1.942

graphical observations; the correlation coefficients diminish markedly as skewness caused by outliers increases, and also the semivariances increase dramatically.

The results in Figures 6.1 and 6.3 show how different the effects are from different causes of asymmetry. They add strength to the statement above that different solutions are likely to be required.

Kerry and Oliver (2007b) computed variograms as before by Matheron's method of moments and also by three robust variogram estimators proposed by Cressie and Hawkins (1980), Dowd (1984) and Genton (1998). The estimator of Cressie and Hawkins (1980),  $\hat{\gamma}_{CH}(\mathbf{h})$ , essentially damps the effect of outliers from the secondary process. It is based on the fourth root of the squared differences and is given by

$$2\hat{\gamma}_{CH}(\mathbf{h}) = \frac{\left\{ \frac{1}{m(\mathbf{h})} \sum_{i=1}^{m(\mathbf{h})} |z(\mathbf{x}_i) - z(\mathbf{x}_i + \mathbf{h})|^2 \right\}^{\frac{1}{2}}}{0.457 + \frac{0.494}{m(\mathbf{h})} + \frac{0.045}{m^2(\mathbf{h})}}. \quad (6.1)$$

The denominator in equation (6.1) is a correction based on the assumption that the underlying process to be estimated has normally distributed differences over all lags.

Dowd's (1984) estimator,  $\hat{\gamma}_D(\mathbf{h})$ , and Genton's (1998),  $\hat{\gamma}_G(\mathbf{h})$ , estimate the variogram for a dominant intrinsic process in the presence of outliers. Dowd's estimator is given by

$$2\hat{\gamma}_D(\mathbf{h}) = 2.198 \{ \text{median}|y_i(\mathbf{h})|\}^2, \quad (6.2)$$

where  $y_i(\mathbf{h}) = z(\mathbf{x}_i) - z(\mathbf{x}_i + \mathbf{h})$ ,  $i = 1, 2, \dots, m(\mathbf{h})$ . The term within the braces of equation (6.2) is the median absolute pair difference (MAPD) for lag  $\mathbf{h}$ , which is a scale estimator only for variables where the expectation of the differences is zero. The constant in the equation is a correction for consistency that scales the MAPD to the standard deviation of a normally distributed population.

Genton's (1998) estimator is based on the scale estimator,  $Q_N$ , of Rousseeuw and Croux (1992, 1993), where in the general case  $N$  is the number of data. The quantity  $Q_N$  is given by

$$Q_N = 2.219 \{ |X_i - X_j|; i < j \}_{\binom{H}{2}}, \quad (6.3)$$

where the constant 2.219 is a correction for consistency with the standard deviation of the normal distribution, and  $H$  is the integral part of

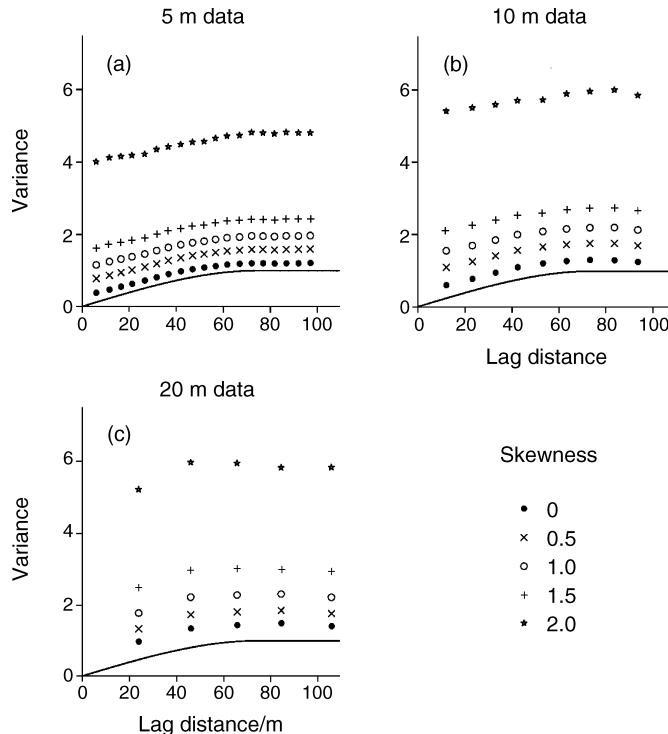
$(N/2) + 1$ . Genton's estimator uses equation (6.3) as an estimator of scale applied to the differences at each lag; it is given by

$$2\hat{\gamma}_G(\mathbf{h}) = \left[ 2.219 \{ |y_i(\mathbf{h}) - y_j(\mathbf{h})|; i < j \} \binom{H}{2} \right]^2, \quad (6.4)$$

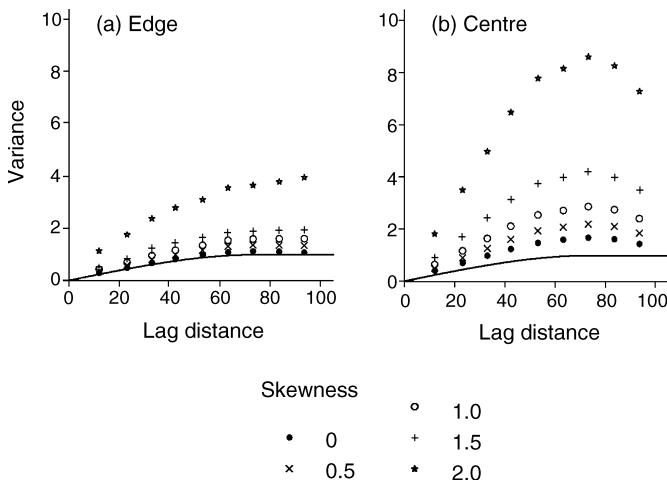
but now with  $H$  being the integral part of  $\{m(\mathbf{h})/2\} + 1$ .

The underlying assumption of robust variogram estimators is that the data have a contaminated normal distribution. Lark (2000) showed that these estimators should be used for such distributions and not for those where the primary process has a simple underlying asymmetry.

Figure 6.4 shows the experimental variograms for the three sizes of field with randomly located outliers and for several coefficients of skewness. As the



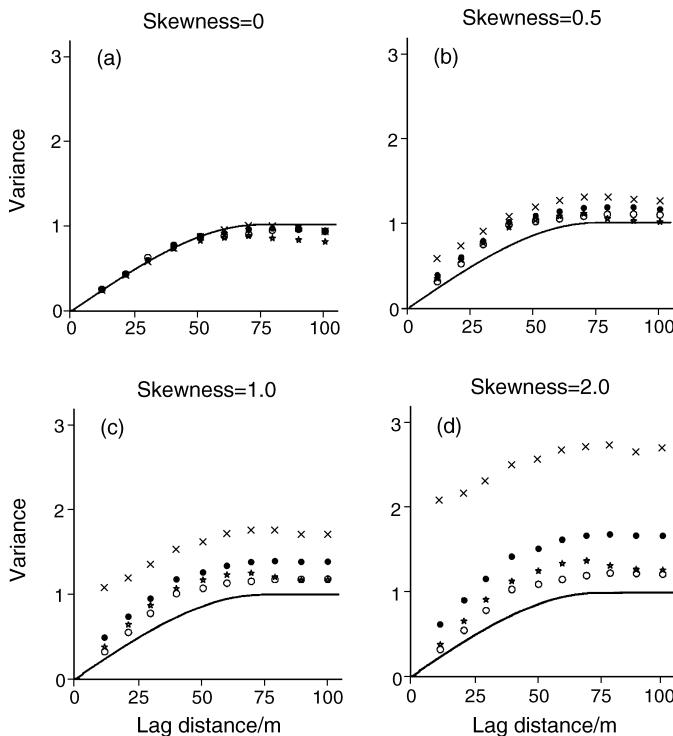
**Figure 6.4** Experimental variograms (shown by point symbols) computed from simulated fields of various sizes: (a) 1600 points (5-m grid); (b) 400 points (10-m grid); (c) 100 points (20-m grid), with skewness caused by randomly located outliers. The solid lines represent the spherical generating function.



**Figure 6.5** Experimental variograms (shown by point symbols) computed from fields of 400 points (10-m grid) contaminated by spatially aggregated outliers (a) at the edges; (b) at the centre of fields, and for five coefficients of skewness. The solid lines represent the spherical generating function.

skewness increases, the nugget and sill variances increase markedly. The nugget:sill ratios increase as skewness increases, even though the fields were generated by a variogram function with zero nugget. The size of field has far less effect than that observed where asymmetry was caused by a long tail in the distribution, as in Figure 6.2. For skewness coefficients up to 1.5 the variograms tend to retain their shape, but for  $g_1 = 3.0$  (not shown) the variograms were almost pure nugget. Variograms computed from fields with aggregated outliers are very different from those where the outliers were randomly located. Figure 6.5(a) shows the variograms computed from the fields on a 10 m grid (400 values) with outliers aggregated near the edge and for several coefficients of skewness. The nugget variances are close to zero, but the sill variances increase with increasing skewness. Nevertheless, these sill variances are less than those for the randomly located outliers, Figure 6.4(b). For outliers grouped near the centres of the fields on the 10-m grid, shown in Figure 6.5(b), the variograms have small or zero nugget variances. The sill variances, however, increase more with increasing skewness than do those in Figure 6.5(a); they are more similar to those for the randomly located outliers, Figure 6.4(b).

Kerry and Oliver (2007b) transformed the data to square roots and logarithms, but these transformations were not as effective in dealing with the observed effects of the outliers as the robust variogram estimators. Figure 6.6(a) shows that Matheron's method-of-moments estimator and the three robust estimators described above result in similar experimental variograms for the normal field. The method-of-moments variogram is closest to the generating



**Figure 6.6** Experimental variograms computed by Matheron's method-of-moments estimator ( $\times$ ) and Cressie and Hawkins's ( $\bullet$ ), Dowd's ( $\circ$ ) and Genton's ( $\star$ ) robust estimators from fields of 400 points (10-m grid) with skewness coefficients of (a) 0, (b) 1.0, (c) 1.5, (d) 2.0, caused by randomly located outliers. The solid lines are of the spherical generating function.

model, however. Figure 6.6(b)–(d) shows that as skewness increases Matheron's variogram departs much more from the generating function than the robust variograms. Of the latter, Dowd's and Genton's variograms remain closer to the original model than does that of Cressie and Hawkins.

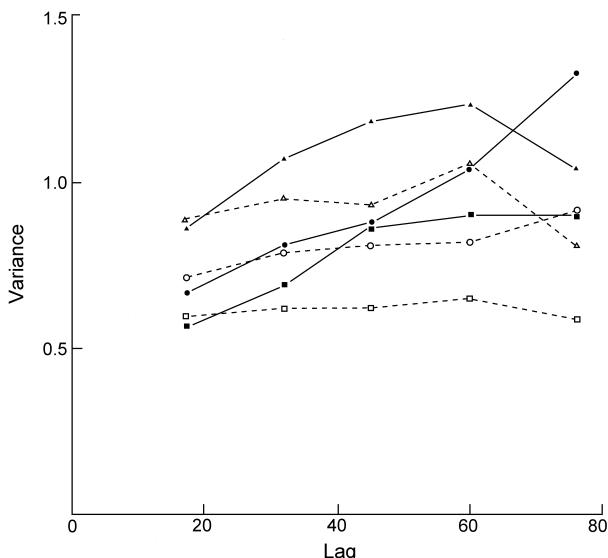
Kerry and Oliver (2007b) concluded from their results that skewness caused by outliers must be dealt with regardless of the number of data. Furthermore, the results suggested practitioners should act when the skewness,  $g_1$ , exceeds 0.5. Robust estimators provide a solution, but they did not perform equally well in all the situations examined. The current 'best practice' approach of removing outliers before computing the variogram appears to be the most appropriate where they are randomly located and will not be returned to the data for kriging. However, where outliers are crucial in an investigation, as on contaminated sites, practitioners should compute several robust variograms and compare them by cross-validation (see Chapter 8).

### 6.1.2 Sample size and design

The reliability of the experimental variogram is affected not only by the statistical distribution of the data but also by the size of the sample (or its inverse, the density of data), and the configuration or design of the sample.

Hundreds, if not thousands, of experimental variograms are now displayed in published papers, reports, theses and books. They are derived from samples of as few as 24 individual measurements up to several thousands, though typically they are computed from 100–200 data. Those based on fewer than 50 data are often erratic sequences of experimental values with little or no evident structure. Figure 6.7 shows some examples. As the size of sample is increased such scatter decreases and the form of the variogram becomes clearer: the plotted points tend to be closer to an increasing line. Evidently the larger is the sample from which the variogram is computed the more precisely it is estimated. In most instances, however, the precision is unknown, and we cannot determine how large a sample to take to achieve some desired precision. The classical formulae for determining confidence intervals cannot be applied unless the sampling itself is designed for the purpose, as suggested by Brus and de Gruijter (1994). Practitioners who attempt to assign error to their estimates based on these formulae are misguided. There are several reasons why:

- (i) the same data are used more than once in each estimate;
- (ii) the estimates are correlated;
- (iii) the sampling is not sufficiently randomized.



**Figure 6.7** Graphs of sample variograms with 49 data. Reproduced with permission *Journal of Soil Science*, Vol. 43, © Blackwell Publishing.

Before we proceed further we must be clear which variogram we are attempting to estimate from the experimental one. In Chapter 4 we identified two distinct functions, one the theoretical variogram and the other the local or regional variogram. The first is the variogram of the underlying stochastic process, whereas the local variogram is that of the particular realization in the region and called the non-ergodic variogram by Brus and de Gruijter (1994). An experimental variogram may contain error deriving from different realizations of the random function or from different samples of the particular realization, or both. In the first case the error arises from fluctuation in the generator, whereas in the second the error arises from the sampling. We take the view here that for most practical purposes we are concerned with just one realization in the region, so we should try to estimate the sampling error expressed in the estimation variance or confidence limits.

Matheron (1965) gave a formula to provide a first approximation to the estimation variances of the local variogram:

$$\text{var}[\hat{\gamma}_R(\mathbf{h})] \approx \frac{1}{N'} 4\gamma(\mathbf{h})\sigma_D^2, \quad (6.5)$$

where  $\gamma(\mathbf{h})$  is the value of the theoretical variogram at lag  $\mathbf{h}$ ,  $\hat{\gamma}_R(\mathbf{h})$  is the estimate of the regional semivariance at that lag, and  $\sigma_D^2$  is the total variance in the region, i.e. the dispersion variance. Matheron describes  $N'$  as the number of points effectively used, i.e. the number of points that are superposed in the intersection of the region with itself when translated by the vector  $\mathbf{h}$ . For a regular transect of length  $M$  it is the number of paired comparisons  $(M - h)$  contributing to the estimate of  $\gamma_R(\mathbf{h})$ . It is from this that confusion has arisen about the number of observations needed to estimate the variogram reliably and in particular a suggested minimum of 30–50 paired comparisons for any one  $\hat{\gamma}_R(\mathbf{h})$  (Journel and Huijbregts, 1978). The advice seems to have been intended for one dimension, but unfortunately it has been applied widely in two dimensions and has given practitioners a false sense of security when computing variograms from small samples.

Muñoz-Pardo (1987) pursued Matheron's idea for estimating the estimation variances for variograms with a sill (bounded). He derived the following expression for the estimation variance of a semivariance:

$$\begin{aligned} \text{var}[\hat{\gamma}_R(\mathbf{h})] &= \frac{1}{2S'^2} \int_S' \int_S f(\mathbf{x}, \mathbf{y}, \mathbf{h}) d\mathbf{x} d\mathbf{y} + \frac{1}{2N'^2(\mathbf{h})} \sum_{i=1}^{N'(\mathbf{h})} \sum_{j=1}^{N'(\mathbf{h})} f(\mathbf{x}_i, \mathbf{x}_j, \mathbf{h}) \\ &\quad - \frac{1}{N'(\mathbf{h})S'} \sum_{i=1}^{N'(\mathbf{h})} \int_S' f(\mathbf{x}_i, \mathbf{x}, \mathbf{h}) d\mathbf{x}, \end{aligned} \quad (6.6)$$

where

$$f(\mathbf{x}, \mathbf{y}, \mathbf{h}) = \{\gamma(\mathbf{x} - \mathbf{y} + \mathbf{h}) + \gamma(\mathbf{x} - \mathbf{y} - \mathbf{h}) - 2\gamma(\mathbf{x} - \mathbf{y})\}^2 \quad (6.7)$$

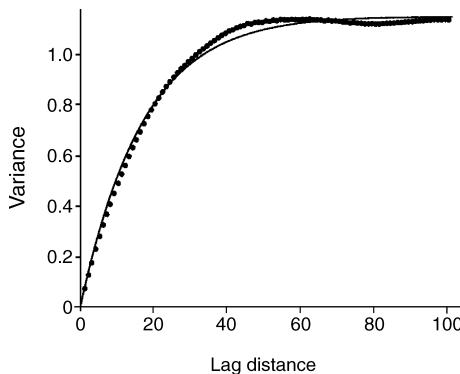
for any value of  $i$  and  $j$ . In equation (6.6)  $\hat{\gamma}_R(\mathbf{h})$  denotes the estimated value of the regional variogram at lag  $\mathbf{h}$ ,  $S'$  is the area of intersection when the region is translated by the vector  $\mathbf{h}$ ,  $N'$  is the number of sampling points in the intersection, and  $\mathbf{x}$  and  $\mathbf{y}$  are two points that describe the region independently. Muñoz-Pardo solved the equation by numerical integration. He showed that the estimation variance depended on the effective range of the variogram in relation to the size of the region as well as on the size of the sample.

One way of obtaining confidence limits on variograms is by Monte Carlo methods (Webster and Oliver, 1992). There are two possible approaches, depending on which variogram (theoretical or regional) we are concerned with. If it is the first then one simulates many realizations from a particular model and computes the experimental variogram of each, as did McBratney and Webster (1986), Taylor and Burrough (1986) and Shafer and Varljen (1990). The result will show the fluctuation arising from the generator, and the quantiles of the observed values for each lag would be reasonable estimates of the confidence intervals for new realizations.

Environmental scientists are more often concerned with single particular realizations, which they must sample, and so they are interested in the sampling fluctuation. Here the Monte Carlo approach is to generate a single large field of 'data' from a plausible model of the variation in the region, sample repeatedly from it, and for each sample compute the sample variogram. The variation in the variograms thereby obtained will be sampling fluctuation, and the quantiles of the semivariances may be used as confidence limits on the regional variogram.

Webster and Oliver (1992) explored this approach by simulating large autocorrelated random fields, which they then sampled on grids and transects of varying size and density with random starting points. We illustrate the approach here with one of their examples.

A field of 65 536 random values on a  $256 \times 256$  square grid with unit interval was generated by sequential Gaussian simulation (Deutsch and Journel, 1992) and an exponential variogram (equation (5.26)) with distance parameter  $r = 16$  units. It is displayed in Figure 5.8(b). We can imagine it as one of exchangeable K in the soil. There are distinct patches with large and small values showing that spatial dependence extends on average to about 50 units, which is about 3 times the distance parameter (as explained in Chapter 5). The variogram from the exhaustive data is close to the generating function (Figure 6.8). The field was then sampled on regular square grids with the sample sizes and sampling intervals in Table 6.3. No position was used more than once, and so no comparisons were duplicated.



**Figure 6.8** The exhaustive variogram computed from Figure 5.5(a).

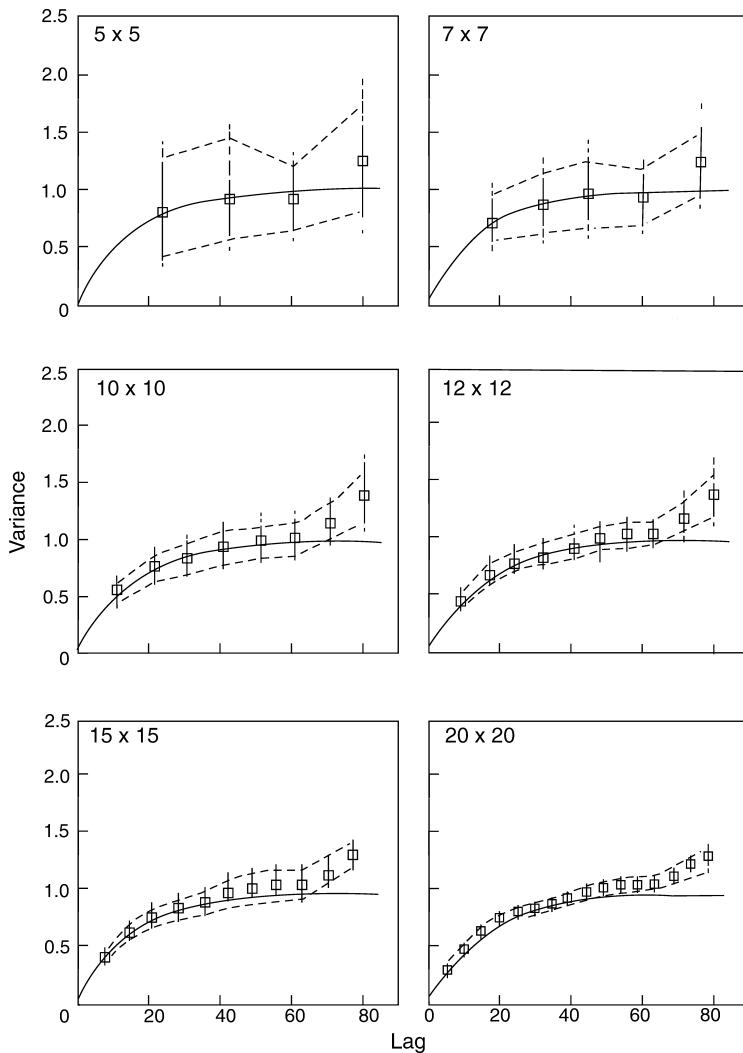
We computed the variograms for all the samples, and in Figure 6.9 we plot the results on the same set of axes. The variograms of the samples of 25 show the wide spread of estimates around the variogram of the generating function. The dotted lines are the 90th percentiles, i.e. the symmetrical 90% confidence limits. The other graphs in the sequence show how the confidence intervals narrow as the size of the sample increases. A sample of 100 points appears to give moderate confidence, but to attain satisfaction at least 144 measurements seem necessary. Increasing the sample to 225 points provides rather little improvement, whereas 400 data enable the variogram to be estimated with great precision.

The results can be summarized in a graph of the standard deviation of the observed semivariances against the size of the sample (Figure 6.10). The standard deviation decreases, and the 90th percentiles narrow, with increasing sample size. We can also judge from Figure 6.10 approximately what size of sample to use to achieve some particular confidence.

The results show clearly that sample variograms from only 25 and 49 data have wide confidence intervals, and are therefore imprecise. Samples of 100 might be acceptable in some circumstances, and ones of 144 are likely to be adequate, at least with normally distributed isotropic data as in the generated fields. Variograms computed from samples of 225 will almost certainly be reliable, and samples of 400 seem extravagant. Based on this evidence we recommend that you have no fewer than 100 sampling points and ideally 150 to estimate the variogram reliably in two dimensions if the variation is isotropic.

**Table 6.3** Sample size, spacing and number of iterations.

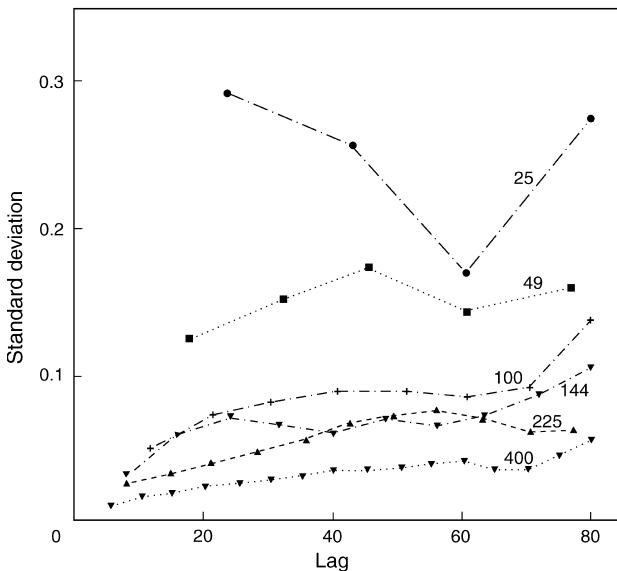
Size	25	49	199	144	225	400
Interval	20	15	10	8	7	5
Iterations	100	100	100	64	49	25



**Figure 6.9** Semivariances computed on samples of various sizes from Figure 5.5(a) and 90% confidence limits obtained. Reproduced with permission *Journal of Soil Science*, Vol. 43, © Blackwell Publishing.

For anisotropic variation we recommend at least 250 sample points because of the need to compute variograms in several directions.

The results also show the importance of interpreting  $N'$  of equation (6.5) correctly. Forty-nine points on a square grid gave us 84–240 paired comparisons, which would have seemed more than enough against the 30–50 comparisons regarded by some authorities as adequate. With 100 points there were 180–774 comparisons, and still the variograms were somewhat erratic.



**Figure 6.10** Standard deviations of semivariances for the several grid samplings from Figure 5.5(a). Reproduced with permission *Journal of Soil Science*, Vol. 43, © Blackwell Publishing.

Brus and de Gruijter (1994) viewed the problem differently. They pointed out that until you know the variogram accurately you cannot simulate realistic fields of values from which to sample. If the variogram used for the simulation has been estimated from few data then the realization generated might represent the true situation in the region poorly and lead to misleading confidence limits. They proposed a procedure based on classical sampling theory. For each lag,  $\mathbf{h}$ , they repeatedly selected pairs of points, the first of each pair at random and the second determined by  $\mathbf{h}$ . In the simplest design, simple random sampling (Chapter 2), the first point is chosen without regard to any other, and all points have an equal chance of inclusion. If the variation is isotropic then the second point can be chosen at distance  $h = |\mathbf{h}|$  from the first but in a random direction. The mean of the individual squared differences obtained with equation (4.40),  $\hat{\gamma}(\mathbf{h})$ , is an unbiased estimate of  $\gamma(\mathbf{h})$ .

Since the pairs of points are chosen independently of one another the calculated squared differences are independent, and so the sampling variance of  $\hat{\gamma}(\mathbf{h})$  can be estimated by the classical formula. If we denote the squared difference at lag  $\mathbf{h}$  by  $d^2(\mathbf{h})$  then, following Cochran (1977), we can write the variance of the semivariance as

$$\begin{aligned}\text{var}[\hat{\gamma}(\mathbf{h})] &= \text{var}[0.5d^2(\mathbf{h})]/m(\mathbf{h}) \\ &= \frac{\frac{1}{4}\sum_{i=1}^{m(\mathbf{h})}\{d_i^2(\mathbf{h}) - \bar{d}^2(\mathbf{h})\}^2}{m(\mathbf{h})\{m(\mathbf{h}) - 1\}},\end{aligned}\quad (6.8)$$

where  $\bar{d}^2(\mathbf{h})$  is the mean of the squared difference at lag  $\mathbf{h}$ . Further, choosing fresh pairs of points for each  $\mathbf{h}$  or  $h$  provides independent estimates for the different lags.

As we saw in Chapter 2, simple random sampling is inefficient, and the precision or efficiency can be improved by better design. Brus and de Gruijter (1994) elaborate the procedure for stratified sampling and give the formulae for the estimator and the estimation variance. The formulae can be modified for other designs.

The estimation variance has still to be converted into confidence limits, and for this one must assume a distribution. It is not immediately evident what that distribution should be. One might expect the individual  $d^2(\mathbf{h})$  to be distributed as  $\chi^2$ . Their means, however, are likely to approach normality with increasing  $m(\mathbf{h})$  in accordance with the central limit theorem. Brus and de Gruijter calculated limits on this assumption but found that it was not entirely satisfactory for the fairly small  $m(\mathbf{h})$  in their study: they obtained several negative lower limits at the 90% level, suggesting that the confidence interval is not symmetric, at least for the small samples they took. This contrasts with our finding, with larger samples, that limits were approximately symmetrical.

Despite this weakness, the method proposed by Brus and de Gruijter gives sound unbiased estimates of the sampling variance of  $\gamma(\mathbf{h})$ , but large samples are needed to obtain precise estimates. In addition, the sampling scheme with pairs of points scattered irregularly and unevenly is inefficient for subsequent kriging (Chapter 8).

Although the above approaches to the problem differ, both show that the confidence intervals are very wide with small samples: you need a large sample to estimate the variogram by Matheron's method of moments reliably.

Pardo-Igúzquiza (1998) suggested that 'a few dozen data may suffice' to estimate variogram parameters by residual maximum likelihood (REML) because of the efficiency of the method; see Section 9.2 for more detail. In this approach the model parameters are estimated directly from the generalized increments of a covariance matrix of the full data. As a consequence there is no smoothing of the spatial structure because there is no *ad hoc* definition of lag classes. Kerry and Oliver (2007c) compared variograms computed by the method of moments and REML as described by Pardo-Igúzquiza (1997) for various numbers of empirical data. Their results show that where there are fewer than 100 data, but more than 50, the REML variograms gave more accurate predictions as assessed by cross-validation (see Chapter 8) than did the method-of-moments variograms. Nevertheless, even with REML variograms the accuracy of prediction decreased when there were fewer than 100 sites, and practitioners should still aim for at least 100 data for accurate predictions.

Practitioners might wonder why computing variograms by REML is not a standard approach. There are several drawbacks to the method:

- the need for second-order stationarity;
- the very limited range of variogram functions that can be fitted by the readily available software;

- it is computationally intensive;
- the limitations to the methods for maximizing the log-likelihood function—see equation (9.25).

In spite of these, Kerry and Oliver (2007c) concluded that the REML variogram is valuable where it is impractical to obtain as many as 100 data.

### **6.1.3 Sample spacing**

In general, as the size of sample increases so the spacing between sampling points decreases for a given region. Nevertheless, we cannot simply allow the sample spacing to be dictated by the size of sample. The spacing must relate to the scale or scales of variation in the region. Otherwise we might sample too sparsely to identify correlation. We should therefore know roughly the spatial scale of variation in  $Z$  so as to choose a sensible sampling density.

Some variables, such as vegetation, have visible patterns, and their spatial scales are obvious. Many properties of soil, rocks, atmosphere and water, on the other hand, are invisible, and so one cannot judge the spatial scales on which they vary without first sampling. They can also vary on scales that differ by several orders of magnitude simultaneously, as described in Chapter 4. In some instances an approximate scale of variation can be judged from that of other features, such as landform or vegetation, but often it is more elusive.

Let us consider the following situations.

1. *Terra incognita.* If we know nothing of the pattern or scale of the variation then it is difficult to choose a sampling interval rationally. A large interval might be too large to capture the autocorrelation. If we choose a small interval then we might have to restrict the area sampled to stay within a budget and fail to estimate long-range variation. If we were to sample a whole region densely and the variation turned out to be entirely long-range then we should have wasted money trying to estimate short-range variation. We want some means of estimating, even roughly, the spatial scale of variation effectively and economically.
2. We have data from a previous survey, but their experimental variogram(s) seem(s) flat, or *pure nugget*, i.e. there is no evident spatial correlation. If the variables are continuous then we can assume that the correlation range is less than the smallest sampling interval. We can know no more than that.
3. We have variograms with apparent ‘structure’, but feel that some parts of the region are undersampled and others oversampled, and that some sampling points could be positioned more effectively to optimize estimation.

The problems faced in 1 and 2 can be resolved by starting with a nested survey and analysis, which we now describe.

## 6.2 THEORY OF NESTED SAMPLING AND ANALYSIS

The model of nested variation is based on the notion that a population can be divided into classes at two or more categoric levels in a hierarchy. The population can then be sampled with a multi-stage (multi-level) or nested scheme to estimate the variance at each level. The population is divided initially into classes at stage 1, and these are subdivided at stage 2 into subclasses, which in turn can be subdivided further at stage 3 to give finer classes, and so on, to form a nested or hierarchical classification with  $m$  stages. In each case the class at the lower level is contained completely within the one immediately above it, and each sampling point is contained in one and only one class at each and every level. The system is a strict hierarchy, and a single observation embodies variation contributed by each of the stages, including an unresolved variance within the classes at the finest level of resolution. We can estimate these contributions to the variance by a hierarchical analysis of variance (ANOVA).

Youden and Mehlich (1937) saw that for an attribute distributed in space the stages could be represented by a hierarchy corresponding to different distances. They adapted classical multi-stage sampling so that each stage in the hierarchy represented a distance between sampling points. They sampled at random, with only the distances between pairs fixed, and so the random effects model, model II of Marcuse (1949), is appropriate for the ANOVA.

For a design with  $m$  stages the data are modelled as

$$Z_{ijk\dots m} = \mu + A_i + B_{ij} + \dots + \varepsilon_{ijk\dots m}, \quad (6.9)$$

where  $Z_{ijk\dots m}$  is the value of the  $m$ th unit in  $\dots$ , in the  $k$ th class at stage 3, in the  $j$ th class at stage 2, and in the  $i$ th class at stage 1. The general mean is  $\mu$ ;  $A_i$  is the difference between  $\mu$  and the mean of class  $i$  in the first category;  $B_{ij}$  is the difference between the mean of the  $j$ th subclass in class  $i$  and the mean of class  $i$ ; and so on. The final quantity  $\varepsilon_{ijk\dots m}$  represents the deviation of the observed value from its class mean at the last stage of subdivision. The quantities  $A_i, B_{ij}, C_{ijk}, \dots, \varepsilon_{ijk\dots m}$  are assumed to be independent random variables associated with stages 1, 2, 3,  $\dots$ ,  $m$ , respectively, having means of zero and variances  $\sigma_1^2, \sigma_2^2, \sigma_3^2, \dots, \sigma_m^2$ . The latter are the components of variance for the respective stages. They are estimated according to the scheme in Table 6.4. The quantities  $n_1, n_2, n_3, \dots, n_m$ , in the table are the numbers of subdivisions of each class at the several levels. If for each stage, say  $j$ ,  $n_j$  is constant then the design is balanced. All the  $n_j, j = 1, 2, \dots, n_m$ , are known for any particular design, and so we can determine the components of variance for all stages in the classification and the residual variance from the right-hand column of Table 6.4.

**Table 6.4** Hierarchical analysis of variance: parameters estimated.

Source	Degrees of freedom	Parameters estimated
Stage 1	$n_1 - 1$	$\sigma_m^2 + n_m \sigma_{m-1}^2 + \dots + n_m n_{m-1} \dots n_3 \sigma_2^2 + n_m n_{m-1} \dots n_3 n_2 \sigma_1^2$
Stage 2	$n_1(n_2 - 1)$	$\sigma_m^2 + n_m \sigma_{m-1}^2 + \dots + n_m n_{m-1} \dots n_3 \sigma_2^2$
Stage 3	$n_1 n_2 (n_3 - 1)$	$\sigma_m^2 + n_m \sigma_{m-1}^2 + \dots + n_m n_{m-1} \dots n_4 \sigma_3^2$
:	:	:
Stage $m - 1$	$n_1 n_2 n_3 \dots (n_{m-1} - 1)$	$\sigma_m^2 + n_m \sigma_{m-1}^2$
Stage $m$ (residual)	$n_1 n_2 n_3 \dots n_{m-1} (n_m - 1)$	$\sigma_m^2$
Total	$n_1 n_2 n_3 \dots n_{m-1} n_m - 1$	

The individual component for a given stage measures the variation attributable to that stage, and together they sum to the total variance:

$$\sigma^2 = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \dots + \sigma_m^2. \quad (6.10)$$

The components of variance for each spacing from this analysis reveal over what part of the spatial scale most of the variation occurs. The particular merit of the method is that a wide range of spatial scales can be covered in a single analysis.

### 6.2.1 Link with regionalized variable theory

Although the hierarchical ANOVA derives from classical statistics, Miesch (1975) showed its links with geostatistics. He also showed that it can provide a first approximation to the variogram if the components of variance are accumulated, starting with the smallest spacing. For the  $m$  stages of subdivision we have the corresponding distances  $d_m, d_{m-1}, \dots, d_1$ , where  $d_m$  is the shortest distance at the  $m$ th stage and  $d_1$  is the largest distance at the first stage. Then the equivalence is given by

$$\begin{aligned} \sigma_m^2 &= \gamma(d_m), \\ \sigma_{m-1}^2 + \sigma_m^2 &= \gamma(d_{m-1}), \\ \sigma_{m-2}^2 + \sigma_{m-1}^2 + \sigma_m^2 &= \gamma(d_{m-2}), \end{aligned} \quad (6.11)$$

and so on. In practice, the distances are chosen in geometric progression in which each is at least 3 times the previous one; in this way the branches of the hierarchy do not overlap on the ground, and it is clear to which each sampling belongs at every stage. The components tend to be fairly imprecise estimates of the true semivariances because each is usually based on few degrees of freedom, at least in the first few stages. We should also bear in mind that they are not

entirely independent of one another and that variation in different directions cannot be distinguished.

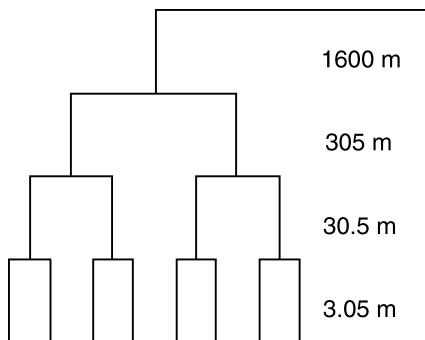
The values  $\gamma(d_i)$  are the equivalent semivariances. When plotted against distance they provide a first approximation to the variogram. The result might be rough, but it indicates how  $Z$  varies in space in the region over several orders of magnitude of distance in a single analysis and with modest sampling. For this reason it is ideal for reconnaissance where little or nothing is known. Once the spatial scale is known then a subsequent survey can be planned to estimate the variogram precisely (Oliver and Webster, 1986) or to plan a more general survey over a larger area. Alternatively, the analysis might show that all the variation occurs over very short distances, and that attempting to measure spatial correlation and map the variable(s) is pointless or would be too costly.

### 6.2.2 Case study: Youden and Mehlich's survey

We illustrate the technique with an example from Youden and Mehlich's (1937) original paper. The authors' sampling scheme to survey the soil in Broome County in New York State had four stages (Table 6.5). They applied it to two soil series: the Culvers and the Sassafras. On each soil type they selected nine primary centres 1.6 km apart forming level 1 in the hierarchy. At the next level (level 2) one subcentre was selected 305 m away from each of the main centres (18 locations). At level 3 two sampling points were chosen 30.5 m from the main centre and the subcentre (36 locations). At level 4 each site was replicated with another 3.05 m away, to give 72 sampling points in all. The survey was fully balanced, so that all classes at a particular level were subdivided equally to form the hierarchy (Figure 6.11). The progression of the sampling intervals was geometric, as above, with a 10-fold multiplication of the distances except at the highest level. As a result the authors felt able to regard the components of variance as independent, thereby allowing confidence intervals to be determined. At each sampling point the pH was determined on soil taken from a depth of 0–15 cm.

**Table 6.5** Components of variance of pH in two soil series sampled by Youden and Mehlich (1937).

Stage	Spacing/m	Degrees of freedom	Culvers series 0–15 cm		Sassafras series 0–15 cm	
			Estimated component	Percentage of variance	Estimated component	Percentage of variance
1	1600.0	8	0.02819	39.7	0	0
2	305.0	9	0.02340	32.9	0.04440	60.3
3	30.5	18	0.00552	7.8	0.00698	9.8
4	3.05	36	0.01391	19.6	0.02225	30.2

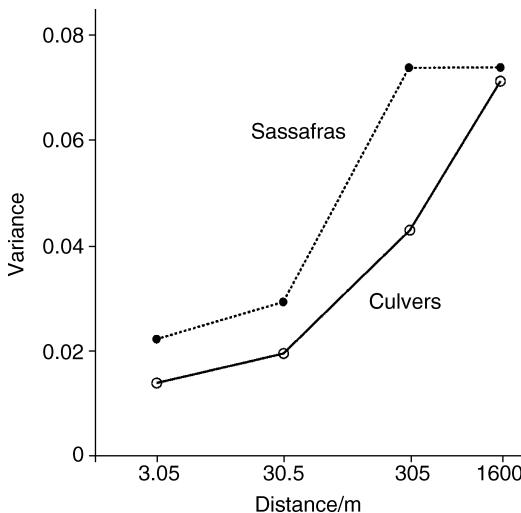


**Figure 6.11** Topological structure of the balanced nested sampling design of Youden and Mehlich (1937).

For each soil series the variation associated with each sampling interval was determined by a nested ANOVA as in Table 6.4. First the sums of squares of the deviations of the means of the classes at level 1 from the general mean were computed, and then each was multiplied by the number of observations that made up the class mean. For each class at level 2, the difference between its mean and the mean of the class to which it belongs in level 1 was squared and multiplied by the number of observations in that class. The sum of these values is the appropriate sum of squares. This was repeated for each stage, and the sums of squares of the individual levels sum to the total sum of squares. The mean squares were obtained by dividing the sums of squares of each stage by the appropriate degrees of freedom (Table 6.4). The mean square at each level, apart from the lowest, contains a unique contribution to the variance from that level, plus contributions from the components at all levels below (Table 6.4). For instance, the unique contribution to the variance at level 2 (Table 6.4) is  $n_m n_{m-1} \dots n_3 \sigma_2^2$ . This enables each component to be determined separately from its mean square. For a balanced design the value of each component can be tested to judge whether it is larger than zero by the  $F$  ratio:

$$F = \frac{\text{mean square at level } m}{\text{mean square at level } m + 1}. \quad (6.12)$$

Table 6.5 gives the results of the analysis. The components of variance can now be accumulated, starting with that for the lowest level, and plotted against sample spacing (Figure 6.12). This gives a first approximation to the variogram, a reconnaissance variogram. Figure 6.12 shows the accumulated components of variance for the Culvers and Sassafras series plotted against distance on a logarithmic scale. The variance for the Culvers series increases substantially as the distance between sampling points increases and without limit (unbounded). The sample variance for the Sassafras series does reach a maximum (bounded),



**Figure 6.12** Variograms of pH from Youden and Mehlich's surveys of pH in Culvers and Sassafras soil series by accumulation of the components of variance. Note the logarithmic scale for distance on the abscissa.

and we might therefore treat the variation as second-order stationary. If we project the variances to spacings less than the smallest, 3 m, then both seem to approach limits larger than 0. Such limits are examples of what we now recognize as *nugget variance*.

### 6.2.3 Unequal sampling

The sampling designs described above are fully balanced in the sense that all classes at each particular stage are subdivided equally. For the particular design in Broome County the sample size doubles with each additional stage in the hierarchy after the first. To achieve good spatial resolution might require many stages and result in prohibitively expensive sampling for reconnaissance. It would to some extent defeat the object whereby one is trying to obtain preliminary information for modest effort. As it happens, full replication at each stage is unnecessary because the mean squares for the lower stages are estimated more precisely than those for the higher ones. Economy can be achieved by replicating only a proportion of the sampling centres in the lower stages. Oliver and Webster (1986) used five stages, but in more recent applications, Webster and Boag (1992), Badr *et al.* (1993) and Oliver and Badr (1995) have used seven. The resulting schemes are unbalanced, and this makes estimating the components somewhat more complex because the

**Table 6.6** Hierarchical analysis of variance: unbalanced design.

Source	Degrees of freedom	Parameters estimated
Level 1	$D_1 = f_1 - 1$	$u_{1,1}\sigma_1^2 + u_{1,2}\sigma_2^2 + u_{1,3}\sigma_3^2 + \sigma_4^2$
Level 2	$D_2 = f_2 - f_1$	$u_{2,2}\sigma_2^2 + u_{2,3}\sigma_3^2 + \sigma_4^2$
Level 3	$D_3 = f_3 - f_2$	$u_{3,3}\sigma_3^2 + \sigma_4^2$
Residual	$D_4 = N - f_3$	$\sigma_4^2$
Total	$N - 1$	

coefficients of the components are no longer simple multiples of the number of divisions in the levels as they are in Table 6.4, which must be replaced by a table such as Table 6.6 for a sample of size  $N$ .

Gower (1962) provided formulae for calculating the coefficients,  $u_{i,j}$ , and they were included in the sixth edition of Snedecor and Cochran's (1967) standard text (but not in the later editions). They can all be expressed in the following formulae:

$$q_{i,j} = \sum_{k=1}^{C_i} \sum_{p=1}^{c_{jk}^i} \frac{n_{jp}^2}{N_{ik}} \quad \text{and} \quad u_{i,j} = \frac{1}{D_i} \{q_{i,j} - q_{i-1,j}\}. \quad (6.13)$$

In these equations  $u_{i,j}$  is the coefficient at stage  $i$  for the  $j$ th component;  $C_i$  is the number of groups at the  $i$ th stage;  $c_{jk}^i$  is the number of subgroups in stage  $j$  within the  $k$ th group at level  $i$ ;  $n_{jp}$  is the number of individual sampling points in the  $p$ th subgroup in stage  $j$  (within group  $k$  at stage  $i$ ), with  $i \leq j$ , and  $D_i$  is the number of degrees of freedom at stage  $i$  (see Table 6.6).

One consequence of the lack of balance is that the coefficients for a given component in the expected values for the mean squares are in general different at different levels, as Table 6.6 shows. As a result one cannot use a simple  $F$  ratio to test whether a component,  $\sigma_j^2$ , is significantly greater than 0.

### Residual maximum likelihood estimation

For balanced designs the estimates of the components provided by ANOVA are the same as those one obtains by computing the sample variogram (Miesch, 1975), but for unbalanced designs the estimators will in general differ (Pettitt and McBratney, 1993). Further, several methods of constructing ANOVA tables have been invented for the general unbalanced analysis, and although the estimators are unbiased they are not necessarily the same (see Searle *et al.*, 1992). If one assumes that the random effects are normally distributed then one can calculate maximum likelihood estimates of the variance components from equation (6.9). This puts the estimation for all designs in a coherent framework.

Maximum likelihood estimation from equation (6.9) calculates the likelihood of the data,  $z$ , in terms of the variance components and then uses the estimators of those components that maximize the log-likelihood. With small samples, however, the estimators are biased; they underestimate the true values because the fixed degrees of freedom are not removed before the components are estimated. Patterson and Thompson (1971) developed the method of residual maximum likelihood (REML), sometimes called ‘restricted maximum likelihood’, that adjusts for the fixed degrees of freedom before estimating the variance components. In the present context there is only one fixed effect, namely the grand mean,  $\mu$ , and so the differences between the estimates from REML and ANOVA are fairly small. Webster *et al.* (2006) describe the method in full; here we provide a summary.

The set of data with  $N$  points can be considered as a set of  $N$  orthogonal contrasts. If there are  $p$  fixed degrees of freedom then  $p$  contrasts will have expectations that are functions of the fixed effects, and the remaining  $N - p$  contrasts have zero expectation. By maximizing the likelihood of the contrasts with zero expectation we can obtain (REML) estimates of the variance components that take account of the degrees of freedom used in estimating fixed effects. The contrasts with zero expectation are directly related to the contrasts that contribute to the residual sums of squares, and hence the estimated variance components, in ANOVA. In the balanced case, the REML estimates of variance components are the same as those from ANOVA.

To determine the REML log-likelihood for the data one must define the full variance–covariance matrix of the data. We do this using design matrices to indicate the structure of the sampling scheme. The design matrix  $\mathbf{U}_k$  at stage  $k$  has  $p$  columns, where  $p$  is the total number of sampling points at stage  $k$ . The rows of the matrix correspond to measurements. Each row of  $\mathbf{U}_k$  takes the value 1 in column  $j$  if the measurement arose from sampling point  $j$  at stage  $k$ , and zero otherwise. Then the variance–covariance matrix of the data can be expressed as

$$\mathbf{V} = \sum_{i=1}^m \sigma_i^2 \mathbf{U}_i \mathbf{U}_i^T = \sum_{i=1}^{m-1} \sigma_i^2 \mathbf{U}_i \mathbf{U}_i^T + \sigma_m^2 \mathbf{I}_N, \quad (6.14)$$

as there is no further sub-sampling at stage  $m$ . The matrix  $\mathbf{I}_N$  is the  $N$ -dimensional identity matrix. The logarithm of the residual likelihood,  $\ln(\text{RL})$ , is then

$$-2 \times \ln(\text{RL}) = \ln |\mathbf{V}| + \ln |\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}| + \mathbf{y}^T \mathbf{P} \mathbf{y}, \quad (6.15)$$

where  $\mathbf{X} = \mathbf{1}_N$  is the design matrix for the fixed effects in the model, and  $\mathbf{P}$  is defined as

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1} \mathbf{X} (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1}. \quad (6.16)$$

One cannot usually maximize  $\ln(RL)$  analytically; rather one must use an iterative algorithm. Lark and Cullis (2004) describe REML estimation in some detail in a geostatistical context, and further information can be found in Searle *et al.* (1992). Most general-purpose statistical packages, such as GenStat and SAS, have facilities for REML estimation in linear mixed models.

The variance components are defined as the variances of a set of random effects, and so  $\sigma_i^2 \geq 0$  for  $i = 1, 2, \dots, m$ . However, when considered as a composite variance model  $\mathbf{V}$ , as above, it is necessary only that  $\mathbf{V}$  is positive definite (and therefore permissible), i.e.  $\mathbf{a}^T \mathbf{V} \mathbf{a} > 0$  for all real vectors  $\mathbf{a} \neq \mathbf{0}$ . In this case individual components might be negative. This might arise in practice if there is some underlying regular feature in the landscape such as an ancient ploughing pattern. More usually, the true value of a variance component for which the estimate is negative is positive but close to zero, in which case the negative estimate can easily arise by chance.

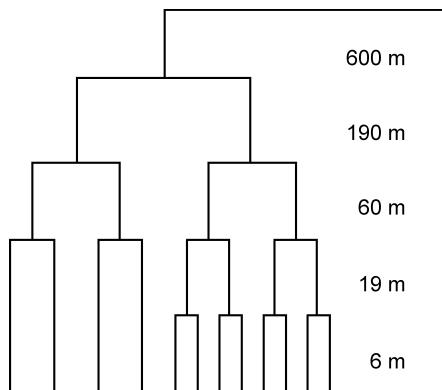
#### **6.2.4 Case study: Wyre Forest survey**

Our survey of the soil in the Wyre Forest, in the English Midlands (Oliver and Webster, 1987), illustrates the unbalanced nested design. The sample variograms from an earlier survey were flat; all the variation in the properties examined appeared to occur within 167 m, the average distance between sampling sites in that survey. The nested survey was designed to discover how the variation is distributed over distances less than 167 m. The scheme had five stages covering a range of sampling intervals from 6 m increasing in a geometrical progression of approximately threefold increments (Table 6.7) to 600 m. This design was expected to encompass most of the spatial variation, and to ensure that there was no overlap between different branches of the hierarchy, as above. The 600-m interval was incorporated in case there were long-range spatial structures.

Nine main centres were located at the nodes of a 600-m square grid oriented randomly over the region. All other points were then placed on random orientations from these as follows to comply with the random effects model.

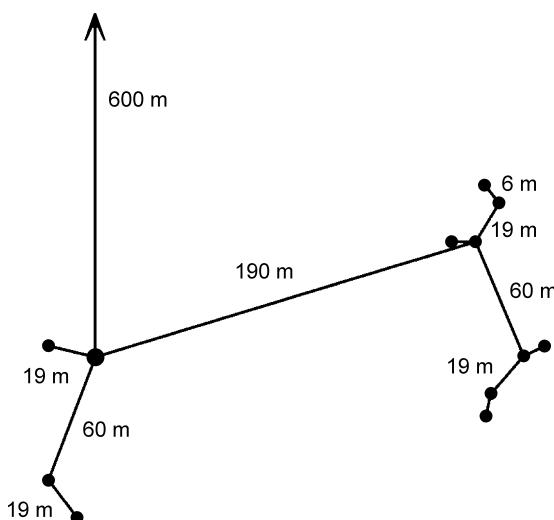
**Table 6.7** Nested sampling design for determining the scale of spatial variation in the soil of the Wyre Forest.

Stage	Sampling interval/m	Number of sampling points
1	600	8
2	190	18
3	60	36
4	19	72
5	6	108

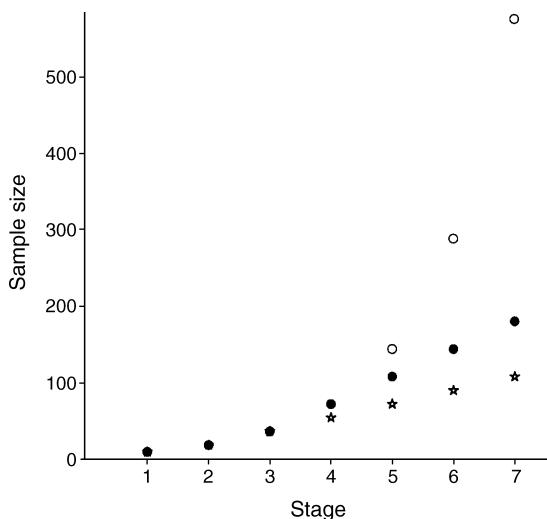


**Figure 6.13** Topology for one main centre of the unbalanced nested sampling as implemented in the soil survey of the Wyre Forest by Oliver and Webster (1987).

From each grid node a second site was chosen 190 m away to provide the second stage. From each of the now 18 sites another point was chosen 60 m away (stage 3). The procedure was repeated at stage 4 to locate points 19 m away from those of stage 3, giving 72 points. At the fifth stage just half of the fourth-stage points were replicated by sampling 6 m away. This gave a sample of 108 points rather than 144 for a fully balanced survey. Table 6.7 summarizes the design, Figure 6.13 illustrates the hierarchical structure used for one centre, and Figure 6.14 shows the configuration of sampling points for



**Figure 6.14** Sampling plan for one main centre of the Wyre Forest survey (not strictly to scale).



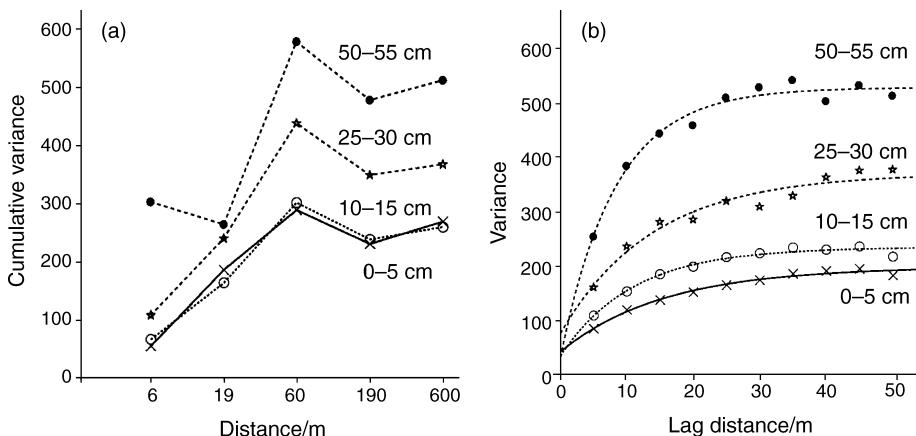
**Figure 6.15** Graph showing the economy achieved by not doubling the sampling at every level in the hierarchy. The symbols are ○ balanced design, ● Wyre Forest scheme including extension to stages 6 and 7, \* scheme used by Webster and Boag (1992) in their surveys of nematode infestations.

one first-stage centre. The design achieved a 25% economy in effort compared with a fully balanced scheme. Figure 6.15 shows the economy possible with even more stages. At each sampling point several properties of the soil were recorded at four fixed depths in the soil profile: 0–5 cm (1), 10–15 cm (2), 25–30 cm (3), and 50–55 cm (4).

Each variable was analysed by ANOVA according to the scheme outlined in Table 6.8 and by REML with and without the components' being constrained to be non-negative. The estimated components of variance for sand content at the four depths are listed in Table 6.8 for ANOVA. Figure 6.16(a) shows the accumulated components of variance for each depth in the profile plotted

**Table 6.8** Components of variance of sand content of the soil at four depths in the survey of the Wyre Forest estimated by analysis of variance.

Stage	Component of variance			
	Depth/cm			
	0–5	10–15	25–30	50–55
1 (600 m)	32.44	17.54	27.95	32.63
2 (190 m)	-51.90	-64.82	-81.97	-103.44
3 (60 m)	88.77	141.79	172.02	316.28
4 (19 m)	139.44	101.10	135.21	-45.79
5 (6 m)	55.58	68.42	116.33	309.73



**Figure 6.16** Variograms of soil properties in the Wyre Forest: (a) obtained by accumulation of the components of variance estimated by REML, with the lag distance on a logarithmic scale; (b) estimated from subsequent transect sampling at 5 m intervals.

against separating distance on a logarithmic scale to give first approximations to the variograms. The graph shows that at least 80% of the variation at the four depths for sand content occurs within 60 m; this was the case for all of the other properties. Stages 1 and 2, i.e. distances of 190–600 m and 60–190 m, respectively, account for less than 20% of the variation. The estimated components of variance for stage 2 were generally negative. This suggests that either there is some repetition in soil character at that distance, or that the components estimate zero because there is no contribution to the variance at this stage. The confidence limits of the components are wide, and so we cannot be sure how to interpret these negative values. Even at stage 5 there is still a considerable contribution to the total variance. This represents the unresolved variation within 6 m plus errors of measurement.

Let us now look at the estimates of the variance components by REML. Table 6.9 lists the results, first without any constraint and then with the estimates constrained to be non-negative. The unconstrained estimates are somewhat different from those from ANOVA, but they show the same general pattern. Constraining the estimates to a minimum of 0 caused little change in the positive components in the lowest stage, but appreciable changes in all stages above those that were negative in the unconstrained analysis.

As described above, the experimental variogram depends on the spatial scale over which we measure it. If a large extent is covered with wide sampling intervals then all of the variance might appear as nugget. Alternatively, if small intervals are chosen to resolve the short-range variance then the sampling required to estimate the contributions to the larger distances would be too costly. A nested survey identifies the scale at which most of the variation occurs

**Table 6.9** Components of variance of sand content of the soil at four depths in the survey of the Wyre Forest, estimated by REML without constraints and constrained to be non-negative.

Stage	Component of variance				Constrained variance			
	Depth/cm				Depth/cm			
	0–5	10–15	25–30	50–55	0–5	10–15	25–30	50–55
1 (600 m)	38.12	21.07	18.68	33.75	19.70	0.44	0	0
2 (190 m)	−58.03	−63.17	−90.02	−100.19	0	0	0	0
3 (60 m)	102.50	137.58	198.51	314.81	63.58	96.06	125.26	235.89
4 (19 m)	131.50	97.65	131.95	−38.89	131.43	97.65	131.03	0
5 (6 m)	54.97	66.54	108.56	303.26	54.88	66.27	109.87	276.06

at the level of our investigation. The reconnaissance variograms for the soil properties of the Wyre Forest showed that most of the spatial variation occurred over distances less than 60 m.

From this information we could design a survey to estimate the variogram more precisely by linear sampling. We did so using ten transects each 100 m long and one of 500 m with a sampling interval of 5 m. The conventional variograms that resulted, Figure 6.16(b), showed correlation extending to little more than about 40 m.

We could have used the results of the nested survey to design an overall survey with a maximum sampling interval equal to half the correlation range identified. This would have been 30 m. In the event, having estimated the variograms more precisely along transects and established an effective range of 40 m, we sampled at a 20-m interval from which to interpolate for mapping. You can read a full account in Oliver and Webster (1987).

## 6.2.5 Summary

Nested survey and analysis can reveal the spatial scale(s) of variation in a region with modest sampling effort. The data can be analysed by straightforward analysis of variance for balanced designs or, preferably, by residual maximum likelihood for unbalanced ones. We recommend it as a first step in the description of variation in a hitherto little known region. Armed with the results, one can plan a second stage of survey to estimate the variogram precisely over the range that matters. The results from nested survey could be used to plan a regional survey if one particular component proved dominant.

# *Spectral Analysis*

In some places the land varies laterally in a regular fashion. The most obvious regular patterns are man-made. They include the characteristic ridge and furrow of the English clay lands, and orchards and plantations in which fruit trees and other crops are arranged in lines with constant intervals between them. The dynamic properties of the soil are likely to vary in tune with them and so also have a regularity. Forest is established on peaty soil by planting young trees on the upturned sod after deep ploughing in lines. Less obvious are the long-lasting patterns of former ploughing on crop yield, revealed by McBratney and Webster (1981), and the effects of earlier drainage schemes on the present-day soil described by Burrough *et al.* (1985). In all these the regularity is deliberate.

Natural features may also seem regular. Examples are the frost polygons of the Arctic region and their fossil relics in the Northern temperate zone (e.g. Hodge and Seale, 1966), the patterns of termite mounds in Africa, especially evident on some of the East African plains (e.g. Scott *et al.*, 1971) and in the miombo woodland of Zambia and Congo and the gilgai of Australia (e.g. Hallsworth *et al.*, 1955; Webster, 1977).

The experimental variograms of such patterns fluctuate with evident periodicity, as Webster (1977) discovered. Other periodic patterns can arise from cultivation and land management (McBratney and Webster, 1981; Burrough *et al.*, 1985). Chapter 5 mentioned basic periodic functions that might be used to describe the fluctuation, but we deferred illustration until now so that we can deal with it and spectral analysis together.

## **7.1 LINEAR SEQUENCES**

More often than not we encounter periodicity in linear, i.e. one-dimensional, sequences of data comprising records made at regular intervals in either time or

space (see, for example, Oliver *et al.*, 1997). Spatial examples include:

- photographic and radiometric survey by aircraft;
- bathymetric and sonar survey from ships;
- electric logs of boreholes for oil exploration;
- pollen counts through peat;
- isotope measurements through polar ice;
- transect surveys of soil.

In some instances each line is one of several or many in  $\mathbb{R}^2$  or  $\mathbb{R}^3$ . In others the lines are isolated representatives of two-dimensional scenes. Variables, such as temperature, may also be recorded at regular intervals in time, and in that instance there is only one dimension. We can analyse the data by all of the standard geostatistical methods described above. However, if there is periodicity then it is often profitable to express the variation in relation to frequency rather than space or time, and this takes us into the realm of spectral analysis.

## 7.2 GILGAI TRANSECT

To illustrate an analysis of periodic variation we use the data from a survey by Webster (1977) of salinity on the Bland Plain of eastern Australia. This virtually flat plain is part of the Murray–Darling Basin. Its soil is dominantly clay, but with a more sandy surface horizon of variable thickness, alkaline and locally saline. One of its most remarkable features is its patterns of gilgai. The gilgaits are small, almost circular depressions from a few centimetres to as much as 50 cm deep in the plain and several metres across. The soil in their bottoms is usually clay and wetter than that elsewhere.

A paddock at Caragabal, NSW, was sampled at regular 4-m intervals along a transect almost 1.5 km long. At each of 365 sampling points a core of soil, 75 mm in diameter, was taken to 1 m, and segments of it were analysed in the laboratory. For present purposes we shall concern ourselves with just one variable, the electrical conductivity at 30–40 cm. Table 7.1 summarizes the data, which were strongly skewed and therefore transformed to logarithms for further analysis. Figure 7.1 shows the logarithm of conductivity plotted against position as the fine line. The bold line is a smoothing spline fitted to the data to filter out the short-range variation and reveal a fluctuation of longer range that appears regular.

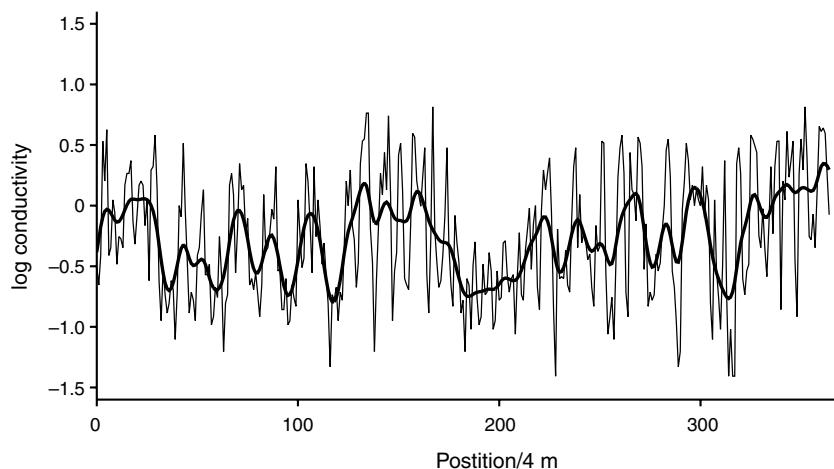
The experimental variogram of the data is shown in Figure 7.2 as the plotted points, to which we have fitted a model with a periodic component. The full model is given by

$$\gamma(h) = c_0 + wh + c\{\text{sph}(a)\} + c_1 \cos\left(\frac{2\pi h}{\omega}\right) + c_2 \sin\left(\frac{2\pi h}{\omega}\right). \quad (7.1)$$

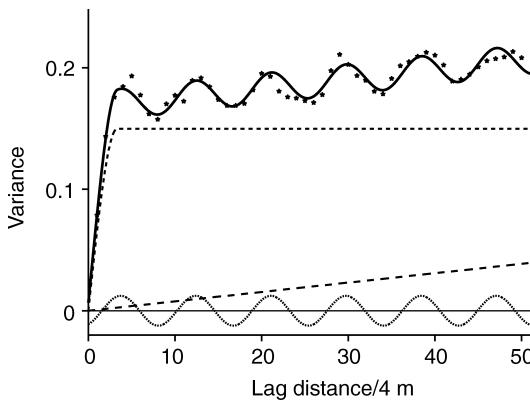
**Table 7.1** Summary statistics of electrical conductivity in the soil at 30–40 cm at Caragabal.

	Electrical conductivity	
	$\text{mS cm}^{-1}$	$\log_{10}(\text{mS cm}^{-1})$
Minimum	0.06	-1.214
Maximum	5.10	0.707
Mean	0.958	-0.2298
Median	0.54	-0.2668
Variance	0.95948	0.19205
Standard deviation	0.975	0.438
Skewness	1.642	0.101

It comprises a small nugget,  $c_0$ , a linear component,  $wh$ , a spherical function with a short range,  $c\{\text{sph}(a)\}$ , and a sine wave,  $c_1 \cos(2\pi h/\omega) + c_2 \sin(2\pi h/\omega)$ . The values of these parameters are given in Table 7.2. The spherical component contributes most to the variance, with a sill of approximately  $0.15 \log^2 (\text{mS cm}^{-1})$ . It represents repetitive variation of a kind that is not periodic. For present purposes the periodic component, though representing less of the variance with an amplitude of only 0.012, is of most interest. Its wavelength is 8.67 sampling units or 35 m. This is approximately equal to the average distance between the centres of the gilgai on the transect. It has a phase shift of  $-0.43 \text{ rad}$  (about  $-25^\circ$ ). The linear component has only a very gentle gradient; it is of little practical consequence, and we may regard the underlying variation



**Figure 7.1** Trace of common logarithm of electrical conductivity at Caragabal (fine line) with a smoothing spline (bold) added as an aid to see the suspected periodicity.



**Figure 7.2** Variogram of log electrical conductivity. The points are the sample values, the heavy line is the fitted model comprising the four components shown with the lighter lines. The parameter values are listed in Table 7.2.

**Table 7.2** Parameter values of model fitted to variogram of log electrical conductivity in the soil at 30–40 cm at Caragabal. Distances are in sampling intervals of 4 m, and angles are in radians.

Component	Parameter	Value
Nugget	constant, $c_0$	0.01760
Linear	gradient, $w$	0.000772
Spherical	sill, $c$	0.1498
	range, $a$	3.323
Periodic	amplitude, $W$ ,	0.01230
	wavelength, $\omega$	8.667
	phase, $\phi$	-0.435
	$c_1$	-0.01116
	$c_2$	0.005181

as second-order stationary. The nugget variance is also very small. In passing, we note that the periodic component does not damp, and so the model is valid in one dimension only.

### 7.3 POWER SPECTRA

Let us now consider how to examine this variation in the frequency domain. We start by assuming that the underlying variable,  $Z(\mathbf{x})$ , is random, spatially correlated, and second-order stationary. Since we are dealing with only one

dimension for the time being, we can replace  $\mathbf{x}$  by  $x = |\mathbf{x}|$  and  $\mathbf{h}$  by  $h = |\mathbf{h}|$ . Its covariance function, in the notation of Chapter 4, is

$$C(h) = E[\{Z(x) - \mu\}\{Z(x + h) - \mu\}] = E[Z(x)Z(x + h) - \mu^2], \quad (7.2)$$

where  $\mu$  is the mean of the process.

The covariance function in the spatial domain has an equivalent in the frequency domain where the variance, instead of being a function of distance (or time), is distributed as a function of frequency,  $f$ . This function, denoted by  $R(f)$ , is the *spectrum*, or *power spectrum*. It is the Fourier transform of the covariance function defined for the interval from positions  $-X/2$  to  $X/2$ , i.e.  $-X/2 \leq Z(x) \leq X/2$ :

$$R(f) = \lim_{X \rightarrow \infty} \frac{1}{2\pi} \int_{-2X}^{2X} \{1 - (|h|/2X)\} \exp(-jfh) C(h) dh, \quad (7.3)$$

where  $j$  is  $\sqrt{-1}$ . Provided  $C(h)$  approaches 0 as  $h$  approaches  $\infty$ , the limiting value of  $R(f)$  is given by

$$R(f) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-jfh) C(h) dh. \quad (7.4)$$

The covariance function is symmetric; it is an ‘even’ function of  $h$ , i.e.  $C(h) = C(-h)$ ; see Chapter 4. As a consequence, the complex term in the integral in equation (7.4) can be replaced by a simple cosine, and  $R(f)$  reduces to

$$R(f) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \cos(fh) C(h) dh. \quad (7.5)$$

Just as the spectrum,  $R(f)$ , is the Fourier transform of the covariance function, the latter,  $C(h)$ , is the Fourier transform of  $R(f)$ :

$$C(h) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \cos(fh) R(f) df. \quad (7.6)$$

In other words, the relation is invertible.

We can equally well transform the autocorrelation function,  $\rho(h) = C(h)/C(0)$ , to obtain the *normalized spectrum*:

$$r(f) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \cos(fh) \rho(h) dh. \quad (7.7)$$

This relation too is invertible.

### 7.3.1 Estimating the spectrum

Equations (7.4) and (7.5) above define the spectrum of a real continuous second-order stationary random process in  $\mathbb{R}^1$ . We want now to estimate the spectrum. As in the example of the gilgai transect, we have data,  $z(x_1), z(x_2), \dots, z(x_N)$ , at regular intervals on a line. The value  $N$  is the length of the series, and replaces  $X$  to accord with geostatistical convention. From the data we compute

$$\hat{C}(h) = \frac{1}{N-h} \sum_{i=1}^{N-h} \{z(i) - \bar{z}\} \{z(i+h) - \bar{z}\}, \quad (7.8)$$

where the  $z(i)$  and  $z(i+h)$  are observed values, and  $\bar{z}$  is the average of the data in the sequence, and by incrementing  $h$  one step at a time we obtain the experimental covariance function. Thus the lag,  $h$ , is in units of the sampling interval.

This set of covariances can be transformed to the corresponding experimental spectrum by

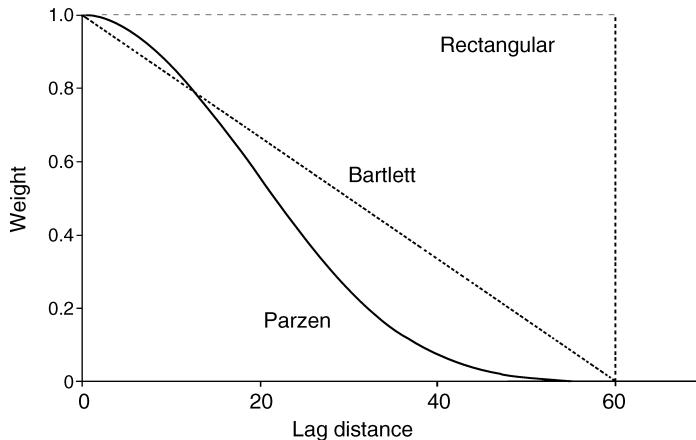
$$\hat{R}(f) = \frac{1}{2\pi} \left\{ \hat{C}(0) + 2 \sum_{k=1}^{L-1} \hat{C}(k) w(k) \cos(\pi fk) \right\} \quad (7.9)$$

for frequency,  $f$ , in the range 0 to  $\frac{1}{2}$  cycle. In this equation  $L$  is the maximum lag from which the transform is computed and  $k$  is the lag.

The quantity  $L$  can be regarded as the width of a ‘window’ through which the covariance is viewed for transformation, and it is for us to choose it. We could set it to the maximum possible from the data. We know from experience that as the lag increases so the experimental covariances become increasingly unreliable, and in Chapter 4 we suggested that the covariance be computed to a lag of no more than about one-fifth of the total length of a series. If we incorporate the uncertainty in estimating  $C(h)$  at long lags in equation (7.9) then we shall obtain detail in the computed spectrum that is untrustworthy. In fact, the longer is  $L$ , the more detailed is the spectrum and the less reliable is that detail. On the other hand, if we choose too small a value of  $L$  then we shall lose detail that might be significant. The window is effectively a smoothing function, and the narrower it is in the spatial domain the more precise are the estimates at the expense of greater bias and loss of detail. So the choice of  $L$  is always a compromise.

Some of the fluctuation in the spectrum that arises from choosing a large  $L$  can be diminished by changing the ‘shape’ of the window. The window in equation (7.9) is rectangular (see Figure 7.3). If  $|k| \leq L$  then  $C(k)$  carries weight  $1/L$ , otherwise its weight is 0:

$$w_R(k) = \begin{cases} 1/L & \text{for } 0 \leq |k| \leq L, \\ 0 & \text{for } |k| > L. \end{cases} \quad (7.10)$$

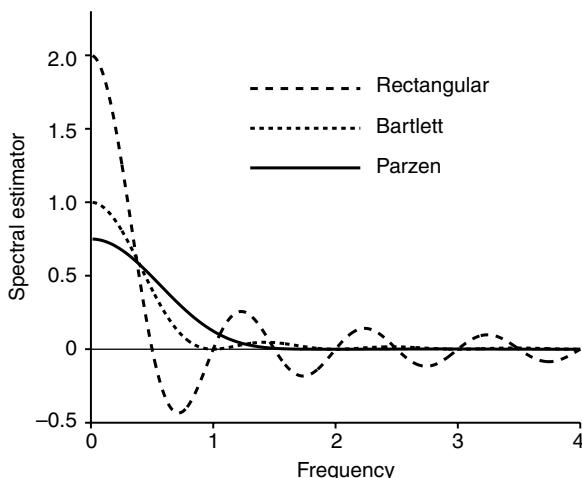


**Figure 7.3** Rectangular, Bartlett and Parzen lag windows, with a basal width of 60 sampling intervals.

It is symmetric about the ordinate, and so we show only the positive half. Its Fourier transform is given by

$$W_R(f) = 2L \left( \frac{\sin(2\pi f L)}{2\pi f L} \right) \quad \text{for } -\infty \leq f \leq \infty. \quad (7.11)$$

The transform of the rectangular lag window fluctuates as the frequency increases with a period of  $1/L$ ; the power takes a long while to damp. This is shown in Figure 7.4 in which there are several peaks of decreasing height. In the jargon of



**Figure 7.4** Rectangular, Bartlett and Parzen spectral windows. These are the Fourier transforms of the lag windows shown in Figure 7.3.

spectral analysis, the rectangular window is ‘leaky’, and it is generally regarded as unsatisfactory. The top corners of the rectangles tend to contribute most of the leakage. This leakage can be diminished substantially by cutting the corners.

Much research has been devoted to finding an optimal shape, ‘window carpentry’ as Jenkins and Watts (1968) called it. The simplest solution is due to Bartlett (1966) and is known as the Bartlett window. It is defined in the spatial domain as follows:

$$w_B(k) = \begin{cases} 1 - (|k|/L) & \text{for } 0 \leq |k| \leq L, \\ 0 & \text{for } |k| > L. \end{cases} \quad (7.12)$$

The Bartlett lag window may be envisaged as an isosceles triangle with its peak at its centre and its height decaying linearly to its lower corners where  $|k|$  of equation (7.9) equals  $L$ . It is shown in Figure 7.3 for  $0 \leq k \leq L$ . Like the spectral window, the lag window is symmetrical about the ordinate, and so again only the positive half is shown. It is incorporated in the transformation equation as

$$\hat{R}(f) = \frac{1}{2\pi} \left\{ \hat{C}(0) + 2 \sum_{k=1}^{L-1} \hat{C}(k) w_B(k) \cos(\pi fk) \right\}. \quad (7.13)$$

The Fourier transform of the Bartlett lag window is

$$W_B(f) = L \left( \frac{\sin(2\pi fL)}{2\pi fL} \right)^2 \quad \text{for } -\infty \leq f \leq \infty, \quad (7.14)$$

and this is shown in Figure 7.4. It fluctuates rather less than the rectangular window, but nevertheless is not entirely satisfactory because of its leakage. Two other popular windows are those defined by J. W. Tukey (see Blackman and Tukey, 1958) and Parzen (1961), and these too are referred to by their authors’ names. A shortcoming of Tukey’s window is that it can return negative estimates of the spectral density, which must be positive. Parzen’s window is more reliable. Its definition is

$$w_P(k) = \begin{cases} 1 - 6\left(\frac{k}{L}\right)^2 + 6\left(\frac{|k|}{L}\right)^3 & \text{for } 0 \leq |k| \leq L/2, \\ 2\left(1 - \frac{|k|}{L}\right)^3 & \text{for } L/2 < |k| \leq L, \\ 0 & \text{for } |k| > L. \end{cases} \quad (7.15)$$

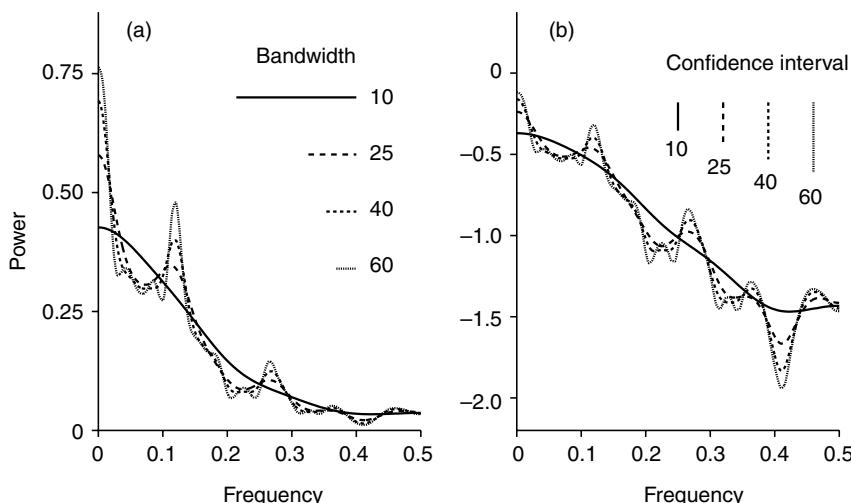
and its Fourier transform is

$$W_P(f) = \frac{3}{4} L \left( \frac{\sin(\pi fL/2)}{\pi fL/2} \right)^4 \quad \text{for } -\infty \leq f \leq \infty. \quad (7.16)$$

In the transformation equation (7.13), for the Parzen window  $w_B(k)$  is replaced by  $w_P(k)$ . Figure 7.4 shows that this transform does not fluctuate, but decays to 0 at a frequency of approximately  $2/L$ . Although Parzen's window seems the most attractive theoretically, it generally requires a more reliable set of covariances and therefore more data in the first place.

To estimate the spectrum from a series of data we compute the experimental covariance function to the maximum lag that is likely to be of interest. This is the initial width,  $L$ , of the lag window. We then choose the shape of the window (we recommend the Parzen window as a start), and we compute the spectral density at frequencies between 0 and  $\frac{1}{2}$  cycle. We then plot the results and join the points. The steps by which  $f$  is incremented need bear no relation to the lag increments, as some authorities claim. In fact, it is better to choose numerous short steps for  $f$  so as to produce a smooth figure for the spectrum, which is a continuous function. We then shorten  $L$  and repeat the procedure. Figure 7.5 shows results of using this procedure with 100 frequencies.

An alternative method for computing spectra from data is the fast Fourier transform (see Brigham, 1974). Cooley and Tukey (1965) devised an algorithm for its computation in the days when computers were orders of magnitude slower than they are now, and code for it is included in Press *et al.* (1992).



**Figure 7.5** Spectrum of log electrical conductivity smoothed with Parzen lag windows of width 10, 25, 40 and 60 sampling intervals on (a) arithmetic scale, and (b) logarithmic scale. Frequency is the reciprocal of sampling interval. The confidence intervals in (b) are for 90%.

### 7.3.2 Smoothing characteristics of windows

The windows used to estimate the spectrum are effectively smoothing functions. The estimates have smaller variances than those of the full sample spectrum. To express this quantitatively, we first integrate over the window to obtain a quantity  $I$ :

$$I = \int_{-\infty}^{\infty} w^2(k) dk. \quad (7.17)$$

So, for example,  $I$  for Bartlett's window is the integral from  $-L$  to  $L$ :

$$I_B = \int_{-L}^L \left(1 - \frac{|k|}{L}\right)^2 dk = \frac{2}{3}L. \quad (7.18)$$

The equivalent integral of the Parzen window of equation (7.15) is  $I_P = L/1.86$ .

To distinguish the various spectral estimates we use  $R(f)$  to denote the full spectrum, as in equation (7.9), and  $\bar{R}$  subscripted with the name of the window and its width for the smoothed estimates. For example,  $\bar{R}_{P,L=25}(f)$  means the estimate of  $R(f)$  at frequency  $f$  smoothed with a Parzen window of width 25 lag intervals. The variance of  $R(f)$  is simply  $R^2(f)$ .

We are interested in the reduction in variance brought about by the smoothing, i.e. the ratio of the variance of the smoothed estimate,  $\text{var}[\bar{R}(f)]$ , to  $R^2(f)$ . It turns out that this is simply

$$\frac{\text{var}[\bar{R}(f)]}{R^2(f)} = \frac{1}{L} \int_{-\infty}^{\infty} w^2(k) dk = \frac{I}{L}. \quad (7.19)$$

So, for example, combining this equation with equation (7.18), we find that the variance ratio for a Bartlett window of  $L/N$  is  $2L/3N = 0.667L/N$ . For the Parzen window it is  $L/1.86N = 0.538L/N$ . Typically  $L/N$  is of the order of 0.1, and so the variance ratio is around 0.067 for the Bartlett window and 0.054 for the Parzen window—these are big gains in precision.

### **Bandwidth**

As above, each window in the spatial domain has its equivalent in the frequency domain. For a given shape, the wider is the window in the spatial domain the narrower is its transform. Also, because the weight of the lag windows of the same basal width is increasingly concentrated in the order rectangular < Bartlett < Parzen, they behave as if they were increasingly wide in the frequency domain—compare Figures 7.3 and 7.4. It is as though one were viewing the spectrum through a slit of increasing width. The spectral windows do not have strict bounds, however, and it is helpful to have some

measure of width for comparison. One approach to this is to define the width of a rectangular window in the frequency domain such that

$$W(f) = \frac{1}{m}, \quad \text{for } -\frac{1}{2}m \leq f \leq \frac{1}{2}m. \quad (7.20)$$

If we denote its bandwidth by  $b$  then  $b = m$ . The variance of the spectral estimator is

$$\text{var}[\bar{R}(f)] = \frac{R^2(f)}{N} \frac{1}{m} = \frac{R^2(f)}{Nb}. \quad (7.21)$$

The bandwidths of the other windows are then defined as those widths that give the same variance as that of the rectangular window,

$$\text{var}[\bar{R}(f)] \approx \frac{R^2(f)}{N} \frac{1}{b} = \frac{R^2(f)}{N} \int_{-\infty}^{\infty} w^2(k) k, \quad (7.22)$$

and so the bandwidth is  $b = 1/I$ . The bandwidth for the rectangular window is thus  $0.5/L$ , and that of the Bartlett window is  $1.5/L$ . The Parzen window's bandwidth is  $1.86/L$ . Evidently the Bartlett and Parzen windows are substantially wider than the rectangular windows.

### 7.3.3 Confidence

Confidence intervals for the spectral densities can be determined. We first define a quantity  $v$ , which is effectively the degrees of freedom:

$$v = \frac{2N}{\int_{-\infty}^{\infty} w^2(k) dk} = \frac{2N}{I}, \quad (7.23)$$

where  $N$  is the total number of observations in the series. Notice that it depends on the width and shape of the window,  $w(k)$ . The quantity  $v\hat{R}(f)/R(f)$  is distributed as  $\chi_v^2$ , and so

$$\text{Prob}\left[\chi_{v,\alpha/2}^2 < \frac{v\hat{R}(f)}{R(f)} \leq \chi_{v,1-\alpha/2}^2\right] = 1 - \alpha, \quad (7.24)$$

where Prob stands for the probability and  $1 - \alpha$  is the confidence level at which one wants to work. The  $100(1 - \frac{1}{2}\alpha)\%$  and  $100(\frac{1}{2}\alpha)\%$  confidence limits for  $R(f)$  are then

$$\frac{v\hat{R}(f)}{\chi_v^2(1 - \frac{1}{2}\alpha)} \quad \text{and} \quad \frac{v\hat{R}(f)}{\chi_v^2(\frac{1}{2}\alpha)}. \quad (7.25)$$

The integral in equation (7.23) can be worked out for the particular size and shape of window, and, since the length of the sequence,  $N$ , is known,  $v$  can be determined. The values of  $\chi^2$  for  $v$  and for  $\frac{1}{2}\alpha$  and  $1 - \frac{1}{2}\alpha$  can be obtained readily from tables, such as those by Fisher and Yates (1963), or with a statistical program. This ability to calculate confidence limits gives the spectrum a substantial advantage over the covariance function and variograms.

## **7.4 SPECTRAL ANALYSIS OF THE CARAGABAL TRANSECT**

The spectrum for the log of electrical conductivity has been computed with a Parzen lag window for four widths: 10, 25, 40 and 60 sampling intervals (Figure 7.5). It is evident that the more covariances are included in the window the more detail there is in the spectrum. One might think there is too much detail with the window set to 60, but with  $L$  set to 10 almost all detail has been lost, and only the general decline in power with increasing frequency is evident. Choosing  $L = 40$  seems to show the principal features of the spectrum most clearly.

Let us now interpret the spectrum in Figure 7.5. The most prominent feature is the marked decrease in power at the low-frequency end of the spectrum. This corresponds to the spherical and linear components in the variogram. The other striking feature is the peak at around 0.12 cycles. It corresponds to a wavelength of 8.4 sampling intervals or 34 m, which is very close to the wavelength (35 m) of the model fitted to the variogram. Evidently, the spectrum and the variogram are complementary ways of viewing the periodicity and estimating the period.

There is a smaller peak at 0.23 cycles. This is almost certainly a harmonic of the main peak at twice its frequency and may be disregarded. When the spectrum is viewed through a wide window (i.e. computed with the narrowest lag window,  $L = 10$  in Figure 7.5) the spectral peak is lost. In this example the bandwidth of the spectral window is much wider than the peak, as Figure 7.5 shows. Therefore, the spectral window must be narrower than the features that one wishes to reveal.

### **7.4.1 Bandwidths and confidence intervals for Caragabal**

In addition to the smoothed spectral estimates, Figure 7.5 shows the bandwidths by the length of the line corresponding to the lag windows 10, 25, 40 and 60. These are calculated for the Parzen windows simply by division of these widths into 1.86 (Figure 7.5). They are listed in Table 7.3.

The corresponding degrees of freedom, from equation (7.23), are  $3N/L$  for the Bartlett window and  $3.71N/L$  for the Parzen window, and Table 7.3 also lists their values for the transect.

**Table 7.3** Bandwidths and degrees of freedom for the smoothed spectrum of log electrical conductivity at Caragabal.

Lag window	Bartlett window		Parzen window	
	Bandwidth	Deg. freedom	Bandwidth	Deg. freedom
10	0.1500	109.5	0.1860	135.4
25	0.0600	43.8	0.0744	54.2
40	0.0375	27.4	0.0465	33.8
60	0.0250	18.3	0.0310	22.6

We can now obtain the confidence limits on the spectral density for any particular frequency. Let us take the Parzen lag window 10. With 365 sampling points this gives  $3.71 \times 365/10 = 135.4$  degrees of freedom. If we choose to work at the 90% confidence level, equivalent to  $\alpha = 0.1$ , then we need  $\chi^2$  for  $1 - \frac{1}{2}\alpha$  and  $\frac{1}{2}\alpha$ . These are 109.5 and 163.6, respectively. We now apply equation (7.25). If, for example, we want the confidence limits on our spectral estimate at frequency 0.15, which is  $\hat{R}_P(0.15) = 0.2286$ , then we calculate

$$c_{\text{lower}} = \frac{135.4}{163.6} \times 0.2286 = 0.189, \quad c_{\text{upper}} = \frac{135.4}{109.5} \times 0.2286 = 0.282.$$

These could be drawn on Figure 7.5(a), but if you are especially interested in the confidence of spectral estimates it is better to express the intervals on a logarithmic scale. Equation (7.25) becomes

$$\log \hat{R}(f) + \log \frac{\nu}{\chi^2_{\nu}(1 - \frac{1}{2}\alpha)} \quad \text{and} \quad \log \hat{R}(f) + \log \frac{\nu}{\chi^2_{\nu}(\frac{1}{2}\alpha)}. \quad (7.26)$$

The interval is constant and symmetric about the logarithm of the estimate. Taking the example above, we compute the logarithm (to base 10) of 0.2286 (which is  $-0.636$ ) and of the 90% confidence limits. The latter are  $-0.723$  and  $-0.589$ , giving a confidence interval of width 0.134 in the logarithms. Therefore if the spectrum itself is drawn on a logarithmic scale then the confidence interval can be represented as a single vertical line on the graph.

In Figure 7.5(b) the estimates of Figure 7.5(a) are converted to logarithms, and the results for the 90% confidence intervals are shown by the lengths of the vertical lines. The width of a confidence interval clearly depends on the width of the corresponding lag window. The wider is that window, and the narrower the bandwidth, the wider is the interval.

## 7.5 FURTHER READING ON SPECTRAL ANALYSIS

The theory of spectral analysis is extensive and complex, and it has numerous applications in many branches of science and engineering. Its principal merits in soil and environmental science are where there is periodicity. It is possible to detect periodicity in variograms and to model it. However, it is often easier to see the periodicity and to estimate it in the spectrum. If periodic variation is suspected from the variogram then spectral analysis can be used to confirm that it is present. Oliver *et al.* (1997) used geostatistics and spectral analysis in such a complementary way.

Two books that deal with spectral analysis at not too advanced a level are by Jenkins and Watts (1968) and by Priestley (1981). The first is intended for engineers, and numerate soil scientists should be able to cope with it. The second, though more mathematical, emphasizes the ideas.

## ***Local Estimation or Prediction: Kriging***

Most properties of the environment could be measured at any of an infinite number of places, but in practice they are measured at rather few, mainly for reasons of economy. If we wish to know their values elsewhere then we must estimate them from the data that we can obtain. The same holds if we want estimates over larger areas for which it has not been possible to measure or observe the properties directly. In Chapter 3 we considered the general problem of estimating values at unsampled places using either a discrete model of spatial variation and classification or a continuous model with deterministic interpolators. Many statisticians prefer to call the procedure *prediction* to distinguish it from estimating parameters of a distribution. In geostatistics, however, it is almost always called *estimation* for reasons explained by Matheron (1989); we shall use the two terms interchangeably unless there is any risk of misunderstanding. Estimation is the task for which geostatistics was developed initially, and it is generally called *kriging* after D. G. Krige, a mining engineer in the gold fields of South Africa (see Krige, 1966). The term was coined originally as *krigeage* by P. Carlier, but Matheron (1963) brought it into the English language in recognition of Krige's contribution to improving the precision of estimating concentrations of gold and other metals in ore bodies and recoverable reserves. Although much of the credit for formalizing the technique goes to Matheron and his colleagues at the Paris School of Mines, the mathematics of simple kriging had been worked out by A. N. Kolmogorov in the 1930s (Kolmogorov, 1939, 1941; see also Gandin, 1965), by Wold (1938) for time-series analysis, and only a little later by Wiener (1949). You can read a brief history of the subject in Cressie (1990).

## 8.1 GENERAL CHARACTERISTICS OF KRIGING

Kriging provides a solution to the problem of estimation based on a continuous model of stochastic spatial variation. It makes the best use of existing knowledge by taking account of the way that a property varies in space through the variogram model. In its original formulation a kriged estimate at a place was simply a linear sum or weighted average of the data in its neighbourhood. Since then kriging has been elaborated to tackle increasingly complex problems in mining, petroleum engineering, pollution control and abatement, and public health. The term is now generic, embracing several distinct kinds of kriging, both linear and non-linear. In this chapter we deal with the simpler linear methods, and in Chapter 11 we consider non-linear ones. In linear kriging the estimates are weighted linear combinations of the data. The weights are allocated to the sample data within the neighbourhood of the point or block to be estimated in such a way as to minimize the estimation or kriging variance, and the estimates are unbiased. Kriging is optimal in this sense.

### 8.1.1 Kinds of Kriging

Kriging covers a range of least-squares methods of spatial prediction.

- *Ordinary kriging* of a single variable, as described in Section 8.2, is the most robust method and the one most used.
- *Simple kriging* (Section 8.9) is rather little used as it stands because we usually do not know the mean. It finds application in other forms such as indicator and disjunctive kriging in which the data are transformed to have known means.
- *Lognormal kriging* (Section 8.10) is ordinary kriging of the logarithms of the measured values. It is used for strongly positively skewed data that approximate a lognormal distribution.
- *Kriging with drift* (Chapter 9), also known as *universal kriging*, recognizes both non-stationary deterministic and random components in a variable, estimates the trend in the former and the variogram of the latter, and recombines the two for prediction. This introduces residual maximum likelihood into the kriging procedure (see Section 9.2).
- *Factorial kriging* or *kriging analysis* (Chapter 9) is of particular value where the variation is nested, i.e. more than one scale of variation is present. Factorial kriging estimates the individual components of variation separately, but in a single analysis.
- *Ordinary cokriging* (Chapter 10) is the extension of ordinary kriging of a single variable to two or more variables. There must be some coregionalization among the variables for it to be profitable. It is particularly useful if some property that can be measured cheaply at many sites is spatially

correlated with one or more others that are expensive to measure and are measured at many fewer sites. It enables us to estimate the more sparsely sampled property with more precision by cokriging using the spatial information from the more intensely measured one.

- *Indicator kriging* (see Chapter 11) is a non-linear, non-parametric form of kriging in which continuous variables are converted to binary ones (indicators). It is becoming popular because it can handle distributions of almost any kind, and empirical cumulative distributions of estimates can be computed and thereby provide confidence limits on them. It can also accommodate 'soft' qualitative information to improve prediction.
- *Disjunctive kriging* (see Chapter 11) is also a non-linear method of kriging, but it is strictly parametric. It is valuable for decision-making because the probabilities of exceeding or not exceeding a predefined threshold are determined in addition to the kriged estimates.
- *Probability kriging* (not described further in this book) was proposed by Sullivan (1984) because indicator kriging does not take into account the proximity of a value to the threshold, but only its position. It uses the rank order for each value,  $z(\mathbf{x})$ , normalized to 1 as the secondary variable to estimate the indicator by cokriging. Chilès and Delfiner (1999) and Goovaerts (1997) describe the method briefly.
- *Bayesian kriging* (not described further in this book) was introduced by Omre (1987) for situations in which there is some prior knowledge about the drift. It is intermediate between simple kriging, used when there is no drift, and universal kriging where there is known to be drift. The kriging equations are those of simple kriging, but with non-stationary covariances (Chilès and Delfiner, 1999).

## 8.2 THEORY OF ORDINARY KRIGING

The aim of kriging is to estimate the value of a random variable,  $Z$ , at one or more unsampled points or over larger blocks, from more or less sparse sample data on a given support, say  $z(\mathbf{x}_1), z(\mathbf{x}_2), \dots, z(\mathbf{x}_N)$ , at points  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ . The data may be distributed in one, two or three dimensions, though applications in the environmental sciences are usually two-dimensional.

Ordinary kriging is by far the most common type of kriging in practice, and for this reason we focus on its theory here. It is based on the assumption that we do not know the mean. If we consider punctual estimation first, then we estimate  $Z$  at a point  $\mathbf{x}_0$  by  $\hat{Z}(\mathbf{x}_0)$ , with the same support as the data, by

$$\hat{Z}(\mathbf{x}_0) = \sum_{i=1}^N \lambda_i z(\mathbf{x}_i), \quad (8.1)$$

where  $\lambda_i$  are the weights. To ensure that the estimate is unbiased the weights are made to sum to 1,

$$\sum_{i=1}^N \lambda_i = 1,$$

and the expected error is  $E[\hat{Z}(\mathbf{x}_0) - Z(\mathbf{x}_0)] = 0$ . The estimation variance is

$$\begin{aligned} \text{var}[\hat{Z}(\mathbf{x}_0)] &= E[\{\hat{Z}(\mathbf{x}_0) - Z(\mathbf{x}_0)\}^2] \\ &= 2 \sum_{i=1}^N \lambda_i \gamma(\mathbf{x}_i, \mathbf{x}_0) - \sum_{i=1}^N \sum_{j=1}^N \lambda_i \lambda_j \gamma(\mathbf{x}_i, \mathbf{x}_j), \end{aligned} \quad (8.2)$$

where  $\gamma(\mathbf{x}_i, \mathbf{x}_j)$  is the semivariance of  $Z$  between the data points  $\mathbf{x}_i$  and  $\mathbf{x}_j$ , and  $\gamma(\mathbf{x}_i, \mathbf{x}_0)$  is the semivariance between the  $i$ th data point and the target point  $\mathbf{x}_0$ .

In the more general case we may wish to estimate  $Z$  in a block  $B$ , which may be a line, an area or a volume depending on whether it is in one, two or three spatial dimensions. The kriged estimate in  $B$  is still a simple weighted average of the data,

$$\hat{Z}(B) = \sum_{i=1}^N \lambda_i z(\mathbf{x}_i), \quad (8.3)$$

but with  $\mathbf{x}_0$  of equation (8.1) replaced by  $B$ . Its variance is

$$\begin{aligned} \text{var}[\hat{Z}(B)] &= E[\{\hat{Z}(B) - Z(B)\}^2] \\ &= 2 \sum_{i=1}^N \lambda_i \bar{\gamma}(\mathbf{x}_i, B) - \sum_{i=1}^N \sum_{j=1}^N \lambda_i \lambda_j \gamma(\mathbf{x}_i, \mathbf{x}_j) - \bar{\gamma}(B, B). \end{aligned} \quad (8.4)$$

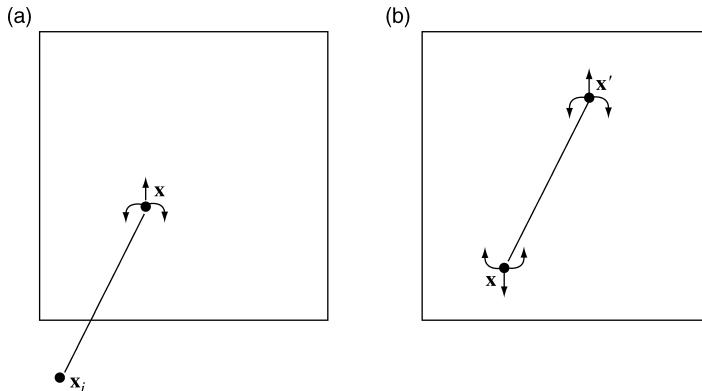
The quantity  $\bar{\gamma}(\mathbf{x}_i, B)$  is the average semivariance between the  $i$ th sampling point and the block  $B$  and is the integral

$$\bar{\gamma}(\mathbf{x}_i, B) = \frac{1}{|B|} \int_B \gamma(\mathbf{x}_i, \mathbf{x}) d\mathbf{x}, \quad (8.5)$$

where  $\gamma(\mathbf{x}_i, \mathbf{x})$  denotes the semivariance between the sampling point  $\mathbf{x}_i$  and a point  $\mathbf{x}$  describing the block, Figure 8.1(a). The third term on the right-hand side of equation (8.4) is the double integral

$$\bar{\gamma}(B, B) = \frac{1}{|B|^2} \int_B \int_B \gamma(\mathbf{x}, \mathbf{x}') d\mathbf{x} d\mathbf{x}', \quad (8.6)$$

where  $\gamma(\mathbf{x}, \mathbf{x}')$  is the semivariance between two points  $\mathbf{x}$  and  $\mathbf{x}'$  that sweep independently over  $B$ , Figure 8.1(b). It is the within-block variance. In punctual



**Figure 8.1** Integration of the variogram: (a) between a sampling point and a block; (b) within a block.

kriging  $\bar{v}(B, B)$  becomes  $v(\mathbf{x}_0, \mathbf{x}_0) = 0$ , which is why equation (8.2) has two terms rather than three.

For each kriged estimate there is an associated kriging variance, which we can denote by  $\sigma^2(\mathbf{x}_0)$  and  $\sigma^2(B)$  for the point and block, respectively, and which are defined in equations (8.2) and (8.4). The next step in kriging is to find the weights that minimize these variances, subject to the constraint that they sum to 1. We achieve this using the method of Lagrange multipliers.

We define an auxiliary function  $f(\lambda_i, \psi)$  that contains the variance we wish to minimize plus a term containing a Lagrange multiplier,  $\psi$ . For punctual kriging it is

$$T(\lambda_i, \psi) = \text{var}[\hat{Z}(\mathbf{x}_0) - z(\mathbf{x}_0)] - 2\psi \left\{ \sum_{i=1}^N \lambda_i - 1 \right\}. \quad (8.7)$$

We then set the partial derivatives of the function with respect to the weights to 0:

$$\begin{aligned} \frac{\partial f(\lambda_i, \psi)}{\partial \lambda_i} &= 0, \\ \frac{\partial f(\lambda_i, \psi)}{\partial \psi} &= 0, \end{aligned} \quad (8.8)$$

for  $i = 1, 2, \dots, N$ . This leads to a set of  $N + 1$  equations in  $N + 1$  unknowns:

$$\begin{aligned} \sum_{i=1}^N \lambda_i v(\mathbf{x}_i, \mathbf{x}_j) + \psi(\mathbf{x}_0) &= v(\mathbf{x}_j, \mathbf{x}_0) \quad \text{for all } j, \\ \sum_{i=1}^N \lambda_i &= 1. \end{aligned} \quad (8.9)$$

This is the ordinary kriging system for points. Its solution provides the weights,  $\lambda_i$ , which are entered into equation (8.1), and from which the estimation variance (prediction variance or specifically kriging variance) can be obtained as

$$\sigma^2(\mathbf{x}_0) = \sum_{i=1}^N \lambda_i \gamma(\mathbf{x}_i, \mathbf{x}_0) + \psi(\mathbf{x}_0). \quad (8.10)$$

If a target point,  $\mathbf{x}_0$ , happens to be one of the data points, say  $\mathbf{x}_j$ , then  $\sigma^2(\mathbf{x}_0)$  is minimized when  $\lambda(\mathbf{x}_j) = 1$  and all of the other weights are 0. In fact,  $\sigma^2(\mathbf{x}_0) = 0$ , and by inserting the weights into equation (8.1) we obtain the recorded value,  $z(\mathbf{x}_j)$ , as our estimate of  $z(\mathbf{x}_0)$ . Punctual kriging is thus an exact interpolator.

The equivalent kriging system for blocks is

$$\begin{aligned} \sum_{i=1}^N \lambda_i \gamma(\mathbf{x}_i, \mathbf{x}_j) + \psi(B) &= \bar{\gamma}(\mathbf{x}_j, B) \quad \text{for all } j, \\ \sum_{i=1}^N \lambda_i &= 1, \end{aligned} \quad (8.11)$$

with the associated variance obtained as

$$\sigma^2(B) = \sum_{i=1}^N \lambda_i \bar{\gamma}(\mathbf{x}_i, B) + \psi(B) - \bar{\gamma}(B, B). \quad (8.12)$$

The kriging equations can be represented in matrix form. For punctual kriging they are

$$\mathbf{A}\boldsymbol{\lambda} = \mathbf{b} \quad (8.13)$$

where

$$\begin{aligned} \mathbf{A} &= \begin{bmatrix} \gamma(\mathbf{x}_1, \mathbf{x}_1) & \gamma(\mathbf{x}_1, \mathbf{x}_2) & \cdots & \gamma(\mathbf{x}_1, \mathbf{x}_N) & 1 \\ \gamma(\mathbf{x}_2, \mathbf{x}_1) & \gamma(\mathbf{x}_2, \mathbf{x}_2) & \cdots & \gamma(\mathbf{x}_2, \mathbf{x}_N) & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma(\mathbf{x}_N, \mathbf{x}_1) & \gamma(\mathbf{x}_N, \mathbf{x}_2) & \cdots & \gamma(\mathbf{x}_N, \mathbf{x}_N) & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix}, \\ \boldsymbol{\lambda} &= \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_N \\ \psi(\mathbf{x}_0) \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} \gamma(\mathbf{x}_1, \mathbf{x}_0) \\ \gamma(\mathbf{x}_2, \mathbf{x}_0) \\ \vdots \\ \gamma(\mathbf{x}_N, \mathbf{x}_0) \\ 1 \end{bmatrix}. \end{aligned}$$

Matrix  $\mathbf{A}$  is inverted, and the weights and the Lagrange multiplier are obtained as

$$\boldsymbol{\lambda} = \mathbf{A}^{-1} \mathbf{b}. \quad (8.14)$$

The kriging variance is given by

$$\hat{\sigma}^2(\mathbf{x}_0) = \mathbf{b}^T \boldsymbol{\lambda}. \quad (8.15)$$

For block kriging the only differences are that

$$\mathbf{b} = \begin{bmatrix} \bar{\gamma}(\mathbf{x}_1, B) \\ \bar{\gamma}(\mathbf{x}_2, B) \\ \vdots \\ \bar{\gamma}(\mathbf{x}_N, B) \\ 1 \end{bmatrix}$$

and

$$\hat{\sigma}^2(B) = \mathbf{b}^T \boldsymbol{\lambda} - \bar{\gamma}(B, B). \quad (8.16)$$

### 8.3 WEIGHTS

When the kriging equations are solved to obtain the weights,  $\lambda_i$ , in general the only large weights are those of the points near to the point or block to be kriged. The nearest four or five might contribute 80% of the total weight, and the next nearest ten almost all of the remainder. The weights also depend on the configuration of the sampling. We can summarize the factors affecting the weights as follows.

1. Near points carry more weight than more distant ones. Their relative proportions depend on the positions of the sampling points and on the variogram: the larger is the nugget variance, the smaller are the weights of the points that are nearest to target point or block.
2. The relative weights of points also depend on the block size: as the block size increases, the weights of the nearest points decrease and those of the more distant points increase (Figure 8.9), until the weights become nearly equal.
3. Clustered points carry less weight individually than isolated ones at the same distance (Figure 8.12).

4. Data points can be screened by ones lying between them and the target (Figure 8.12).

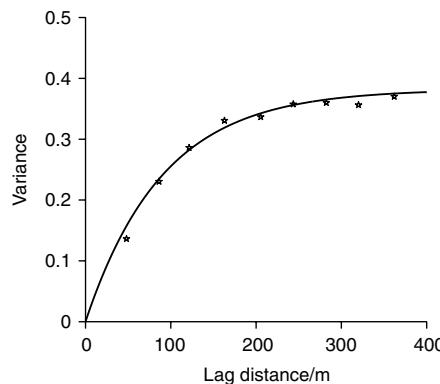
These effects are all intuitively desirable, and the first shows that kriging is local. They will become evident in the examples below. They also have practical implications. The most important for present purposes is that because only the nearest few data points to the target carry significant weight, matrix  $A$  in the kriging system need never be large and its inversion will be swift. We can replace  $N$  in equations (8.9) and (8.11) by some much smaller number, say  $n \ll N$ . We shall reiterate this below after the examples in which we set  $n$  to 16.

## 8.4 EXAMPLES

This section shows the effects of a changing variogram, target point and sampling intensity on the weights in a way analogous to the kriging exercises in GSLIB (Deutsch and Journel, 1992). It uses the data on pH from Broom's Barn Farm for the purpose. We have chosen pH because it is easy to appreciate changes in the estimated values and because we can start with a simple isotropic exponential model without nugget (Figure 8.2), which is the best-fitting model:

$$\gamma(h) = c\{1 - \exp(-h/r)\}, \quad (8.17)$$

with  $c = 0.382$  and  $r = 90.53$  m, i.e. an effective range ( $a' = 3r$ ) of approximately 272 m (Table 8.1). We have also selected  $n = 16$  points on a  $4 \times 4$  lattice from the full set of data (Figure 8.3). There are also three separate target points, one at the centre of the lattice, Figure 8.3(a), one off-centre, Figure 8.3(b), and a



**Figure 8.2** Variogram of pH at Broom's Barn Farm. The points are the experimental semivariances, and the solid line is the best fitting exponential model, the parameters of which are given in the text.

**Table 8.1** Model parameters with changing ratio of nugget:sill variance and fixed distance parameter,  $r = 90.53$  m, equivalent to an effective range of 271.6 m.

Model	$c_0$	$c$
Exponential N1	0	0.3820
Exponential N2	0.1	0.2820
Exponential N3	0.3	0.0820
Exponential N4	0.382	0
(Pure nugget)	1	0

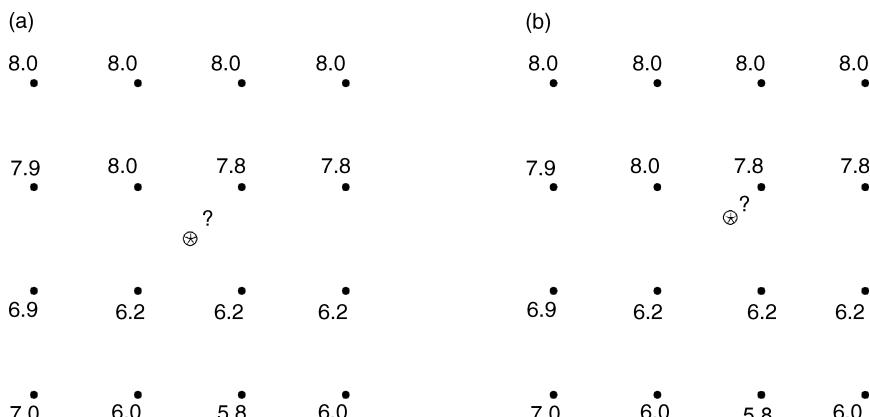
third coinciding with one of the sampling points, Figure 8.11(c). Using equation (8.9) and the 16 points we estimated the values at the target points as follows.

### 8.4.1 Kriging at the centre of the lattice

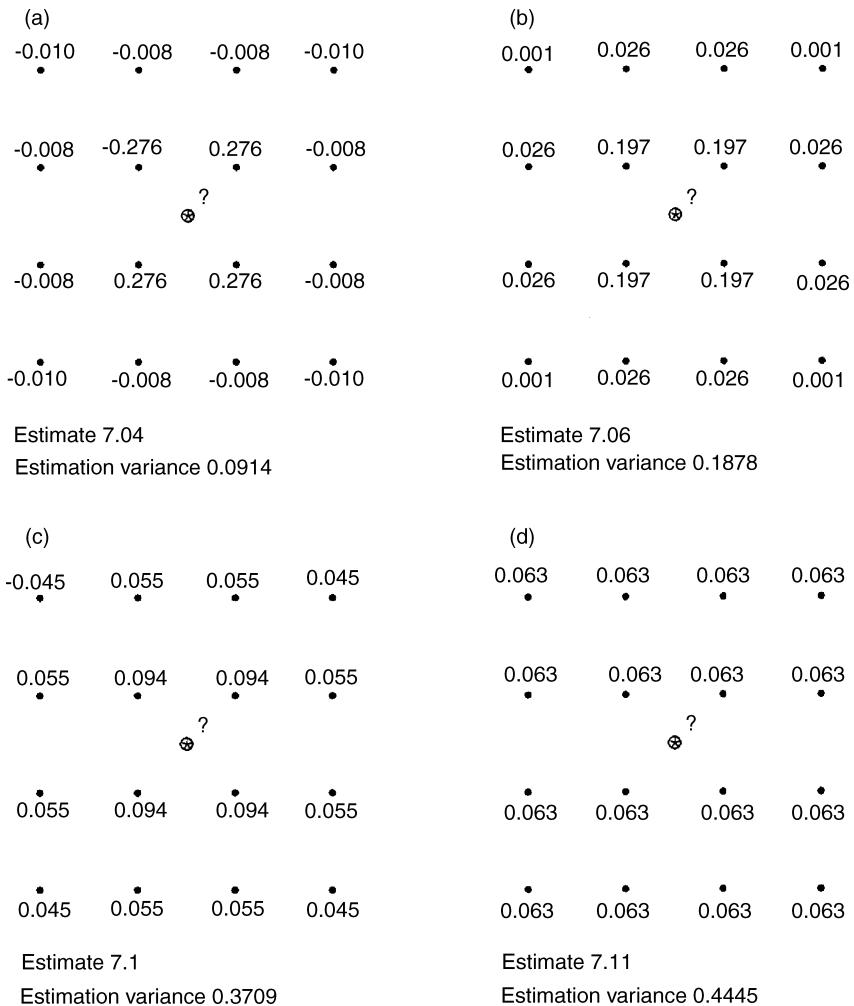
## Changing the ratio of nugget:sill variance

Figure 8.4(a) shows the weights derived using the best-fitting model to pH (exponential N1, Table 8.1; and exponential R2, Table 8.2). The weights of the four points nearest to the target point are large and positive, and their sum exceeds 1. To ensure unbiasedness the sum of all the weights must be 1, and hence the weights of the outer points are negative. In this case the outer points are close to 0 and so have little influence on the estimate.

We now change the variogram by introducing a nugget variance,  $c_0 = 0.1$  (the model parameters are those of exponential N2 in Table 8.1). The resulting



**Figure 8.3** The grid of 16 sample values selected from Broom's Barn Farm with the pH values given for each sampling location. The point to be estimated is located: (a) centrally; (b) off-centre.



**Figure 8.4** Kriging weights from punctual kriging of pH with an exponential function with the distance parameter  $r = 90.53$  m, and changing the nugget:sill variance: (a)  $c_0 = 0$ ,  $c = 0.382$ ; (b)  $c_0 = 0.1$ ,  $c = 0.282$ ; (c)  $c_0 = 0.3$ ,  $c = 0.082$ ; (d)  $c_0 = 0.382$ ,  $c = 0$ .

**Table 8.2** Model parameters with changing distance parameter,  $r$ , for exponential model.

Model	$c_0$	$c$	$r/m$	Effective range/m
Exponential R1	0	0.382	133.3	400.00
Exponential R2	0	0.382	90.53	271.59
Exponential R3	0	0.382	26.67	80.00
Exponential R4	0	0.382	6.67	20.00

weights are shown in Figure 8.4(b): those of the inner four points have decreased somewhat, whilst those of the outer points have increased and are now all positive. The weights of the corner points of the lattice are the smallest because they are the furthest from  $\mathbf{x}_0$ .

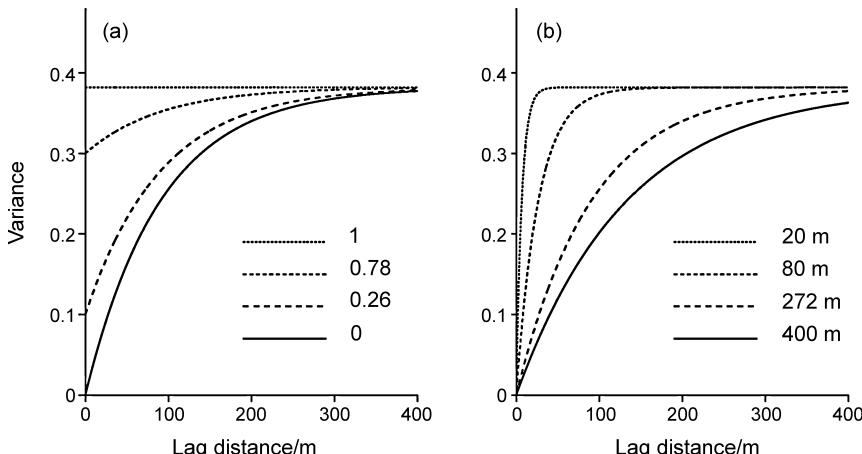
Figure 8.4(c) shows the weights obtained by increasing the proportion of nugget more substantially so that it dominates (the model parameters are those of exponential N3 in Table 8.1). The weights of the inner points have decreased considerably, and those of the outer ones have increased correspondingly.

For a pure nugget variogram, with parameters exponential (N4) in Table 8.1, the weights are all the same (Figure 8.4(d)). The result is the same as if we had sampled at random in classical estimation; the kriging variance is the variance of the process,  $c_0$ , plus the variance of the mean, given by  $\psi(\mathbf{x}_0)$ . The solution of equation (8.15) is

$$c_0 + \psi(\mathbf{x}_0) = 0.382 + 0.0625 = 0.4445. \quad (8.18)$$

Figure 8.5(a) summarizes the shapes of the exponential variogram models that resulted from changing the ratio of nugget:sill variance and keeping the distance parameter constant.

The estimated value for pH and kriging variance are given for each of the above examples (Figure 8.4). The estimated value changes each time: we can assume that 7.04 is the optimal estimate because it was derived from the best-fitting model. The average pH of the 16 values is 7.11, which is also the estimate returned with a pure nugget variogram and for which the kriging variance is the largest. The kriging variance increases as the nugget variance



**Figure 8.5** (a) Exponential variograms used to obtain the weights in Figure 8.4 with the distance parameter  $r = 90.53$  m, and changing the nugget:sill variance: 0:0.382; 0.1:0.282; 0.3:0.082; 0.382:0. (b) Effect on the exponential variogram of changing the effective range ( $a' = 3r$ ) with  $c_0 = 0$  and  $c = 0.382$ :  $a' = 400, 271.59, 80, 20$ .

increases, as we should expect, because the greater the variance that remains unresolved the more uncertain is the estimate. The estimates and their associated variances illustrate two points:

- (i) A nugget variance increases the kriging variance, and for punctual kriging it sets a lower limit to that variance (see Figures 8.17 and 11.8(b)).
- (ii) It is important to fit the model correctly to the sample semivariances because of the effect of the model on both the estimates and their variances.

Although the kriging variances are smaller for a smaller nugget variance, the model must represent the nugget effect realistically. If it does not then the estimates could be judged to be more or less reliable than they really are.

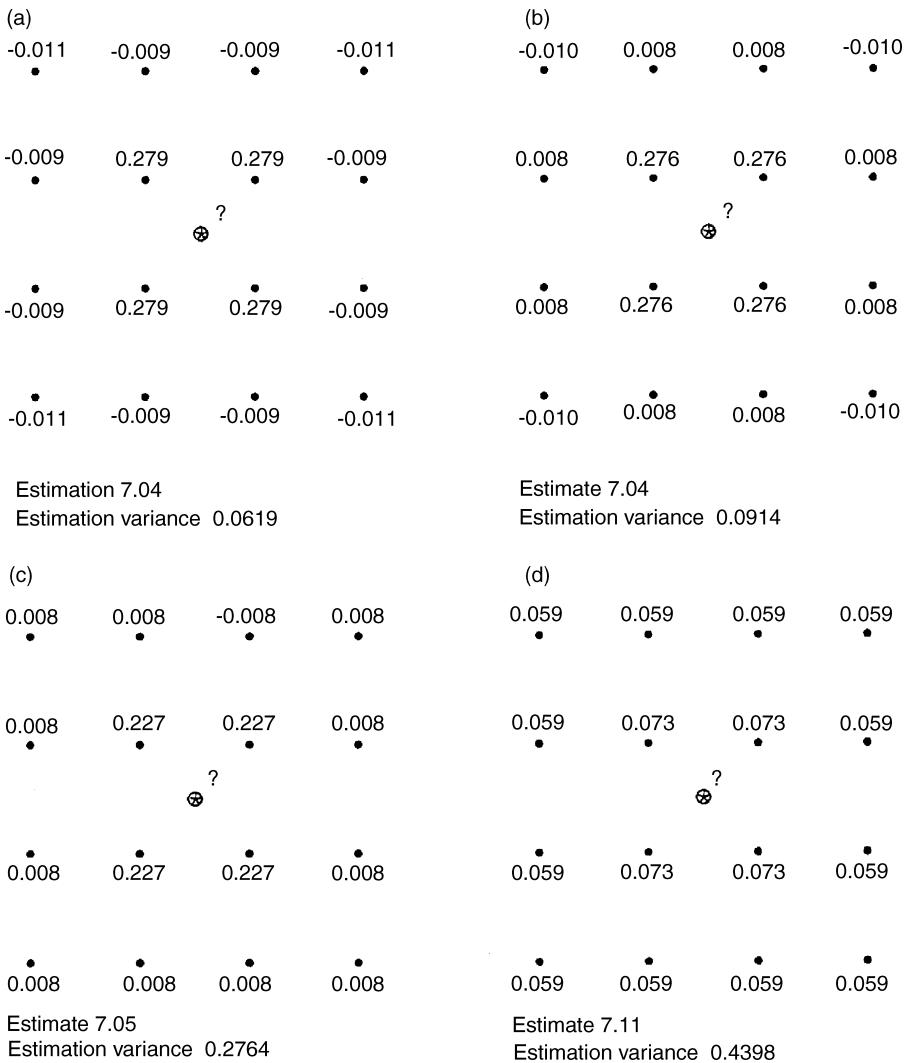
### **Changing the range or sampling intensity**

We now explore the effect of decreasing the range of spatial dependence, or, what amounts to the same thing, decreasing the sampling density. The nugget variance and the sill of the spatially dependent component,  $c$ , were kept constant and we changed the distance parameter, as shown in Table 8.2. Figure 8.5(b) shows the effect on the shape of the exponential variogram, and Figure 8.6(a) shows the weights for exponential R1, where the effective range of dependence ( $a' = 3r$ ) is 400 m. The weights of the inner four points are the largest, and the outer ones contribute little or nothing. If we compare this with Figure 8.6(b) for the best-fitting exponential model R2, it is clear that they are similar. As the effective range lengthens, however, the inner points gain weight in accordance with the increase in spatial continuity in the variation. If we reduce the effective range substantially to 80 m (exponential (R3)), then the weights of the inner points decrease and those of the outer ones increase (Figure 8.6(c)). When we reduce the effective range to half the sampling interval, i.e. 20 m (exponential R4), the variogram is effectively pure nugget. Figure 8.6(d) shows the weights which are now small for all of the points, though they are not all the same: the inner ones are somewhat larger than the outer ones, because with the exponential model the distance parameter does not disappear completely. Nevertheless, the estimate is the mean of the data as in the previous example, Figure 8.4(a), but the kriging variance is a little less because of the effect of the differences in the weights.

Since changing the distance parameter of a spherical model has a different effect, we include the results of changing the range of the best-fitting spherical function to the 16 points. The spherical function is given by

$$\gamma(h) = c_0 + c \left\{ \frac{3h}{2a} - \frac{1}{2} \left( \frac{h}{a} \right)^3 \right\}, \quad (8.19)$$

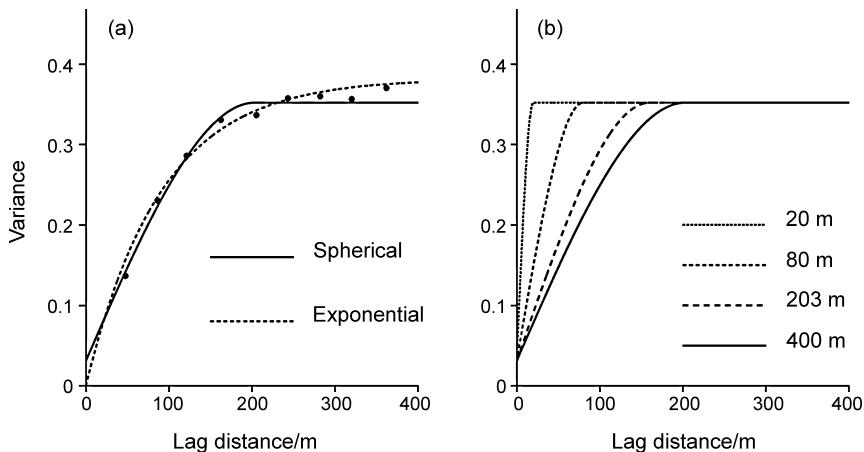
with the parameter values  $c_0 = 0.0309$ ,  $c = 0.3211$  and  $a = 203.2$  m for the best-fitting spherical function.



**Figure 8.6** Kriging weights from punctual kriging of pH with an exponential function with  $c_0 = 0$  and  $c = 0.382$ , and changing the effective range ( $a' = 3r$ ): (a)  $a' = 400$ ; (b)  $a' = 203.2$ ; (c)  $a' = 80.0$ ; (d)  $a' = 20.0$ .

There are several interesting differences between the results of these models. Figure 8.7(a) shows the best-fitting spherical and exponential models fitted to pH, and Figure 8.7(b) shows the effect of changing the range on the shape of the spherical model.

To compare the weights with those for the exponential model we started with spherical A1 of Table 8.3, with a range of 400 m. The weights of the inner



**Figure 8.7** (a) The best-fitting spherical (solid line) and exponential (dashed line) fitted to the experimental variogram of pH. (b) Spherical variograms used to obtain the weights in Figure 8.8 with  $c_0 = 0.031$ ,  $c = 0.321$ , and range  $a = 203.2$  m, 160 m, 80 m, 20 m.

points are smaller, and those of the outer ones larger, than those for the exponential model, Figure 8.8(a). This is because the spherical model has a small nugget variance, whereas the exponential had none, and there is a difference in the curvature of these two models (Figure 8.7(a)). Figure 8.8(b) shows the weights obtained when using the best-fitting spherical function, spherical A2; the inner weights are larger and the outer ones slightly smaller. It is a reversal of the effect with the exponential model. When the range is reduced to 80 m, spherical (A3), the inner weights, Figure 8.8(c), are much larger than for the equivalent exponential model, Figure 8.6(c), again because of the effect of the model's curvature. Finally, when the range is 20 m the weights are all the same, Figure 8.8(d), and the observed effect is the same as that for the pure nugget variogram. In this situation all of the variation occurs within the sampling interval. It illustrates clearly that if the distance over which most of the variation occurs is less than the sampling interval then the simple formula for random sampling gives the best estimate for an unsampled point, which is the mean of the data. It also shows the importance of sampling sufficiently densely to estimate the variogram at the spatial scale of the investigation.

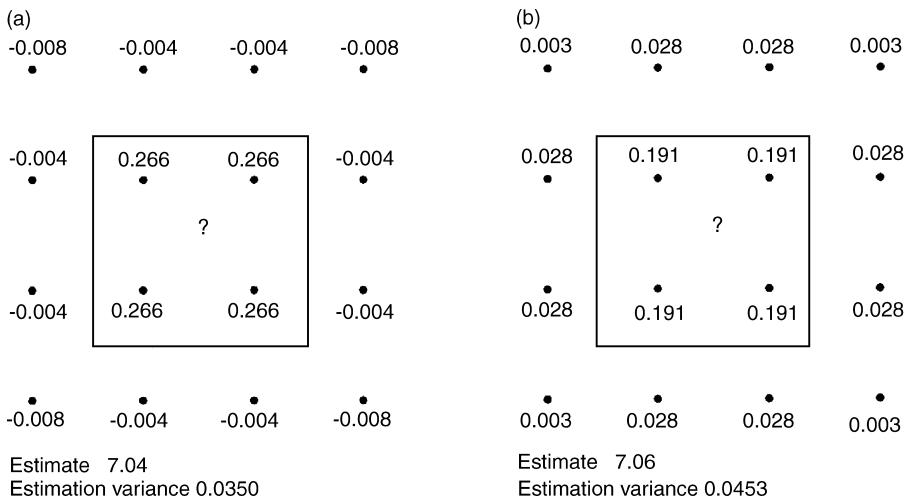
**Table 8.3** Parameters with the range changing for spherical model.

Model	$c_0$	$c$	Range/m
Spherical A1	0.0309	0.3211	400.0
Spherical A2	0.0309	0.3211	203.2
Spherical A3	0.0309	0.3211	80.0
Spherical A4	0.0309	0.3211	20.0

(a)				(b)			
-0.010	0.021	0.021	-0.010	-0.017	0.010	0.010	-0.017
0.021	0.219	0.219	0.021	0.010	0.248	0.248	0.010
	?				?		
⊗				⊗			
0.021	0.219	0.219	0.021	0.010	0.248	0.248	0.010
-0.010	0.021	0.021	-0.010	-0.017	0.010	0.010	-0.017
Estimate	7.05			Estimate	7.04		
Estimation variance	0.0580			Estimation variance	0.0872		
(c)				(d)			
0.009	-0.027	-0.027	0.009	0.063	0.063	0.063	0.063
-0.027	0.294	0.294	-0.027	0.063	0.063	0.063	0.063
	?				?		
⊗				⊗			
-0.027	0.294	0.294	-0.027	0.063	0.063	0.063	0.063
0.009	-0.027	-0.027	0.009	0.063	0.063	0.063	0.063
Estimate	7.05			Estimate	7.11		
Estimation variance	0.1953			Estimation variance	0.4445		

**Figure 8.8** Kriging weights from punctual kriging of pH with a spherical function with  $c_0 = 0.031$ ,  $c = 0.321$ , and changing the distance parameter (range): (a)  $a = 400$  m; (b)  $a = 203.2$  m; (c)  $a = 80$  m; (d)  $a = 20$  m.

In summary, Figure 8.8(a)–(c) shows that as the distance parameter decreases the weights of the inner points increase and those of the outer ones decrease. Apart from Figure 8.8(a), the estimates are sensibly the same as those for the exponential model, but the kriging variances for the spherical model are smaller in every case.



**Figure 8.9** Kriging weights from block kriging of pH over a centrally located block of 80 m × 80 m: (a) for the best-fitting exponential model with  $c_0 = 0$ ,  $c = 0.382$  and  $r = 90.53$  m; (b) with  $c_0 = 0.1$ ,  $c = 0.282$  and  $r = 90.53$  m.

### Kriging over a block

The results for block kriging over a centrally located 80 m × 80 m block with the parameters of the best-fitting exponential model (N1 in Table 8.1), are shown in Figure 8.9(a), and those with exponential (N2) in Figure 8.9(b). A comparison of the weights in Figures 8.4(a) and 8.9(a) shows that by increasing the block size the inner weights decrease and the outer ones increase. The differences between the two figures are small, but the kriging variance for block kriging with this model is only a little more than a third of that for punctual kriging. With a modest nugget variance, exponential N2, the relative decrease in the inner weights for block kriging, Figure 8.9(b), is somewhat less than in Figure 8.4(b), but the decrease in the kriging variance over that of punctual kriging is even more marked; it is now less than a quarter. Nevertheless, the estimated values are the same in each case.

This comparison shows two effects of the nugget variance as follows:

1. The nugget variance sets a lower limit to the punctual kriging variance.
2. The nugget variance disappears from the block-kriging variance; see equations (8.4) and (8.12). Therefore, the larger is the proportion of the nugget variance, which is taken out of consideration, the smaller is the block-kriging variance and the greater is the difference between it and the punctual kriging variance.

It also raises an important issue of confidence. When practitioners fit models to variograms, whether by eye or by minimizing some function of the residuals, they

project their models towards the ordinate with the least change in curvature. They know nothing about the shape of the variogram at distances less than the shortest lag interval, and the practice may be regarded as prudent. The intercept gives them a nugget variance that is almost certainly larger than any error of measurement or short-range spatial component. When the model is used for punctual kriging the errors will, therefore, tend to be on the large side; the estimates are conservative. However, when the same model is used for block kriging, if the nugget variance is exaggerated then the kriging variance will be too small for the reasons given above, and the practitioner will obtain a false sense of confidence.

### The effect of anisotropy

We examine the effect of geometric anisotropy on the weights by punctual kriging with the exponential model

$$\gamma(h, \vartheta) = c\{1 - \exp(-h/\Omega)\}, \quad (8.20)$$

where

$$\Omega = \sqrt{A^2 \cos^2(\vartheta - \varphi) + B^2 \sin^2(\vartheta - \varphi)}, \quad (8.21)$$

in which  $A = 271.6$ ,  $B = 90.5$  and  $\varphi = \pi/2 = 0.7854$  rad or  $45^\circ$ . The angle  $\varphi$  is the direction of maximum continuity, i.e. largest effective range, as in Figure 5.13. The ratio  $A/B$  is the anisotropy ratio, and is  $3 = 271.6/90.5$ . Figure 8.10(a) shows the weights for the isotropic variogram and Figure 8.10(b) those for the anisotropic function. The largest weights are at the points adjacent to the target point along the  $45^\circ$  diagonal.

There is a marked decrease in the weights of the adjacent points at right angles. The changes in the weights of the outer points are far less marked. If we change  $\varphi$  to  $0.2618$  rad, or  $15^\circ$  ( $75^\circ$  in geographical notation) then Figure 8.10(c) ensues; the distribution of the weights has changed. The increase in the weights of the nearest points is less dramatic, but the outer weights close to the ( $15^\circ$ ) line have increased substantially to 0.153.

#### 8.4.2 Kriging off-centre in the lattice and at a sampling point

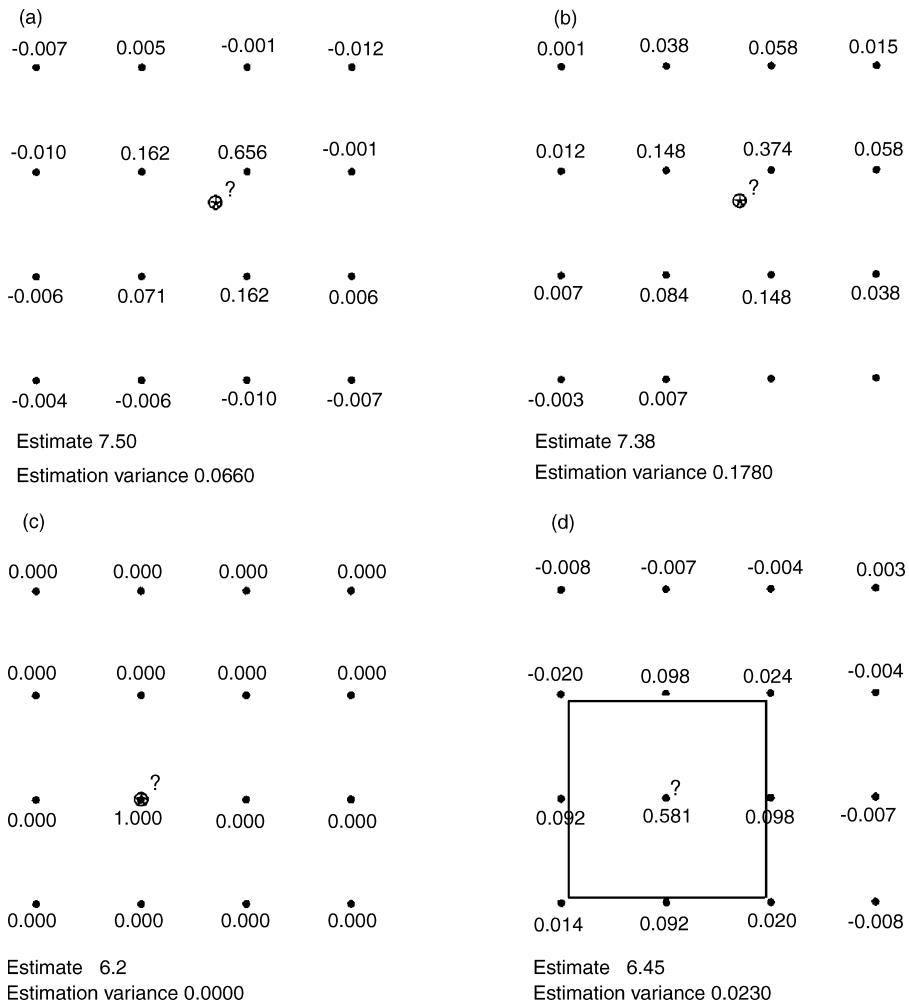
Let us now use the exponential models (Table 8.1), one with no nugget N1 and the other with a nugget variance of 0.1 N2, to estimate the value at a target point that is off-centre but on a diagonal of the grid (Figure 8.11). In both Figure 8.11(a) and 8.11(b) the point closest to the target has the largest weight, and the point diagonally opposite has the smallest weight of the four inner points. The other two inner points have the same weights because they are equidistant from the target. The weights of the outer points now show the effect

(a)				(b)			
-0.010 •	-0.008 •	-0.008 •	-0.010 •	0.001 •	-0.003 •	0.009 •	-0.011
-0.008 •	0.276 •	0.276 •	-0.008 •	-0.003 •	0.043 •	0.457 •	0.009 •
?	⊗	?	⊗	?	⊗	?	⊗
-0.008 •	0.276 •	0.276 •	-0.008 •	0.009 •	0.457 •	0.043 •	-0.003
-0.010 •	-0.008 •	-0.008 •	-0.010 •	-0.011 •	0.009 •	-0.003 •	0.001
Estimate 7.06 Estimation variance 0.1878				Estimate 7.00 Estimation variance 0.1151			
(c)							
0.001 •	-0.002 •	-0.007 •	-0.021 •	0.005 •	0.083 •	0.289 •	0.153 •
?	⊗	?	⊗	?	⊗	?	⊗
0.153 •	0.289 •	0.083 •	0.005 •	-0.021 ■	-0.007 ■	-0.002 ■	0.001 ■
Estimate 7.10 Estimation variance 0.1759							

**Figure 8.10** Kriging weights from punctual kriging of pH: (a) for the best-fitting exponential model with  $c_0 = 0$ ,  $c = 0.382$ , and  $r = 90.53$  m; (b) for an anisotropic exponential model with the direction of maximum variation  $\pi/2$  radians and an anisotropy ratio of 3; (c) with the direction of maximum variation  $1.309$  rad.

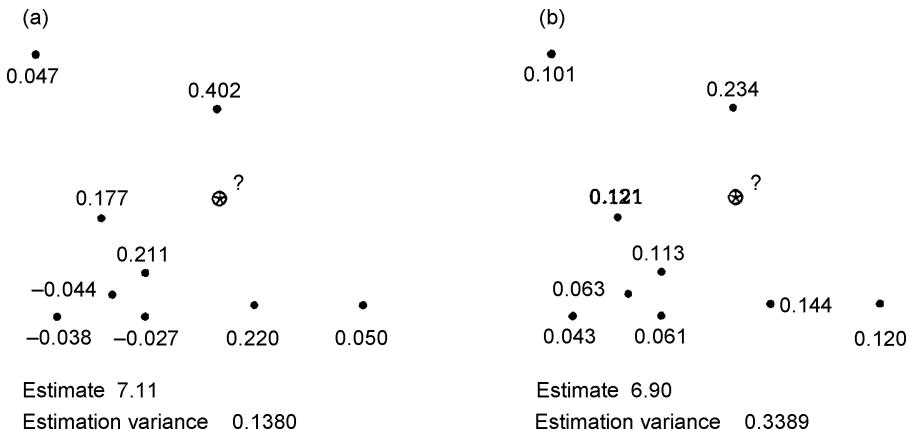
of screening. Figure 8.11(a) shows that the unscreened outer points have positive weights, whereas those that are screened are negative.

The weights in Figure 8.11(c) were obtained by solution of the kriging system for punctual kriging at the target point coinciding with the sampling location indicated. They are 1 at the sampling point and 0 elsewhere, which is what we should expect from theory. The estimate is the sample value, and the estimation variance is 0, so illustrating that ordinary punctual kriging is an exact interpolator. The weights in Figure 8.11(d) were obtained with the same model



**Figure 8.11** Kriging weights of pH with the exponential function: punctual kriging with the point to be estimated off-centre and the model: (a)  $c_0 = 0$ ,  $c = 0.382$ ,  $r = 90.53$  m; (b)  $c_0 = 0.1$ ,  $c = 0.282$ ,  $r = 90.53$  m; (c) with the point to be estimated at a sampling location with  $c_0 = 0$ ,  $c = 0.382$ ,  $r = 90.53$  m; (d) block kriging with an  $80 \text{ m} \times 80 \text{ m}$  block centred at a sampling point with  $c_0 = 0$ ,  $c = 0.382$  and  $r = 90.53$  m.

and kriging over a  $80 \text{ m} \times 80 \text{ m}$  block centred at the same sampling point. The weight at the sampling location is an order of magnitude larger than those of the surrounding nearest points, while those of the outer edges are negative. The estimate is substantially different from the measured value at the centre of the block and shows the smoothing effect of block kriging. The kriging variance is also very small, but not zero.



**Figure 8.12** Kriging weights from punctual kriging of pH with an exponential model and irregularly scattered sampling points: (a)  $c_0 = 0$ ,  $c = 0.382$  and  $r = 90.53$  m; (b)  $c_0 = 0.2$ ,  $c = 0.182$  and  $r = 90.53$  m.

### 8.4.3 Kriging from irregularly spaced data

Figure 8.12 shows an irregular configuration of nine sampling points plus a target point; the nine are a selection from the 16 values used previously, but some of the locations were changed. The weights in Figure 8.12(a) were obtained with the best-fitting exponential model N1, and punctual kriging. Those in Figure 8.12(b) were derived with exponential N2. The two diagrams show more clearly than those for the grid the effect of the data configuration on the weights. Points that are clustered carry less weight relative to isolated ones. The point to the north of the target carries almost twice the weight of the next most important point because it is far from any other point. The points that are screened by others have negative weights.

## 8.5 NEIGHBOURHOOD

The notion of the neighbourhood embodies the local nature of kriging, and it confers advantages on the method, as follows.

1. Only the nearest few points to the target point or block carry significant weight, therefore the kriging system need never be large and inverting matrix  $\mathbf{A}$  will be swift. We can replace  $N$  in equations (8.9) and (8.11) by a much smaller  $n \ll N$ . This might not matter when kriging only one point or block, but for mapping in which many estimates are needed it can make a big difference because the time required to invert a matrix is

approximately proportional to the cube of its order. It also avoids instability that can arise with large matrices.

2. If only the points near to the target carry significant weight then the variogram need be estimated and modelled well only at short lag distances, and in fact this is usually where the variogram is best estimated. The widening of the confidence intervals on the experimental variogram is somewhat less serious than it might appear from Chapter 5. This adds to the desirability of giving most weight to the experimental semivariances at the short lags when modelling the variogram.
3. The local nature of ordinary kriging means that what happens over large distances is of little consequence for the estimates. We can accept the notion of quasi-stationarity, i.e. local stationarity (Chapter 4), compute the variogram over only short distances, and apply it without taking account of long-range fluctuations in  $E[Z(\mathbf{x})]$ . The assumptions underpinning the method are not violated. It is perhaps this feature that has made ordinary kriging the 'workhorse' of geostatistics.

There are no strict rules for defining the neighbourhood, but we suggest some guidelines as follows:

1. If the variogram is bounded and has a small nugget variance and the data are dense then the radius of the neighbourhood can be set close to the range or effective range. Any data beyond the range will have negligible weights.
2. If data are sparse, however, points beyond the range from the target might carry sufficient weight to be important, and the neighbourhood should be such as to include them.
3. If the nugget variance is large, then again distant points are likely to carry significant weight.
4. As an alternative, the user may choose the nearest  $n$  data points, and effectively let this number limit the neighbourhood. If the sampling configuration is irregular then the size of the neighbourhood will vary more or less as the target point is moved. We have found that a maximum of  $n \approx 20$  is usually enough.
5. If you set a maximum radius for the neighbourhood then you may also need to set a minimum for  $n$ , especially to cater for targets near the boundary of a region. A value of  $n \approx 7$  is likely to be satisfactory.
6. Where the scatter is very uneven good practice is to divide the space around the target point into octants and take the nearest two points in each. Several kriging programs do this as a matter of course.

We recommend that when you start to analyse new data you examine what happens to the kriging weights as you change the neighbourhood. This is especially important in mapping where you move the neighbourhood. In these circumstances the most distant points should have zero weight so that the estimated surface appears seamless; see Laslett *et al.* (1987) for an illustrated discussion.

## 8.6 ORDINARY KRIGING FOR MAPPING

Kriging was developed in mining originally to estimate the amounts of metal in blocks of rock, and it is still used in this way. In these circumstances every block of rock is potentially of interest, and its metal content will be estimated. The miner may then decide whether the rock contains sufficient metal to be mined and sent for processing. Environmental scientists, and pedologists in particular, have used kriging in a rather different way, namely optimal interpolation at many places for mapping. The earliest examples are those by Burgess and Webster (1980a, 1980b) and Burgess *et al.* (1981), who used ordinary kriging. There have been many since, for example Mulla (1997), Frogbrook (1999) and Frogbrook *et al.* (1999) in precision agriculture.

To map a variable the values are kriged at the nodes of a fine grid. Isarithms are then threaded through this grid, and there are now many programs and packages, such as Surfer (Golden Software, 2002) and Gsharp, and geographical information systems, such as Arc/Info, that will do this with excellent graphics. Computing the isarithms involves another interpolation which is rarely optimal in the kriging sense, but if the kriged grid is fine enough this lack of optimality is immaterial. In most instances kriging at intervals of 2 mm on the finished map will be adequate.

The kriging variances and their square roots, the kriging errors, can be mapped similarly, and these maps give an idea of the reliability of the maps of estimates.

Creating a grid of kriged values to make a map can involve heavy computation. In principle all the estimates and their variances could be found from a single inversion of matrix  $\mathbf{A}$  in equation (8.13) that contains all of the semivariances between the sampling sites. As above, however, this is unwise or even impossible when the matrices are large. In practice, therefore, one enters into  $\mathbf{A}$  only the semivariances for some  $n$  data points, i.e. within the neighbourhood, near each grid node. This keeps the matrix small, but increases the number of inversions needed. Inversion can be accelerated if you work with the covariances instead of the semivariances because in the usual method of matrix inversion the largest element in each row, which serves as a pivot, is always in the diagonal of the covariance matrix and need not be sought.

For variables that are second-order stationary all the formulae for finding the weights from the variogram also apply to the covariance function with only changes of sign. For variables that are intrinsic only, the technique can still be used if you take some arbitrary large value for the covariance at  $\mathbf{h} = \mathbf{0}$ .

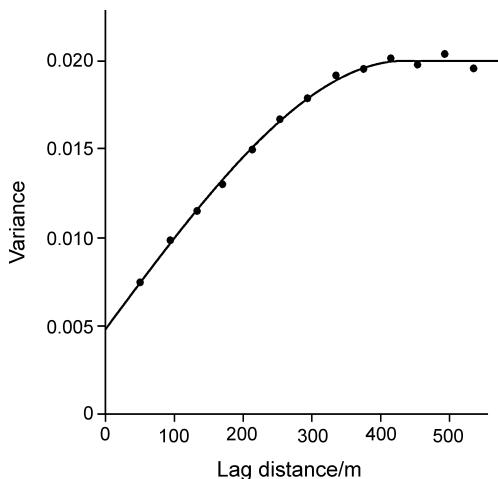
Other economies can be made depending on the location of the sampling points. If they are irregularly scattered then the same few data will often be used to estimate  $Z(\mathbf{x})$  at several grid nodes within a small area. Furthermore, the finer the interpolation grid the more nodes can be interpolated from the same observations. Matrix  $\mathbf{A}$  remains the same and needs inverting only once. Much

larger economies are possible where the data are on a regular grid because the same configuration recurs many times. Not only does the variogram remain constant, but so also does matrix  $\mathbf{A}$  for any given configuration. Each configuration requires only one matrix inversion. If sampling has been done on a square grid and the interpolation grid fits on to it with interval  $1/r$  times that of the sampling grid then there are only  $r^2$  possible configurations except near to the edge of the map. Where variation is isotropic the spatial relations have a fourfold symmetry, so even fewer solutions are needed.

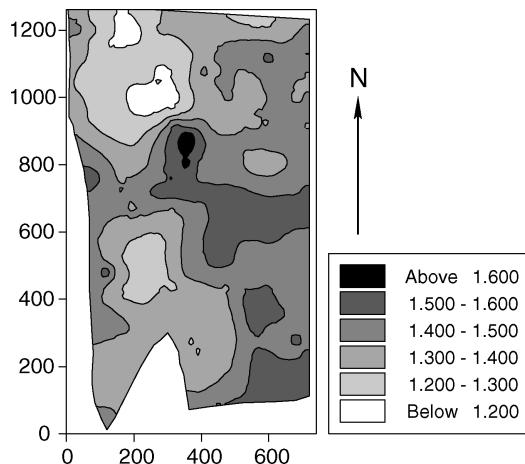
## 8.7 CASE STUDY

To illustrate the application of kriging to mapping we return to the analysis of exchangeable potassium (K) from Broom's Barn Farm. Since the distribution of K is skewed (skewness 2.04, Table 2.1) we transformed the values to common logarithms ( $\log_{10}K$ ) which reduced the skewness to 0.39.

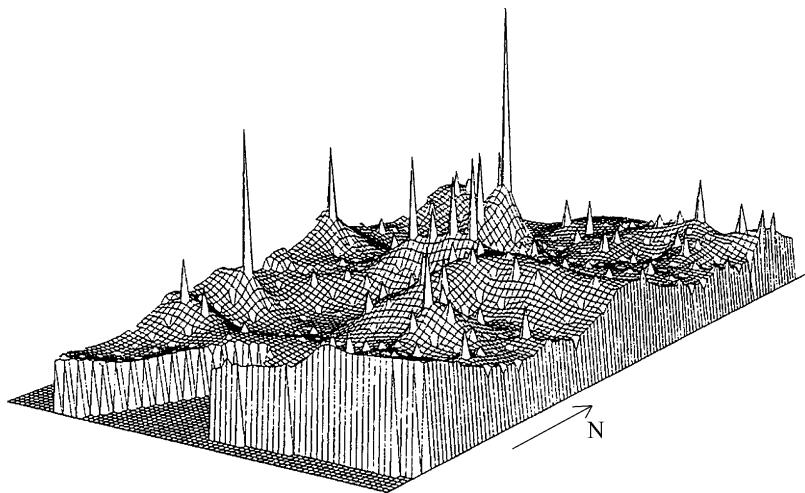
The variogram was computed on the transformed data, and the experimental semivariances were fitted best by a spherical function, equation (8.19), by weighted least-squares approximation as described in Chapter 5. The resulting coefficients are  $c_0 = 0.0048$ ,  $c = 0.01519$  and  $a = 439.2$  m. Figure 8.13 shows the experimental variogram (symbols) and the fitted spherical model (solid line). This function was then used for the kriging. We set the maximum radius of the neighbourhood to 400 m, and we set the minimum number of



**Figure 8.13** Variogram of exchangeable potassium at Broom's Barn Farm transformed to common logarithms. The points are the experimental semivariances, and the solid line is the best-fitting spherical model, the parameters of which are given in the text.



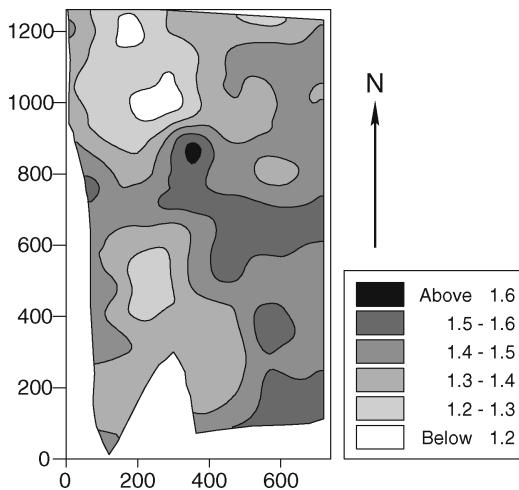
**Figure 8.14** Map of exchangeable potassium, transformed to common logarithms, at Broom's Barn Farm made by punctual kriging on a  $10\text{ m} \times 10\text{ m}$  grid that coincided with the sampling grid. The units are  $\log_{10}(\text{mg K l}^{-1})$ .



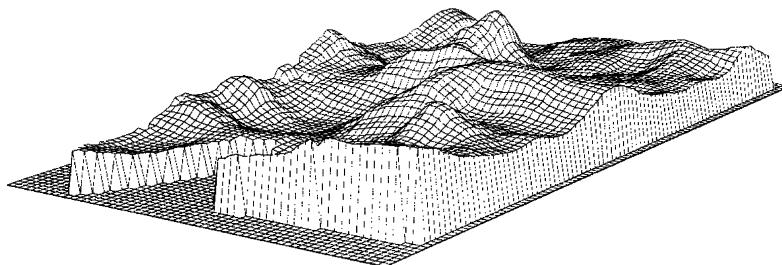
**Figure 8.15** Perspective diagram of exchangeable potassium transformed to common logarithms at Broom's Barn Farm made by punctual kriging on a  $10\text{ m} \times 10\text{ m}$  grid that coincided with the sampling grid.

points to seven and the maximum to 20. We kriged at intervals of 10 m, and for the block kriging our blocks were  $50\text{ m} \times 50\text{ m}$ . The estimated values and kriging variances have been mapped with Gsharp.

Figure 8.14 is a map of the punctual estimates. For it we deliberately placed the kriging grid over the sampling grid so that the sampling points lay on it to



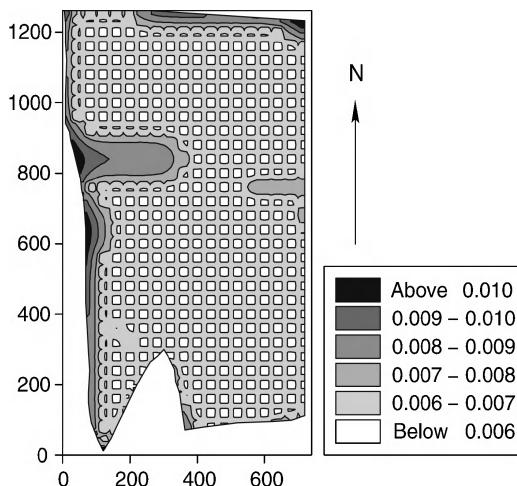
**Figure 8.16** Map of  $\log_{10}(\text{mg K l}^{-1})$  at Broom's Barn Farm made by kriging 50 m × 50 m blocks on a 10 m × 10 m grid.



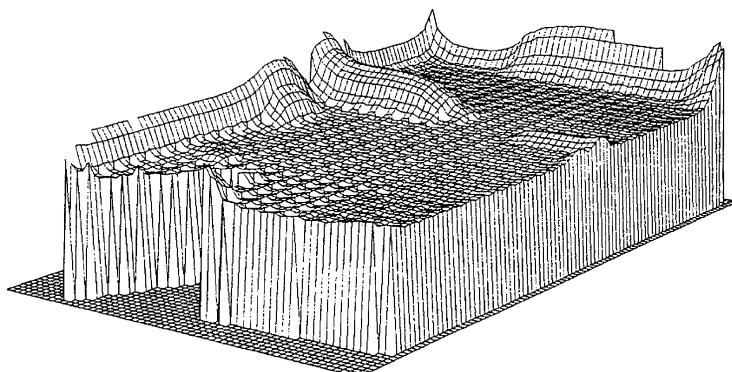
**Figure 8.17** Perspective diagram of exchangeable potassium at Broom's Barn made by kriging 50 m × 50 m blocks on a 10 m × 10 m grid.

illustrate the nugget effect. The map is somewhat ‘spotty’ because we have kriged at the data points. The spatial pattern of  $\log_{10}K$  is distinctly patchy, as we should expect from the spherical variogram; there are patches of large values and patches of small ones. The average extent of the patches is about 400 m.

In the alternative representation as a perspective diagram (Figure 8.15), the spots now appear as spectacular spikes, both above and below the surface. The reason is that at the sampling points punctual kriging returns the measured values there, whereas elsewhere it forms weighted averages of the data. The nugget variance in the variogram represents a discontinuity (Chapter 5), and this continues through to the kriging. Another way of viewing the effect is to consider the estimate as comprising two parts: the nugget variance and the continuous autocorrelated variation. Combining these two components produces the effect. The larger is the nugget variance as a proportion of the total



**Figure 8.18** Map of the estimation variances of  $\log_{10}(\text{mg K l}^{-1})$  at Broom's Barn Farm for punctual kriging.

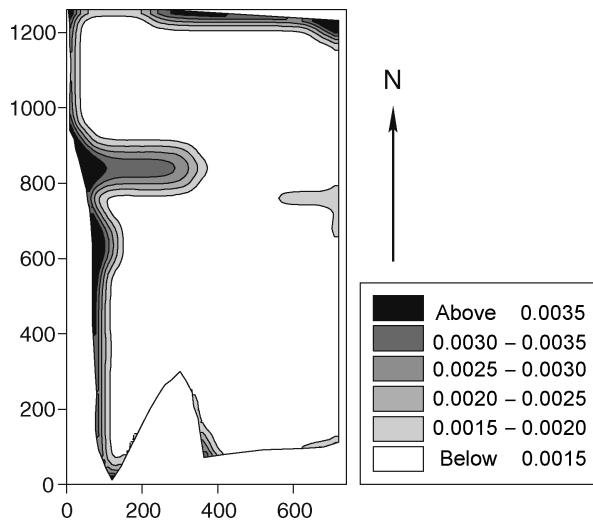


**Figure 8.19** Perspective diagram of the estimation variances of  $\log_{10}(\text{mg K l}^{-1})$  at Broom's Barn Farm for punctual kriging.

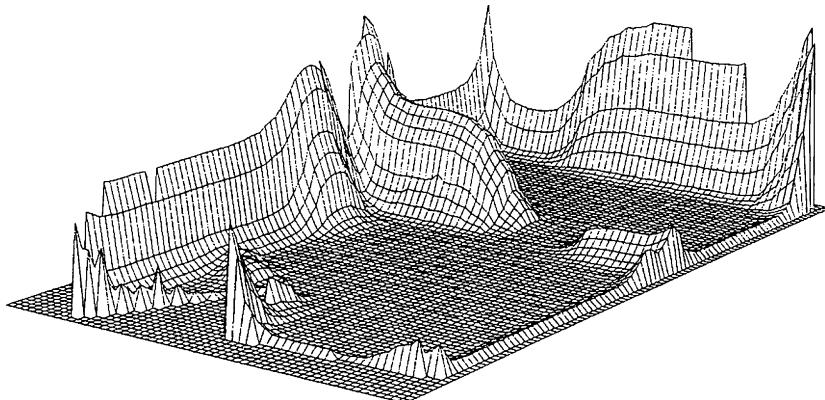
variance the more pronounced this effect becomes; when all of the variance is nugget the surface becomes flat between the sampling points.

Figure 8.16 is a map of the block estimates which has lost the 'spotty' appearance of Figure 8.14. Nevertheless, the same broad pattern in the distribution of  $\log_{10}K$  is evident. The block-kriged surface is smoother, and this is evident in the perspective diagram of this surface, shown in Figure 8.17.

The punctual kriging variances are shown in Figure 8.18. In general they are much larger than those for the block kriging (see Figure 8.20), and the technique therefore appears much less precise. At the data points, however,



**Figure 8.20** Map of the estimation variances of  $\log_{10}(\text{mg K l}^{-1})$  at Broom's Barn Farm



**Figure 8.21** Perspective diagram of the estimation variances of exchangeable potassium at Broom's Barn Farm for block kriging over 50 m  $\times$  50 m blocks on a 10 m  $\times$  10 m grid.

they are zero. The perspective diagram (Figure 8.19), shows both features. Between the sampling points the nugget variance sets a minimum to the kriging variance, and at the sampling points the surface descends to zero.

The block-kriging variances (Figures 8.20 and 8.21) are in general small, but they increase rapidly near the boundaries of the farm, beyond which there are no data, and similarly around the farm buildings (left above centre). The small ridge on the right of Figure 8.21 is an access road, again without any data along it, and the small hump on the lower left-hand side is where two data were lost.

### 8.7.1 Kriging with known measurement error

Throughout the above description of kriging and in the examples, we have proceeded as if there were no errors in the measurements. We have treated the nugget variance as if it were purely short-range spatial variation. Yet in Chapter 5 we recognized that the nugget variance was likely to include measurement error in addition to short-range variation. Like many practitioners, we tend to ignore the former because it is usually much smaller than the spatial component of the nugget, and often we do not know it. We should recognize, however, that practitioners would like to estimate the true values at unsampled places, not the values there plus measurement error. To do this, we proceed as follows.

First, we distinguish the two sources of variance as

$$c_0 = c_s + c_m, \quad (8.22)$$

in which  $c_s$  is the limit of the spatial component of  $\gamma(\mathbf{h})$  as  $\mathbf{h}$  approaches  $\mathbf{0}$ , and  $c_m$  is the variance of the measurement error. We can then use this decomposition in kriging, as follows. In the punctual kriging system (equation (8.9)), we inserted 0 in the right-hand side where a target point,  $\mathbf{x}_0$ , coincides with a data point,  $\mathbf{x}_j$ , on the assumption that there is no difference between the true value and the observed one. If, however, we know  $c_m$  then we insert that value instead. The rest of the kriging system and the kriging systems for other points remain as we give them in equation (8.9). Incorporating the measurement error affects only estimates at data points, which are no longer the same as the observed values. In these circumstances punctual kriging is no longer an exact interpolator. Finally, all the kriging variances are diminished by  $c_m$ :

$$\sigma_m^2(\mathbf{x}_0) = \sum_{i=1}^N \lambda_i \gamma(\mathbf{x}_i, \mathbf{x}_0) + \psi(\mathbf{x}_0) - c_m. \quad (8.23)$$

### 8.7.2 Summary

In practice exact interpolation might not be as attractive as one imagines, because of the nugget effect. Nevertheless, we can avoid this effect of the nugget variance either by offsetting the kriging grid so that estimates are not made at any data points or by omitting any data point when it coincides with a target point.

We can use the maps or diagrams of the estimation variance as a guide to the reliability of our estimates, but with caution. The reliability of kriging depends on how accurately the variation is represented by the chosen spatial model. If the nugget variance is overestimated then so will be the punctual kriging

variances, and our estimates will be more reliable than they appear to be. With block kriging the reverse can be the case, and we might imagine our estimates to be more reliable than they are. The block estimation variance comprises three terms, one of which is the within-block variance. The latter is estimated by integrating the variogram from  $|\mathbf{h}| = 0$  to the limit of the block; see Figure 8.1(b). If the semivariance is overestimated at short lags then the within-block variance will also be overestimated, at least for small blocks the sides of which are less than the shortest sampling interval of the variogram. The estimates might therefore be less reliable than they appear. For larger blocks estimates should be reliable because the contribution to the within-block variance from the short lags will be a small proportion of the whole.

## 8.8 REGIONAL ESTIMATION

In the limit we can think of the whole region,  $R$ , of interest as a single large block for which we could estimate the mean of  $Z$ ,  $\hat{Z}(R)$ , by including all the data. In classical estimation this is precisely what we do, giving all data the same weight; see equation (2.34). The solution takes no account of known spatial correlation, and kriging should do better by assigning differential weights.

We assume first that  $Z(\mathbf{x})$  is second-order stationary with mean  $\mu$  and variance  $\sigma^2$ . As  $R$  increases so the average distance between pairs of points in it increases, and the average semivariances,  $\bar{\gamma}(\mathbf{x}_i, B)$ , in equation (8.11) approach  $\sigma^2$ , the sill of the variogram. If the distance across  $R$  is much larger than the effective range of the variogram then the  $\bar{\gamma}(\mathbf{x}_i, B)$  will be so close to  $\sigma^2$  that the two can be taken as equal. The kriging system (8.11) can therefore be rewritten as

$$\begin{aligned} \sum_{i=1}^N \lambda_i \gamma(\mathbf{x}_i, \mathbf{x}_j) + \psi(R) &= \hat{\sigma}^2 \quad \text{for all } j, \\ \sum_{i=1}^N \lambda_i &= 1. \end{aligned} \tag{8.24}$$

The kriging weights are found in the usual way, and the kriging variance, from equation (8.16), is

$$\begin{aligned} \sigma^2(R) &= \mathbf{b}^T \boldsymbol{\lambda} - \sigma^2 \\ &= \sigma^2 \sum_{i=1}^N \lambda_i + \psi(R) - \bar{\gamma}(R, R). \end{aligned} \tag{8.25}$$

The sum  $\sum_{i=1}^N \lambda_i = 1$ , and so we have that

$$\sigma^2(R) = \psi(R). \tag{8.26}$$

Since  $Z(\mathbf{x})$  must be second-order stationary, the covariances exist, and the kriging system is usually expressed in terms of covariances:

$$\begin{aligned} \sum_{i=1}^N \lambda_i C(\mathbf{x}_i, \mathbf{x}_j) - \psi(R) &= 0 \quad \text{for all } j, \\ \sum_{i=1}^N \lambda_i &= 1, \end{aligned} \tag{8.27}$$

from which it follows immediately that  $\sigma^2(R) = \mathbf{b}^T \boldsymbol{\lambda} + \psi(R)$ .

Kriging the mean is undoubtedly attractive from a theoretical point of view. Unfortunately there are reasons why the approach cannot or should not be pursued.

1. It is unwise to assume that a property, which is locally stationary in the mean and semivariances, maintains that stationarity throughout a large region.
2. The experimental variogram is usually known accurately only for the first few lags; it almost certainly will not be well estimated for lags approaching the distance across a large region.
3. A large sample could produce kriging matrices that are too large to invert or that become unstable.

A practical alternative that avoids the difficulties is to divide the region into small rectangular blocks or strata, estimate the mean in each by kriging, and then compute the average of the estimates. If for some reason the blocks are not all of the same size then their estimates can be weighted according to their areas. For a region,  $R$ , divided into  $n$  blocks,  $B_i, i = 1, 2, \dots, n$ , of area  $H_i$ , the global mean,  $Z(R)$ , is estimated by

$$\hat{Z}(R) = \left( \sum_{i=1}^n H_i \hat{Z}(B_i) \right) / \left( \sum_{i=1}^n H_i \right) \tag{8.28}$$

where  $\hat{Z}(B_i)$  is the kriged estimate of  $Z$  within the  $i$ th block. If the blocks are of equal size then the  $H_i$  cancel, and  $\hat{Z}(R) = \sum_{i=1}^n \hat{Z}(B_i)/n$ .

A problem arises in calculating the estimation variance. The error in the global average equals the sum of the errors in the local estimates:

$$\hat{Z}(R) - Z(R) = \left( \sum_{i=1}^n H_i \{ \hat{Z}(B_i) - Z(B_i) \} \right) / \left( \sum_{i=1}^n H_i \right). \tag{8.29}$$

The estimation variance,  $\sigma^2(R) = E[\{\hat{Z}(R) - Z(R)\}^2]$ , cannot be estimated without bias by a simple sum, however, because the estimates in the neigh-

bouring blocks are not independent; some of the data from which they are computed are common. We can solve the problem by considering the error that results from using the value at a sampling point to estimate the average value over the portion of the region that is nearer to it than to any other, i.e. for its Thiessen polygon or Dirichlet tile. For a rectangular grid each polygon is a rectangle with an observation at its centre,  $\mathbf{x}_c$ , and sides equal to the sampling intervals along the principal axes of the grid. The variance of the estimate of its average is

$$\sigma^2(B) = 2\bar{\gamma}(\mathbf{x}_c, B) - \bar{\gamma}(B, B), \quad (8.30)$$

where  $\bar{\gamma}(\mathbf{x}_c, B)$  is the average semivariance between the centre and all other points in the rectangle, and  $\bar{\gamma}(B, B)$  is the variance within the polygon. Since the estimated values for these rectangles are  $\hat{Z}(B_i), i = 1, 2, \dots, n$ , the average for the region is approximately

$$\hat{Z}_B(R) = \frac{1}{n} \sum_{i=1}^n \hat{Z}(B_i). \quad (8.31)$$

The error of this estimate is approximately  $Z(R) - \hat{Z}_B(R)$ , and the corresponding variance of the regional mean is

$$\begin{aligned} E[\{Z(R) - \hat{Z}_B(R)\}^2] &\approx \frac{1}{n^2} \sum_{i=1}^n E[\{Z(B_i) - Z(\mathbf{x}_i)\}^2] \\ &= \frac{1}{n} \sigma^2(B). \end{aligned} \quad (8.32)$$

The approximation improves as  $n$  increases.

Thus the error of the regional estimate depends on the variances within small rectangular blocks, and these are likely to be much smaller than the variance within the entire region.

## 8.9 SIMPLE KRIGING

Sometimes we know or can assume the mean of a random variable from the nature of the problem. In these circumstances we should use that knowledge to improve our estimates, and we can do so by ‘simple kriging’. Our kriged estimate is still a linear sum, but now incorporating the mean,  $\mu$ , of the process, which must be second-order stationary. Prediction by simple kriging is

not an option for processes that are intrinsic only, a variogram with an upper bound is needed. For punctual kriging the equation is

$$\hat{Z}_{\text{SK}}(\mathbf{x}_0) = \sum_{i=1}^N \lambda_i z(\mathbf{x}_i) + \left\{ 1 - \sum_{i=1}^N \lambda_i \right\} \mu. \quad (8.33)$$

The  $\lambda_i$  are the weights, as before, but they are no longer constrained to sum to 1. The unbiasedness is assured by inclusion of the second term on the right-hand side of equation (8.33). Also, because the weights no longer sum to 1 we have to work with the covariances,  $C$ , instead of the semivariances,  $\gamma$ . We write the simple kriging system as

$$\sum_{N=1}^N \lambda_i C(\mathbf{x}_i, \mathbf{x}_j) = C(\mathbf{x}_0, \mathbf{x}_j) \quad \text{for } j = 1, 2, \dots, N. \quad (8.34)$$

There is no Lagrange multiplier: there are only  $N$  equations in  $N$  unknowns. The kriging variance is given by

$$\sigma_{\text{SK}}^2(\mathbf{x}_0) = C(\mathbf{0}) - \sum_{i=1}^N \lambda_i C(\mathbf{x}_i, \mathbf{x}_0), \quad (8.35)$$

where  $C(\mathbf{0})$  is the variance of the process.

As with ordinary kriging the technique can be generalized for blocks,  $B$ , larger than the supports of the sample by replacing the  $C(\mathbf{x}_0, \mathbf{x}_j)$  on the right-hand side of equation (8.34) by the averages  $\bar{C}(B, \mathbf{x}_j)$ . Also,  $N$ , the total size of the sample, can usually be replaced by  $n \ll N$  data in close proximity to  $\mathbf{x}_0$  or  $B$ .

In general, the variances obtained by simple kriging are somewhat smaller than those from ordinary kriging, and we might think that we could improve the predictions by introducing the mean estimated from the data,  $\hat{\mu}$ . Wackernagel (2003) shows that if we use the kriged mean, i.e. by putting  $\hat{\mu} = \hat{Z}(R)$ , we obtain the ordinary kriging predictor with variance

$$\sigma_{\text{OK}}^2(\mathbf{x}_0) = \sigma_{\text{SK}}^2(\mathbf{x}_0) + \left\{ 1 - \sum_{i=1}^N \lambda_i^{\text{SK}} \right\}^2 \psi(R). \quad (8.36)$$

In words, the ordinary kriging variance is the sum of the simple kriging variance plus the variance arising from the estimate of the mean. There is nothing to be gained by taking this approach because there is no more information. If the mean is estimated from many data, as will usually be the case, then  $\psi(R)$  will be small in relation to  $\sigma_{\text{SK}}^2(\mathbf{x}_0)$ , and provided the sum of the simple kriging weights is close to 1 the second term on the right-hand side of equation (8.36) is likely to be very small indeed.

## 8.10 LOGNORMAL KRIGING

A more common situation in the environmental sciences, and in mining and petroleum engineering too, is that the data are markedly skewed and non-normal. As mentioned in Chapter 6, the variogram is sensitive to strong positive skewness because a few exceptionally large values contribute to so many squared differences. Such skewness can often be removed and the variances stabilized by taking logarithms. If by transforming to logarithms the distribution is made near-normal then it is said to be lognormal. This leads to lognormal kriging.

The data  $z(\mathbf{x}_1), z(\mathbf{x}_2), \dots$  are transformed to their corresponding natural logarithms, say  $y(\mathbf{x}_1), y(\mathbf{x}_2), \dots$ , which represent a sample from the random variable  $Y(\mathbf{x}) = \ln Z(\mathbf{x})$ , which is assumed to be second-order stationary. The variogram of  $Y(\mathbf{x})$  is computed and modelled and then used with the transformed data to estimate  $Y$  at the target points or blocks by either ordinary or simple kriging. The estimated values are in logarithms.

For some purposes, as for example at Broom's Barn Farm where an index of soil fertility is wanted, the logarithms can serve well. However, in many other disciplines, such as mining, exploration geochemistry, and pollution monitoring, surveyors want estimates expressed in the original units, and the logarithms must be transformed back to concentration.

The back-transformation of a punctual estimate is fairly straightforward. If we denote the kriged estimate of the natural logarithm at  $\mathbf{x}_0$  as  $\hat{Y}(\mathbf{x}_0)$  and its variance as  $\sigma^2(\mathbf{x}_0)$  then the formulae for the back-transformation of the estimates are, for simple kriging,

$$\hat{Z}_{\text{SK}}(\mathbf{x}_0) = \exp\{\hat{Y}_{\text{SK}}(\mathbf{x}_0) + \sigma_{\text{SK}}^2(\mathbf{x}_0)/2\}, \quad (8.37)$$

and for ordinary kriging,

$$\hat{Z}_{\text{OK}}(\mathbf{x}_0) = \exp\{\hat{Y}_{\text{OK}}(\mathbf{x}_0) + \sigma_{\text{OK}}^2(\mathbf{x}_0)/2 - \psi(\mathbf{x}_0)\}, \quad (8.38)$$

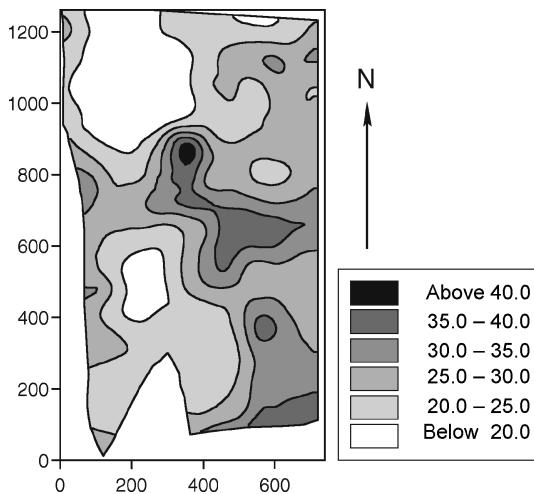
where  $\psi$  is the Lagrange multiplier in ordinary kriging. The estimation variance of  $Z(\mathbf{x}_0)$  for simple kriging is

$$\text{var}_{\text{SK}}[\hat{Z}(\mathbf{x}_0)] = \mu^2 \exp(\sigma_{\text{SK}}^2)[1 - \exp\{-\sigma_{\text{SK}}^2(\mathbf{x}_0)/2\}], \quad (8.39)$$

where  $\mu$  is the mean of  $Z(\mathbf{x})$ . We cannot obtain an unbiased back-transform of the ordinary kriging variance because the mean,  $\mu$ , is not known.

In many fields of application people prefer to work with common logarithms. The variogram of  $\log_{10}Z(\mathbf{x})$  replaces that of  $\ln Z(\mathbf{x})$ , and the back-transform for ordinary kriging is given by

$$\hat{Z}(\mathbf{x}_0) = \exp\left\{\hat{Y}(\mathbf{x}_0) \times \ln 10 + 0.5\sigma_Y^2(\mathbf{x}_0) \times (\ln 10)^2 - \psi(\mathbf{x}_0) \times (\ln 10)^2\right\}. \quad (8.40)$$



**Figure 8.22** Map of block-kriged estimates of potassium at Broom's Barn Farm after back-transformation.

Journel and Huijbregts (1978) point out that the expression in equation (8.37) for the back-transformation is sensitive to departures from lognormality and that in consequence the estimates of  $Z$  can be biased. They suggest a check for bias by comparing the mean of the estimates,  $\hat{Z}$ , with the mean of the data,  $z(\mathbf{x}_i)$ ,  $i = 1, 2, \dots, N$ . If we denote the ratio of the means,  $\text{mean}[\hat{Z}] : \bar{z}$ , by  $Q$  then we modify equation (8.37) to

$$\hat{Z}_{\text{SK}}(\mathbf{x}_0) = Q \exp\{\hat{Y}_{\text{SK}}(\mathbf{x}_0) + \sigma_{\text{SK}}^2(\mathbf{x}_0)/2\}, \quad (8.41)$$

or equation (8.38) in like manner if we have used ordinary kriging. In our experience  $Q$  has always been so close to 1 that we have not needed the elaboration. Figure 8.22 shows the back-transformed values of the block-kriged estimates of  $\log_{10}K$ .

You can find an up-to-date review of the problems associated with back-transformation and solutions for several situations in Cressie (2006).

## 8.11 OPTIMAL SAMPLING FOR MAPPING

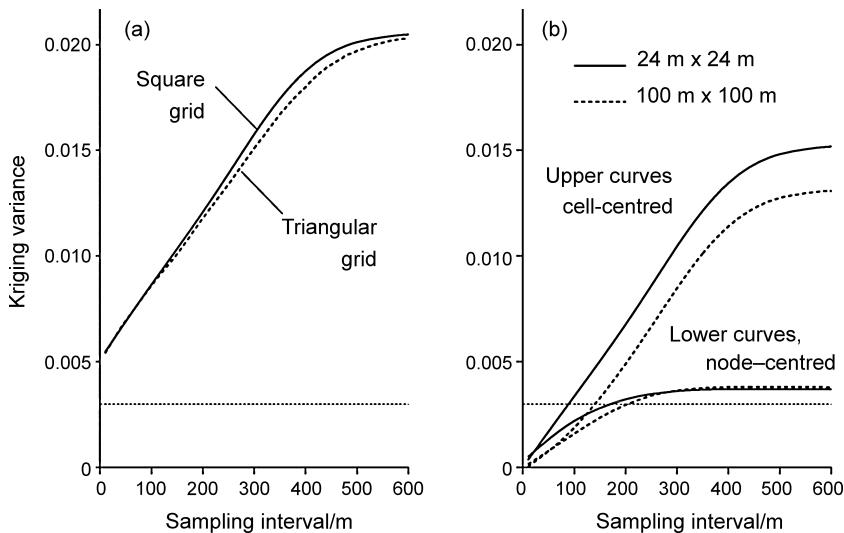
From equations (8.2) and (8.4) it is evident that the kriging weights depend on the configuration of the sampling points in relation to the target point or block and on the variogram. They do not depend at all on the observed values at those points. The same applies to the kriging variances, see equation (8.2).

Therefore, if we know the variogram then we can determine the kriging errors for any sampling configuration *before* doing the sampling, and we can design a sampling scheme to meet a specified tolerance or precision.

In general, mapping is most efficient if survey is done on a regular grid in the sense that the maximum kriging error is minimized. Where there is spatial dependence the information from an observation pertains to an area surrounding it, and specifically to the neighbourhood within its range if the variable is second-order stationary. If the neighbourhoods of two observations overlap then information is duplicated to some extent. Any kind of clustering of points, such as arises with random sampling, means that information can be replicated while elsewhere there is underrepresentation or even big gaps. We can minimize redundancy by placing the sampling points as far away from their neighbours as possible for a given sampling density. This approach also minimizes the area that is underrepresented. Triangular configurations are the most efficient in this respect. For a grid with one node per unit area neighbouring sampling points are 1.0746 units of distance apart, and no point is more than 0.6204 units away from another. We denote this maximum distance  $d_{\max}$ . Rectangular grids have some neighbours that are closer and others that are further away. For a square grid with one node per unit area the sampling interval is 1, and  $d_{\max} = 1/\sqrt{2} = 0.7071$ . For a hexagonal grid with unit sampling density  $d_{\max} = 0.8772$ . From this we should expect triangular sampling configurations to be the most efficient. Matérn (1960) and Dalenius *et al.* (1961) showed that where the variogram is exponential the triangular grid is optimal for estimating the mean of a region, and in most circumstances with bounded variograms that have finite ranges. The same is also true if the variogram is unbounded. In certain restricted circumstances with variograms with a finite range, a hexagonal grid can be the most efficient (Yfantis *et al.*, 1987). In general, however, rectangular grids are preferred because they are easier to work with in the field. Figure 8.23(a) shows that the difference in precision between a triangular configuration and a square one is small, and that we can choose the type of grid that we prefer to work with.

The variogram then enables us to optimize the sampling interval to estimate both the regional mean and local values for mapping. For estimation by kriging, or indeed any other method of interpolation, the distances between neighbouring sampling points should be well within the correlation range. As we have seen above, if they are beyond the range then kriging simply returns the mean of the points in the neighbourhood.

The kriging errors are not the same everywhere. With punctual kriging there is no error at the sampling points, see Figure 8.11(c), and, in general, the further a target point is from the data the larger the error. If we sample on a regular grid we minimize  $d_{\max}$ , which is the distance between a target point at the centre of a grid cell and its nearest sampling point on the grid node. We also minimize the maximum kriging error, except near the margins of the map.



**Figure 8.23** Graph of kriging variance against sampling interval to map exchangeable potassium at Broom's Barn Farm: (a) for punctual kriging on a square grid and a triangular grid; (b) for block kriging with  $24\text{ m} \times 24\text{ m}$  and  $100\text{ m} \times 100\text{ m}$  blocks (lower lines are the variances centred on grid nodes and the upper ones centred in grid cells).

### 8.11.1 Isotropic variation

Burgess *et al.* (1981) and McBratney *et al.* (1981) showed how the kriging equations can be solved to design an optimal sampling scheme. For punctual kriging we solve equations (8.9), and determine the kriging variances and errors by equation (8.10) at the centres of grid cells for a range of sampling intervals. The variances are then plotted against the grid spacing. If we have in mind a maximum variance or error that we can tolerate then we can draw a horizontal line across the graph until it meets the maximum kriging variance. A perpendicular from this point gives the optimal sample spacing.

To illustrate the procedure we use the variogram  $\log_{10}K$  for Broom's Barn Farm (Figure 8.13 and Table 5.1). Figure 8.23(a) shows the maximum punctual kriging variance for square and triangular grids. Note that the difference between the curves for the square and triangular grids is not nearly as large as the 12% difference in  $d_{\max}$  for the two grids. The line drawn across the graph at 0.003 is the kriging variance on the logarithmic scale that is approximately equivalent to a 90% confidence interval of  $10\text{ mg l}^{-1}$  at the deficiency threshold of  $25\text{ mg l}^{-1}$ . The kriging variances are large, and all exceed this tolerance.

Figure 8.23(a) illustrates two other features of punctual kriging. If we set the maximum tolerable kriging variance at 0.003 then it is impossible to design a satisfactory sampling scheme because we cannot diminish  $\sigma_{\max}^2$  to less than the nugget variance, 0.00478. Second,  $\sigma_{\max}^2$  increases to a maximum at which it flattens. This maximum is somewhat larger than the sill of the variogram; in fact it is the sill plus the Lagrange multiplier,  $\psi$ , of equation (8.9). Once  $d_{\max}$  exceeds the range of the variogram, 439 m in this case, all the semivariances in the kriging system are equal, as are the weights, as we saw in the example above. The additional quantity  $\psi$  represents the additional uncertainty of predicting the value at a place from only local data.

The same reasoning and procedure apply to block kriging, equation (8.11). However, it is less straightforward, and the result depends on the block size. For blocks of side much smaller than the sampling interval the kriging variance will be largest when the blocks are in the centres of grid cells. As the block is increased in size the kriging variance decreases—contrast the 24 m  $\times$  24 m blocks with the 100 m  $\times$  100 m blocks in Figure 8.23(b). Consider, however, a block centred on a grid node. If the block is no larger than the sample support this is effectively punctual kriging and the kriging variance is zero. As the block size increases, its kriging variance initially increases because the dominant effect of the observation at its centre declines. Only when it is big enough for the nearest neighbours to be more influential does the kriging variance start to decline. This difference in the configuration has another important effect. As the block increases in size the weights of the sampling points nearest its centre decrease, whereas the weights of those further away increase (see Figure 8.9). A block size is eventually reached at which its estimation variance equals that for a block centred in a grid cell. If the block size becomes larger still the kriging variance can be greater than that of a block of the same size centred in a grid cell. Therefore, for block kriging one must decide where to determine the kriging variances, i.e. whether for blocks centred on grid cells or ones centred on grid nodes. The position at which the kriging variance is greatest for a given block size is the one to choose. Burgess *et al.* (1981) describe these effects in detail.

In Figure 8.23(b) the kriging variances for blocks centred at the cell centres and grid nodes are plotted against distance for a square grid for blocks of side 24 m and 100 m. At the chosen tolerance the horizontal line intersects the graph of the variances for blocks about 80 m apart for blocks of side 24 m and about 130 m apart for 100 m blocks.

For block kriging of potassium at Broom's Barn Farm the results suggest that sampling might have been denser than necessary for mapping.

Using the variogram and the kriging equations one can design a new survey to be optimal in the sense that sampling is just sufficiently intense to meet the specified tolerance. Near the margins of the region some modifications might be needed if sampling cannot be extended outside it because the variance increases at the margin (see Figures 8.19 and 8.21); sampling would need to be increased near the margin to keep within the tolerance.

We can also use this approach if we feel that part of a region is undersampled. We can see whether adding further points will increase the precision before sampling more. Also, if we have a network of stations for monitoring rainfall or pollutants in ground water the effect of adding stations, moving them or removing them can be assessed. This is what McCullagh (1976) did with the Trent telemetry network. Barnes (1989) used different strategies to optimize the placement of a new sampling station—depending on whether the need was to improve the worst situation or to diminish the estimation variance on average.

This approach allows sampling to be optimized in the sense of minimizing effort.

### 8.11.2 Anisotropic variation

One can take anisotropy into account when planning sampling. The grid spacing is adjusted so that the sampling is more intense in the direction of minimum continuity, i.e. the direction with the maximum rate of spatial change, than in other directions. The problem is to keep within the specified tolerable error for least effort. The optimum solution depends on the form of the anisotropy. The one that we illustrate is for strict geometric anisotropy (Burgess *et al.*, 1981).

Consider the linear variogram

$$\gamma(h, \vartheta) = \Omega(\vartheta)|\mathbf{h}|, \quad (8.42)$$

in which  $\Omega(\vartheta)$  is the sinusoidal function

$$\Omega(\vartheta) = \sqrt{A^2 \cos^2(\vartheta - \varphi) + B^2 \sin^2(\vartheta - \varphi)}. \quad (8.43)$$

In this equation  $\varphi$  is the direction of maximum variation,  $A$  is the gradient of the variogram in that direction, and  $B$  is the gradient in the perpendicular direction,  $\varphi + \pi/2$ . When  $\vartheta = \varphi$ , equation (8.42) reduces to

$$\gamma_1(h) = Ah, \quad (8.44)$$

and when  $\vartheta = \varphi + \pi/2$  it becomes

$$\gamma_2(h) = Bh. \quad (8.45)$$

As above, we can define an anisotropy ratio  $R$ :

$$R = A/B = \gamma_1(h)/\gamma_2(h). \quad (8.46)$$

The semivariance in direction  $\varphi$  at any lag  $h$  is thus equal to the semivariance at lag  $Rh$  in the direction  $\varphi + \pi/2$ :

$$\gamma_1(h) = \gamma_2(Rh). \quad (8.47)$$

Using equation (8.47) we find the most economical sampling scheme as follows. We treat the problem as though variation were isotropic with the variogram  $\gamma_1(h)$ . The sampling interval  $d$  is found in exactly the same way as for the square grid. This then becomes the sampling interval in direction  $\varphi$ . We take the anisotropy into account by making the sampling interval in the perpendicular direction,  $\varphi + \pi/2$ , equal to  $Rd$ .

## 8.12 CROSS-VALIDATION

In Chapter 5 we fitted models by minimizing the deviations between the observed semivariances and the ones expected from the model, and we chose finally from among different kinds of model those for which the squared deviations were least on average. We weighted the experimental values in proportion to the numbers of pairs contributing to them, but we paid no regard to the lag except incidentally when we refined the weighting as a function of the expected value. This is not necessarily the best for kriging because points near to the target point or block get more weight than more distant ones. So we should really like the variogram to be accurate at short lags, if necessary at the expense of less accuracy at longer lags. But how should we choose?

One way of choosing between competing models is to use them for kriging and see how well they perform. We can do this rigorously by having a separate set of sample data against which to compare kriged estimates. Except in research studies this would waste information, and validation usually is done by a process known as ‘cross-validation’. It works as follows.

1. An experimental variogram is computed from the whole set of sample data, and plausible models are fitted to it.
2. For each model,  $Z$  is estimated from the data and the model by kriging at each sampling point in turn after excluding the sample value there. The kriging variance is also calculated.
3. Three diagnostic statistics are calculated from the results:
  - (a) the mean deviation or mean error, ME, given by

$$ME = \frac{1}{N} \sum_{i=1}^N \{z(\mathbf{x}_i) - \hat{Z}(\mathbf{x}_i)\}; \quad (8.48)$$

- (b) the mean squared deviation or mean squared error, MSE:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N \{z(\mathbf{x}_i) - \hat{Z}(\mathbf{x}_i)\}^2; \quad (8.49)$$

- (c) and the mean squared deviation ratio, MSDR, computed from the squared errors and kriging variances,  $\hat{\sigma}^2(\mathbf{x})$ , by

$$\text{MSDR} = \frac{1}{N} \sum_{i=1}^N \frac{\{z(\mathbf{x}_i) - \hat{Z}(\mathbf{x}_i)\}^2}{\hat{\sigma}^2(\mathbf{x}_i)}. \quad (8.50)$$

The mean error should ideally be 0 because kriging is unbiased. The calculated ME, however, is a weak diagnostic because kriging is insensitive to inaccuracies in the variogram. We want the MSE to be small, of course. If the model for the variogram is accurate then the MSE should equal the kriging variance; and so the MSDR should be 1.

Let us see how the models for  $\log_{10}K$  at Broom's Barn Farm compare in this test. The three test criteria are listed in Table 8.4 for the five models summarized in Table 5.1, from which we have transferred the mean square residuals for comparison.

The first three models in the table, the circular, spherical and pentaspherical, have similar values for each of the three diagnostics. The MSDRs suggest that the kriging variances progressively underestimate the true estimation variances in that sequence, though not seriously. The MSE for the exponential model looks a little worrying, and we see that its mean squared residual is substantially larger than that of the first three models. The power function clearly performs poorly on the cross-validation with an MSDR of only 0.18. The kriging variance grossly exaggerates the true estimation

**Table 8.4** Mean error (ME), mean squared error (MSE), and mean squared deviation ratio (MSDR) for ordinary kriging of  $\log_{10}K$  with five models. The mean squared residuals are added for comparison.

Model	ME	MSE	MSDR	Mean squared residual
Circular	0.000321	0.007739	1.010	0.000172
Spherical	0.000327	0.007639	1.044	0.000155
Pentaspherical	0.000346	0.007584	1.081	0.000248
Exponential	0.000682	0.007314	1.232	0.001054
Power function	0.000726	0.007465	0.184	0.003295

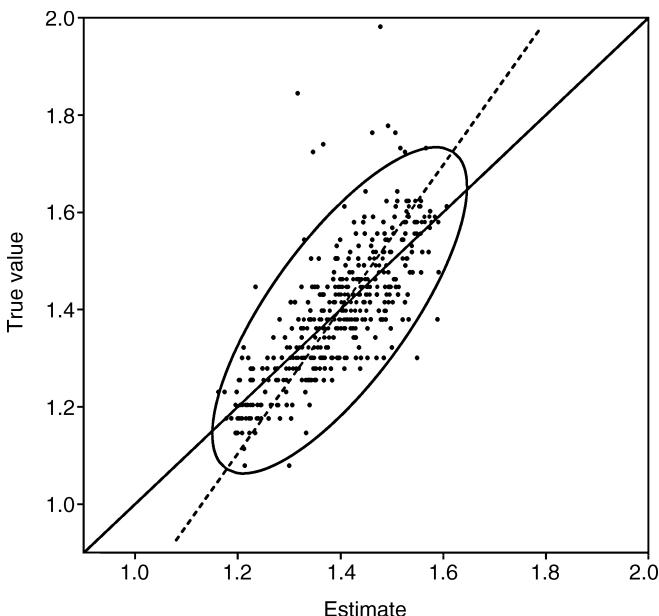
variance. The mean squared residual tells the same story; that of the power function is quite the largest. Figure 5.1 suggests that its MSDR is so small because the model values exceed the observed ones at the short lags between the data and the target points, which are the ones that dominate the kriging systems.

### 8.12.1 Scatter and regression

Another way of examining the behaviour of kriging is to plot the scatter of the true values against their estimates. We should like the two to be the same, but perfection of this kind is elusive in nature. The best we can expect is that our estimator is conditionally unbiased, by which we mean

$$E[Z(\mathbf{x}_0)|\hat{Z}(\mathbf{x}_0)] = \hat{Z}(\mathbf{x}_0). \quad (8.51)$$

From this it follows that the regression of  $Z(\mathbf{x}_0)$  on  $\hat{Z}(\mathbf{x}_0)$  is 1, therefore the covariance between the true values and their estimates must equal the variance of the estimates.



**Figure 8.24** Scatter diagram of the true  $\log_{10}K$  for Broom's Barn Farm plotted against the punctually kriged estimates. The ellipse is a probability contour, the dashed line is its longer diameter, and the solid diagonal line is the regression of  $z(\mathbf{x}_0)$  on  $\hat{Z}(\mathbf{x}_0)$ .

Armstrong (1998) shows that the above hold for simple kriging. For ordinary kriging, however, the variance of the estimates includes the Lagrange multiplier, and so the regression coefficient is somewhat less than 1.

Figure 8.24 illustrates the situation in which the true values,  $z(\mathbf{x}_0)$  are plotted against their estimates for  $\log_{10}K$  at Broom's Barn Farm. The scatter forms an elliptical cloud with a few points lying outside it. The ellipse itself is a probability 'contour' (see Chapter 2) drawn to include all but a few of the points. Its diameters are proportional to the standard deviations along the principal axes, the longer of which is drawn with a dashed line. They and the orientation have been estimated by a principal component analysis. The regression of  $z(\mathbf{x}_0)$  on  $\hat{Z}(\mathbf{x}_0)$  is the 1:1 line, the diagonal joining the corners of the frame and passing through the points where the vertical tangents touch the ellipse. The actual regression coefficients for simple and ordinary kriging estimated in this way are 1.035 and 1.024, respectively. They are barely distinguishable from 1. Like the mean error, this regression is a poor diagnostic because the kriged estimates are so insensitive to the model.

Figure 8.24 shows another feature of kriging. The long axis of the ellipse is oriented at about  $56^\circ$  from the horizontal; it is substantially more than  $45^\circ$ . The variance of the estimates, 0.009 79 on the abscissa, is less than that of the true values, 0.018 00 on the ordinate. In other words, kriging has lost variance; kriging smooths. It underestimates the larger values and overestimates the smaller ones, as in the simpler forms of regression.

# **Kriging in the Presence of Trend and Factorial Kriging**

## **9.1 NON-STATIONARITY IN THE MEAN**

The several kinds of kriging described in Chapter 8 are for realizations of stationary processes. They are based on the simple model given in equation (4.10) and repeated here:

$$Z(\mathbf{x}) = \mu + \varepsilon(\mathbf{x}), \quad (9.1)$$

in which  $\mu$  is the mean, which is constant, and  $\varepsilon(\mathbf{x})$  is a random variable with mean zero and variogram  $\gamma(\mathbf{h})$ . If the process is second-order stationary then  $\varepsilon(\mathbf{x})$  also has a covariance function  $C(\mathbf{h})$ , given in equation (4.11). We now turn our attention to spatial processes in which  $\mu$  varies.

As we mentioned in Chapter 4 some spatial processes include trend, or ‘drift’ as it is commonly known in geostatistics; they are not stationary in the mean. The variation in  $Z(\mathbf{x})$  then contains a systematic component in addition to the random one. Equation (4.21) expressed this by

$$Z(\mathbf{x}) = u(\mathbf{x}) + \varepsilon(\mathbf{x}), \quad (9.2)$$

where  $u(\mathbf{x})$ , which varies smoothly and is deterministic, replaces the mean,  $\mu$ , in equation (9.1). In these circumstances  $E[\{Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})\}^2]$  does not equal  $E[\{\varepsilon(\mathbf{x}) - \varepsilon(\mathbf{x} + \mathbf{h})\}^2]$ , and the raw semivariances computed by equation (4.40) will be biased estimates of  $\gamma(\mathbf{h})$ , the variogram of the residuals from the trend, i.e. of

$$\varepsilon(\mathbf{x}) = Z(\mathbf{x}) - u(\mathbf{x}). \quad (9.3)$$

To estimate  $\gamma(\mathbf{h})$ , or equivalently  $C(\mathbf{h})$ , without bias we must separate  $u(\mathbf{x})$  from  $\varepsilon(\mathbf{x})$ . We know neither; all we have are data,  $z(\mathbf{x}_i)$ ,  $i = 1, 2, \dots, N$ .

The trend,  $u(\mathbf{x})$ , can usually be expressed as a simple functional form

$$u(\mathbf{x}) = \sum_{k=0}^K \beta_k f_k(\mathbf{x}), \quad (9.4)$$

in which  $\beta_k, k = 0, 1, \dots, K$ , are unknown coefficients, and the  $f_k(\mathbf{x})$  are known functions of  $\mathbf{x}$  of our choosing. If we combine equations (9.3) and (9.4) then we can represent a process with trend by the model

$$Z(\mathbf{x}) = u(\mathbf{x}) + \varepsilon(\mathbf{x}) = \sum_{k=0}^K \beta_k f_k(\mathbf{x}) + \varepsilon(\mathbf{x}). \quad (9.5)$$

Typically a spatial trend can be modelled as a low-order polynomial in the geographical coordinates. So, in the simplest case of linear trend we can expand equation (9.5) to

$$Z(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon(\mathbf{x}), \quad (9.6)$$

in which  $x_1$  and  $x_2$  are the spatial coordinates, and for which  $K + 1 = 3$ .

If  $K = 0$  then  $f_0 = 1$ ,  $u(\mathbf{x}) = \beta_0 = \mu$ , and we have a stationary process as represented by equation (9.1) with the usual variogram, which is unbiased and which we can use for ordinary or simple kriging. If  $K > 0$  then we have a more complex problem to which we must find a solution. Nevertheless, ultimately our task is to estimate  $Z(\mathbf{x})$  at unsampled places as in ordinary or simple kriging.

### 9.1.1 Some background

The problem outlined above has been recognized for many years. Matheron (1969) solved the prediction part of the problem with his universal kriging. A punctual estimate of  $Z$  at  $\mathbf{x}_0$  from  $N$  data is still a linear sum:

$$\hat{Z}(\mathbf{x}_0) = \sum_{i=1}^N \lambda_i f_k(\mathbf{x}_i). \quad (9.7)$$

Its expectation is

$$E[\hat{Z}(\mathbf{x}_0)] = \sum_{k=0}^K \sum_{i=1}^N \beta_k \lambda_i f_k(\mathbf{x}_i), \quad (9.8)$$

and the estimator is unbiased if

$$\sum_{i=1}^N \lambda_i f_k(\mathbf{x}_i) = f_k(\mathbf{x}_0) \quad \text{for all } k = 0, 1, \dots, K. \quad (9.9)$$

Matheron elaborated the ordinary kriging system to take into account the fixed effects of the trend in addition to the random component as

$$\begin{aligned} \sum_{i=1}^N \lambda_i \gamma(\mathbf{x}_i, \mathbf{x}_j) + \psi_0 + \sum_{k=0}^K \psi_k f_k(\mathbf{x}_j) &= \gamma(\mathbf{x}_0, \mathbf{x}_j) \quad \text{for all } j = 1, 2, \dots, N, \\ \sum_{i=1}^N \lambda_i &= 1, \\ \sum_{i=1}^N \lambda_i f_k(\mathbf{x}_i) &= f_k(\mathbf{x}_0) \quad \text{for all } k = 0, 1, \dots, K. \end{aligned} \tag{9.10}$$

The values  $\gamma(\mathbf{x}_i, \mathbf{x}_j)$  are the semivariances of the residuals between the data points  $\mathbf{x}_i$  and  $\mathbf{x}_j$ , and the  $\gamma(\mathbf{x}_0, \mathbf{x}_j)$  are the semivariances between the target point and the data points. The functions  $f_k(\mathbf{x})$  refer to an origin  $\mathbf{x} = \mathbf{0}$  for the target point  $\mathbf{x}_0$ . For a linear drift there are three functions, i.e.  $K + 1 = 3$ , with values

$$f_0 = 1, \quad f_1 = x_1, \quad f_2 = x_2.$$

For quadratic drift there are three additional functions:

$$f_3 = x_1^2, \quad f_4 = x_1 x_2, \quad f_5 = x_2^2.$$

In addition there are now three Lagrange multipliers,  $\psi_0$ ,  $\psi_1$  and  $\psi_2$ , for the linear drift and three more,  $\psi_3$ ,  $\psi_4$  and  $\psi_5$ , for quadratic drift.

The universal kriging system, like that for ordinary kriging, is a set of linear equations which we can represent in matrix notation by

$$\mathbf{A}\boldsymbol{\lambda} = \mathbf{b}, \tag{9.11}$$

as in equation (8.13). Now, however, the matrix  $\mathbf{A}$  and the vectors  $\boldsymbol{\lambda}$  and  $\mathbf{b}$  are augmented with functions of the spatial positions of the data points and of the target:

$$\mathbf{A} = \begin{bmatrix} \gamma(\mathbf{x}_1, \mathbf{x}_1) & \gamma(\mathbf{x}_1, \mathbf{x}_2) & \cdots & \gamma(\mathbf{x}_1, \mathbf{x}_N) & 1 & f_1(\mathbf{x}_1) & f_2(\mathbf{x}_1) & \cdots & f_K(\mathbf{x}_1) \\ \gamma(\mathbf{x}_2, \mathbf{x}_1) & \gamma(\mathbf{x}_2, \mathbf{x}_2) & \cdots & \gamma(\mathbf{x}_2, \mathbf{x}_N) & 1 & f_1(\mathbf{x}_2) & f_2(\mathbf{x}_2) & \cdots & f_K(\mathbf{x}_2) \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ \gamma(\mathbf{x}_N, \mathbf{x}_1) & \gamma(\mathbf{x}_N, \mathbf{x}_2) & \cdots & \gamma(\mathbf{x}_N, \mathbf{x}_N) & 1 & f_1(\mathbf{x}_N) & f_2(\mathbf{x}_N) & \cdots & f_K(\mathbf{x}_N) \\ 1 & 1 & \cdots & 1 & 0 & 0 & 0 & \cdots & 0 \\ f_1(\mathbf{x})_1 & f_1(\mathbf{x}_2) & \cdots & f_1(\mathbf{x}_N) & 0 & 0 & 0 & \cdots & 0 \\ f_2(\mathbf{x})_1 & f_2(\mathbf{x}_2) & \cdots & f_2(\mathbf{x}_N) & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ f_K(\mathbf{x})_1 & f_K(\mathbf{x}_2) & \cdots & f_K(\mathbf{x}_N) & 0 & 0 & 0 & \cdots & 0 \end{bmatrix},$$

$$\boldsymbol{\lambda} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_N \\ \psi_0 \\ \psi_1 \\ \psi_2 \\ \vdots \\ \psi_K \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} \gamma(\mathbf{x}_1, \mathbf{x}_0) \\ \gamma(\mathbf{x}_2, \mathbf{x}_0) \\ \vdots \\ \gamma(\mathbf{x}_N, \mathbf{x}_0) \\ 1 \\ f_1(\mathbf{x}_0) \\ f_2(\mathbf{x}_0) \\ \vdots \\ f_K(\mathbf{x}_0) \end{bmatrix}.$$

As in ordinary kriging,  $\mathbf{A}$  is inverted, and the weights and the Lagrange multipliers are obtained as

$$\boldsymbol{\lambda} = \mathbf{A}^{-1} \mathbf{b}. \quad (9.12)$$

The weights are inserted into equation (9.7), and the kriging variance is given by

$$\sigma_{UK}^2 = \mathbf{b}^T \boldsymbol{\lambda}. \quad (9.13)$$

Also as in ordinary kriging, we can usually work within a window with many fewer data than the whole set of size  $N$ .

Thus, universal kriging looks remarkably like ordinary kriging, and like ordinary kriging the procedure is automatic once you have a satisfactory function for the variogram. The difficult task is obtaining such a function. In fact, it is the biggest impediment to kriging in the presence of drift.

Olea (1975) spelled out in detail steps by which one could estimate the variogram of the random component from data by a structural analysis, but only where those data are at regular intervals on transects or grids. Matheron's (1973) intrinsic random functions, which we mentioned as a solution on page 59, see equation (4.24), are similarly constrained. Environmental scientists typically do not have such data; their data more often come from observations irregularly distributed over the land or in the sea. If they can recognize some simple long-range trend then one legitimate way forward has been to compute and model the variogram in the direction perpendicular to the trend, as advocated by Goovaerts (1997) and applied to soil, for example, by Meul and Van Meirvenne (2003). Their model will represent the random process free of trend on the assumption that the process is isotropic, and it can be used for kriging. A weakness of the method is that there might be few pairs of data in that direction, so that the variogram is estimated poorly, and the drift might be more complex.

Another way of dealing with drift has been to model it first, as in trend surface analysis (Section 3.1.5), and remove it from the data. The residuals are treated as realizations of stationary correlated random variables, the variogram is computed and modelled and then used to krig. Finally the trend is added back to the kriged estimates. The method is attractive, especially if the trend is interesting in its own right, as was the conformation of the Chalk (Cenomanian) strata beneath the Chiltern Hills in southern England investigated by Moffat *et al.* (1986). Since then it has become popular in earth sciences under the title 'regression kriging' (e.g. Knotters *et al.*, 1995; Odeh *et al.*, 1994, 1995). The estimates, both of the trend and of the random residuals are unbiased provided that the data are unbiased in the first place. The method is equivalent to universal kriging for a given variogram provided that all the data are used in the kriging system and not only those in a local window.

There are two disadvantages of regression kriging. First, the trend is generally estimated by ordinary least squares (OLS), which is unbiased, but does not yield estimates of minimum variance unless the sampling sites have been selected independently at random. Such selection is rare in resource surveys, and so other methods of analysis should be used.

The second disadvantage is that the estimates of the semivariances obtained from residuals from the trend are biased. This is because they depend in a non-linear way on the trend parameters, which are themselves estimated with error. As a result the variogram is underestimated, and the bias increases with increasing lag distance (Cressie, 1993). Lark *et al.* (2006) illustrate this effect well.

One proposed solution to these problems is to use generalized least squares to estimate the trend parameters. The generalized least-squares method itself requires a variogram for the residuals, so an iterative procedure is followed. The OLS estimates are obtained, and a variogram is fitted to the residuals. This variogram is then used in generalized least squares to re-estimate the trend parameters, and the procedure is repeated until the estimates stabilize (e.g. Hengl *et al.*, 2004). This approach reduces the error variance of the trend parameters, but it does not remove the bias from the estimates in the variogram because these still depend on the trend parameters (Gambolati and Galeati, 1987). This bias might not matter where data are dense because it is typically very small at short lag distances, and we have seen above that only data at such short distances from target points or blocks carry appreciable weight in the kriging systems.

Finally, even if we ignore the bias of the prediction variances of both the trend and the kriging from the residuals, regression kriging does not allow us to combine them into a valid prediction variance for the kriging estimate, although we could compute the universal kriging variance, as did Hengl *et al.* (2004).

In summary, to predict values of environmental variables that have both pronounced spatial trend and spatially dependent random variation requires us to obtain minimum-variance estimates of the trend, to estimate the variogram

of the residuals from the trend without bias and to estimate the sum of the trend and the random variation at unsampled sites with known variance. A practical way of doing this is to compute the empirical best linear unbiased predictor (E-BLUP) with a variogram estimated by residual maximum likelihood (REML). The method was recommended by Stein (1999), and now with ever increasing computing power and sophisticated software it is becoming feasible in practice. Lark and Webster (2006) used it to re-estimate the heights of the Cenomanian surface beneath the Chiltern Hills of England.

## **9.2 APPLICATION OF RESIDUAL MAXIMUM LIKELIHOOD**

### **9.2.1 Estimation of the variogram by REML**

From here on we shall use matrix notation for compactness. We start by writing equation (9.6) as

$$Z(\mathbf{x}) = \mathbf{w}\boldsymbol{\beta} + \varepsilon(\mathbf{x}), \quad (9.14)$$

in which the vector  $\mathbf{w}$ , with  $K + 1$  columns, contains the  $K + 1$  elements, 1,  $x_1, \dots, x_k$  of the trend function, and the vector  $\boldsymbol{\beta}$  contains the coefficients. Statisticians call this a linear mixed model of fixed and random effects; these are the  $\mathbf{w}\boldsymbol{\beta}$  and  $\varepsilon(\mathbf{x})$ , respectively, in the above equation.

Now let us represent a set of data in a similar way:

$$\mathbf{z}(\mathbf{X}_d) = \mathbf{W}_d\boldsymbol{\beta} + \boldsymbol{\varepsilon}(\mathbf{X}_d), \quad (9.15)$$

in which the subscript d denotes data points. For  $N$  data the vectors  $\mathbf{z}$  and  $\boldsymbol{\varepsilon}$  have  $N$  rows. The vector  $\mathbf{w}$  of equation (9.14) is replaced by matrix  $\mathbf{W}_d$ , known as a ‘design matrix’, with  $N$  rows and  $K + 1$  columns. Matrix  $\mathbf{X}_d$  denotes the positions of the points. We assume that the random components are second-order stationary and jointly normally distributed with zero means and a covariance matrix  $\mathbf{C}_{dd}$ .

The covariance matrix is obtained from the covariance function  $C(\mathbf{h})$  which, since the process is second-order stationary, has its equivalence in the variogram:

$$C(\mathbf{h}) = C(\mathbf{0}) - \gamma(\mathbf{h}) = \sigma^2 - \gamma(\mathbf{h}). \quad (9.16)$$

As we have seen in Chapter 5, most apparently bounded experimental variograms are readily fitted by simple functions with three parameters, namely a nugget variance,  $c_0$ , a sill of the correlated structure,  $c$ , and a distance parameter,  $a$ . We shall find it convenient to denote these by the vector  $\boldsymbol{\theta} = [c_0, c, a]$ .

The parameters  $\boldsymbol{\theta}$  must be estimated from the data. To do this we must separate the random component from the trend, otherwise they will be biased because they depend non-linearly on  $\boldsymbol{\beta}$ . The solution involves the transformation of the non-stationary data,  $\mathbf{z}$ , into stationary increments,  $\mathbf{s}$ .

We first define what is technically known as a projection matrix,  $\mathbf{P}$  (see Kitanidis, 1983; Pardo-Igúzquiza, 1997):

$$\mathbf{P} = \mathbf{I} - \mathbf{W}_d(\mathbf{W}_d^T \mathbf{W}_d)^{-1} \mathbf{W}_d^T. \quad (9.17)$$

We can use this matrix to transform the data in  $\mathbf{z}$  into generalized stationary increments,  $\mathbf{y}$ , by

$$\mathbf{y} = \mathbf{P}\mathbf{z}(\mathbf{X}_d). \quad (9.18)$$

Matrix  $\mathbf{P}$  has the property that

$$\mathbf{P}\mathbf{W}_d = \mathbf{0}. \quad (9.19)$$

So

$$\begin{aligned} \mathbf{P}\mathbf{z}(\mathbf{X}_d) &= \mathbf{P}\mathbf{W}_d\boldsymbol{\beta} + \mathbf{P}\boldsymbol{\varepsilon}(\mathbf{X}_d) \\ &= \mathbf{P}\boldsymbol{\varepsilon}(\mathbf{X}_d). \end{aligned} \quad (9.20)$$

In words, pre-multiplying the data by  $\mathbf{P}$  filters out the fixed effects, the trend, whatever the (unknown) coefficients of  $\boldsymbol{\beta}$  are.

As there are  $K + 1$  terms in the trend function,  $K + 1$  of the stationary increments depend linearly on the others, and so we can remove any  $K + 1$  rows from matrix  $\mathbf{P}$  and still retain all the information. We denote this matrix by  $\mathbf{H}$ , and then define the reduced set of  $m = N - K - 1$  stationary increments as

$$\mathbf{s} = \mathbf{H}\mathbf{z}(\mathbf{X}_d). \quad (9.21)$$

Further, as for matrix  $\mathbf{P}$ ,

$$\mathbf{H}\mathbf{W}_d = \mathbf{0}. \quad (9.22)$$

So we have that

$$\mathbf{E}[\mathbf{s}] = \mathbf{0} \quad (9.23)$$

and

$$\mathbf{E}[\mathbf{s}\mathbf{s}^T] = \mathbf{H}\mathbf{E}[\mathbf{z}\mathbf{z}^T]\mathbf{H}^T = \mathbf{H}\mathbf{C}_{dd}\mathbf{H}^T. \quad (9.24)$$

The increments are assumed to be normally distributed. Note also that the vector  $\mathbf{s}$  is of length  $m = N - K - 1$  and that  $\mathbf{s}\mathbf{s}^T$  has dimensions  $m \times m$ .

The parameters of the variogram model,  $\boldsymbol{\theta}$ , are obtained by maximizing the log-likelihood of the residuals. Given the data, this is

$$\begin{aligned} L[\boldsymbol{\theta} \mid \mathbf{z}(\mathbf{X}_d)] &= \frac{1}{2} \ln(m) - \frac{1}{2} m \ln(2\pi) - \frac{1}{2} m \\ &\quad - \frac{1}{2} \ln |\mathbf{H}\mathbf{C}_{dd}\mathbf{H}^T| - \frac{1}{2} m \ln \left\{ \mathbf{s}^T (\mathbf{H}\mathbf{C}_{dd}\mathbf{H}^T)^{-1} \mathbf{s} \right\}. \end{aligned} \quad (9.25)$$

Those values of  $c_0$ ,  $c$  and  $a$  in  $\boldsymbol{\theta}$  that maximize the log-likelihood,  $L$ , are found numerically, and, knowing these, we can proceed with the estimation.

### The prediction

From the values we have obtained for the parameters in  $\boldsymbol{\theta}$  we compute the estimated covariance matrix,  $\hat{\mathbf{C}}_{dd}$ . We then use this to obtain estimates of  $\boldsymbol{\beta}$  by generalized least squares:

$$\hat{\boldsymbol{\beta}} = (\mathbf{W}_d^T \hat{\mathbf{C}}_{dd}^{-1} \mathbf{W}_d)^{-1} \mathbf{W}_d^T \hat{\mathbf{C}}_{dd}^{-1} \mathbf{z}(\mathbf{X}_d). \quad (9.26)$$

We can now predict  $Z$  at  $\mathbf{x}_0$ , the target point, by

$$\hat{Z}(\mathbf{x}_0) = \{\mathbf{w}(\mathbf{x}_0) - \hat{\mathbf{c}}_{d0}^T \hat{\mathbf{C}}_{dd}^{-1} \mathbf{W}_d\} \hat{\boldsymbol{\beta}} + \hat{\mathbf{c}}_{d0}^T \hat{\mathbf{C}}_{dd}^{-1} \mathbf{z}(\mathbf{X}_d). \quad (9.27)$$

Here vector  $\mathbf{w}(\mathbf{x}_0)$  is the ‘design matrix’ of the target point,  $\mathbf{x}_0$ , with one row and  $K + 1$  columns, and vector  $\hat{\mathbf{c}}_{d0}$  contains the estimated covariances between the data points and  $\mathbf{x}_0$ .

Equation (9.27) represents the E-BLUP of Lark and Cullis (2004); empirical because it is derived empirically from sample data. It has two distinct parts. The first term on the right-hand side is the generalized least-squares estimate of the trend component at  $\mathbf{x}_0$ . Notice that it is more elaborate than the OLS estimate given by equations (3.9) and (3.10). The second term is the simple kriging estimate, simple because the mean of the residuals is 0 by definition. The two terms added together give us our final prediction.

The prediction variance is given by

$$\begin{aligned} \sigma_{E-BLUP}^2(\mathbf{x}_0) &= \left\{ \mathbf{W}(\mathbf{x}_0) - \hat{\mathbf{c}}_{d0}^T \hat{\mathbf{C}}_{dd}^{-1} \mathbf{W}_d \right\} \mathbf{U}^{-1} \left\{ \mathbf{W}(\mathbf{x}_0) - \hat{\mathbf{c}}_{d0}^T \hat{\mathbf{C}}_{dd}^{-1} \mathbf{W}_d \right\}^T \\ &\quad + \left\{ \hat{c}(\mathbf{x}_0) - \hat{\mathbf{c}}_{d0}^T \hat{\mathbf{C}}_{dd}^{-1} \hat{\mathbf{c}}_{d0} \right\}, \end{aligned} \quad (9.28)$$

where  $\mathbf{U} = \mathbf{W}_d^T \hat{\mathbf{C}}_{dd}^{-1} \mathbf{W}_d$ . Note that  $\hat{\mathbf{C}}_{dd}$  contains the nugget variance comprising both measurement error and very short-range spatial variation which are separated by Lark *et al.* (2006).

### 9.2.2 Practicalities

Although the principles of REML estimation in geostatistics have been recognized for some 20 years—see, for example, Kitanidis (1983, 1987) and Zimmermann and Zimmermann (1991)—practitioners have only recently started to apply them. One reason is that the full covariance matrices must be held and inverted in the computer's memory. Twenty years ago few computers were big enough or fast enough for such tasks. The size and power of modern computers now makes the method feasible with reasonably large sets of data. The other handicap has been the lack of readily available software for geostatistical applications. Pardo-Igúzquiza's (1997) MLREML Fortran program is in the public domain and provides options for three variogram models, the exponential, spherical and Gaussian. The program also has five options for minimizing negative log-likelihood functions. Of these the author considers the simplex method of Nelder and Mead (1965) to be the most effective; Kerry and Oliver (2007c) also found this to work well. The few options for the variogram in general packages seriously limit what can be done. ASReml (Gilmour *et al.*, 2002) is exceptional; its AI algorithm can handle very large sets of data. Another problem arises with functions such as the spherical model, for which the algorithm can all too readily converge to a local optimum. Lark and Cullis (2004) recognized this shortcoming and wrote a program that uses simulated annealing for the purpose, but it is still in the research phase.

### 9.2.3 Kriging with external drift

In presenting kriging in the presence of trend we have shown how to separate the deterministic trend from the random component and how to estimate the contributions from the two components by REML. The case we considered in which the trend in the target variable is a function of the spatial coordinates is one of a more general class of linear mixed models. In other cases the deterministic, or fixed, effect might be another variable, say  $y$ , or several variables,  $y_1, y_2, \dots$ , related linearly to  $Z$ , and we might be able to measure or calculate it at both target sites and where we know  $Z$ . In these circumstances we can modify equations (9.4) and (9.5) to

$$\begin{aligned} Z(\mathbf{x}) &= \sum_{k=0}^K \beta_k y_k(\mathbf{x}) + \varepsilon(\mathbf{x}) \\ &= \beta_0 + \beta_1 y_1(\mathbf{x}) + \beta_2 y_2(\mathbf{x}) + \cdots + \beta_K y_K(\mathbf{x}) + \varepsilon(\mathbf{x}). \end{aligned} \tag{9.29}$$

What we have done is to replace the functions of the coordinates by the values of one or more other variables at those places. The  $y_1(\mathbf{x}), y_2(\mathbf{x}), \dots, y_K(\mathbf{x})$  are known and the  $\beta_k$  are unknown coefficients to be determined.

The  $y_k, k = 1, 2, \dots$ , are ‘external’ variables, as distinct from the internal  $Z(\mathbf{x})$ , and kriging with them is known as ‘kriging with external drift’ (KED). The term is used to contrast it with universal kriging in which the drift is in the ‘internal’ variable.

The KED estimator is

$$\hat{Z}_{\text{KED}}(\mathbf{x}_0) = \sum_{i=1}^N \lambda_i^{\text{KED}} z(\mathbf{x}_i). \quad (9.30)$$

Its expectation is

$$E[\hat{Z}_{\text{KED}}(\mathbf{x}_0)] = \sum_{k=0}^K \sum_{i=1}^N \beta_k \lambda_i^{\text{KED}} y_k(\mathbf{x}_i), \quad (9.31)$$

and the estimator is unbiased if

$$\sum_{i=1}^N \lambda_i^{\text{KED}} y_k(\mathbf{x}_i) = y_k(\mathbf{x}_0) \quad \text{for all } k = 0, 1, \dots, K. \quad (9.32)$$

The kriging weights,  $\lambda_i^{\text{KED}}$ , are obtained by solution of the following system of equations:

$$\begin{aligned} \sum_{i=1}^N \lambda_i^{\text{KED}} \gamma(\mathbf{x}_i, \mathbf{x}_j) + \psi_0 + \sum_{k=1}^K \psi_k y_k(\mathbf{x}_j) &= \gamma(\mathbf{x}_0, \mathbf{x}_j) \quad \text{for all } j, j = 1, 2, \dots, N, \\ \sum_{i=1}^N \lambda_i^{\text{KED}} &= 1, \\ \sum_{i=1}^N \lambda_i^{\text{KED}} y_k(\mathbf{x}_i) &= y_k(\mathbf{x}_0) \quad \text{for } k = 1, 2, \dots, K, \end{aligned} \quad (9.33)$$

where  $\gamma(\mathbf{x}_i, \mathbf{x}_j)$  are the semivariances of the residuals between the data points  $\mathbf{x}_i$  and  $\mathbf{x}_j$ , the  $\gamma(\mathbf{x}_0, \mathbf{x}_j)$  are the semivariances between the target point and the data points, and the  $\psi_k, k = 0, 1, \dots, K$ , are Lagrange multipliers. The system is solved to provide the weights, which are then inserted in equation (9.30) for the prediction, and the kriging variance is obtained by vector multiplication as in equation (9.13).

The form of the equations is the same as for universal kriging; all we have done is to replace the functions of the spatial coordinates,  $f(\mathbf{x})$ , by the  $y_k(\mathbf{x})$ . In this way KED incorporates secondary drift variables into the kriging system; it combines information from the deterministic  $y_k(\mathbf{x})$  with that of the random  $Z(\mathbf{x})$ . Typically it is used to predict  $Z(\mathbf{x})$  more precisely than from

measurements of  $Z$  alone. It was developed in the petroleum and gas exploration industries where boreholes from which accurate measurements can be obtained are few but can be supplemented with seismic data from many sites. Delhomme (1978) combined seismic data ( $y$ , the external drift variable) with measurements of height of the top of a reservoir ( $Z$  in our terms above) to map the latter. The method has since been applied, for example, to estimate transmissivity of an aquifer (Ahmed and de Marsily, 1987) with specific capacity as the external drift variable, and regional temperature (Hudson and Wackernagel, 1994) and soil mineral nitrogen (Baxter and Oliver, 2005) with height of the land as external drift.

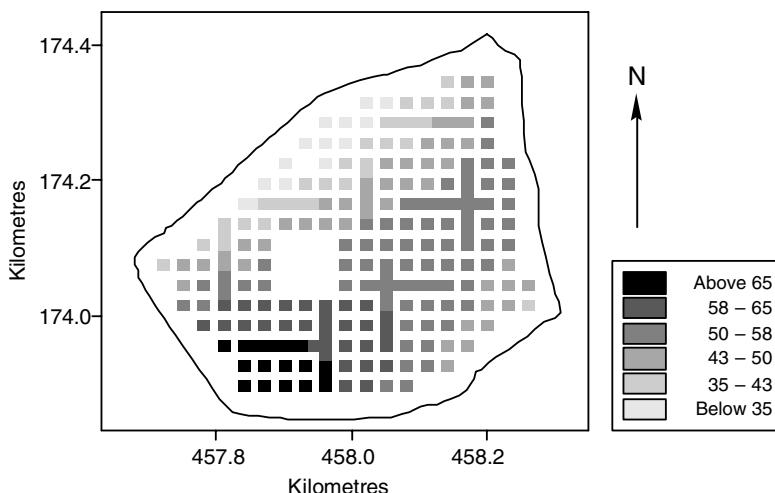
Before we illustrate the kriging with external drift we emphasize three points.

- The subsidiary variable(s),  $y$ , should vary smoothly at the scale of the survey. If any  $y$  does not then it should be treated as a random variable and used as a covariate in cokriging (Chapter 10).
- One must know or be able to calculate the values of the subsidiary variable(s) at all target points and all points for which the primary variable  $Z$  has been recorded.
- The variogram from which the entries in the kriging system are drawn is that of the residuals  $\varepsilon(\mathbf{x})$  from the external drift  $y$ . If we have an independent estimate of it then we may use it. Usually we do not, and we have to separate its effect from the fixed effect in our data  $z(\mathbf{x}_i)$ , just as in universal kriging, and again we can now use REML to solve the problem this poses.

### 9.3 CASE STUDY

We illustrate kriging in the presence of trend with the results from recent research on precision farming for the British Home-Grown Cereals Authority (HGCA) by Oliver and Carroll (2004). The study was done in a 23-ha field, National Grid reference SU 458174, on the Yattendon Estate in Berkshire, England. It is on the Chalk downland of southern England and has the typical undulating topography of the region. The soil, which is moderately to well drained, varies from sandy loam to clay loam, and it is its sand content in the topsoil that we use for this illustration.

Samples of topsoil (0–15 cm) were taken at the nodes of a 30 m × 30 m grid. At each node ten cores of soil were taken with an auger of 3 cm diameter from a support of 5 m × 2 m and bulked. Additional observations were made at 15-m intervals along short transects from randomly selected grid nodes. The sand content was measured by laser diffraction grain sizing. Figure 9.1 is a scatter map showing the sampling scheme and the percentages of sand at the sampling points on a grey scale. There is evidently a trend across the field. Table 9.1 summarizes the statistics. The sand content varies widely from 14% to 83%, with a symmetric distribution that is less peaked than normal.



**Figure 9.1** Scatter map of the sand content of the topsoil at Yattendon.

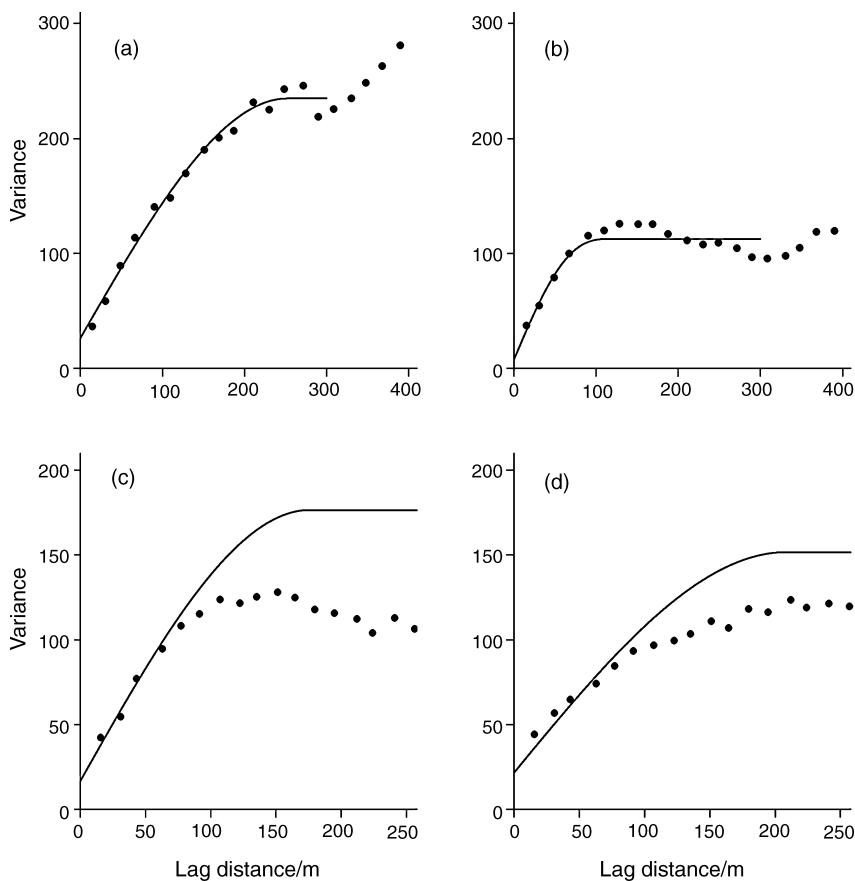
The experimental variogram of the data computed by the usual method of moments, equation (4.40), is shown as the plotted points in Figure 9.2(a). The values appear to reach a sill and then increase again; the latter increase is symptomatic of regional trend. The solid line is a model fitted to lag 300 m, a matter to which we return below.

To explore the data we fitted trend surface models on the coordinates by OLS. The linear trend surface (an inclined plane) accounts for 28% of the variance and the quadratic accounts for more than 46%. Therefore, we assume the trend to have a quadratic form in further analyses.

Figure 9.2(b) shows the experimental variogram of the OLS residuals from the quadratic trend (symbols) and the fitted pentaspherical function (solid line). Table 9.2 gives the parameters of the model that we used to krig the quadratic

**Table 9.1** Summary statistics for topsoil sand content (%) at Yattendon.

Minimum	14.00
Maximum	83.00
Mean	50.84
Median	51.00
Standard deviation	14.40
Variance	207.4
Skewness	0.02
Kurtosis	-0.60
Number of observations	230

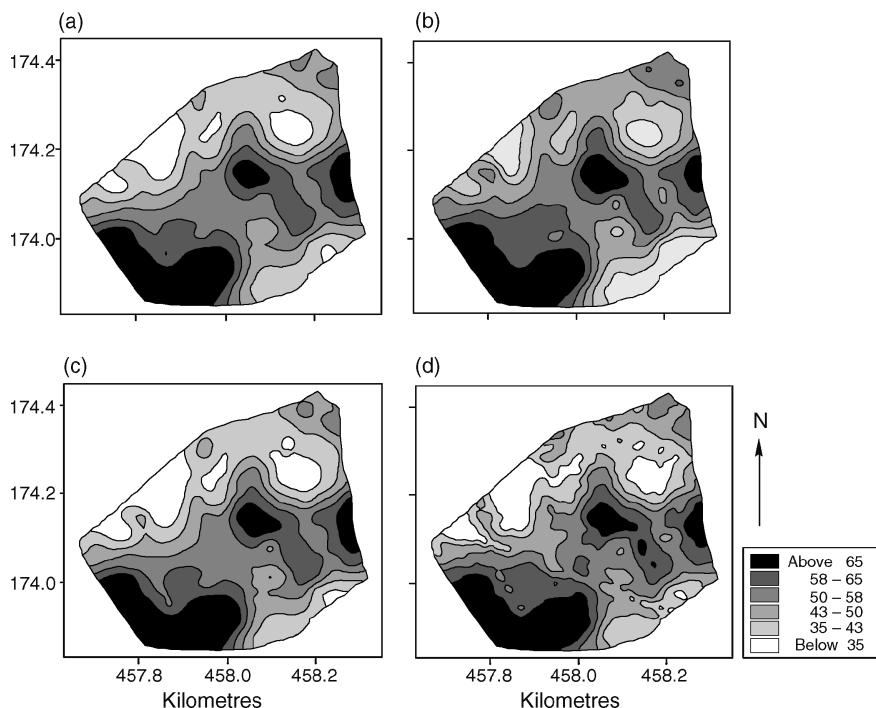


**Figure 9.2** Variograms of the sand content: (a) experimental variogram computed from the raw data by the method of moments and spherical model fitted to lag 300 m; (b) experimental variogram of the OLS residuals from a quadratic trend surface with pentaspherical model fitted; (c) variogram of the REML residuals from the quadratic trend (solid line) with the experimental semivariances of the REML residuals plotted as points; (d) variogram of the REML residuals after the EC<sub>a</sub> has been fitted as an external drift (solid line) with the experimental semivariances for the OLS residuals from the external drift plotted as points.

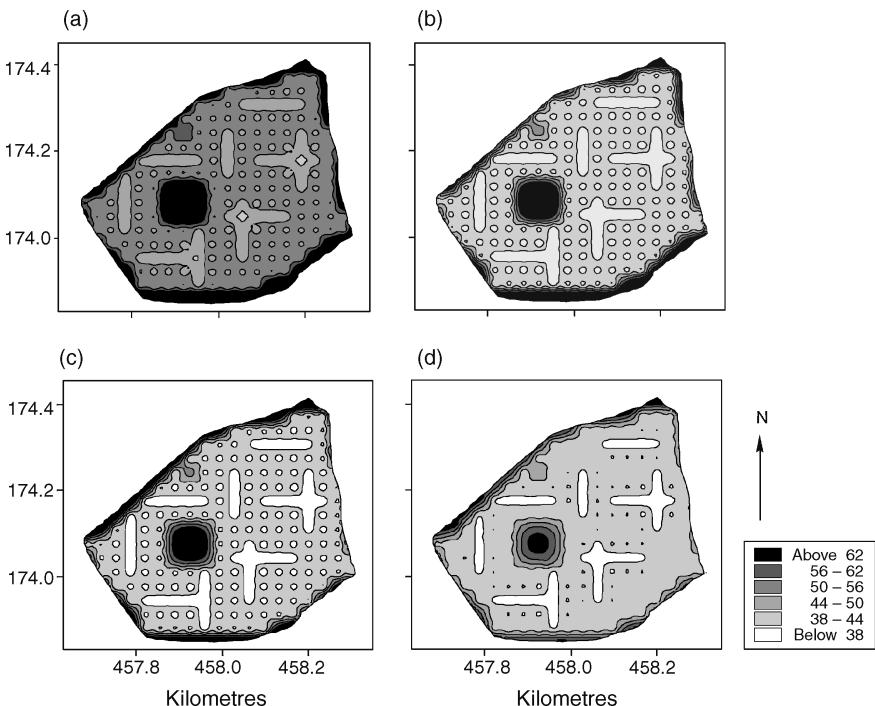
residuals. We then added back the trend to the predictions of the residuals to give the final estimates. These estimates are shown as a map in Figure 9.3(b), and Figure 9.4(b) is that of the associated kriging variances. The latter are for the residuals only; they show that the estimation variances are smallest in the region of the short transects and also at the sampling points and become large only near the field boundary.

**Table 9.2** Model parameters of the variograms computed on the topsoil sand content at Yattendon. The symbols are the familiar  $c_0$  for the nugget variance,  $c$  for the sill of the autocorrelated variance, and  $a$  for the range.

Variogram	Model	$c_0$	$c$	$c_0 + c$	$a/m$	$c_0/(c_0 + c)$
Raw data OLS	Spherical	27.5	208.1	235.6	254.9	0.117
residuals	Pentaspherical	15.6	110.2	125.8	146.6	0.124
REML	Spherical	16.6	159.8	176.4	175.8	0.104
REML with EC <sub>a</sub>	Spherical	21.7	129.9	151.6	208.7	0.167



**Figure 9.3** Maps of punctually kriged estimates of sand content: (a) made by ordinary kriging of raw data; (b) made by kriging of OLS residuals from a quadratic trend and adding back the trend; (c) made by REML estimation, taking into account the quadratic trend (universal kriging); (d) made by REML estimation with EC<sub>a</sub> as external drift.



**Figure 9.4** Maps of punctual kriging variances of sand content: (a) ordinary kriging variances of raw data; (b) ordinary kriging variances of OLS residuals from a quadratic trend; (c) universal kriging variances for REML estimation, taking into account the quadratic trend; (d) variances for kriging by REML with  $EC_a$  as external drift.

As above, dealing with trend by OLS can no longer be regarded as best practice, and we now illustrate how do it with a REML analysis. The experimental variogram of the quadratic residuals (the plotted points) and the model estimated by REML for the residuals (the solid line) are shown in Figure 9.2(c). We remind readers that there are no experimental semivariances for the REML variogram. The model parameters are given in Table 9.2.

Let us now compare these models. The most marked difference is in the sill variances,  $c_0 + c$ ; that for the OLS residuals is substantially smaller than that estimated by REML. The range of the former is the shorter of the two by some 30 m.

With the parameters of the REML variogram we can now predict the sand content by E-BLUP taking into account the quadratic trend, which is equivalent to universal kriging, as above. Figure 9.3(c) is the resulting map, and Figure 9.4(c) shows the associated variances.

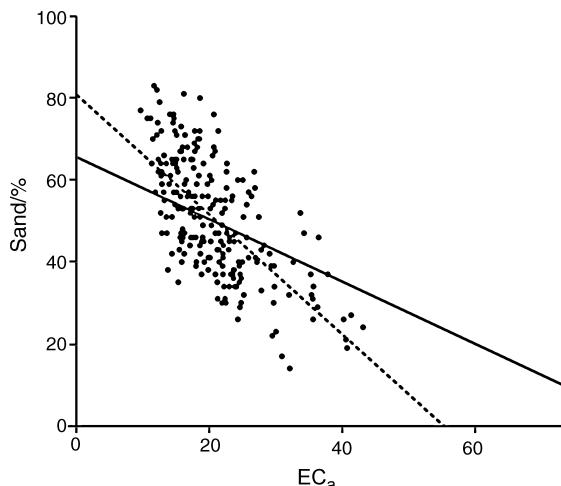
Another variable with a strong quadratic form of trend in this field is the soil's apparent electrical conductivity ( $EC_a$ ). This is also strongly associated with the

soil's particle size distribution. Carroll and Oliver (2005) measured the  $\text{EC}_a$  by an electromagnetic induction sensor, an EM38 (Geonics<sup>®</sup> Ltd—see McNeill, 1990) in the vertical position (i.e. with the coils aligned vertically), at the nodes of a dense grid and at all the positions where sand content was recorded. The Pearson correlation coefficient computed between it and sand from collocated data at the time of the survey was  $-0.8$ .

This strong correlation, the dense grid on which  $\text{EC}_a$  was measured and the marked trend in it make it a potentially useful external drift variable for kriging. Our aim is to use it to improve the accuracy of the predictions of the sand content, the primary variable.

The  $\text{EC}_a$  was estimated by ordinary punctual kriging at the nodes of the  $5 \text{ m} \times 5 \text{ m}$  grid as for sand. The relation between sand content and  $\text{EC}_a$  is linear, and Figure 9.5 shows both the OLS regression (dotted) and the weighted least-squares regression (solid) of sand on  $\text{EC}_a$ . The REML variogram for sand was computed on the random residuals from the relation. Figure 9.2(d) shows this variogram, together with the experimental variogram of the OLS residuals, and Table 9.2 gives the model parameters. These parameters were then used for kriging sand with  $\text{EC}_a$  on the  $5 \text{ m} \times 5 \text{ m}$  grid as the external drift. Figure 9.3(d) is the resulting map of estimated sand content from this analysis, and Figure 9.4(d) shows the associated E-BLUP kriging variances.

The four variograms (Figure 9.2) appear substantially different from one another, and their model parameters confirm this impression. The variogram of the raw data, Figure 9.2(a), has the largest sill variance. The variogram of the OLS quadratic residuals, Figure 9.2(b), has the smallest sill variance, showing



**Figure 9.5** Scatter diagram of sand content and  $\text{EC}_a$ . The solid line is the weighted least-squares regression of sand on  $\text{EC}_a$ ,  $\text{sand} = 65.7 - 0.76 \times \text{EC}_a$ , from which the experimental semivariances in Figure 9.2(d) are computed.

that removal of the trend has lost more of the variance than REML has done. The REML variogram of the residuals from the quadratic trend, Figure 9.2(c), has a larger sill variance as has the REML variogram computed on the residuals from the relation between sand and EC<sub>a</sub>, Figure 9.2(d).

We mentioned above that there is a dearth of software in the public domain for spatial prediction by REML. Readers, having seen Figure 9.2(a) in which the experimental sequence of semivariances follows what looks like a spherical form to about 300 m, and having also seen in Chapter 8 that only data close to a target point carry significant weight, might wonder whether in this case they could use ordinary kriging with the raw data. Let us see what happens if we take this approach. A spherical function, the solid line in Figure 9.2(a), fits the experimental variogram well to 300 m. Table 9.2 gives its parameters. We used this model with the raw data to estimate values at the nodes of the 5 m × 5 m grid by ordinary punctual kriging. Figure 9.3(a) shows the resulting map, which looks little different from those made with REML. The map of the kriging variances, however, shows that the kriging errors are greater.

We can summarize the four outcomes of the procedures. All four estimators are unbiased, and because kriging is so robust the estimates themselves are similar. They differ substantially in their variances, which are summarized in Table 9.3. Ordinary kriging is the least precise, with a median variance of 53.1.

Universal kriging by REML reduces the variance to a median of 41.6. Kriging with external drift has a very similar median variance, 41.3. In both, making use of the additional information, either in the trend or in the subsidiary correlated variable, EC<sub>a</sub>, improves the precision of the predictions. Note, however, that the kriging variances from the KED have a smaller standard deviation; they are less variable than those from universal kriging, and from the other two techniques. The median kriging variance of the OLS residuals of 41.6 is remarkably similar to those of the REML predictions in this instance. Note, however, that the OLS kriging variances underestimate the true kriging variances of that method because the errors arising from the OLS fitting of the trend are not taken into account. The fact that they are so similar to the REML variances is fortuitous.

**Table 9.3** Summary statistics of variances for four forms of kriging.

Kriging	Mean	Median	Std dev.
Raw data	63.2	53.1	21.0
OLS residuals	52.0	41.6	21.7
REML, universal	53.5	41.6	26.3
REML, external drift	48.2	41.3	14.6

## 9.4 FACTORIAL KRIGING ANALYSIS

### 9.4.1 Nested variation

Spatial variation in the environment can occur on scales that differ by several orders of magnitude simultaneously. This is because the physical processes responsible for the variation operate and interact at different spatial scales. In any region there may be several sources and scales of variability present. For example, variation in the soil could arise from the effects of microbial activity, worms, roots, tree-throw, relief, or geology. Variation of this kind is widespread in the environment; for example, Serra (1968) found seven different scales of spatial variation in the Lorraine iron ore deposit, ranging from  $15\ \mu\text{m}$  to several hundred metres. The result is a nested structure in the variation that we have already observed through the nested variogram functions in Chapter 5.

Although nested variation has been recognized for some time, there is now a greater need than before to investigate it. This need has arisen from the increasingly rich sets of data emerging from new technology, such as satellite imagery, ground-penetrating radar, and sensors that measure electrical conductivity. Such data often cover large areas at an intermediate spatial resolution of about  $30\text{ m} \times 30\text{ m}$ , for example, or are very intensive, such as the 1-m pixel resolution Hymap imagery. Because such sources of data provide full cover of the areas of interest, nested variation is often evident (Oliver *et al.*, 2000).

### 9.4.2 Theory

We can formalize nested variation in a geostatistical framework as follows. A particular random process,  $Z(\mathbf{x})$ , may be treated as a combination of several independent processes, one nested within another and acting at different characteristic spatial scales. In these circumstances the variogram of  $Z(\mathbf{x})$  is itself a nested combination of two or more, say  $S$ , individual variograms:

$$\gamma(\mathbf{h}) = \gamma^1(\mathbf{h}) + \gamma^2(\mathbf{h}) + \cdots + \gamma^S(\mathbf{h}), \quad (9.34)$$

where the superscripts refer to the separate variograms (not powers).

If we assume that the processes are uncorrelated then we can represent equation (9.34) by the sum of  $S$  basic variograms:

$$\gamma(\mathbf{h}) = \sum_{k=1}^S b^k g^k(\mathbf{h}), \quad (9.35)$$

where  $g^k(\mathbf{h})$  is the  $k$ th basic variogram function, and  $b^k$  is a coefficient that measures the relative contribution of the variance of  $g^k(\mathbf{h})$  to the sum. The

nested variogram comprises the  $S$  variograms with different coefficients  $b^k$ . This is known as the linear model of regionalization. It reflects the real world in which factors such as relief, geology, the soil, vegetation, and fauna each operate on their own characteristic spatial scale(s) and each with its particular form and parameters,  $b^k$ , for  $k = 1, 2, \dots, S$ .

### 9.4.3 Kriging analysis

The aim of kriging analysis is to estimate separately the independent components of  $Z(\mathbf{x})$ . Matheron (1982) devised the technique of kriging analysis or factorial kriging to do this in a single operation from the data and the variogram. For this the random function  $Z(\mathbf{x})$  with a nested variogram is regarded as the sum of  $S$  orthogonal random functions, each with its particular contributory variogram,  $b^k g^k(\mathbf{h})$  in equation (9.35). Provided  $Z(\mathbf{x})$  is second-order stationary this sum can be represented as

$$Z(\mathbf{x}) = \sum_{k=1}^S Z^k(\mathbf{x}) + \mu, \quad (9.36)$$

in which  $\mu$  is the mean of the process. Each  $Z^k(\mathbf{x})$  has expectation 0, and the squared differences are

$$\frac{1}{2} E[\{Z^k(\mathbf{x}) - Z^k(\mathbf{x} + \mathbf{h})\}\{Z^{k'}(\mathbf{x}) - Z^{k'}(\mathbf{x} + \mathbf{h})\}] = \begin{cases} b^k g^k(\mathbf{h}) & \text{if } k = k', \\ 0 & \text{otherwise.} \end{cases} \quad (9.37)$$

It is possible that the last component,  $Z^S(\mathbf{x})$ , is intrinsic only, so that  $g^S(\mathbf{h})$  in equation (9.35) is unbounded with gradient  $b^S$ . For two components equation (9.36) reduces to

$$Z(\mathbf{x}) = Z^1(\mathbf{x}) + Z^2(\mathbf{x}) + \mu. \quad (9.38)$$

Relation (9.37) expresses the mutual independence of the  $S$  random functions  $Z^k(\mathbf{x})$ . With this assumption, the nested model in equation (9.35) is easily retrieved from the relation in equation (9.36).

We recall that in ordinary kriging we usually estimate  $Z$  at any place  $\mathbf{x}_0$  as a linear combination of the  $n$  observations in the neighbourhood of  $\mathbf{x}_0$ :

$$\hat{Z}(\mathbf{x}_0) = \sum_{i=1}^n \lambda_i z(\mathbf{x}_i). \quad (9.39)$$

The weights,  $\lambda_i$ ,  $i = 1, 2, \dots, N$ , are obtained by solution of the kriging system

$$\begin{aligned} \sum_{j=1}^n \lambda_j \gamma(\mathbf{x}_i, \mathbf{x}_j) - \psi(\mathbf{x}_0) &= \gamma(\mathbf{x}_i, \mathbf{x}_0) \quad \text{for all } i = 1, 2, \dots, n, \\ \sum_{j=1}^n \lambda_j &= 1, \end{aligned} \tag{9.40}$$

in which  $\psi(\mathbf{x}_0)$  is a Lagrange multiplier introduced to ensure that the estimation variance is minimized.

In kriging analysis we can estimate each spatial component  $Z^k(\mathbf{x})$  separately as a linear combination of the observations  $z(\mathbf{x})$ ,  $i = 1, 2, \dots, n$ :

$$\hat{Z}^k(\mathbf{x}_0) = \sum_{i=1}^n \lambda_i^k z(\mathbf{x}_i). \tag{9.41}$$

The  $\lambda_i^k(\mathbf{x}_0)$  are the weights assigned to the observations as before; but now they must sum to 0, not to 1, to ensure that the estimate is unbiased and to accord with equation (9.36). Subject to this condition, they are again chosen so that the estimation variance is minimal. This leads to the kriging system

$$\begin{aligned} \sum_{j=1}^n \lambda_j^k \gamma(\mathbf{x}_i, \mathbf{x}_j) - \psi^k(\mathbf{x}_0) &= b^k g^k(\mathbf{x}_i, \mathbf{x}_0) \quad \text{for all } i = 1, 2, \dots, n, \\ \sum_{j=1}^n \lambda_j^k &= 0, \end{aligned} \tag{9.42}$$

where  $\psi^k(\mathbf{x}_0)$  is the Lagrange multiplier for the  $k$ th component. This system is solved for each spatial component,  $k$ , to find the weights,  $\lambda_i^k$ , which are then inserted into equation (9.41) for that component. In general the weights for the different components will be different, and as a result we can extract from data the individual components of the spatial variation that we identified from the experimental variogram. Estimates are made for each spatial scale, i.e. each  $k$ , by solving equations (9.42).

In many instances data contain long-range trend. This need not complicate the analysis because the kriging is usually done in fairly small moving neighbourhoods centred on  $\mathbf{x}_0$ , as for ordinary kriging (Chapter 8). Thus it is necessary only that  $Z(\mathbf{x})$  is locally stationary, or *quasi-stationary*. Equation (9.36) may then be rewritten as

$$Z(\mathbf{x}) = \sum_{k=1}^S Z^k(\mathbf{x}) + \mu(\mathbf{x}), \tag{9.43}$$

where  $\mu(\mathbf{x})$  is a local mean which can be considered as a long-range spatial component. Matheron (1982) showed that this relation is also verified in terms of estimators, i.e.

$$\hat{Z}(\mathbf{x}) = \sum_{k=1}^S \hat{Z}^k(\mathbf{x}) + \hat{\mu}(\mathbf{x}). \quad (9.44)$$

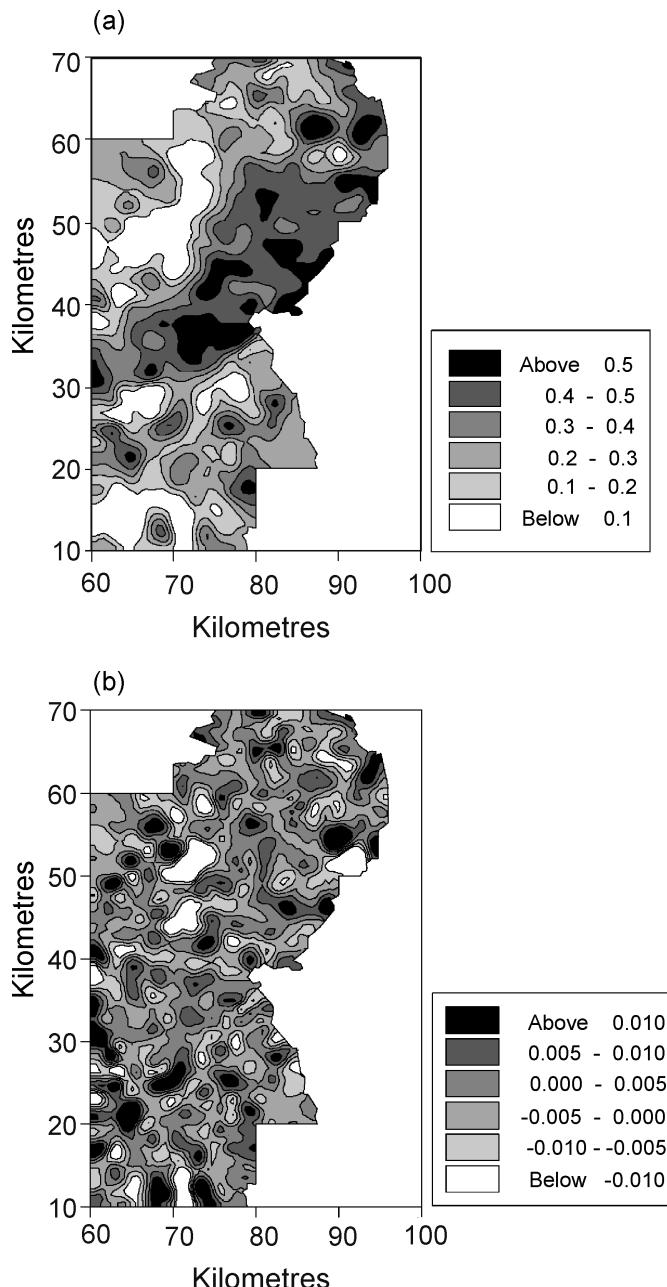
We have also to krig the local mean, which is again a linear combination of the observations  $z(\mathbf{x}_i)$ :

$$\hat{\mu}(\mathbf{x}_0) = \sum_{j=1}^n \lambda_j z(\mathbf{x}_j). \quad (9.45)$$

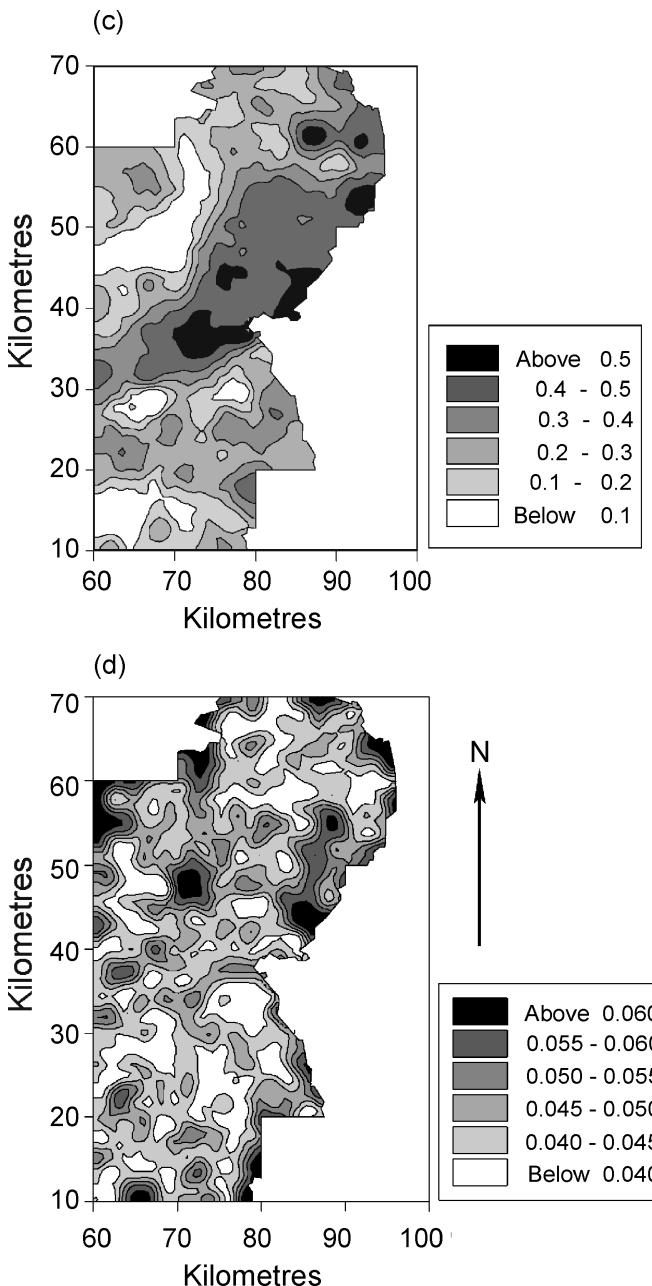
The weights are obtained by solving the kriging system

$$\begin{aligned} \sum_{j=1}^n \lambda_j \gamma(\mathbf{x}_i, \mathbf{x}_j) - \psi(\mathbf{x}_0) &= 0 \quad \text{for all } i = 1, 2, \dots, n, \\ \sum_{j=1}^n \lambda_j &= 1. \end{aligned} \quad (9.46)$$

Estimation of the long-range component, i.e. the local mean  $\mu(\mathbf{x})$  and the spatial component with the largest range, can be affected by the size of the moving neighbourhood (Galli *et al.*, 1984). In fact, to estimate a spatial component with a given range the distance across the neighbourhood should be at least equal to that range. It happens frequently when the sampling density and the range are large that there are so many data within the chosen neighbourhood that only a small proportion of them is retained. Although modern computers can handle many data at a time, the number of data used must be limited to avoid instabilities when inverting large variance matrices. Further, even if all the data could be retained, only the nearest ones contribute to the estimate because they screen the more distant data. Consequently, the neighbourhood actually used is smaller than the neighbourhood specified, which means that the range of the estimated spatial component is smaller than the range apparent from the structural analysis. Galli *et al.* (1984) recognized this, and where data lie on a regular grid they proposed using only every second or every fourth point to cover a large enough area, but still with sufficient data. Such selection is somewhat arbitrary, and we recommend an alternative proposed by Jaquet (1989) and used by Goovaerts and Webster (1994) which involves adding to the long-range spatial component the estimate of the local mean.



**Figure 9.6** Maps of copper in the topsoil of the Borders Region of Scotland: (a) ordinary kriged estimates; (b) estimates of short-range component; (c) estimates of long-range component; (d) kriging variances.



**Table 9.4** Summary statistics and variogram parameters of available copper in the topsoil of the Borders Region of Scotland. Original measurements were in mg kg<sup>-1</sup> soil.

	Mean	Median	Variance	Std dev.	Skewness
Measurements, Cu	2.22	1.85	2.135	1.46	2.52
$\log_{10}\text{Cu}$	0.271	0.267	0.06502	0.255	0.06
Double spherical variogram parameters					
	$c_0$	$c_1$	$c_2$	$a_1/\text{km}$	$a_2/\text{km}$
	0.02767	0.02585	0.01505	2.7	20.5

#### 9.4.4 Illustration

We illustrate factorial kriging with the available copper data in the topsoil of the Borders Region of Scotland. These data were described in Chapter 6 to illustrate the application of the Akaike information criterion (AIC). Table 9.4 lists the summary statistics, which show that these data have a large skewness coefficient of 2.52. After transformation to common logarithms,  $\log_{10}\text{Cu}$ , the distribution becomes almost normal. The experimental variogram was computed on the transformed  $\log_{10}\text{Cu}$  values, and a double spherical function fitted the experimental values best both in terms of the mean squared residual and the AIC (Figure 5.15 and Table 5.2). This function was then used for factorial punctual kriging. We kriged at intervals of 500 m with the maximum radius of the neighbourhood set to 20 km, the range of the long-range spatial component. The minimum and maximum numbers of points in the neighbourhood were set to seven and 20, respectively.

Figure 9.6(a) is a map of the punctually kriged estimates of  $\log_{10}\text{Cu}$ . There is a band of large values that extend across the region from southwest to northeast, with smaller concentrations of  $\log_{10}\text{Cu}$  to the north and south. The extent of these patches of large and small concentrations represents the long-range component of the variation of 20.5 km. These larger areas embrace many distinct small patches of larger or smaller values which represent the short-range component of variation of 2.7 km. Figure 9.6(d) shows the kriging variances; these are large at the margins of the study area and in other areas where sampling was also sparse. Figure 9.6(b) shows the short-range predictions that have been extracted by factorial kriging. They show the more intricate local variation that is superimposed on the broader-scale variation shown in Figure 9.6(c). It is possible to match several of the small patches of large and small values in Figure 9.6(a) with those in Figure 9.6(c). The long-range variation is associated with the major geological units and soil parent materials in the region. The areas with small copper concentrations are on the sedimentary rocks of the Old Red Sandstone, whereas on the other rocks concentrations are generally larger. Some of the smaller patches of large Cu values are around the towns and others are associated with outcrops of volcanic rocks.

# 10

## ***Cross-Correlation, Coregionalization and Cokriging***

### **10.1 INTRODUCTION**

In this chapter we develop the ideas of spatial correlation in individual variables for use in situations in which two or more environmental variables interest us simultaneously. We shall assume that each variable individually can be treated as if it were random, so that all of the statistical theory and techniques of Chapters 4–7 apply. We shall use the data from two surveys to illustrate the development. One set comprises the exchangeable potassium (K), available phosphorus (P) and yield of barley in the topsoil of a 6.4 ha field in southeast England (CEDAR Farm, Centre for Dairy Research); and the other comprises the concentrations of potentially toxic trace metals in the Swiss Jura. Table 10.1 summarizes the data for the Farm, and Table 10.5 that of the Jura.

There are now two additional features of the variation to consider. One comprises the relations between variables, regardless of space, as expressed in the ordinary product-moment correlation,  $r$ , of equation (2.11). The correlation matrix for CEDAR Farm is given in Table 10.2. Evidently K, P and yield are related, though not strongly. In the Swiss Jura the correlations among the trace metals in the soil are stronger (Table 10.6), and we might wish to consider them all together in assessing the risk of pollution. The other feature concerns the spatial aspects of this correlation: one variable may be spatially related with another in the sense that its values at places are correlated with the values of the other variable. For example, the potassium and phosphorus in the soil at CEDAR Farm might be spatially correlated with the crop yield, and the cadmium with the zinc in the Jura. In these circumstances we might be able to take advantage of the correlation and the information contained in the several variables to predict any one of them.

**Table 10.1** Summary statistics of K, P and yield of barley at CEDAR Farm based on a sample of  $N = 160$ .

	K	P	Yield
Minimum	101.0	16.8	1.28
Maximum	243.0	89.0	4.43
Mean	155.7	49.4	3.03
Median	151.0	50.9	3.11
St. dev.	28.7	14.4	0.50
Variance	825.65	206.84	0.249
Skewness	0.91	-0.24	-0.72

We formalize these ideas under the general heading of *coregionalization*. We start by considering two regionalized variables,  $Z_u(\mathbf{x})$  and  $Z_v(\mathbf{x})$ , which we shall denote  $u$  and  $v$ , both obeying the intrinsic hypothesis. Thus for variable  $u$  we have, from Chapter 4,

$$E[Z_u(\mathbf{x}) - Z_u(\mathbf{x} + \mathbf{h})] = 0.$$

The variable will also have a variogram, specifically an *autovariogram*:

$$\gamma_{uu}(\mathbf{h}) = \frac{1}{2}E[\{Z_u(\mathbf{x}) - Z_u(\mathbf{x} + \mathbf{h})\}^2]. \quad (10.1)$$

The reason for the double  $uu$  will become apparent presently. Similarly for  $v$ , the expected differences are 0, and its autovariogram is  $\gamma_{vv}(\mathbf{h})$ , consisting of the expected squared differences in  $v$ .

The two variables will also have a *cross-variogram*,  $\gamma_{uv}(\mathbf{h})$ , defined as

$$\gamma_{uv}(\mathbf{h}) = \frac{1}{2}E[\{Z_u(\mathbf{x}) - Z_u(\mathbf{x} + \mathbf{h})\}\{Z_v(\mathbf{x}) - Z_v(\mathbf{x} + \mathbf{h})\}]. \quad (10.2)$$

This function describes the way in which  $u$  is related spatially to  $v$ .

If both variables are second-order stationary with means  $\mu_u$  and  $\mu_v$ , then both will have covariance functions:

$$C_{uu}(\mathbf{h}) = E[\{Z_u(\mathbf{x}) - \mu_u\}\{Z_u(\mathbf{x} + \mathbf{h}) - \mu_u\}] \quad (10.3)$$

**Table 10.2** Correlation matrix for K, P and yield at CEDAR Farm with 158 degrees of freedom.

K	1		
P	0.585	1	
Yield	-0.329	-0.395	1
K		P	Yield

and analogously for  $C_{vv}(\mathbf{h})$ . They will also have a cross-covariance function:

$$C_{uv}(\mathbf{h}) = E[\{Z_u(\mathbf{x}) - \mu_u\}\{Z_v(\mathbf{x} + \mathbf{h}) - \mu_v\}]. \quad (10.4)$$

As in the univariate case, there is a spatial cross-correlation coefficient,  $\rho_{uv}(\mathbf{h})$ , which is given by

$$\rho_{uv}(\mathbf{h}) = \frac{C_{uv}(\mathbf{h})}{\sqrt{C_{uu}(\mathbf{0})C_{vv}(\mathbf{0})}}. \quad (10.5)$$

Equation (10.5) is the extension of the ordinary Pearson product-moment correlation coefficient (Chapter 2) into the spatial domain for  $Z_u(\mathbf{x})$  and  $Z_v(\mathbf{x} + \mathbf{h})$ . When  $\mathbf{h} = \mathbf{0}$  it is the Pearson coefficient. Note, however, that  $\rho_{uv}(\mathbf{0}) = 0$ , i.e. no linear correlation in the usual sense, does not mean no correlation at lag distances greater than zero.

Equation (10.4) contains another new feature, namely asymmetry, for in general

$$E[\{Z_u(\mathbf{x}) - \mu_u\}\{Z_v(\mathbf{x} + \mathbf{h}) - \mu_v\}] \neq E[\{Z_v(\mathbf{x}) - \mu_v\}\{Z_u(\mathbf{x} + \mathbf{h}) - \mu_u\}]. \quad (10.6)$$

In words, the cross-covariance between  $u$  and  $v$  in one direction is in general different from that in the opposite direction; the function is asymmetric:

$$C_{uv}(\mathbf{h}) \neq C_{uv}(-\mathbf{h}) \quad \text{or equivalently} \quad C_{uv}(\mathbf{h}) \neq C_{vu}(\mathbf{h}),$$

since

$$C_{uv}(\mathbf{h}) = C_{vu}(-\mathbf{h}).$$

Asymmetry in time is common. The temperature of the air during the day reaches its maximum after the sun has reached its zenith and its minimum occurs after midnight, and the air's mean daily temperature has maxima and minima after the solstices. There is a delay between the elevation of the sun and the temperature of the air. Analogous asymmetry in one dimension in space is easy to envisage. The topsoil might be related asymmetrically to the subsoil on a slope as a result of soil creep, and irrigation by periodic flooding from the same end of a field might redistribute salts differentially down the profile. However, unless the evidence for asymmetry is strong or there is some physical rationale for spatial asymmetry, one might treat differences in estimates, equation (10.10) below, as sampling effects and proceed as though the cross-correlation is symmetric.

The cross-variogram and the cross-covariance function (if it exists) are related, and as in the univariate case the variogram can be obtained from the covariance function by extension of equation (4.5), as follows:

$$\gamma_{uv}(\mathbf{h}) = C_{uv}(\mathbf{0}) - \frac{1}{2}\{C_{uv}(\mathbf{h}) + C_{uv}(-\mathbf{h})\}. \quad (10.7)$$

However, this conversion does not retain all of the information, as we can see by splitting the cross-covariance into an even and an odd term:

$$C_{uv}(\mathbf{h}) = \frac{1}{2}\{C_{uv}(+\mathbf{h}) + C_{uv}(-\mathbf{h})\} + \frac{1}{2}\{C_{uv}(+\mathbf{h}) - C_{uv}(-\mathbf{h})\}. \quad (10.8)$$

The odd term, the second term on the right-hand side of equation (10.8), does not appear in equation (10.7). Unlike the cross-covariance, therefore, the cross-variogram is an even function, i.e. it is symmetric:

$$\gamma_{uv}(\mathbf{h}) = \gamma_{vu}(\mathbf{h}) \quad \text{for all } \mathbf{h}.$$

The cross-variogram cannot express asymmetry, and it should not be used where asymmetry is thought to be significant.

Another way of expressing the spatial relations between the two variables is by the codispersion coefficient. For a lag  $\mathbf{h}$ , this is

$$\nu_{uv}(\mathbf{h}) = \frac{\gamma_{uv}(\mathbf{h})}{\sqrt{\gamma_{uu}(\mathbf{h})\gamma_{vv}(\mathbf{h})}}. \quad (10.9)$$

This coefficient may be thought of as the correlation between the spatial differences of  $u$  and  $v$ . Its merit is that it is symmetric, and so its estimate might be preferred to the cross-correlogram (10.5) for describing the cross-correlation. For second-order stationarity,  $\nu_{uv}(\mathbf{h})$  approaches  $\rho_{uv}(\mathbf{0})$  as  $|\mathbf{h}|$  approaches infinity.

## 10.2 ESTIMATING AND MODELLING THE CROSS-CORRELATION

Providing there are sites where both  $u$  and  $v$  have been measured,  $\gamma_{uv}(\mathbf{h})$  can be estimated in a way similar to that for autosemivariances by

$$\hat{\gamma}_{uv}(\mathbf{h}) = \frac{1}{2m(\mathbf{h})} \sum_{i=1}^{m(\mathbf{h})} \{z_u(\mathbf{x}_i) - z_u(\mathbf{x}_i + \mathbf{h})\} \{z_v(\mathbf{x}_i) - z_v(\mathbf{x}_i + \mathbf{h})\}. \quad (10.10)$$

The result is an experimental cross-variogram for  $u$  and  $v$ .

The cross-variogram can be modelled in the same way as the autovariogram, and the same restricted set of functions is available. To describe the coregionalization there is an added condition. Any linear combination of the variables is itself a regionalized variable, and its variance must be positive or zero: it may not be negative. This is ensured as follows.

We adopt what is called the linear model of coregionalization. In it we assume that each variable  $Z_u(\mathbf{x})$  is a linear sum of orthogonal, i.e. independent, random

variables  $Y_j^k(\mathbf{x})$ , each with mean 0 and variance 1, and in which the superscript  $k$  is simply an index, not a power:

$$Z_u(\mathbf{x}) = \sum_{k=1}^K \sum_{j=1}^2 a_{uj}^k Y_j^k(\mathbf{x}) + \mu_u. \quad (10.11)$$

In this expression

$$\begin{aligned} E[Z_u(\mathbf{x})] &= \mu_u, \\ E[Y_j^k(\mathbf{x})] &= 0 \quad \text{for all } k \text{ and } j, \end{aligned}$$

and

$$\begin{aligned} \frac{1}{2}E[\{Y_j^k(\mathbf{x}) - Y_j^k(\mathbf{x} + \mathbf{h})\}\{Y_{j'}^{k'}(\mathbf{x}) - Y_{j'}^{k'}(\mathbf{x} + \mathbf{h})\}] \\ = \begin{cases} g_k(\mathbf{h}) > 0 & \text{for } k = k' \text{ and } j = j', \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Then the variogram for any pair of variables  $u$  and  $v$  is

$$\gamma_{uv}(\mathbf{h}) = \sum_{k=1}^K \sum_{j=1}^2 a_{uj}^k a_{vj}^k g_k(\mathbf{h}). \quad (10.12)$$

We can replace the products in the second summation by  $b_{uv}^k$  to obtain

$$\gamma_{uv}(\mathbf{h}) = \sum_{k=1}^K b_{uv}^k g_k(\mathbf{h}). \quad (10.13)$$

These  $b_{uv}^k$  are the variances and covariances, i.e. nugget and sill variances, for the independent components if they are bounded. The result might look like the set of spherical-plus-nugget functions in Figure 10.3 below. The intercepts are the three nugget variances,  $b^1$ , and the differences between these and the maxima are the sills of the correlated variances,  $b^2$ . For unbounded variograms the  $b_{uv}^k$  are the nugget variances and gradients. The coefficients  $b_{uv}^k = b_{vu}^k$  for all  $k$ , and for each  $k$  the matrix of coefficients

$$\begin{bmatrix} b_{uu}^k & b_{uv}^k \\ b_{vu}^k & b_{vv}^k \end{bmatrix}$$

must be positive definite. Since the matrix is symmetric, it is sufficient that  $b_{uu}^k \geq 0$  and  $b_{vv}^k \geq 0$  and that its determinant is positive or zero:

$$|b_{uv}^k| = |b_{vu}^k| \leq \sqrt{b_{uu}^k b_{vv}^k}.$$

This is Schwarz's inequality.

For  $V$  coregionalized variables the full matrix of coefficients,  $[b_{ij}]$ , will be of order  $V$ , and its determinant and all its principal minors must be positive or zero.

Schwarz's inequality has the following consequences for each pair of variables:

1. Every basic structure,  $g_k(\mathbf{h})$ , represented in a cross-variogram must also appear in the two auto-variograms, i.e.  $b_{uu}^k \neq 0$  and  $b_{vv}^k \neq 0$  if  $b_{uv}^k \neq 0$ . As a corollary, if a basic structure  $g_k(\mathbf{h})$  is absent from either auto-variogram, then it may not be included in the cross-variogram.
2. The reverse is not so:  $b_{uv}^k$  may be zero when  $b_{uu}^k > 0$ , and structures may be present in the auto-variograms without their appearing in the cross-variogram.

In practice, fitting an optimal model to the coregionalization with these constraints seems formidable. Nevertheless, Goulard and Voltz (1992) have provided an algorithm that converges swiftly. One chooses a suitable combination of basic variogram functions, say nugget plus spherical, and for the autocorrelated function(s) one provides the distance parameters. These can be approximated in advance by fitting models independently to the experimental variograms. Starting with reasonable values for the coefficients,  $b_{uv}^k$ , the computer fits the model and then iterates to minimize the residual sum of squares, checking at each step that the solution is CNSD.

As a check on the validity of a model of coregionalization one can plot the cross experimental variogram for any pair of variables and the model for them plus the limiting values that would hold if correlation were perfect. This last gives what Wackernagel (2003) calls the 'hull of perfect correlation', and for any pair of variables  $u$  and  $v$  it is obtained from the coefficients  $b_{uu}^k$  and  $b_{vv}^k$  by

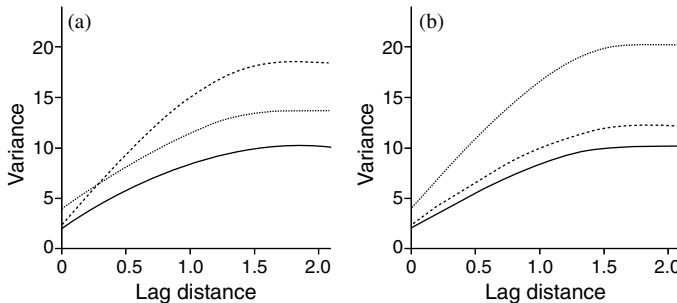
$$\text{hull}[\gamma_{uv}(\mathbf{h})] = \pm \sum_{k=1}^K \sqrt{b_{uu}^k b_{vv}^k} g_k(\mathbf{h}). \quad (10.14)$$

The proximity of the line of the model to the experimental points shows the goodness of fit, as before (Chapter 5). The line must also lie within the hull to be acceptable. But perhaps most revealing is the proximity of the cross-variogram to the hull. If the two are close then the cross-correlation is strong. If, in contrast, the cross-variogram lies far from the bounds then the correlation is weak. This feature may be appreciated by examining Figure 10.3, and we shall discuss it in the first example below.

### 10.2.1 Intrinsic coregionalization

In general, the ratios of the coefficients to one another vary from one basic function to another. In Figure 10.1(a), for example, we have a simple nugget-plus-spherical variogram,

$$\gamma(h) = 2 + 8 \text{sph}(1.7),$$



**Figure 10.1** Spherical variograms with constant range, 1.7: (a) of differing shapes and differing nugget: sill ratios in the general case; (b) of constant nugget: sill ratios in the intrinsic case. See text for further explanation.

shown as a solid line. If we multiply the nugget, 2, by 1.2 and the spherical component by 2 we obtain the dashed line, which has a different shape from the solid line. If our two multipliers are 2 and 1.2 then we obtain the dotted line, which is of a similar shape to the first, but with a different nugget: sill ratio. The range is the same, but the proportions of nugget to sill are all different.

It sometimes happens, however, that all the auto- and cross-variograms are proportional to a single variogram function, so that in terms of equation (10.13) all the coefficients  $b_{uv}^k$  are the same for all  $k$  for each combination of  $u$  and  $v$ , thus:

$$\gamma_{uv}(\mathbf{h}) = \sum_{k=1}^K b_{uv} g_k(\mathbf{h}), \quad (10.15)$$

in which we replace the  $b_{uv}^k$ ,  $k = 1, 2, \dots, K$ , by the single coefficient  $b_{uv}$ . They are simply multiples of one another with the same basic shape. As an example, Figure 10.1(b) shows the basic spherical variogram. If we multiply the two original components in turn by 1.2 and 2, representing  $b_{uu}$ ,  $b_{vv}$  and  $b_{uv}$ , then we obtain the two additional variograms, represented by the dashed and dotted lines, respectively. These are the same apart from the vertical scale.

Where the variables are second-order stationary,

$$\gamma_{uv}(\mathbf{h}) = C_{uv}(\mathbf{0})g(\mathbf{h}), \quad (10.16)$$

with  $g(\mathbf{h}) \rightarrow 1$  as  $|\mathbf{h}| \rightarrow \infty$ . Spatial cross-correlation of this kind is said to be *intrinsic*. The term is somewhat unfortunate in that this usage of ‘intrinsic’ differs from that in the ‘intrinsic hypothesis’.

Where spatial correlation is intrinsic the codispersion coefficient  $\nu_{uv}(\mathbf{h})$  remains constant for all  $\mathbf{h}$ , i.e.

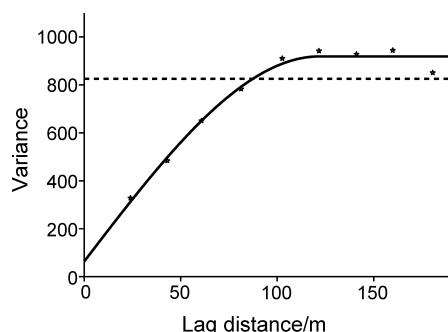
$$\nu_{uv}(\mathbf{h}) = \frac{b_{uv}}{\sqrt{b_{uu}b_{vv}}} = \frac{C_{uv}(\mathbf{0})}{\sqrt{C_{uu}(\mathbf{0})C_{vv}(\mathbf{0})}} = \nu_{uv}(\mathbf{0}).$$

### 10.3 EXAMPLE: CEDAR FARM

We illustrate the procedure and some of the features of coregionalization using the survey data of a field on CEDAR Farm in southeast England. They derive from an original study of precision farming by Dr Z. L. Frogbrook, who kindly provided them and to whom we are grateful. The field covers approximately 6.4 ha of fairly flat land on clay and is cultivated to produce cereals. Its topsoil (0–15 cm) was sampled at 160 places on 5 m × 2 m supports at 20 m intervals on a square grid. The yield of barley was measured on the same supports in 1998. The principal plant nutrients, exchangeable potassium (K) and available phosphorus (P), were measured. The data are summarized in Table 10.1. Potassium and yield are somewhat skewed, but not so seriously as to warrant transformation. The three variables are correlated, though not strongly, as Table 10.2 shows.

By applying equation (10.10) and treating the variation as isotropic, we obtain the experimental auto- and cross-variograms. The experimental variograms have simple forms. Figure 10.2 shows an example; it is the autovariogram of K with a spherical model fitted to it by weighted least squares, as described in Chapter 5. The model's coefficients are listed in Table 10.3. The other experimental variograms appear in Figure 10.3 as the point symbols. Four of them, namely the autovariogram of yield (Figure 10.3(e)), and the three cross-variograms, Figures 10.3(b), 10.3(d) and 10.3(f), are evidently bounded, and again the spherical model fitted them well. The coefficients of fitting the model independently are listed in Table 10.3. The autovariogram of P does not reach a bound within the field.

The five bounded variograms have approximately the same range, and so we can reasonably fit the linear model of coregionalization with two basic components,  $g_1(0)$ , i.e. nugget, and  $g_2(|\mathbf{h}|)$ . We set the range of  $g_2(|\mathbf{h}|)$  to 144 m, the average of the five bounded variogram models. The resulting coefficients  $b_{uv}^k$  are listed in Table 10.4, and the solid lines in Figure 10.3 are those of the linear model of coregionalization. We can see by comparing Figure 10.2 with Figure 10.3(a) for K that the model of coregionalization fits



**Figure 10.2** Variogram of potassium at CEDAR Farm with experimental values plotted as point symbols and the spherical model shown as the solid line.

**Table 10.3** Coefficients of spherical model fitted independently to auto- and cross-variograms of K, P and barley yield at CEDAR Farm. The parameters  $c_0$ ,  $c$  and  $a$  are the nugget, sill of the correlated variance and the range, respectively.

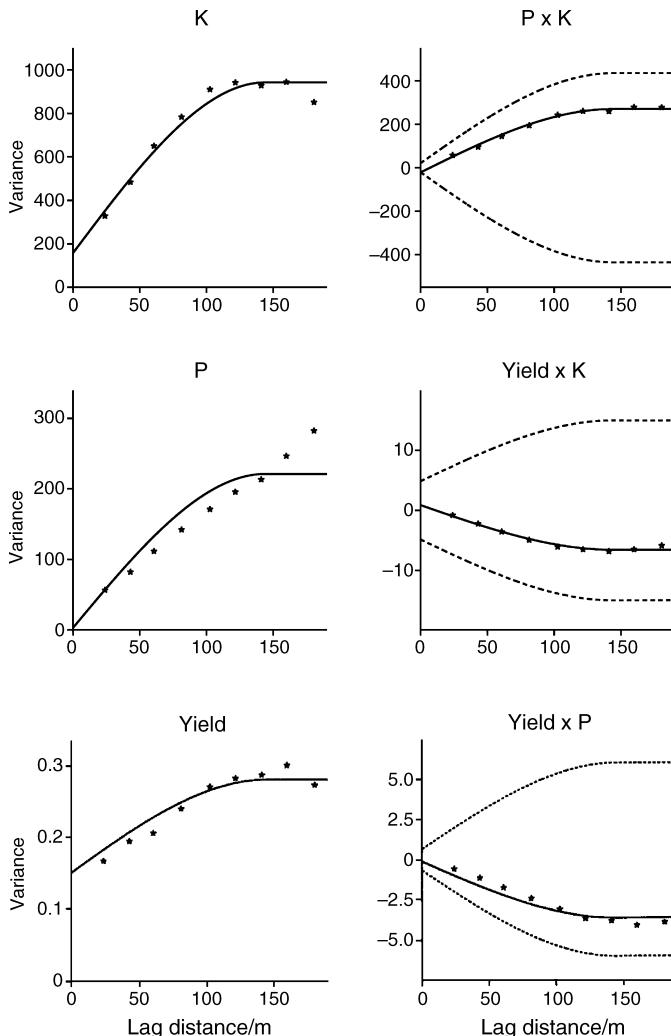
	$c_0$	$c$	$a/\text{m}$
K	63.7	855.4	121.9
P	25.2	—	—
Yield	0.121	0.167	151.0
K × P	-20.48	294.6	150.2
K × yield	1.51	-7.94	130.1
P × yield	0.550	-4.548	168.5

somewhat less well than the model fitted to K alone, and we can imagine that we could fit the other variograms better if we treated them individually. Nevertheless, the fit is generally good, and even for P the model fits fairly well over most of the working lag distance. The dashed lines in the graphs in the right-hand column of Figure 10.3 are the hulls of perfect correlation for the cross-variograms. In all three the model is some way from the hulls, showing that the spatial cross-correlations are at best moderate and like the ordinary simple correlations (Table 10.2).

Finally, we comment on the practical meaning of the coregionalization. The correlation between K and P is positive, and all the cross-semivariances are positive. The two variables characterize the nutrient status of the soil. The correlations between yield and K and P are all negative, and this might come as a surprise. Farmers and their advisers have been used to thinking that more K and P in the soil would result in greater yield. Now that yield can be recorded automatically at harvest, they are discovering that large yields deplete the soil locally and that they should fertilize differentially to maintain sufficient K and P over the whole of each field. Large concentrations of K and P indicate small off-take of these nutrients by the crop and thus smaller yields. Webster, in Lake *et al.* (1997, pp. 74–77), has shown another example from the same district.

**Table 10.4** Coefficients  $b_{uv}^k$  of the linear coregionalization with nugget plus spherical for K, P and yield at CEDAR Farm.

	K	P	Yield
Nugget, $b_{uv}^1$			
K	157.28		
P	-21.34	2.8960	
Yield	0.8683	-0.11305	0.15027
Correlated variance, $b_{uv}^2$			
K	785.15		
P	291.39	218.1996	
Yield	-7.4239	-3.5149	0.13027



**Figure 10.3** Autovariograms (left) and cross-variograms (right) of K, P and barley yield at CEDAR Farm. The experimental values are plotted as points and the solid lines are of the model of coregionalization. The dashed lines in the right-hand graphs are the hulls of perfect correlation.

## 10.4 COKRIGING

Having learned how to model the coregionalization, we can use our knowledge of the spatial relations between two or more variables to predict their values by cokriging. Typically the aim is to estimate just one variable, which we may regard as the principal or target variable, at a point  $\mathbf{x}_0$  or in a block  $B$ , from

data on it plus those of one or more other variables, which we regard as subsidiary variables. Cokriging is simply an extension of autokriging in that it takes into account additional correlated information in the subsidiary variables. It appears more complex because the additional variables increase the notation.

Let there be  $V$  variables,  $l = 1, 2, \dots, V$ , and let us denote the one we wish to predict as  $u$ ; this will usually have been less densely sampled than the others. In ordinary cokriging we form the linear sum

$$\hat{Z}_u(B) = \sum_{l=1}^V \sum_{i=1}^{n_l} \lambda_{il} z_l(\mathbf{x}_i), \quad (10.17)$$

where the subscript  $i$  refers to the sites, of which there are  $n_l$  where the variable  $l$  has been measured. The  $\lambda_{il}$  are weights, satisfying

$$\sum_{i=1}^{n_l} \lambda_{il} = \begin{cases} 1 & l = u, \\ 0 & l \neq u. \end{cases} \quad (10.18)$$

These are the non-bias conditions, and subject to them the estimation variance of  $\hat{Z}_u(B)$  for a block,  $B$ , is minimized by solution of the kriging system, which, in full, is

$$\begin{aligned} \sum_{l=1}^V \sum_{i=1}^{n_l} \lambda_{il} \gamma_{lv}(\mathbf{x}_i, \mathbf{x}_j) + \psi_v &= \bar{\gamma}_{uv}(\mathbf{x}_j, B), \\ \sum_{i=1}^{n_l} \lambda_{il} &= \begin{cases} 1 & l = u, \\ 0 & l \neq u. \end{cases} \end{aligned} \quad (10.19)$$

for all  $v = 1, 2, \dots, V$  and all  $j = 1, 2, \dots, n_v$ . The quantity  $\gamma_{lv}(\mathbf{x}_i, \mathbf{x}_j)$  is the (cross-)semivariance between variables  $l$  and  $v$  at sites  $i$  and  $j$ , separated by the vector  $\mathbf{x}_i - \mathbf{x}_j$ ;  $\bar{\gamma}_{uv}(\mathbf{x}_j, B)$  is the average (cross-)semivariance between a site  $j$  and the block  $B$ ; and  $\psi_v$  is the Lagrange multiplier for the  $v$ th variable. We print ‘cross’ in parentheses because if  $l = v$  or  $u = v$  the semivariances are the autosemivariances. This set of equations is the extension of the autokriging system, equations (8.11).

Solving equations (10.19) gives the weights,  $\lambda$ , which are inserted into equation (10.17) to estimate  $Z_u(B)$ , and the estimation variance, the cokriging variance, is obtained from

$$\sigma_u^2(B) = \sum_{l=1}^V \sum_{j=1}^{n_l} \lambda_{jl} \bar{\gamma}_{ul}(\mathbf{x}_j, B) + \psi_u - \bar{\gamma}_{uu}(B, B), \quad (10.20)$$

where  $\bar{\gamma}_{uu}(B, B)$  is the integral of  $\gamma_{uu}(\mathbf{h})$  over  $B$ , i.e. the within-block variance of  $u$ .

The equations can be represented in matrix form. For simplicity consider two variables,  $u$  and  $v$ , only. The matrices are easily extended to more. Let  $\Gamma_{uv}$  denote a matrix of semivariances (including cross-semivariances where  $u \neq v$ ) between sampling points in a neighbourhood. Let there be  $n_u$  places at which variable  $u$  was measured and  $n_v$  where  $v$  was measured. The order of the matrix is  $n_u \times n_v$ :

$$\Gamma_{uv} = \begin{bmatrix} \gamma_{uv}(\mathbf{x}_1, \mathbf{x}_1) & \gamma_{uv}(\mathbf{x}_1, \mathbf{x}_2) & \cdots & \gamma_{uv}(\mathbf{x}_1, \mathbf{x}_{n_v}) \\ \gamma_{uv}(\mathbf{x}_2, \mathbf{x}_1) & \gamma_{uv}(\mathbf{x}_2, \mathbf{x}_2) & \cdots & \gamma_{uv}(\mathbf{x}_2, \mathbf{x}_{n_v}) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{uv}(\mathbf{x}_{n_u}, \mathbf{x}_1) & \gamma_{uv}(\mathbf{x}_{n_u}, \mathbf{x}_2) & \cdots & \gamma_{uv}(\mathbf{x}_{n_u}, \mathbf{x}_{n_v}) \end{bmatrix}.$$

We denote by  $\mathbf{b}_{uu}$  and by  $\mathbf{b}_{uv}$  the vectors of autosemivariances for variable  $u$  and cross-semivariances:

$$\mathbf{b}_{uu} = \begin{bmatrix} \bar{\gamma}_{uu}(\mathbf{x}_1, B) \\ \bar{\gamma}_{uu}(\mathbf{x}_2, B) \\ \vdots \\ \bar{\gamma}_{uu}(\mathbf{x}_{n_u}, B) \end{bmatrix}, \quad \mathbf{b}_{uv} = \begin{bmatrix} \bar{\gamma}_{uv}(\mathbf{x}_1, B) \\ \bar{\gamma}_{uv}(\mathbf{x}_2, B) \\ \vdots \\ \bar{\gamma}_{uv}(\mathbf{x}_{n_v}, B) \end{bmatrix}.$$

The matrix equation is then

$$\begin{bmatrix} \Gamma_{uu} & \Gamma_{uv} & \begin{matrix} 10 \\ 10 \\ \vdots \\ 10 \\ 01 \\ 01 \end{matrix} \\ \Gamma_{vu} & \Gamma_{vv} & \begin{matrix} \vdots \\ 01 \\ 11 \dots 1 & 00 \dots 0 & 00 \\ 00 \dots 0 & 11 \dots 1 & 00 \end{matrix} \end{bmatrix} \cdot \begin{bmatrix} \lambda_{1u} \\ \lambda_{2u} \\ \vdots \\ \lambda_{n_u u} \\ \lambda_{1v} \\ \lambda_{2v} \\ \vdots \\ \lambda_{n_v v} \\ \psi_u \\ \psi_v \end{bmatrix} = \begin{bmatrix} \mathbf{b}_{uu} \\ \mathbf{b}_{uv} \\ 1 \\ 0 \end{bmatrix}.$$

If we denote the augmented matrix of  $\Gamma$ 's by  $\mathbf{G}$ , the vector of weights and Lagrange multipliers by  $\boldsymbol{\lambda}$ , and the right-hand side vector by  $\mathbf{b}$ , then we can write the solution of the equation succinctly as

$$\boldsymbol{\lambda} = \mathbf{G}^{-1} \mathbf{b}. \quad (10.21)$$

The cokriging (prediction) variance is given by

$$\hat{\sigma}_u^2(B) = \mathbf{b}^T \boldsymbol{\lambda} - \bar{\gamma}_{uu}(B, B). \quad (10.22)$$

As in autokriging, the block  $B$  may be of any reasonable size and shape, and it may be reduced to a point,  $\mathbf{x}_0$ , having the same dimensions as the support on which the data were obtained. In these circumstances the averages  $\bar{\gamma}_{uv}(\mathbf{x}_j, B)$  become  $\gamma_{uv}(\mathbf{x}_j, \mathbf{x}_0)$ , and  $\bar{\gamma}_{uu}(B, B)$  is zero and hence disappears, thus:

$$\mathbf{b}_{uu} = \begin{bmatrix} \gamma_{uu}(\mathbf{x}_1, \mathbf{x}_0) \\ \gamma_{uu}(\mathbf{x}_2, \mathbf{x}_0) \\ \vdots \\ \gamma_{uu}(\mathbf{x}_{n_u}, \mathbf{x}_0) \end{bmatrix}, \quad \mathbf{b}_{uv} = \begin{bmatrix} \gamma_{uv}(\mathbf{x}_1, \mathbf{x}_0) \\ \gamma_{uv}(\mathbf{x}_2, \mathbf{x}_0) \\ \vdots \\ \gamma_{uv}(\mathbf{x}_{n_v}, \mathbf{x}_0) \end{bmatrix},$$

and

$$\hat{\sigma}_u^2(\mathbf{x}_0) = \mathbf{b}^T \boldsymbol{\lambda}. \quad (10.23)$$

Myers (1982) presents the equations for cokriging somewhat differently and comprehensively.

#### 10.4.1 Is cokriging worth the trouble?

Cokriging is more complex than autokriging, and the practitioner can and should ask whether the extra complexity improves the results: are the estimates better in any sense?

We distinguish two situations. First consider the *undersampled* case. By undersampling we mean that the variable to be estimated, the primary variable  $u$  in the kriging equations, is sampled less intensely than the others, usually at a subset of the sampling points. In this case the spatial correlation in the other variables and their relation to  $u$  add information that is lacking in that of  $u$  alone. As a result cokriging increases the precision, i.e. it reduces the estimation variance. By how much depends on the degree of undersampling. In general, the smaller the sampling intensity of  $u$  in relation to that of the other(s) the greater is the benefit of cokriging. We illustrate this below.

In the *fully sampled* case, all variables are recorded at all sampling points. Here the principal advantage is *coherence*. Kriging is coherent when the kriged estimate of the sum of a set of variables, say  $\hat{S}$ , equals the sum of their individually kriged estimates:

$$\hat{S}(B) = \sum_{i=1}^{n_l} \lambda_i \sum_{l=1}^V z_l(\mathbf{x}_i) = \sum_{k=1}^V \sum_{l=1}^V \sum_{i=1}^{n_l} \lambda_{il} z_l(\mathbf{x}_i). \quad (10.24)$$

Cokriging ensures coherence. Otherwise the equality depends on the nature of the coregionalization.

As an example, consider estimating the thickness of a soil horizon. A field surveyor might have recorded the depths from the surface to the top and bottom of the horizon in question at the sampling points. At each point the thickness is simply the difference between the two. We could krig the thickness to estimate it at unrecorded positions. Alternatively, we could krig the depths to the top and the bottom of the horizon and compute their differences. If we were to do that for each variable independently then we should find that in general the differences between the kriged estimates were not the same as the kriged differences, i.e. the kriged thickness. If, however, we were to cokrige the depths to top and bottom then the differences between the kriged estimates would equal the kriged thickness.

Where the variables are intrinsically coregionalized, i.e. all the variograms are related linearly to a single basic model, autokriging of any variable gives the same result as cokriging. The spatial information for the one variable is all there is in the data, there is no more in that of the others, and there is no merit in the more complex procedure.

Where the variogram of the primary variable,  $u$ , is linearly related to the cross-variogram(s), autokriging  $u$  again gives results identical with cokriging. The cross-correlation adds nothing.

In other situations the results are in general different. However, with full sampling the differences are likely to be small, and experience suggests that the differences are usually so small that they can be ignored unless coherence is essential.

#### **10.4.2 Example of benefits of cokriging**

We can see something of the benefits of cokriging by following the same logic as in Chapter 8, where we calculated the kriging variances for various sample spacings from a model variogram. We placed sampling points on regular grids, and we computed the variances at the centres of the grid cells where for punctual kriging they were greatest. For block kriging the maxima can occur when the target block is centred on a grid node, and so we calculated the variance in those positions also. We displayed the results as curves of maximum kriging variance against grid spacing (Figure 8.23).

For cokriging we have two or more variables. We can choose the primary grid for the undersampled target variable,  $u$ , in the same way as for autokriging. We can then superimpose denser grids for the subsidiary variables. With the points at the nodes of these two grids we can set up and solve the cokriging equations. The maximum kriging variances are no longer necessarily at the centres of grid cells or centred over grid nodes, and so their positions must be found by searching. As the density of the subsidiary grid is increased so the

kriging variance of the target variable should decrease, and doing the calculations as above and plotting the results will show just how beneficial cokriging is at various scales. McBratney and Webster (1983) describe the procedure in detail.

We illustrate the approach with the coregionalization at CEDAR Farm. As we mentioned above, measuring nutrients in the soil is expensive in relation to the benefits to be gained from knowing the concentrations. Arable farmers in Britain can afford to sample their soil at a density of approximately one per hectare; not more. Yet they would like to know the nutrient concentration at a much finer resolution to vary their application of fertilizers. Automatic recording of yield as the grain passes through the harvester is now quite feasible and produces abundant dense data. So if the relation between yield and nutrient status is sufficiently strong the farmer might use the dense data on yield to improve his prediction of nutrient concentration. So let us see to what extent we might use this approach in the situation at this farm.

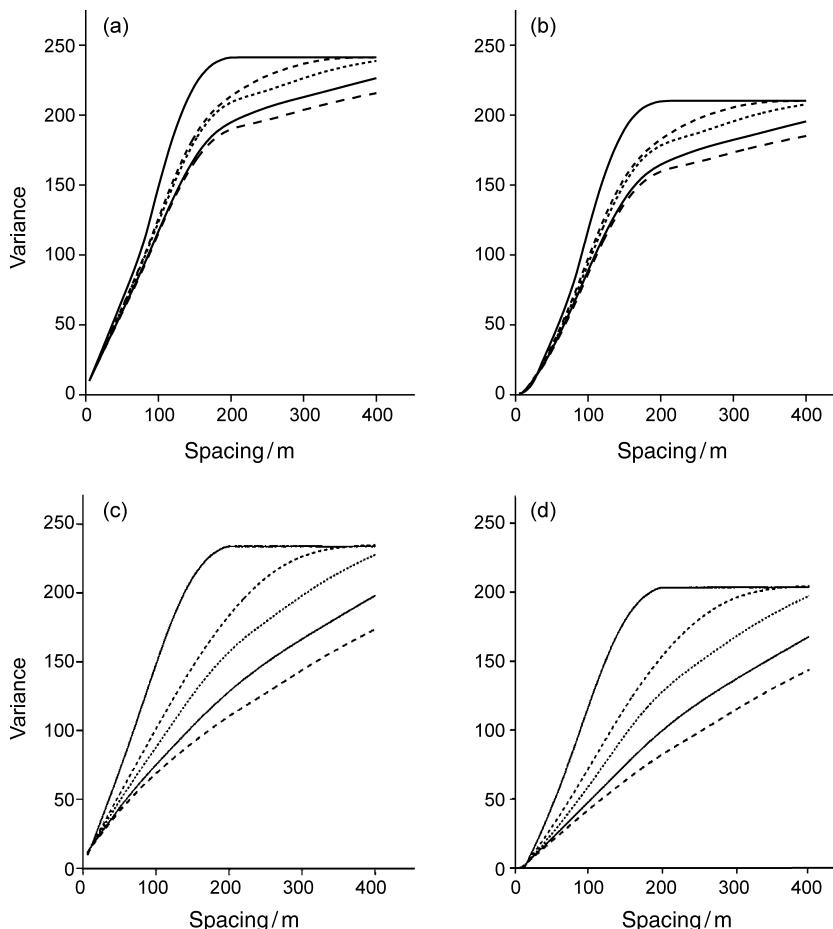
We suppose that available phosphorus (P) is the target variable and we shall use yield as the subsidiary variable. We take the parameters for the cokriging from the coregionalization model (Table 10.4). We have chosen intervals for the primary grid from near 0 to 400 m. We have imposed subsidiary grids with intervals of  $1/2$ ,  $1/3$ ,  $1/4$  and  $1/5$  of the primary grid, giving sampling ratios of 4, 9, 16 and 25. The smallest intervals are impracticable because the cutter bar of a modern harvester is typically 4 m wide on British farms, but we include them to complete the picture and for theoretical interest. We have solved the kriging systems for punctual kriging and also computed the kriging variances for blocks  $24\text{ m} \times 24\text{ m}$ . We choose this size because the standard farm machinery spreads fertilizer in bands this wide.

The results are plotted as graphs of maximum kriging variance against sample spacing in Figure 10.4. In each graph the uppermost solid curve is for autokriging and the ones beneath it are in order from top to bottom for cokriging with the subsidiary grid interval  $1/2$ ,  $1/3$ ,  $1/4$  and  $1/5$  of that of the primary grid.

The upper pair of graphs, Figure 10.4(a) for punctual kriging and Figure 10.4(b) for block kriging, show that with the actual model of coregionalization for this field the reductions in kriging variance from adding yield in the kriging equations to predict P are modest. The reason is that the correlation between the two is itself modest. If the cost of installing a recorder to measure yield and handling the data is much less than that of analysing the soil for P then it might be worth the trouble, but in any event the farmer cannot expect large gains in precision or to save much in soil sampling and analysis.

In passing, we note that the block-kriging variance is less than the variance of punctual kriging with the same sampling configuration by an amount approximately equal to the within-block variance of P.

The outlook might be rosier with stronger association between target and subsidiary variables, and to illustrate this we have repeated the exercise using



**Figure 10.4** Graphs of maximum kriging variance of phosphorus (P) against sample spacings on a primary grid with denser observations of yield on subsidiary grids. (a) Punctual kriging of P using the fitted model of coregionalization (Table 10.4); (b) kriging of  $24\text{ m} \times 24\text{ m}$  blocks with the same model; (c) punctual kriging of P using a model of perfect correlation (Figure 10.3); (d) block kriging ( $24\text{ m} \times 24\text{ m}$  blocks) with the perfect model. In each graph the uppermost curve is that for autokriging and the ones below are in order for spacings on the subsidiary grid of  $1/2$ ,  $1/3$ ,  $1/4$  and  $1/5$  of those on the primary grid.

a model of perfect correlation, the lower part of the hull in Figure 10.3(f). Figure 10.4(c)–(d) displays the results. Now large differences emerge as the density of the subsidiary grid increases. With a sampling ratio of 25 for the subsidiary variable we can reduce the maximum kriging variance to one-third of that from autokriging. The farmer could increase the ratio further and gain even bigger benefits in such circumstances.

## 10.5 PRINCIPAL COMPONENTS OF COREGIONALIZATION MATRICES

The full coregionalization model of  $V$  variables has  $V \times (V + 1)/2$  variograms, and if each has  $K$  basic functions then there are  $K \times V$  coefficients. We have already seen the set of auto- and cross-variograms for CEDAR Farm, with three variables and  $K = 2$  basic functions (Table 10.4 and Figure 10.3). The correlations are moderate, and though we have used the relation between  $P$  and yield to illustrate cokriging there is not a great deal of interest in exploring them further. Where the correlations are stronger, however, it may be worth analysing the coregionalization matrices to see how the correlation varies with scale. To illustrate this we turn to an original investigation of heavy metals in the soil in the Swiss Jura by Atteia *et al.* (1994) and Webster *et al.* (1994).

Some 14.5 km<sup>2</sup> near La Chaux-de-Fonds in the Jura were surveyed to determine the concentrations of seven potentially toxic metals, namely cadmium (Cd), cobalt (Co), chromium (Cr), copper (Cu), nickel (Ni), lead (Pb) and zinc (Zn), in the topsoil. Soil was removed with a cylindrical corer of 5 cm diameter to a depth of 25 cm, which therefore defined the support of the sample. Cores were taken at 214 intersections of a 250 m grid plus an additional 152 points arranged in nests around 38 of the grid nodes. The ‘total’ metal was extracted from each sample by strong acid and measured. Atteia *et al.* (1994) describe the sampling and analytical procedure in detail.

Table 10.5 summarizes the data from the 366 sites. It shows immediately that the frequency distributions of four of the metals—Cd, Cu, Pb and Zn—were

**Table 10.5** Summary statistics for heavy metals at La Chaux-de-Fonds on the original scales (mg kg<sup>-1</sup>) and with Cd, Cu, Pb and Zn transformed to their common logarithms.

	Cd	Co	Cr	Cu	Ni	Pb	Zn
Minimum	0.14	1.55	3.32	3.55	1.98	18.7	25.0
Maximum	5.13	20.6	70.0	242	53.2	382.0	338.0
Mean	1.31	9.45	35.2	24.6	20.2	57.0	78.5
Median	1.11	9.82	34.8	17.4	20.8	46.8	74.0
Variance	0.7598	12.56	118.3	638.8	67.91	1527.9	1147.0
St. dev.	0.87	3.54	10.9	25.3	8.24	39.1	33.9
Skewness	1.43	-0.20	0.34	3.87	0.17	4.22	2.74
Kurtosis	2.44	-0.60	0.33	18.1	0.31	23.6	12.9
Logarithms							
Mean	0.022			1.26		1.70	1.86
Variance	0.0868			0.1046		0.0414	0.0306
St. dev.	0.29			0.34		0.21	0.18
Skewness	-0.30			0.51		1.10	0.07
Kurtosis	-0.30			0.66		2.64	0.96

**Table 10.6** Correlation matrix for seven heavy metals in the soil at La Chaux-de-Fonds in the Swiss Jura with 364 degrees of freedom.

Log cadmium	1						
Cobalt	0.393	1					
Chromium	0.653	0.473	1				
Log copper	0.243	0.271	0.300	1			
Nickel	0.634	0.727	0.717	0.326	1		
Log lead	0.346	0.212	0.335	0.795	0.372	1	
Log zinc	0.677	0.523	0.669	0.700	0.687	0.685	1
	Cd	Co	Cr	Cu	Ni	Pb	Zn

strongly skewed, and so the data for these were transformed to their common logarithms to stabilize their variances.

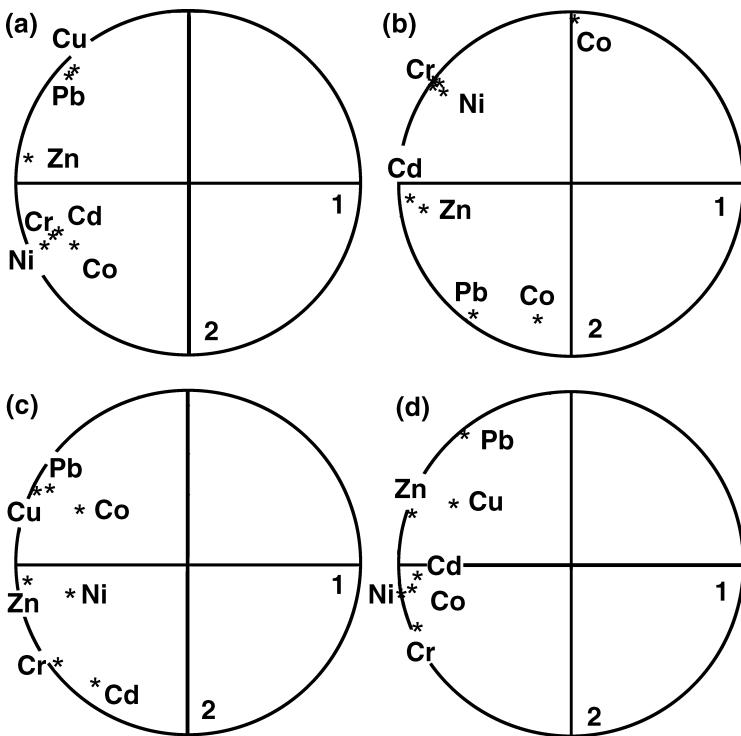
Further, the correlation matrix, Table 10.6, shows some fairly strong correlations—between Co and Ni, and between Cu and Pb, for example. The general strength of correlation in the data may be judged by converting the matrix to principal components. The results are summarized in Table 10.7. The two leading principal components account for 78% of the variance in the matrix, given in the right-hand column. The correlation may then be displayed in the plane of the first two axes by computing the correlation coefficients,  $c_{ij}$ , between the principal component scores and the original variables, as follows:

$$c_{ij} = a_{ij} \sqrt{v_j / \sigma_i^2}, \quad (10.25)$$

where  $a_{ij}$  is the  $i$ th element of the  $j$ th eigenvector,  $v_j$  is the  $j$ th eigenvalue, and  $\sigma_i^2$  is the variance of the  $i$ th original variable. We then plot these coefficients in circles of unit radius in the planes of the leading components. Figure 10.5(a) shows the result. The first axis represents the magnitude of the concentrations: large concentrations of one metal are associated with large concentrations of the others. Axis 2 spreads the metals out, and it is evident that Cu and Pb are

**Table 10.7** Eigenvalues of correlation matrix for seven heavy metals in the soil at La Chaux-de-Fonds.

Order	Eigenvalue	Percentage	Accumulated percentage
1	4.123	58.90	58.90
2	1.342	19.17	78.07
3	0.681	9.72	87.79
4	0.344	4.91	92.70
5	0.229	3.27	95.97
6	0.162	2.31	98.28
7	0.120	1.72	100.00

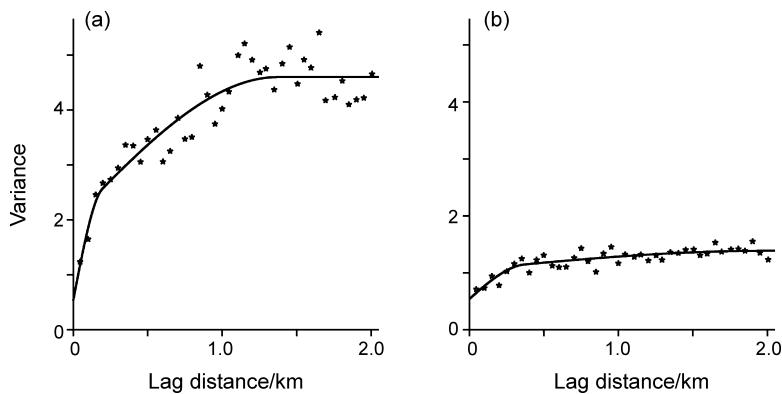


**Figure 10.5** Projections of the correlations between the original (standardized) variables and the principal component scores into unit circles in the plain of the first two principal components for the heavy metals in the Swiss Jura: (a) the ordinary product-moment correlation matrix; (b) the nugget matrix; (c) the short-range matrix; (d) the long-range matrix.

closely associated, as are the transition metals Co, Ni and Cr. Cadmium seems to be related to these metals, while Zn lies about half-way between the two groups. The occurrence of all the points close to the circumference of the circle is one more reflection of there being only little more information in the other dimensions.

The principal component analysis has another advantage: the leading components should concentrate the information on the spatial structure. Figure 10.6 shows the variograms of the first two components. The plotted points are the experimental values and the solid lines are the fitted models. We fitted double spherical functions with nugget variances (see Chapter 5) to both, and the coefficients for the models are listed in Table 10.8. The nested structure of the first component is clear, with two distinct ranges,  $a_1 \approx 0.2$  km and  $a_2 \approx 1.3$  km.

The first principal component contains such a large proportion of the total variation that we have taken its spatial structure and its distance parameters,



**Figure 10.6** Variograms: (a) the first; (b) the second principal components of the heavy metals. The points show the experimental values and the solid lines are of the independently fitted double spherical models, the coefficients of which are listed in Table 10.8.

the two ranges, as typical of the full set of data. We set these ranges as constants, and then found the sills of the seven autovariograms and the 21 cross-variograms iteratively by the Goulard and Voltz algorithm. These sills are listed in Table 10.9. Figure 10.7 shows the autovariograms with the model fitted to them, and readers can see the full set including the cross-variograms in Webster *et al.* (1994).

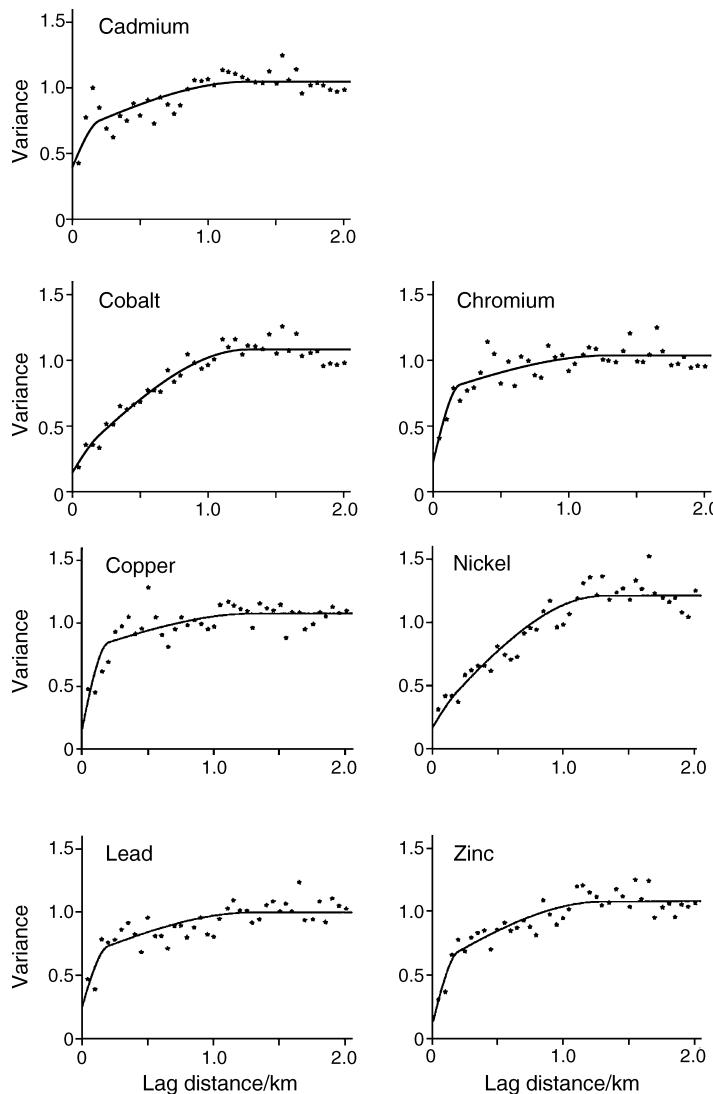
The differences between the metals are striking. The spatial correlation of Cd, Pb and Cu is dominantly of short range, whereas that of Co and Ni is of long

**Table 10.8** Coefficients of double spherical model fitted to principal components for La Chaux-de-Fonds.

	$c_0$	$c_1$	$c_2$	$a_1$	$a_2$
Component 1	0.547	1.473	2.580	0.198	1.376
Component 2	0.543	0.513	0.330	0.377	1.976
Component 3	0.174	0.287	0.239	0.161	1.297

**Table 10.9** Coefficients,  $b^k$ , of double spherical model of coregionalization for standardized autovariograms of the seven heavy metals in the soil at La Chaux-de-Fonds. All of the scales have been standardized to variance equal to 1 for comparison.

	Cd	Co	Cr	Cu	Ni	Pb	Zn
Nugget, $b^1$	0.396	0.146	0.225	0.160	0.168	0.244	0.113
Short range, $b^2$	0.267	0.092	0.524	0.621	0.073	0.408	0.454
Long range, $b^3$	0.384	0.844	0.288	0.291	0.961	0.344	0.508



**Figure 10.7** Experimental autovariograms of the seven heavy metals in the soil of the Swiss Jura shown by point symbols and the fitted model of coregionalization shown by solid lines. All of the scales have been standardized to variance equal to 1 for comparison. The coefficients are listed in Table 10.9.

range. Somewhat surprisingly, the variogram of Cr is dominated by the short-range component. Zinc has an intermediate structure.

We can take the analysis one stage further by finding the principal components of the coregionalization matrices, as follows. The coefficients,  $b_{uv}^k$ , for all  $u = 1, 2, \dots, V$  and all  $v = 1, 2, \dots, V$ , constitute a  $V \times V$  variance-covariance

**Table 10.10** Nugget and structural correlation coefficients in lower triangles for seven heavy metals in the soil at La Chaux-de-Fonds.

Nugget variances						
log Cadmium		1				
Cobalt	-0.051		1			
Chromium	0.399	0.217		1		
log Copper	0.118	-0.253	-0.137		1	
Nickel	0.347	0.197	0.446	-0.040		1
log Lead	0.322	-0.276	0.037	0.249	0.021	
log Zinc	0.521	-0.093	0.249	0.094	0.237	0.233
						1
Short-range components						
log Cadmium		1				
Cobalt	0.167		1			
Chromium	0.358	0.050		1		
log Copper	0.126	0.174	0.292		1	
Nickel	0.097	-0.021	0.579	0.422		1
log Lead	0.028	0.210	0.326	0.637	0.153	
log Zinc	0.316	0.193	0.502	0.609	0.152	0.421
						1
Long-range components						
log Cadmium		1				
Cobalt	0.427		1			
Chromium	0.459	0.786		1		
log Copper	0.237	0.578	0.292		1	
Nickel	0.578	0.766	0.499	0.302		1
log Lead	0.363	0.412	0.127	0.327	0.481	
log Zinc	0.485	0.719	0.398	0.370	0.837	0.461
						1
		Cd	Co	Cr	Cu	Ni
		Pb	Zn			

matrix,  $\mathbf{B}^k$ , and principal components of these can be found in exactly the same way as those of any other variance–covariance or correlation matrix. The elements of the matrix, which are listed in Table 10.10, are converted to correlation coefficients by dividing by the square roots of the variances on the diagonal so that all variables have equal weight. The eigenvalues,  $v$ , and eigenvectors,  $a$ , are then extracted. To explore the relations among the variables we computed the correlations between the original variables and the principal components at each scale using equation (10.25), replacing  $\sigma_i^2$  by the relevant  $b_{uu}^k$ .

Figure 10.5(b)–(d) shows the results for the nugget, short-range and long-range components, respectively. The first two eigenvalues account for more than 85% of the variance in all three matrices (Table 10.11), and in consequence all the points plot near the circumferences. The contributions of the nugget variance appear as a scatter of points to the left of centre in Figure 10.5(b). In Figure 10.5(c), representing the short-range components, Cu and Pb are close neighbours—evidently they are closely correlated at this

**Table 10.11** Eigenvalues of structural variance–covariance matrices for La Chaux-de-Fonds.

Order	Nugget		Short range		Long range	
	Eigenvalue	Accumulated percentage	Eigenvalue	Accumulated percentage	Eigenvalue	Accumulated percentage
1	0.7498	52.20	1.5965	65.44	2.8339	78.27
2	0.4792	85.61	0.4979	85.85	0.3608	88.23
3	0.0974	92.39	0.1930	93.76	0.2546	95.26
4	0.0685	97.16	0.1018	97.93	0.1215	98.62
5	0.0406	99.99	0.0504	100.00	0.0497	99.99
6	0.0001	100.00	0.0000	100.00	0.0003	100.00
7	0.0000	100.00	0.0000	100.00	0.0000	100.00

scale—whereas the uncorrelated Co and Ni make little contribution at this scale and so lie closer to the centre. The reverse is the case at the long range (Figure 10.5(d)), in which the strong correlation between Co and Ni is apparent.

Webster *et al.* (1994) thought that the two distinct patterns of variation might result from two distinct sources of the metals: the lithophile metals Co and Ni deriving from the rocks, and Cu, Pb, Zn and perhaps Cd having been added in manure, fertilizer, sewage sludge or urban waste. They used the results to explore these possibilities.

## 10.6 PSEUDO-CROSS-VARIOGRAM

It will be evident from the computing formula, equation (10.10), that cross-semivariances can be calculated only from points where both variables  $u$  and  $v$  have been measured. In the examples from Broom's Barn Farm and the Jura there are few missing data, and the restriction is of little consequence. There are other situations, however, where it is difficult or even impossible to measure the two variables at the same place, as when sampling is destructive. This happens in soil monitoring. Soil material may be taken away initially for analysis and is not there subsequently, so on later occasions the soil must be measured at different places (see Papritz *et al.*, 1993; Papritz and Webster, 1995a, 1995b). Nevertheless, one may have many observations from which to assess spatial relations and one would like to use them.

Clark *et al.* (1989) recognized the desire, and they proposed a ‘pseudo-cross-variogram’. They introduced it with the following definition:

$$\gamma_{uv}^C(\mathbf{h}) = \frac{1}{2}\mathbb{E}[\{Z_u(\mathbf{x}) - Z_v(\mathbf{x} + \mathbf{h})\}^2]. \quad (10.26)$$

This is unsatisfactory because, unless  $\mu_u = \mu_v$ ,  $\gamma_{uv}^C(\mathbf{h})$  is not equal to half of the variance of the difference. Myers (1991) recognized this shortcoming and redefined the pseudo-cross-variogram as the variance:

$$\gamma_{uv}^P(\mathbf{h}) = \frac{1}{2} \operatorname{var} [Z_u(\mathbf{x}) - Z_v(\mathbf{x} + \mathbf{h})]. \quad (10.27)$$

If the means of  $u$  and  $v$  are equal then  $\gamma_{uv}^C(\mathbf{h}) = \gamma_{uv}^P(\mathbf{h})$ ; otherwise the function defined by Clark *et al.* equals  $\gamma_{uv}^P(\mathbf{h}) + (\mu_u - \mu_v)^2$ .

For second-order stationary processes  $\gamma_{uv}^P(\mathbf{h})$  is related to the cross-covariance function by

$$\gamma_{uv}^P(\mathbf{h}) = \frac{1}{2}\{C_{uu}(\mathbf{0}) + C_{vv}(\mathbf{0})\} - C_{uv}(\mathbf{h}). \quad (10.28)$$

Like the cross-covariance function, it is in general not symmetric in  $\mathbf{h}$ . It is also related to the ordinary cross-variogram by

$$\begin{aligned} \gamma_{uv}^P(\mathbf{h}) + \gamma_{vu}^P(\mathbf{h}) \\ = \gamma_{uu}(\mathbf{h}) + \gamma_{vv}(\mathbf{h}) + 2\gamma_{uv}(\mathbf{h}) \\ - [\{Z_u(\mathbf{x}) - Z_v(\mathbf{x} + \mathbf{h})\}, \{Z_v(\mathbf{x}) - Z_u(\mathbf{x} + \mathbf{h})\}], \end{aligned} \quad (10.29)$$

and for second-order stationary processes with symmetric cross-covariances

$$\gamma_{uv}^P(\mathbf{h}) = \gamma_{uv}(\mathbf{h}) + \frac{1}{2}\{C_{uu}(\mathbf{0}) + C_{vv}(\mathbf{0}) - 2C_{uv}(\mathbf{0})\}. \quad (10.30)$$

Papritz *et al.* (1993) explored the properties of the pseudo-cross-variogram and discovered that it has rather restricted validity, though in the right conditions it can be modelled with the ordinary autovariograms and used for cokriging, and this is likely to be its main attraction. More generally, the inability to estimate the usual cross-variogram for want of comparisons between variables at lag zero is tantalizing. Papritz *et al.* (1993) suggested a way forward for situations in which the pseudo-cross-variogram is valid, but the computational load still seems prohibitive for the size of sample needed for reliable estimation.

At present we leave the reader with the pseudo-cross-variogram as a possible function to describe cross-correlation. It is far from ideal, and it seems to us preferable to plan surveys in such a way that there are always enough sites at which all the variables are or can be measured.

For further details and explanation, see Journel and Huijbregts (1978), Matheron (1979), Myers (1982), McBratney and Webster (1983), Papritz *et al.* (1993) and Wackernagel (1994, 2003).

# 11

## *Disjunctive Kriging*

### 11.1 INTRODUCTION

Ordinary kriging is the most common form of geostatistical estimation. As described in Chapter 8, it estimates the values of regionalized variables at unsampled places, i.e. at the target points or blocks, as simple linear combinations of measured values in the neighbourhoods of those targets. The estimates are the best of their kind in the sense that they are unbiased and the variance, which is also estimated, is the minimum. Sometimes we should like to have more information than this; for instance, we might want to know, given the data, the likelihood or probability that the true values at the target points exceed some threshold. These probabilities are not linear combinations of the data. To estimate them we need more elaborate techniques that depend on the statistical distributions of the variables at the target points. The following examples illustrate where this need arises.

In developed countries, in particular, there is a desire to clean up and protect the environment. In some cases laws have been passed to limit the concentrations of certain materials in the air, water and soil. For example, the European Union has stipulated a permissible maximum for the concentration of nitrate in drinking water of  $50 \text{ mg l}^{-1}$ . This has given local authorities in England and Wales powers to prosecute farmers who cause this to be exceeded in water supplies. The Swiss federal government has specified maxima for the concentrations of heavy metals in the soil of the country (FOEFL, 1987). For cadmium and lead, as examples, they are  $0.8 \text{ mg kg}^{-1}$  and  $50 \text{ mg kg}^{-1}$ , respectively. They are guide values, but if they are exceeded then the cantonal administrations must act appropriately. The quality of the air may be judged on the amount of  $\text{SO}_2$  it contains, and governments may again set limits to what is tolerable. If a limit, denoted  $z_c$ , is exceeded then the law-enforcement agency may order polluters to cut their emissions.

In agriculture there are similar situations. In humid temperate climates the soil tends to be acid, cropping there increases the tendency, and farmers need to apply lime to counteract it. There is often a critical value of pH that signifies the

need for lime. If the soil's pH falls below that value then it is time to act by addition of lime to the land. The farmer would like to know, therefore, whether the pH is less than this threshold for each point on the farm. If it exceeds the critical value then the farmer need do nothing. Farmers in drier regions often have to control salinity and alkalinity. Again there are critical values of electrical conductivity in the soil solution (for salinity) and exchangeable sodium percentage (for alkalinity), and if these are exceeded then the farmer should apply gypsum and try to leach the soluble salt out. Here the thresholds,  $z_c$ , are maxima. In other situations there are minimum recommended concentrations for certain nutrient elements in soil. This is especially true of the trace metals copper and cobalt which are essential in the diets of grazing livestock, and graziers should ensure that the herbage, and therefore the soil on which it grows, contains enough.

What is common to these situations is that true values are known only at sample points. Elsewhere the environmental protection agency, the farmer, the grazier, must estimate or predict the values, and these estimates are subject to error. Decisions, however, must be based on these estimates despite the errors. Where an estimate exceeds a threshold widely or is much less than it the decision-maker can take it at its face value and act or not as is appropriate. Difficulty arises where the estimate is close to the threshold and might result in a misjudgement that could have serious or expensive consequences, or both. For example, if the true concentration of nitrate in the ground water is less than  $50 \text{ mg l}^{-1}$ , the  $z_c$ , and the water authority estimates it as more then farmers might be constrained or fined unnecessarily, whereas if the situation is the reverse then consumers might suffer.

Similarly, if the true pH of the soil is less than 5.5, the relevant threshold for a given crop, and the farmer estimates it to be more then he will not add lime. The likely outcome is a loss of yield and profit. If on the other hand the true value exceeds 5.5 and the farmer's estimate is less then he could spend money unnecessarily on lime. If the grazier overestimates the concentration of cobalt in soil that is deficient and as a result does nothing to correct the deficiency then his sheep will not thrive and may die prematurely. If he underestimates the concentration in soil containing sufficient cobalt then he might add cobalt to the soil or to the animals' diet unnecessarily or, more expensively, have his animals' blood tested.

In such situations the land manager might attempt to remedy a soil condition that did not exist, or an agency could have a false sense of security and fail to deal with a threat that did exist. To avoid unnecessary expenditure or treatment or the risk of losing yield or perpetuating environmental damage and suffering if nothing is done, the land manager or law enforcer needs to know the risks of taking their estimates at face value.

Miners face a similar problem. At any particular time there is a price of metal and the cost of processing its ore, and there is a threshold concentration greater than which it is profitable to extract each block of rock and less than which it is

not. As Journel and Huijbregts (1978) remark, ‘decisions are based on estimates, whereas profits depend on the true values’. Miners take financial risks when treating estimates as if they are true.

If we use linear kriging to estimate  $Z$  at the nodes of a fine grid we could examine the effect of the threshold by threading an isarithm at  $z_c$  through the grid and display the result as a map. This would show two classes: one where the estimates of  $Z$  exceed  $z_c$  and the other where they do not. As with the individual estimates, the map would be more or less in error, and there would be a risk in taking the map at its face value.

In all of these situations we need estimates of the probability, given the data, that the true values exceed (or do not exceed) the threshold,  $z_c$ , at an unsampled location  $\mathbf{x}_0$ . It can be expressed formally by

$$\text{Prob}[Z(\mathbf{x}_0) > z | z(\mathbf{x}_i); i = 1, 2, \dots, N] = 1 - \text{Prob}[Z(\mathbf{x}_0) \leq z_c | z(\mathbf{x}_i)], \quad (11.1)$$

where  $N$  is the number of data points.

To determine the probabilities we need to know the conditional expectation or expected value at each target point, which depends on knowing the probability distribution of  $Z(\mathbf{x})$ . Unfortunately, the full multivariate distribution of  $Z(\mathbf{x})$  is inaccessible, partly because we have only one realization and partly because the actual probability distributions depart more or less from theoretical ones.

Two solutions have been proposed to overcome this difficulty; both involve transformations of data, and both are used in practice. The simpler is indicator kriging (Journel, 1983); it needs no assumption of a theoretical distribution, and in this sense it is non-parametric. It converts a variable that has been measured on a continuous scale to several indicator variables, each taking the values 0 or 1 at the sample sites, and estimating their values elsewhere. It is appealing for these reasons. The other solution, disjunctive kriging, is due to Matheron (1976). It transforms the data to a standard normal distribution using Hermite polynomials and then compares the estimated values with the normal distribution to obtain the required probabilities.

Although indicator and disjunctive kriging are described as non-linear methods, both are linear krigings of non-linear transforms of data. Indicator kriging involves simple or ordinary kriging of indicators, and disjunctive kriging is a simple kriging of Hermite polynomials. Both lead to estimates of the probabilities that the true values exceed (or not) specified thresholds at unknown points or blocks in the neighbourhood of data. In this way they enable us to assess the risk we take by accepting the estimates at their face values.

Many case studies using the techniques have been reported. Examples of indicator kriging in mining include ones by Journel (1983) and Lemmer (1984), and in environmental protection by Bierkens and Burrough (1993a, 1993b), Journel (1988), Goovaerts (1994) and Goovaerts *et al.* (1997). Matheron developed disjunctive kriging specifically for mining, and its potential benefits for that industry are evident (Rendu, 1980; Maréchal, 1976; Rivoirard,

1994). Nevertheless, it is proving well suited for environmental protection and land management. Applications in soil science have been especially successful. Yates *et al.* (1986a, 1986b) set it in the context of soil water, and Yates and Yates (1988) used it to estimate viral contamination of soil by sewage. We have applied it to several case studies in soil science (Webster and Oliver, 1989; Wood *et al.* 1990; Webster, 1991, 1994; Oliver *et al.*, 1996).

In this chapter we describe Gaussian disjunctive kriging, but before going into detail we devote a short section to indicators in general.

## 11.2 THE INDICATOR APPROACH

### 11.2.1 Indicator coding

An indicator variable, often abbreviated to ‘indicator’ in geostatistical parlance, is essentially a binary variable; it is one that takes the values 1 and 0 only. Typically such variables denote presence or absence. In soil science we could score a soil sample with a 1 for earthworms if they were present and 0 if they were not. We might score the presence of roots and stones similarly.

We can also create an indicator,  $\omega(\mathbf{x})$ , from a continuous variable,  $z(\mathbf{x})$ , quite simply by scoring it 1 if  $z(\mathbf{x})$  is less than or equal to a specified threshold or cut-off,  $z_c$ , and 0 otherwise:

$$\omega(\mathbf{x}) = \begin{cases} 1 & \text{if } z(\mathbf{x}) \leq z_c, \\ 0 & \text{otherwise.} \end{cases} \quad (11.2)$$

We thereby dissect the scale of  $z$  into two parts, one for which  $z(\mathbf{x}) \leq z_c$  and one for which  $z(\mathbf{x}) > z_c$ , and assign to them the values 1 and 0, respectively. This is what is meant by disjunctive coding. If  $z(\mathbf{x})$  is a realization of a random process,  $Z(\mathbf{x})$ , then  $\omega(\mathbf{x})$  may be regarded as the realization of the indicator random function,  $\Omega[Z(\mathbf{x}) \leq z_c]$ . This is a new binary random process.

It will be convenient to abbreviate the notation somewhat for these variables to  $\omega(\mathbf{x}; z_c)$  for the realization and  $\Omega(\mathbf{x}; z_c)$  for the random function.

The relevance of this transformation to environmental protection is evident. If we have a threshold,  $z_c$ , for the concentration of a pollutant that may not be exceeded then the continuous random variable,  $Z(\mathbf{x})$ , is converted to an indicator function for which the value 1 means clean or acceptable and 0 means polluted and unacceptable. We have already mentioned examples for nitrate in drinking water in the European Union and heavy metals in the soil of Switzerland.

Converting a continuous variable to an indicator clearly loses much of the information in the original data, and it might seem prodigal to transform quantitative data in this way. There are reasons for doing it, however. In some instances the statistical distribution is such that transforming it to one that is known is difficult. This is often the case where there are many zeros in the record.

There may be outliers, the effects of which we want to retain. Pollutants, for which only the outliers exceed the statutory thresholds, often fall into this class.

If this were all, however, the transformation would be of little practical significance. What makes it of value is that several thresholds can be defined and a new indicator variable created for each. Thus, if we define  $S$  thresholds as  $z_{c(1)}, z_{c(2)}, \dots, z_{c(S)}$ , then we shall obtain  $S$  indicators from the data,  $\omega_1, \omega_2, \dots, \omega_S$ :

$$\begin{aligned}\omega_1(\mathbf{x}) &= 1 && \text{if } z(\mathbf{x}) \leq z_{c(1)}, \quad \text{else } 0, \\ \omega_2(\mathbf{x}) &= 1 && \text{if } z(\mathbf{x}) \leq z_{c(2)}, \quad \text{else } 0, \\ &\vdots \\ \omega_S(\mathbf{x}) &= 1 && \text{if } z(\mathbf{x}) \leq z_{c(S)}, \quad \text{else } 0.\end{aligned}\tag{11.3}$$

These may be regarded as the realizations of the corresponding random functions  $\Omega_s(\mathbf{x})$ ,  $s = 1, 2, \dots, S$ , for which

$$\Omega_s(\mathbf{x}) = 1 \quad \text{if } Z(\mathbf{x}) \leq z_{c(s)}, \quad \text{else } 0.\tag{11.4}$$

The expectation of the indicator,  $E[\Omega[Z(\mathbf{x}) \leq z_c]]$ , is the probability,  $\text{Prob}[z_c]$ , that  $Z(\mathbf{x})$  does not exceed  $z_c$ :

$$\text{Prob}[z_c] = \text{Prob}[Z(\mathbf{x}) \leq z_c] = E[\Omega[Z(\mathbf{x}) \leq z_c]].\tag{11.5}$$

This probability,  $\text{Prob}[Z(\mathbf{x}) \leq z_c]$ , is the cumulative distribution

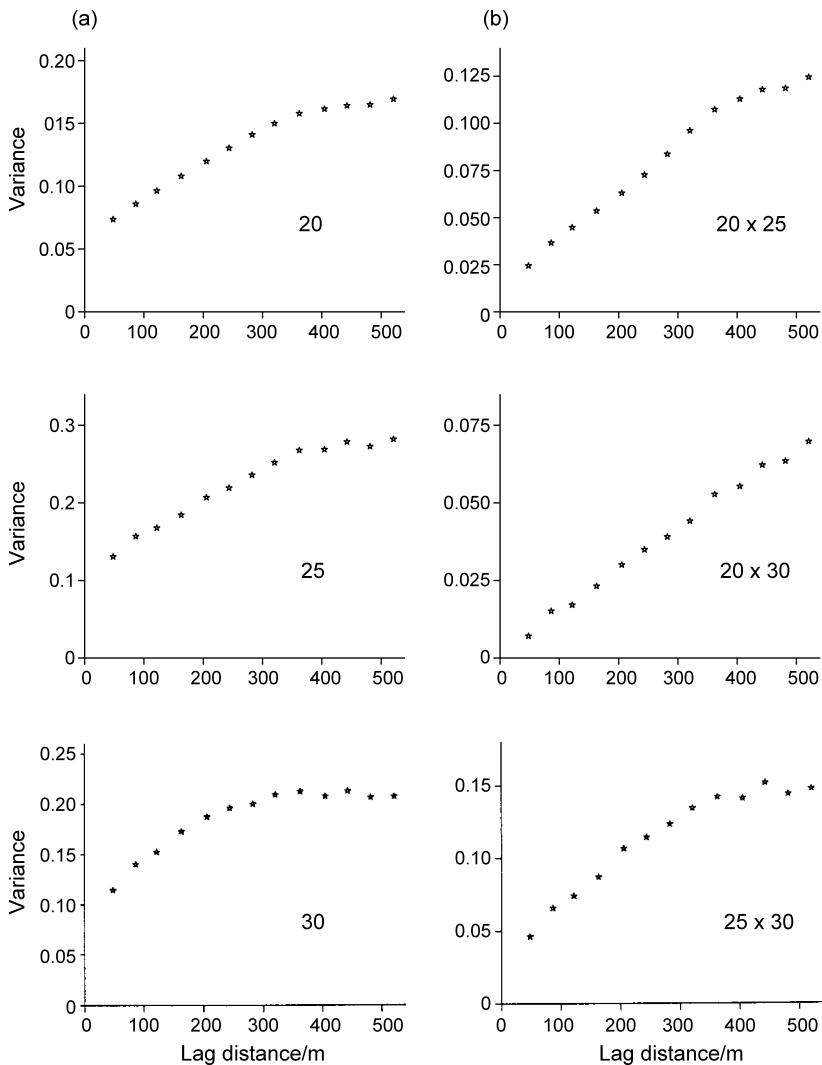
$$\begin{aligned}\text{Prob}[Z(\mathbf{x}) \leq z_c] &= 1 - \text{Prob}[Z(\mathbf{x}) > z_c] \\ &= G[Z(\mathbf{x}; z_c)].\end{aligned}\tag{11.6}$$

Many environmental variables are multi-state characters, such as types of rock, soil and vegetation, that have more than two classes. These can also be converted to indicators by coding each class as present or absent. If we wished to distinguish podzols, brown earths, rendzinas and gleys in a region then we could set up four binary variables, one for each class and code each in turn as 1 or 0. The classes are mutually exclusive, and so in this instance just one of the four would be coded 1 and the other three as 0.

### 11.2.2 Indicator variograms

An indicator random function has a variogram

$$\gamma_z^{\Omega}(\mathbf{h}) = \frac{1}{2} E[\{\Omega(\mathbf{x}; z_c) - \Omega(\mathbf{x} + \mathbf{h}; z_c)\}^2],\tag{11.7}$$



**Figure 11.1** Indicator variograms of potassium at Broom's Barn Farm for thresholds of 20, 25 and 30 mg l<sup>-1</sup>: (a) auto-variograms; (b) cross-variograms.

which is analogous to the variogram of a continuous variable, equation (4.13). The expected semivariances can be estimated from indicator data by

$$\hat{\gamma}_{z_c}^Q(\mathbf{h}) = \frac{1}{2m(\mathbf{h})} \sum_{i=1}^{m(\mathbf{h})} \{\omega(\mathbf{x}_i; z_c) - \omega(\mathbf{x}_i + \mathbf{h}; z_c)\}^2. \quad (11.8)$$

Figure 11.1(a) shows some examples. Further, the ordered sets,  $\hat{\gamma}_{z_c}^Q(\mathbf{h})$ , obtained by applying this formula with changing  $\mathbf{h}$  and for all thresholds,  $z_c$ , can be modelled as described in Chapter 5.

### Cross-indicator variograms

For any two thresholds, say  $z_u$  and  $z_v$ , we can define a cross-indicator variogram and estimate it by elaborating the above formulae:

$$\begin{aligned}\gamma_{uv}^Q(\mathbf{h}) = & \frac{1}{2} E[\{\Omega[Z(\mathbf{x}; z_u)] - \Omega[Z(\mathbf{x} + \mathbf{h}; z_u)]\} \\ & \times \{\Omega[Z(\mathbf{x}; z_v)] - \Omega[Z(\mathbf{x} + \mathbf{h}; z_v)]\}].\end{aligned}\quad (11.9)$$

Examples of cross-variograms of indicators appear in Figure 11.1(b).

### Indicator covariance functions

If the processes are second-order stationary then the spatial correlations among the indicators can all be expressed in terms of covariances:

$$\begin{aligned}C_{z_c}^Q(\mathbf{h}) &= \text{cov}[\Omega[Z(\mathbf{x}; z_c)], \Omega[Z(\mathbf{x} + \mathbf{h}; z_c)]] \\ &= E[\Omega[Z(\mathbf{x}; z_c)]\Omega[Z(\mathbf{x} + \mathbf{h}; z_c)]] - \{E[\Omega[Z(\mathbf{x}; z_c)]]\}^2.\end{aligned}\quad (11.10)$$

Similarly, the cross-covariance at lag  $\mathbf{h}$  of the indicators for thresholds  $z_u$  and  $z_v$  is

$$\begin{aligned}C_{uv}^Q(\mathbf{h}) &= \text{cov}[\Omega[Z(\mathbf{x}; z_u)], \Omega[Z(\mathbf{x} + \mathbf{h}; z_v)]] \\ &= E[\Omega[Z(\mathbf{x}; z_u)]\Omega[Z(\mathbf{x} + \mathbf{h}; z_v)]] \\ &\quad - E[\Omega[Z(\mathbf{x}; z_u)]]E[\Omega[Z(\mathbf{x}; z_v)]].\end{aligned}\quad (11.11)$$

## 11.3 INDICATOR KRIGING

As above, we can krige an indicator variable. So for each target point or block we compute

$$\hat{\Omega}(\mathbf{x}_0; z_c) = \sum_{i=1}^N \lambda_i \omega(\mathbf{x}_i; z_c),\quad (11.12)$$

where the  $\lambda_i$  are the weights as usual. This is the ordinary kriged estimate. The indicator is necessarily bounded, and its sample mean  $(\bar{\omega}; z_c)$ , is usually taken as

its expectation. We can therefore use simple kriging to estimate  $\Omega(\mathbf{x}_0; z_c)$ :

$$\hat{\Omega}(\mathbf{x}_0; z_c) = \sum_{i=1}^N \lambda_i \omega(\mathbf{x}_i; z_c) + \left\{ 1 - \sum_{i=1}^N \lambda_i \right\} (\bar{\omega}; z_c), \quad (11.13)$$

with weights obtained by solving the simple kriging system

$$\sum_{i=1}^N \lambda_i \gamma^{\Omega}(\mathbf{x}_i, \mathbf{x}_j; z_c) = \gamma^{\Omega}(\mathbf{x}_0, \mathbf{x}_j; z_c) \quad \text{for } j = 1, 2, \dots, N, \quad (11.14)$$

where  $\gamma^{\Omega}(\mathbf{x}_i, \mathbf{x}_j; z_c)$  is the indicator semivariance between the  $i$ th and  $j$ th sampling points at threshold  $z_c$  and  $\gamma^{\Omega}(\mathbf{x}_0, \mathbf{x}_j; z_c)$  is the semivariance of the indicator between the target point  $\mathbf{x}_0$  and point  $\mathbf{x}_j$  for the same threshold. As when kriging continuous variables, we can replace  $N$  by  $n \ll N$  in the neighbourhood of  $\mathbf{x}_0$ .

The result is a value lying between 0 and 1 (with exceptions because the kriging minimizes the variance without any constraint on the estimates it returns). Such a value is effectively the probability, given the data, that the true value is 1, i.e.

$$\text{Prob}[\Omega(\mathbf{x}_0; z_c) = 1 | \omega(\mathbf{x}_i), i = 1, 2, \dots, n] = F\{\mathbf{x}_0|(n)\}, \quad (11.15)$$

where we use  $(n)$  to mean all the data in the particular neighbourhood. The quantity  $F\{\mathbf{x}_0|(n)\}$  denotes the conditional or ‘posterior’ probability that  $\Omega(\mathbf{x}_0)$  is 1.

If now we return to our problem, namely to estimate the probability, given data, that the true value of  $Z$  at an unsampled place  $\mathbf{x}_0$  does not exceed  $z_c$ , then we can write

$$\begin{aligned} \text{Prob}[Z(\mathbf{x}_0) \leq z_c | z(\mathbf{x}_i); i = 1, 2, \dots, N] \\ = 1 - \text{Prob}[Z(\mathbf{x}_0) > z_c | z(\mathbf{x}_i); i = 1, 2, \dots, N]. \end{aligned} \quad (11.16)$$

Notice that the two sides of equation (11.16) are complementary.

At first sight it might seem that the way to tackle the problem is to transform the data to indicators determined by the particular threshold. However, we soon see that an individual probability estimated in this way is crude. Much of the rich information in the original data has been lost by dissecting the scale into just two classes.

This loss can be made good to a large extent by repeating the process for several thresholds in the range of  $Z$  and constructing a cumulative distribution function, conditional on the data, for each target point by accumulating the  $\hat{F}(\mathbf{x}_0; z_s)$ ,  $s = 1, 2, \dots, S$ .

The procedure is somewhat tedious because the variograms for all the thresholds must be computed and modelled. Furthermore, because the  $\hat{F}(\mathbf{x}_0; z_s)$  for the different  $z_s$  are computed independently of one another, there is no guarantee that they will sum to 1, or that the cumulative function will increase monotonically, or that the estimated probabilities will lie in the range 0 to 1. Some adjustment of the results may therefore be needed to ensure that the bounds are honoured and the order relations maintained. Nevertheless, an empirical distribution function can be obtained and then used to refine the estimate of the conditional probability that  $Z(\mathbf{x}_0) \leq z_c$ .

Goovaerts (1997) describes the procedure fully and illustrates it with examples using the computer programs in GSLIB (Deutsch and Journel, 1992), while Olea (1999) devotes a section of his book to the topic. We shall not repeat the detail here.

## 11.4 DISJUNCTIVE KRIGING

Disjunctive kriging provides another way of estimating an indicator transform of continuous data. It does so without losing information, though requiring rather stronger assumptions than does indicator kriging as described above. It may take several forms (see Rivoirard, 1994), the most common of which is Gaussian disjunctive kriging and the one we describe.

### 11.4.1 Assumptions of Gaussian disjunctive kriging

The assumptions underlying Gaussian disjunctive kriging are as follows. First,  $z(\mathbf{x})$  is a realization of a second-order stationary process  $Z(\mathbf{x})$  with mean  $\mu$ , variance  $\sigma^2$  and covariance function  $C(\mathbf{h})$ . The underlying variogram must therefore be bounded. Second, the bivariate distribution for the  $n + 1$  variates, i.e. for each target site and the sample locations in its neighbourhood, is known and is stable throughout the region. If the distribution of  $Z(\mathbf{x})$  is normal (Gaussian) and the process is second-order stationary then we can assume that the bivariate distribution for each pair of locations is also normal. Each pair of variates has the same bivariate density, and this density function is determined from the spatial autocorrelation coefficient. These assumptions allow the conditional expectations to be written in terms of the autocorrelation coefficients, as we shall show.

The variable  $Z(\mathbf{x})$  is spatially continuous, so that in going from a small value at one place to a large one elsewhere it must pass through intermediate values *en route*. It is an example of a *Gaussian diffusion process*. One test of this assumption is to compare the variograms of the indicators for several thresholds within the bounds of the measured  $z$ . The cross-indicator variograms should be more ‘structured’ than the autovariograms. Figure 11.1 shows this to be so for potassium at Broom’s Barn Farm.

### 11.4.2 Hermite polynomials

The requirement of normality is a strong one that is rarely met in practice, even though many environmental properties seem approximately normal. The first task therefore is to transform an actual distribution of  $Z(\mathbf{x})$ , which may have almost any form, to a standard normal one,  $Y(\mathbf{x})$ , such that

$$Z(\mathbf{x}) = \Phi[Y(\mathbf{x})]. \quad (11.17)$$

This can be done with Hermite polynomials.

We recall, from Chapter 2, the equation of the standard normal distribution with probability density given by

$$g(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right). \quad (11.18)$$

Hermite polynomials are related to this function and are defined by Rodrigues's formula as

$$H_k(y) = \frac{1}{\sqrt{k!}g(y)} \frac{d^k g(y)}{dy^k}. \quad (11.19)$$

Here  $k$  is the degree of the polynomial, taking values  $0, 1, \dots$ , and  $1/\sqrt{k!}$  is a standardizing factor (Matheron, 1976). The first two Hermite polynomials, i.e. for  $k = 0$  and  $k = 1$ , are

$$H_0(y) = 1, \quad (11.20)$$

$$H_1(y) = -y; \quad (11.21)$$

and thereafter the higher-order polynomials obey the recurrence relation

$$H_k(y) = -\frac{1}{\sqrt{k}} y H_{k-1}(y) - \sqrt{\frac{k-1}{k}} H_{k-2}(y). \quad (11.22)$$

So the polynomials can be calculated up to any order for a standard normal distribution.

The Hermite polynomials are orthogonal with respect to the weighting function  $\exp(-y^2/2)$  on the interval  $-\infty$  to  $+\infty$ . They are independent components of the normal distribution of ever increasing detail.

Almost any function of  $Y(\mathbf{x})$  can be represented as the sum of Hermite polynomials:

$$f\{Y(\mathbf{x})\} = f_0 H_0\{Y(\mathbf{x})\} + f_1 H_1\{Y(\mathbf{x})\} + f_2 H_2\{Y(\mathbf{x})\} + \dots, \quad (11.23)$$

and since the Hermite polynomials are orthogonal

$$\begin{aligned} E[f\{Y(\mathbf{x})\}H_k\{Y(\mathbf{x})\}] &= E\left[H_k\{Y(\mathbf{x})\}\sum_{l=0}^{\infty}f_lH_l\{Y(\mathbf{x})\}\right] \\ &= \sum_{l=0}^{\infty}f_lE[H_l\{Y(\mathbf{x})\}H_k\{Y(\mathbf{x})\}] \\ &= f_k. \end{aligned} \quad (11.24)$$

This enables us to calculate the coefficients  $\phi_k$  of  $\Phi[Y(\mathbf{x})]$  in equation (11.17) as

$$\begin{aligned} Z(\mathbf{x}) &= \Phi[Y(\mathbf{x})] \\ &= \phi_0H_0\{Y(\mathbf{x})\} + \phi_1H_1\{Y(\mathbf{x})\} + \phi_2H_2\{Y(\mathbf{x})\} + \dots \\ &= \sum_{k=0}^{\infty}\phi_kH_k\{Y(\mathbf{x})\}. \end{aligned} \quad (11.25)$$

The transform is also invertible, which means that the results can be expressed in the same units as the original measurements.

### Determining the Hermite coefficients

To determine the coefficients of the transformation,  $\phi_k$ , for a particular set of data we proceed as follows. We arrange the  $N$  data in ascending order:

$$z_1 < z_2 < z_3 < \dots < z_N,$$

and we denote their relative frequencies by

$$q_1, q_2, q_3, \dots, q_N,$$

such that the sum of the frequencies is 1:  $\sum_{i=1}^N q_i = 1$ . Their cumulative frequencies are

$$\begin{aligned} F(z_1) &= \text{Prob}[Z(\mathbf{x}) < z_1] = 0, \\ F(z_2) &= \text{Prob}[Z(\mathbf{x}) < z_2] = q_1, \\ F(z_3) &= \text{Prob}[Z(\mathbf{x}) < z_3] = q_1 + q_2, \\ &\vdots \\ F(z_i) &= \text{Prob}[Z(\mathbf{x}) < z_i] = \sum_{j=1}^{i-1}q_j, \\ &\vdots \\ F(z_N) &= \text{Prob}[Z(\mathbf{x}) < z_N] = 1 - q_N. \end{aligned}$$

The cumulative frequencies have equivalents on the standard normal distribution:

$$F(z_i) = G(y_i). \quad (11.26)$$

Thus

$$F(z_{i+1}) - F(z_i) = G(y_{i+1}) - G(y_i), \quad (11.27)$$

and so

$$\begin{aligned} \text{Prob}[z_i \leq Z(\mathbf{x}) < z_{i+1}] &= \text{Prob}[y_i \leq Y(\mathbf{x}) < y_{i+1}], \\ \text{Prob}[Z(\mathbf{x}) = z_i] &= \text{Prob}[y_i \leq Y(\mathbf{x}) < y_{i+1}]. \end{aligned} \quad (11.28)$$

In words,  $Z(\mathbf{x})$  equals  $z_i$  when the standard normal equivalent lies between  $y_i$  and  $y_{i+1}$ . We can then determine the transformation coefficients as follows:

$$\phi_0 = E[\Phi\{Y(\mathbf{x})\}] = E[Z(\mathbf{x})] = \sum_{i=1}^N q_i z_i, \quad (11.29)$$

and thereafter

$$\begin{aligned} \phi_k &= E[Z(\mathbf{x})H_k\{Y(\mathbf{x})\}] \\ &= \int_{-\infty}^{+\infty} \Phi(y)H_k(y)g(y) dy \\ &= \sum_{i=1}^N \int_{y_i}^{y_{i+1}} z_i H_k(y)g(y) dy \\ &= \sum_{i=1}^N z_i \left[ \frac{1}{\sqrt{k}} H_{k-1}(y_{i+1})g(y_{i+1}) - \frac{1}{\sqrt{k}} H_{k-1}(y_i)g(y_i) \right] \\ &= \sum_{i=2}^N (z_{i-1} - z_i) \frac{1}{\sqrt{k}} H_{k-1}(y_i)g(y_i) \end{aligned} \quad (11.30)$$

because  $g(y_0) = g(-\infty) = 0$ , and  $g(y_{N+1}) = g(+\infty) = 0$  also.

### 11.4.3 Disjunctive kriging for a Hermite polynomial

Since the polynomials are orthogonal any pair of values,  $Y(\mathbf{x})$  and  $Y(\mathbf{x} + \mathbf{h})$ , drawn from a bivariate normal distribution with correlation coefficient  $\rho$  has expectation

$$E[H_k\{Y(\mathbf{x})\}|Y(\mathbf{x} + \mathbf{h})] = \rho^k(\mathbf{h})H_k\{Y(\mathbf{x} + \mathbf{h})\}. \quad (11.31)$$

The covariance between two functions of  $Y$  at  $\mathbf{x}$  and  $\mathbf{x} + \mathbf{h}$  is

$$\begin{aligned}\text{cov}[H_k\{Y(\mathbf{x})\}, H_l\{Y(\mathbf{x} + \mathbf{h})\}] \\ = \text{E}[H_k\{Y(\mathbf{x})\} H_l\{Y(\mathbf{x} + \mathbf{h})\}] \\ = \text{E}[H_k\{Y(\mathbf{x})\} \text{E}[H_l\{Y(\mathbf{x} + \mathbf{h})\}] | Y(\mathbf{x} + \mathbf{h})] \\ = \rho^k(\mathbf{h}) \text{E}[H_k\{Y(\mathbf{x})\} H_l\{Y(\mathbf{x})\}].\end{aligned}\quad (11.32)$$

When  $k = l$  this equation gives the covariance of  $H_k\{Y(\mathbf{x})\}$ , which equals  $\rho^k(\mathbf{h})$ , since  $Y(\mathbf{x})$  is a standard normal variate. The correlation coefficient,  $\rho(\mathbf{h})$ ,  $\mathbf{h} \neq \mathbf{0}$ , must lie between  $-1$  and  $+1$ ; so  $\rho^k(\mathbf{h})$  rapidly approaches 0 as  $k$  increases, and the spatial dependence in  $H_k\{Y(\mathbf{x})\}$  declines to nothing, i.e.  $H_k\{Y(\mathbf{x})\}$  becomes pure nugget.

Any pair of Hermite polynomials is spatially independent, so they are the independent factors of the bivariate normal model. By kriging them separately the estimates have only to be summed to give the disjunctive kriging estimator:

$$\hat{Z}^{\text{DK}}(\mathbf{x}) = \phi_0 + \phi_1 \hat{H}_1^{\text{K}}\{Y(\mathbf{x})\} + \phi_2 \hat{H}_2^{\text{K}}\{Y(\mathbf{x})\} + \dots \quad (11.33)$$

So, if we have  $n$  points in the neighbourhood of  $\mathbf{x}_0$  where we want an estimate, we estimate the Hermite polynomials by

$$\hat{H}_k^{\text{K}}\{Y(\mathbf{x}_0)\} = \sum_{i=1}^n \lambda_{ik} H_k\{Y(\mathbf{x}_i)\}, \quad (11.34)$$

and we insert them into equation (11.33). The  $\lambda_{ik}$  are the kriging weights, which are found by solving the equations for simple kriging because we can assume the mean is known:

$$\begin{aligned}\sum_{i=1}^n \lambda_{ik} \text{cov}[H_k\{Y(\mathbf{x}_j)\}, H_k\{Y(\mathbf{x}_i)\}] \\ = \text{cov}[H_k\{Y(\mathbf{x}_j)\}, H_k\{Y(\mathbf{x}_0)\}]\quad \text{for all } j,\end{aligned}\quad (11.35)$$

or alternatively,

$$\sum_{i=1}^n \lambda_{ik} \rho^k(\mathbf{x}_i - \mathbf{x}_0) = \rho^k(\mathbf{x}_j - \mathbf{x}_0) \quad \text{for all } j, \quad (11.36)$$

from equation (11.31). In particular, the procedure enables us to estimate  $Z(\mathbf{x}_0)$  by

$$\hat{Z}(\mathbf{x}_0) = \Phi\{\hat{Y}(\mathbf{x}_0)\} = \phi_0 + \phi_1 [\hat{H}_1^{\text{K}}\{y(\mathbf{x}_0)\}] + \phi_2 [\hat{H}_2^{\text{K}}\{y(\mathbf{x}_0)\}] + \dots \quad (11.37)$$

#### 11.4.4 Estimation variance

The kriging variance of  $\hat{H}_k\{Y(\mathbf{x})\}$  is

$$\sigma_k^2(\mathbf{x}_0) = 1 - \sum_{i=1}^n \lambda_{ik} \rho^k(\mathbf{x}_i - \mathbf{x}_0), \quad (11.38)$$

and the disjunctive kriging variance of  $\hat{f}[Y(\mathbf{x}_0)]$  is

$$\sigma_{\text{DK}}^2(\mathbf{x}_0) = \sum_{k=1}^{\infty} f_k^2 \sigma_k^2(\mathbf{x}_0). \quad (11.39)$$

#### 11.4.5 Conditional probability

Once the Hermite polynomials have been estimated at a target point we can estimate the conditional probability that the true value there exceeds the critical value,  $z_c$ . The transformation  $Z(\mathbf{x}) = \Phi[Y(\mathbf{x})]$  means that  $z_c$  has an equivalent  $y_c$  on the standard normal scale. Since the two scales are monotonically related their indicators are the same:

$$\Omega[Z(\mathbf{x}) \leq z_c] = \Omega[Y(\mathbf{x}) \leq y_c]. \quad (11.40)$$

For  $\Omega[Y(\mathbf{x}) > y_c]$ , which is the complement of  $\Omega[Y(\mathbf{x}) \leq y_c]$ , the  $k$ th Hermite coefficient is

$$\begin{aligned} f_k &= \int_{-\infty}^{+\infty} \Omega[y \leq y_c] H_k(y) g(y) dy \\ &= \int_{-\infty}^{y_c} H_k(y) g(y) dy. \end{aligned} \quad (11.41)$$

The coefficient for  $k = 0$  is the cumulative distribution to  $y_c$ ,

$$f_0 = G(y_c),$$

and for larger  $k$ ,

$$f_k = \frac{1}{\sqrt{k}} H_{k-1}(y_c) g(y_c). \quad (11.42)$$

The indicator can be expressed in terms of the cumulative distribution and the Hermite polynomials:

$$\Omega[Y(\mathbf{x}) \leq y_c] = G(y_c) + \sum_{k=1}^{\infty} \frac{1}{\sqrt{k}} H_{k-1}(y_c) g(y_c) H_k\{Y(\mathbf{x})\}. \quad (11.43)$$

Its disjunctive kriging estimate is obtained by

$$\hat{\Omega}^{\text{DK}}[y(\mathbf{x}_0) \leq y_c] = G(y_c) + \sum_{k=1}^L \frac{1}{\sqrt{k}} H_{k-1}(y_c) g(y_c) \hat{H}_k^{\text{K}}\{y(\mathbf{x}_0)\}, \quad (11.44)$$

where  $L$  is some small number. The kriged estimates  $\hat{H}_k^{\text{K}}\{y(\mathbf{x}_0)\}$  approach 0 rapidly with increasing  $k$ , and so summation need extend over only a few terms even though the  $(1/\sqrt{k})H_{k-1}(y_c)g(y_c)$  are considerable. Of course, this is the same as  $\hat{\Omega}^{\text{DK}}[z(\mathbf{x}_0) \leq z_c]$ . Conversely, to obtain the probability of excess we can compute

$$\begin{aligned} \hat{\Omega}^{\text{DK}}[z(\mathbf{x}_0) > z_c] &= \hat{\Omega}^{\text{DK}}[y(\mathbf{x}_0) > y_c] \\ &= 1 - G(y_c) - \sum_{k=1}^L \frac{1}{\sqrt{k}} H_{k-1}(y_c) g(y_c) \hat{H}_k^{\text{K}}\{y(\mathbf{x}_0)\}. \end{aligned} \quad (11.45)$$

#### 11.4.6 Change of support

In describing disjunctive kriging above we have treated each target as a ‘point’ with the same support as the data. The simple kriging equations are readily modified to estimate the Hermite polynomials and hence  $Z(\mathbf{x})$  over larger blocks  $B$  by replacing the covariances on their right-hand sides with block averages. The result is a block kriging, i.e. an estimate of the average value of  $Z$  within a target block, say  $Z(B)$ . It will also produce an estimate of the average probability that  $Z(\mathbf{x}) \leq z_c$  in  $B$ , but note that this probability is not same as the probability that the average of  $Z$  in  $B$  is less than or equal to  $z_c$ .

As we saw above, in Chapter 4, the distribution of a spatially correlated variable changes as the support changes. In particular, the variance diminishes as the support increases and this is evident in the regularized variogram of Figure 4.7. If we are to estimate the conditional probabilities that block averages exceed  $z_c$  then we need to take into account the larger support and to model the change of support. Webster (1991) summarizes the theory and illustrates it with an example from agricultural science, and Rivoirard (1994) treats it more didactically, again with an illustration using the same data. The subject is beyond the scope of this book, but you can read about the theory and technique in the two works cited.

### 11.5 CASE STUDY

To illustrate the method and to enable the results of disjunctive kriging to be compared with those of ordinary kriging in Chapter 8, we use the data on exchangeable potassium from the soil survey of Broom’s Barn Farm. Chapter 2

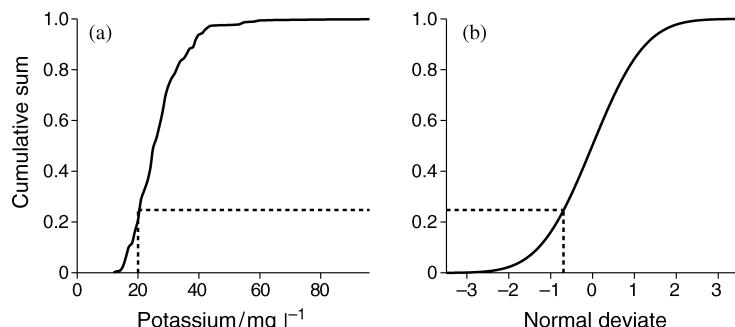
**Table 11.1** Summary statistics.

Statistic	K	$\log_{10} K$	Hermite-transformed K
Mean	26.3	1.40	0.0740
Median	25.0	1.40	0.104
Standard deviation	9.04	0.134	0.974
Variance	81.706	0.018 00	0.9495
Skewness	2.04	0.39	-0.03
Kurtosis	9.45	0.57	0.07
Deficiency threshold	25.0	1.40	0.104

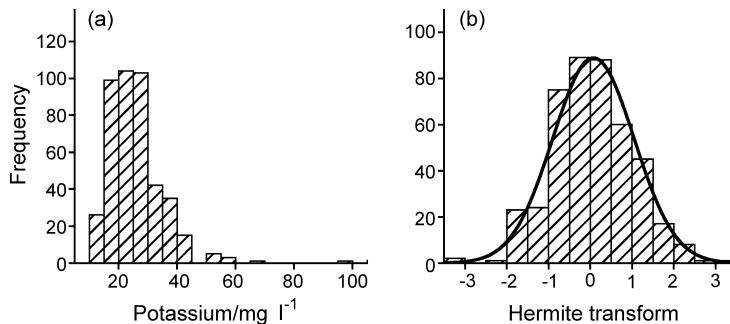
describes them in full, and here we repeat only the most salient features. Table 11.1 summarizes the statistics of the data and the transforms to standard normal deviates using Hermite polynomials. It includes the summary for the common logarithms for comparison.

Figure 11.2 shows the general nature of the problem. Figure 11.2(a) is the cumulative distribution of exchangeable K as observed. For a defined threshold concentration,  $z_c$ , we should like to know its equivalent on the standard normal curve, because then we can calculate confidence limits. We suppose for illustration that it is  $20 \text{ mg l}^{-1}$  of soil. From Figure 11.2(a) we see that the cumulative sum is approximately 0.24. Tracing this value across by the horizontal dashed lines to the standard normal distribution on Figure 11.2(b), we see that its equivalent normal deviate is -0.69, shown by the vertical dashed line there. The first task therefore is to transform the data to a standard normal distribution so that we have the equivalences for all reasonable values of  $z$ .

The distribution on the original scale ( $\text{mg l}^{-1}$ ) is strongly skewed,  $g_1 = 2.04$  (Figure 11.3(a) and Table 11.1). Taking logarithms removes most of the skewness, with  $g_1 = 0.39$ , as shown in Figure 2.1(b).



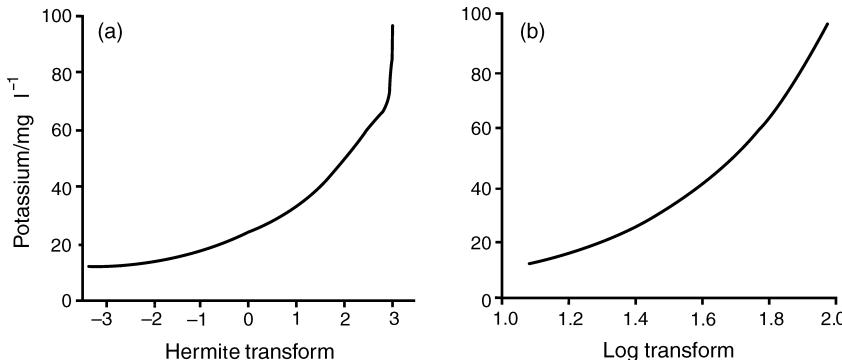
**Figure 11.2** The cumulative distribution: (a) of potassium; (b) of a standard normal distribution. The vertical dashed line in (a) is for a threshold of  $20 \text{ mg l}^{-1}$ , and the others show how it equates in (b); see text.



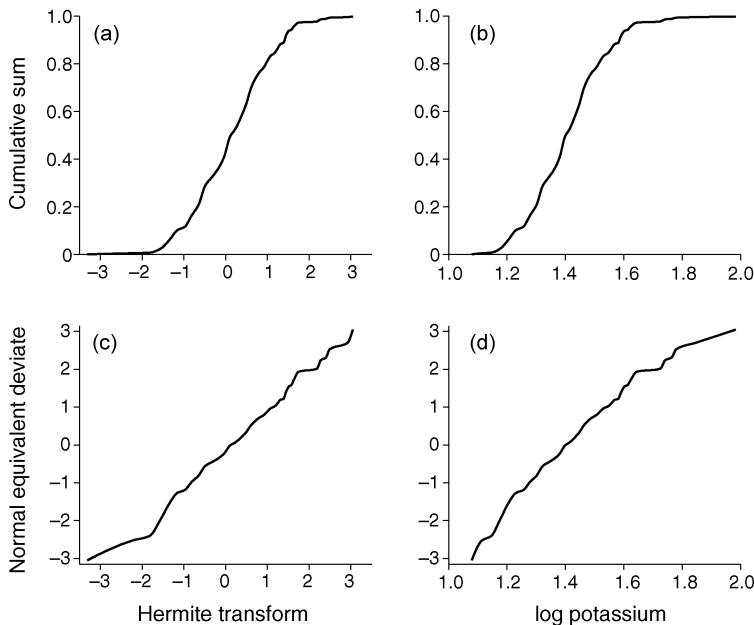
**Figure 11.3** Histograms of potassium: (a) as measured in  $\text{mg l}^{-1}$ ; (b) after transformation by Hermite polynomials with the curve of the normal distribution fitted.

Transforming with Hermite polynomials up to order 7 is more effective, giving approximate standard normal deviates. The mean and variance depart somewhat from 0 and 1, respectively. The skewness is virtually nil ( $-0.03$ ), as is the kurtosis ( $0.07$ ), see Figure 11.3(b) and Table 11.1. Figure 11.4(a) shows the transform function with the measured values plotted against the Hermite transformed ones. The graph is concave upwards, resulting from the positive skewness of the data. For a normal distribution the transform function would be a straight line; the departure from this is a measure of the non-normality. We show the logarithmic transformation function in Figure 11.4(b) for comparison.

Figure 11.5 shows other features of the transformation, again with those of the logarithms alongside for comparison. In Figure 11.5(a) and (b) are the cumulative distributions with  $G(y)$  plotted against  $y$ , the transformed values. Both are characteristically sigmoid, as expected for data from a normal distribution. In Figure 11.5(c) and (d) we have plotted the normal equivalent deviates of the cumulative distributions against  $y$ . The normal equivalent



**Figure 11.4** Transform functions of potassium at Broom's Barn Farm: (a) for Hermite polynomials; (b) for logarithms.



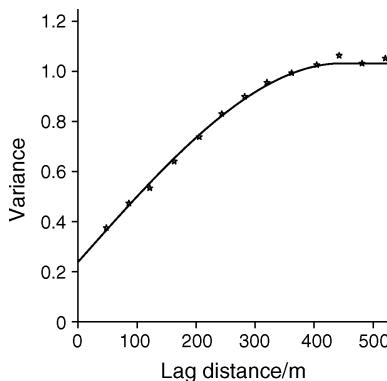
**Figure 11.5** Cumulative distributions of potassium at Broom's Barn Farm: (a) the cumulative sum for the Hermite transform, and (b) for the logarithms; (c) and (d) the cumulants plotted as normal equivalent deviates.

deviate is the area beneath a curve of the standard normal pdf from  $-\infty$  to  $g(y)$ , equivalent to  $G(y)$ . For a normal distribution this function plots as a straight line. For the Hermite transformation of potassium it is straight, apart from local fluctuation. For the logarithms, however, there is still detectable curvature.

We mentioned above our assumption that  $z(\mathbf{x})$  is the outcome of a Gaussian diffusion process for which the cross-variograms of the indicators should be more structured than the autovariograms. To check that the exchangeable  $K$  conforms we computed the relevant variograms for  $K > K_c$  for  $c = 20, 25$  and  $30 \text{ mg l}^{-1}$ , which correspond closely to the quartiles of the cumulative distribution; they are the cumulants 0.24, 0.51 and 0.75, respectively. The results are shown in Figure 11.1 with the autoindicator variograms on the left and the cross-indicator variograms on the right. We have not fitted models to them, but quite evidently the latter are more structured; any curve fitted closely to the experimental values will project on to the ordinate near the origin, whereas all three autovariograms will have substantial nugget variances.

We computed the experimental variogram from the Hermite-transformed values and fitted an isotropic spherical model to it:

$$\hat{\gamma}(h) = 0.216 + 0.784 \text{ sph}(434), \quad (11.46)$$



**Figure 11.6** Variogram of potassium after transformation by Hermite polynomials.

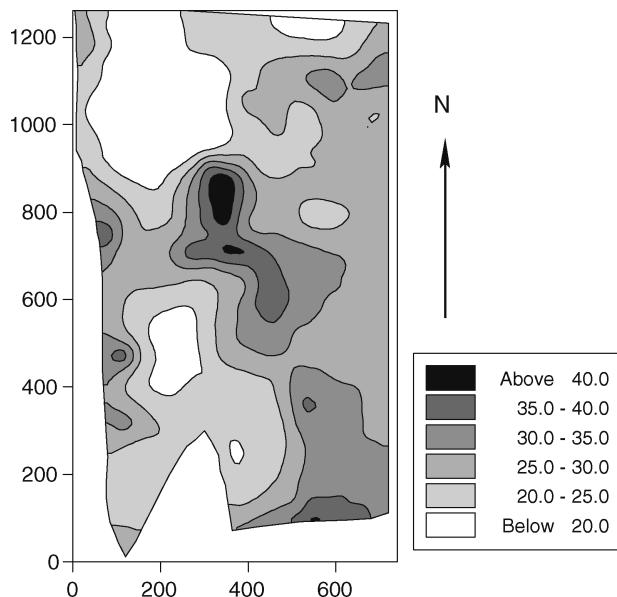
in which  $\text{sph}(434)$  indicates the spherical function with a range of 434 m. Figure 11.6 shows the experimental values as points and the fitted model as a solid line. Using this model and the transformed values, we estimated the concentrations of potassium at the nodes of a square grid at 10-m intervals by punctual kriging of the Hermite polynomials, as described above.

For cereal crops at the time the survey was made, a critical value for readily exchangeable potassium was  $25 \text{ mg l}^{-1}$ ; this was the threshold below which the Ministry of Agriculture, Fisheries and Food (1986) recommended farmers to fertilize cereal crops. We computed the conditional probabilities of the values' being less than or equal to this threshold at the same grid nodes.

Figure 11.7 is the map of the disjunctively kriged estimates of exchangeable K. As it happens in this instance, it is little different from the map made by lognormal kriging (Figure 8.22), because the transform functions are similar. We can see this by plotting the disjunctively and lognormally kriged estimates against each other, as in Figure 11.8(a). There is little scatter in the points from the solid line of perfect correlation on the graph, and the correlation is  $r = 0.994$ . Figure 11.8(b) is the scatter diagram of the disjunctively kriged estimates plotted against the kriging variance. This shows clearly the effect of the nugget variance in punctual kriging. The nugget variance sets a lower limit to the precision of the estimates, and this is evident in the horizontal line at a kriging variance of about  $25 (\text{mg l}^{-1})^2$ .

In addition, disjunctive kriging enables us to map the estimated conditional probabilities of deficiency or excess from the same set of target points. Figures 11.9(a) and 11.9(b) are maps of the probabilities for thresholds of  $25 \text{ mg l}^{-1}$  and  $20 \text{ mg l}^{-1}$ , respectively.

In Figure 11.8(c) we have plotted the conditional probabilities that the exchangeable  $K \leq 25 \text{ mg l}^{-1}$  against the disjunctively kriged estimates. It is evident from this graph that some of the estimates exceeding the threshold have

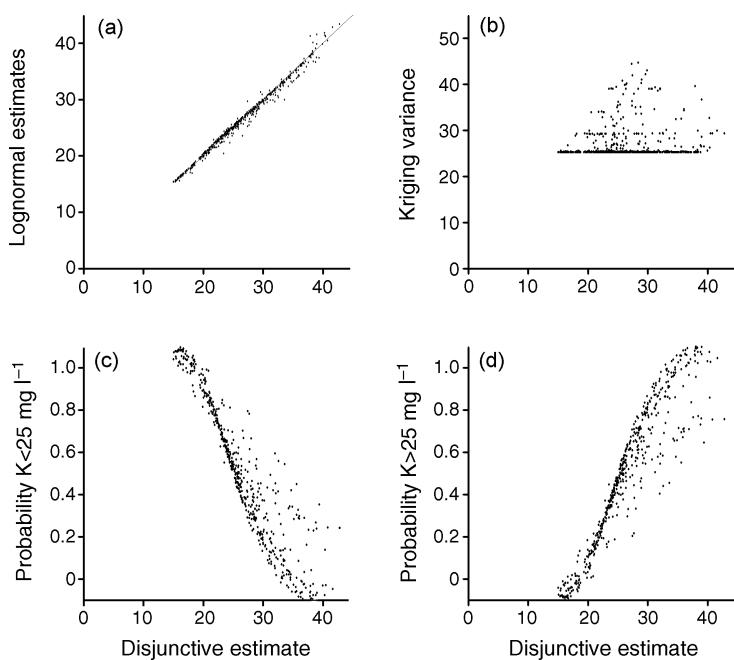


**Figure 11.7** Map of exchangeable potassium at Broom's Barn Farm, estimated by disjunctive kriging.

associated with them fairly large probabilities of deficiency; evidently we should not judge the likelihood of deficiency from the estimates alone. You may also notice that some of the points on the graph lie outside the bounds of 0 and 1 for the probabilities. This is because they are themselves estimates.

In environmental management we are often concerned with the probabilities that some substance exceeds a threshold. If potassium were a pollutant then we might plot the probabilities of its exceeding the threshold of  $25 \text{ mg l}^{-1}$ . We should then obtain Figure 11.8(d), the inverse of Figure 11.8(c).

In a situation concerning deficiency the farmer would fertilize where the map showed exchangeable K to be less than  $25 \text{ mg l}^{-1}$ , the pale grey and white areas of Figure 11.7. However, the farmer would not want to risk losing yield where the estimated concentration of K is more than the threshold and the probability of deficiency is moderate. If he were prepared to set the maximum risk at a probability of 0.3 then he should fertilize the areas in Figure 11.9(a) where the probability is greater, i.e. areas of medium and dark grey and black. The area is considerable. When the map of probabilities is compared with that of the estimates it is clear that the farmer could risk loss of yield by taking the estimates at face value—the

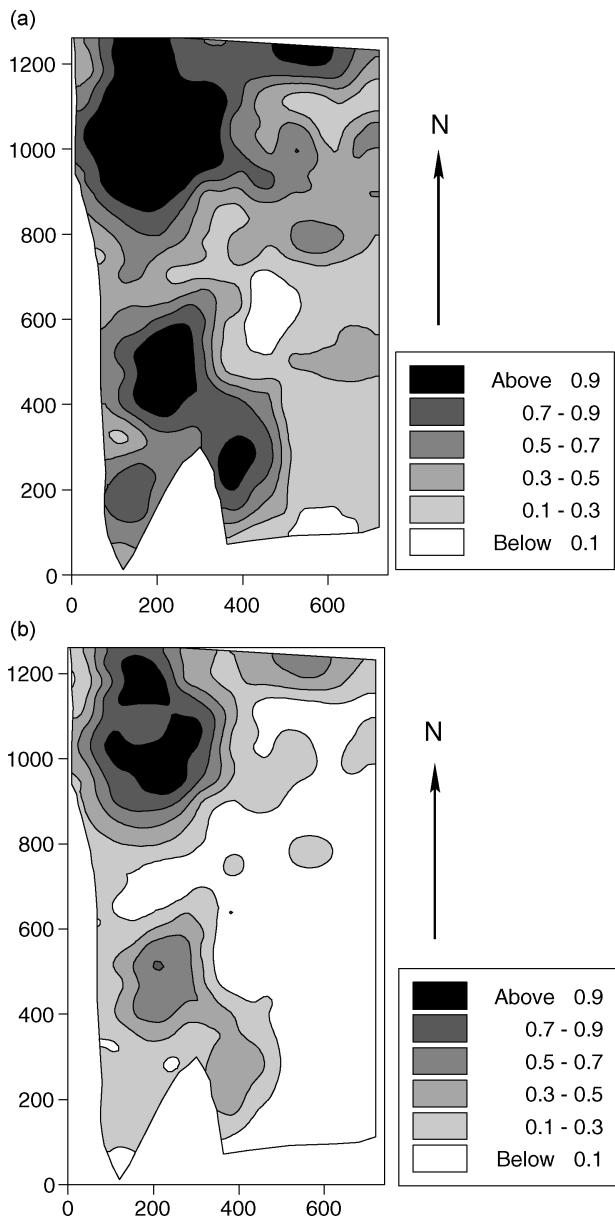


**Figure 11.8** Scatter of: (a) disjunctively kriged estimates against lognormally kriged ones; (b) estimates obtained by disjunctive kriging against their estimation variances; (c) estimated probabilities of deficiency ( $\leq 25 \text{ mg l}^{-1}$ ) against estimates; (d) estimated probabilities of excess ( $> 25 \text{ mg l}^{-1}$ ) against estimates.

area requiring fertilizer is considerably greater than that where  $K \leq 25 \text{ mg l}^{-1}$ .

## 11.6 OTHER CASE STUDIES

Wood *et al.* (1990) described an application of disjunctive kriging to estimating the salinity of the soil in the Bet Shean Valley to the west of the River Jordan in Israel. The combination of climate, irrigation and smectite clay soil has resulted in significant concentrations of sodium salts in the topsoil. In general, salinity limits the range of crops that can be grown as well as reducing the yields of those that can tolerate it. A critical threshold,  $z_c$ , of electrical conductivity (EC) is  $4 \text{ mS cm}^{-1}$ : it is widely recognized as marking the onset of salinization of the soil. The principal crops that are affected by too much salt in the valley are lucerne, wheat and dates. The losses of yield of lucerne and dates become serious when this threshold is exceeded. Winter wheat, however, will still grow, but when the threshold is exceeded it germinates poorly.



**Figure 11.9** Maps of probabilities of potassium deficiency at Broom's Barn Farm with thresholds of (a)  $25 \text{ mg l}^{-1}$ ; (b)  $20 \text{ mg l}^{-1}$ .

The EC of the soil solution in November before the onset of winter rain is the most telling, and it was measured at some 200 points in a part of the valley at that time to indicate salinity. The EC was then estimated at the nodes of a fine grid and mapped. The conditional probabilities of the ECs exceeding  $4 \text{ mS cm}^{-1}$  were also determined and mapped, and in the event they exceeded 0.3 over most of the region. There was a moderate risk of salinity in most of the region. Farmers would find it too costly to remediate the entire area, but they could use the map of probabilities as a guide for deciding on the priority of areas for remediation.

Webster and Oliver (1989), Webster and Rivoirard (1991) and Webster (1994) used the data from the original survey by McBratney *et al.* (1982) of copper and cobalt in the soil of southeast Scotland to study the merit and relevance of disjunctive kriging in agriculture. Deficiencies of copper and cobalt in the soil of the region cause poor health in grazing sheep and cattle there. The critical value,  $z_c$ , for copper in the soil is  $1 \text{ mg kg}^{-1}$  and for cobalt  $0.25 \text{ mg kg}^{-1}$ . Data from some 3500 sampling points were available, and from them they computed the probabilities of the soil's being deficient in the two trace metals by disjunctive kriging. The concentration of copper exceeded the  $1 \text{ mg kg}^{-1}$  threshold almost everywhere. The concentration is near to the threshold in only small parts of the region where the estimated probability of deficiency was typically in the range 0.2–0.3. For cobalt, however, for which the mean concentration was almost exactly equal to the threshold of  $0.25 \text{ mg kg}^{-1}$ , the estimates for approximately half the region were less than the threshold with an estimated probability of deficiency greater than 0.5, and elsewhere most of the computed probabilities exceeded 0.2. The potential loss of thrift in the animals and therefore profit to the farmer is considerable, whereas preventive measures such as supplementary cobalt in the animals' feed or additions in the fertilizer are cheap. In these circumstances the farmer would be advised to take one of these courses of action where the probability of deficiency exceeded 0.2.

Maps of probabilities also help environmental scientists to design programmes of remediation for areas considered to be polluted. Once the users have decided what risks they are prepared to take, the scientist can use such maps to recommend suitable action. If there are strictly limited funds for remediation the map of probabilities enables them to assign priorities for action; the parts of the region where the probabilities are greatest can be tackled first.

Von Steiger *et al.* (1996) estimated the concentrations of heavy metals in polluted soil in part of northeast Switzerland by disjunctive kriging. The soil contained lead in excess of the Swiss Federal Guide value of  $50 \text{ mg kg}^{-1}$ . The probabilities exceeded 0.3 to the north and east of the town of Weinfelden, suggesting that these areas should be monitored to ensure that the burden in the soil does not exceed the existing concentrations.

## 11.7 SUMMARY

The principles described in this chapter can be applied to various substances in the environment, whether they are nutrients that might be deficient or heavy metals and xenobiotics, excesses of which are toxic. The probabilities of exceeding specific thresholds enable the risk of inaction to be assessed quantitatively. Disjunctive kriging, in particular, provides environmental analysts with a useful decision-making tool, especially where failure to act could result in litigation, damage to health or loss of revenue. Assessing this risk is now feasible in an optimal way.

# 12

## ***Stochastic Simulation***

### **12.1 INTRODUCTION**

To introduce stochastic simulation we return to the final section of Chapter 8 and recapitulate on the effect of kriging. Figure 8.24 shows that kriging tends to underestimate values that are larger than average and to overestimate those that are smaller. It behaves in the same way as ordinary regression; its estimates are less variable than the true values. In the example, the variance of the estimates, 0.009 79, is only a little more than half that of the sample values,  $s^2 = 0.018\ 00$ . Thus, kriging has lost variance that was in the original data. In general, when we display the statistical surface we estimate as maps, as in Figures 8.16 and 8.17, we obtain a smoothed representation of reality. The spikes in the punctually kriged surface at the sampling points (Figure 8.15) are due to the nugget component in the variogram. Nevertheless, they show that a smoothed representation can give a misleading picture of reality, and we appreciate this only because for this example we included the kriged estimates at the sampling points where they are equal to the observed values. If we remove the predictions at the data points or krig at nodes of a prediction grid that is offset from the sampling grid, then the surface would be smooth, as in Figures 8.16 and 8.17. Although these figures are from block-kriged estimates, the maps for punctual kriging on a grid offset from the original were little different.

We can formalize the reasons for the smoothing as follows. Starting with sample data,  $z(\mathbf{x}_i)$ ,  $i = 1, 2, \dots, N$ , from a region of interest,  $R$ , we compute the experimental variogram,  $\hat{\gamma}(\mathbf{h})$ . To this we fit a plausible model,  $\gamma(\mathbf{h})$ , which we regard as the underlying variogram of the process. With this model and our data, we estimate values of  $z$  at unsampled places by kriging. These estimates are unbiased and for each estimate the variance is minimized,

$$\sigma_k^2(\mathbf{x}_0) = E\left[\{z(\mathbf{x}_0) - \hat{Z}(\mathbf{x}_0)\}^2\right], \quad (12.1)$$

where  $z(\mathbf{x}_0)$  is the true value at  $\mathbf{x}_0$  and  $\hat{Z}(\mathbf{x}_0)$  is our estimate. To map  $Z$  over the region kriging is repeated at numerous positions on a grid and these kriged estimates can be used to create either pixel or isarithmic ('contour') maps. As mentioned above, the variance of the estimates is less than that of the data,  $s^2$ . It is also less than the dispersion variance,  $\sigma^2(R)$ , in the region, which can be obtained by integration of the variogram model, equation (4.25). This difference is approximately

$$\sigma^2(R) - \sigma_K^2(R) \approx \overline{\sigma_K^2}(\mathbf{x}_0) - 2\overline{\psi}(\mathbf{x}_0). \quad (12.2)$$

In this equation  $\sigma_K^2(R)$  is the dispersion variance of the estimates,  $\overline{\sigma_K^2}(\mathbf{x}_0)$  is the average kriging variance of the estimates, and  $\overline{\psi}(\mathbf{x}_0)$  is the average of Lagrange multipliers.

Usually the  $\psi$  for any  $\mathbf{x}_0$  is much smaller than the corresponding  $\sigma_K^2(\mathbf{x}_0)$ , and if the kriging system embraces all the data then it is negligible. In these circumstances we can rewrite equation (12.2) as

$$\sigma_K^2(R) \approx \sigma^2(R) - \overline{\sigma_K^2}(\mathbf{x}_0). \quad (12.3)$$

This equation shows crucially how variance is lost when we krige over the region, and how kriging *smooths*. The larger is the kriging variance on average the more variance is lost. The kriging variance is large where more of the variance is unexplained, i.e. with a large nugget variance, and where sample sites are sparse. In the limit, when all the variance is nugget it dominates the kriging variance, and if we have a single kriging system then predictions will be uniform, i.e. we are left with no variation.

Although a kriged map shows our best estimates of  $Z$ , it does not represent the variation well; this loss of information and detail in the variation could mislead. To obtain a statistical surface that retains the variation we know or believe to be present, then, we need some other technique. Simulation is such a technique.

## 12.2 SIMULATION FROM A RANDOM PROCESS

In geostatistics the term 'simulation' is used to mean the creation of values of one or more variables that emulate the general characteristics of those we observe in the real world. The variables may be categorical or continuous. Values can be created at positions in one, two or three dimensions that are the outcomes of stochastic processes we choose to represent reality. In Chapter 4 we introduced the idea of treating any particular physical variable,  $z(\mathbf{x})$ , as a realization of a stochastic process,  $Z(\mathbf{x})$ , in  $R$ . If the process is second-order stationary then we can characterize it by its mean and covariance function; if it

is intrinsic only then its variogram characterizes the variation. In principle these functions could give rise to any number of realizations, of which the actuality is but one. We can simulate many equally probable realizations that are as likely as the actuality and have the same statistical characteristics. In this way we can obtain dense fields of values from sparse data, just as we do by kriging, but the variance in the original data is retained.

Stochastic simulation differs from kriging in two ways, as follows.

1. Kriging provides the ‘best’, i.e. minimum variance, local estimates without regard to the resulting statistics of those estimates. In simulation, however, the aim is to reproduce the global statistics and maintain the texture of the variation, and these take precedence over local accuracy.
2. A kriged estimate at any place has associated with it a variance, and hence an uncertainty, that is independent of estimates at all other places. Confidence about it is usually based on an assumed Gaussian distribution with the mean equal to the estimate and a cumulative distribution function. We can modify equation (4.1) to take into account the  $n$  data in the neighbourhoods and in the kriging systems to give

$$\text{Prob}[Z(\mathbf{x}_j) \leq z_c | n_j] = F\{Z(\mathbf{x}_j); z_c | n_j\} \quad \text{for } j = 1, 2, \dots \quad (12.4)$$

This enables us to judge the probability that some threshold is exceeded at each target point. However, we cannot derive from it the probability that the values at two or more, say  $J$ , places in a neighbourhood jointly exceed a threshold,

$$\text{Prob}[Z(\mathbf{x}_j) \leq z_c, j = 1, 2, \dots, J | n_j] \neq \prod_{j=1}^J F\{Z(\mathbf{x}_j); z_c | n_j\}, \quad (12.5)$$

unless the  $Z(\mathbf{x})$  themselves are independent, a situation that is of little interest to us. We can assess the joint probability by simulating numerous, say  $M$ , realizations at the  $J$  locations and averaging the probabilities,

$$\text{Prob}[Z(\mathbf{x}_j) \leq z_c, j = 1, 2, \dots, J | n_j] \approx \frac{1}{M} \sum_{m=1}^M \prod_{j=1}^J \omega_m(\mathbf{x}_j; z_c), \quad (12.6)$$

where  $\omega_m(\mathbf{x}_j; z_c)$  is an indicator taking the value 1 if the simulated value of  $z$  is less than or equal to  $z_c$ , and 0 otherwise. Goovaerts (1997) deals with the matter at length.

The simulation may be ‘unconditional’, meaning that we place no constraints on the values generated other than they should have the mean and variogram that we specify. Alternatively, we may ‘condition’ the simulation to return the known values at sampling points in addition.

### 12.2.1 Unconditional simulation

Unconditional simulation is simply an application of the general Monte Carlo technique whereby values are created with a particular covariance function or variogram. There are several ways of doing it.

Perhaps the simplest to envisage is in two dimensions as follows. Draw values at random and independently of one another from a standard normal distribution and place them at the nodes of a square grid. This will give a set  $y(\mathbf{x})$  in which there is no correlation. Place a circle of diameter  $a$  at each node in turn and average the values inside. This will give a series of autocorrelated means,  $\bar{y}(\mathbf{x})$ . These constitute a realization of a second-order stationary autocorrelated random function  $Z(\mathbf{x})$  which has a circular isotropic variogram with range  $a$  (see Matérn, 1960). The values can be scaled to a desired variance and a mean added to match some reality. If a nugget component is required it can be introduced by drawing further values at random from the distribution, scaling them and adding them to the existing simulation.

The same principle can be used to simulate values on a line, in which case the variogram will be the bounded linear model. The three-dimensional analogue has a spherical variogram. Other combinations of dimensionality and model require more sophisticated techniques (see below).

### 12.2.2 Conditional simulation

In unconditional simulation all we ask is that the result has the correct mean and variogram or covariance function. In conditional simulation, however, the generator must return the data values at places where we know them in addition to creating plausible values of  $Z(\mathbf{x})$  elsewhere. We condition the simulation on the sample data,  $z(\mathbf{x}_i), i = 1, 2, \dots, N$ . Let us denote the conditionally simulated values by  $z_c^*(\mathbf{x}_j), j = 1, 2, \dots, T$ . Where we have data we want the simulated values to be the same:

$$z_c^*(\mathbf{x}_i) = z(\mathbf{x}_i) \quad \text{for all } i = 1, 2, \dots, N. \quad (12.7)$$

Elsewhere  $z_c^*(\mathbf{x})$  may depart from the true but unknown values in accord with the model of spatial dependence adopted.

Consider what happens when we krig  $Z$  at  $\mathbf{x}_0$  where we have no measurement; the true value there,  $z(\mathbf{x}_0)$ , is estimated by  $\hat{Z}(\mathbf{x}_0)$  with an error  $z(\mathbf{x}_0) - \hat{Z}(\mathbf{x}_0)$  which is unknown:

$$z(\mathbf{x}_0) = \hat{Z}(\mathbf{x}_0) + \{z(\mathbf{x}_0) - \hat{Z}(\mathbf{x}_0)\}. \quad (12.8)$$

A characteristic of kriging is that the error is independent of the estimate, i.e.

$$\mathrm{E}[\hat{Z}(\mathbf{y})\{z(\mathbf{x}) - \hat{Z}(\mathbf{x})\}] = 0 \quad \text{for all } \mathbf{x}, \mathbf{y}. \quad (12.9)$$

This feature is used to condition the simulation.

We create a simulated field from the same covariance function or variogram as that of the conditioning data to give values  $z_s^*(\mathbf{x}_j), j = 1, 2, \dots, T$ , that include the sampling points,  $\mathbf{x}_i, i = 1, 2, \dots, N$ . We then krig at  $\mathbf{x}_0$  from the simulated values at the sampling points to give an estimate  $\hat{Z}_s^*(\mathbf{x}_0)$ . Its error,  $z_s^*(\mathbf{x}_0) - \hat{Z}_s^*(\mathbf{x}_0)$ , comes from the same distribution as the kriging error in equation (12.8), yet the two are independent. We can use it to replace the kriging error to give our conditionally simulated value as

$$z_c^*(\mathbf{x}_0) = \hat{Z}(\mathbf{x}_0) + \{z_s^*(\mathbf{x}_0) - \hat{Z}_s^*(\mathbf{x}_0)\}. \quad (12.10)$$

The result has the properties we desire, as below.

1. The simulated values are realizations of a random process with the same expectation as the original:

$$E[Z_s^*(\mathbf{x})] = E[Z(\mathbf{x})] = \mu \quad \text{for all } \mathbf{x}, \quad (12.11)$$

where  $\mu$  is the mean.

2. The simulated values should have the same variogram as the original.
3. At the data points the kriging errors  $z(\mathbf{x}_0) - \hat{Z}(\mathbf{x}_0)$  and  $z_s^*(\mathbf{x}_0) - \hat{Z}_s^*(\mathbf{x}_0)$  are 0, and  $z_c^*(\mathbf{x}_0) = z(\mathbf{x}_0)$ .

Another interesting property of the simulated value at  $\mathbf{x}_0$  is that it is the sum of two independent quantities, namely  $\hat{Z}(\mathbf{x}_0)$  and  $z_s^*(\mathbf{x}_0) - \hat{Z}_s^*(\mathbf{x}_0)$ . The variance of the first is  $E[\{z(\mathbf{x}_0) - \hat{Z}(\mathbf{x}_0)\}^2]$ , but so is the second because we made it so. Consequently,

$$\begin{aligned} E[\{z(\mathbf{x}_0) - Z_c^*(\mathbf{x}_0)\}^2] &= 2E[\{z(\mathbf{x}_0) - \hat{Z}(\mathbf{x}_0)\}^2] \\ &= 2\sigma_K^2(\mathbf{x}_0). \end{aligned} \quad (12.12)$$

The variance of a simulated value is twice that of a kriged value; put another way, kriging is twice as good as conditional simulation at estimation. Therefore, we do not simulate if our purpose is estimation. Conditional simulation is more appropriate than kriging where our interest is in the local variability of the property and too much information would be lost by the smoothing effects of kriging. A suite of conditional simulations also provides a measure of uncertainty about the spatial distribution of the property of interest.

## 12.3 TECHNICALITIES

The simple way of simulating values by averaging data as described above is too restrictive in practice. Typically in environmental science we deal with two dimensions and want to be able to simulate from the models that describe our data best. These include the popular spherical, exponential and power functions, in addition to the circular model and others described in Chapter 5 and

yet others. Analysts have now programmed several simulation techniques that enable practitioners to use these functions. Three are now in common use; they are lower–upper (LU) decomposition, sequential Gaussian simulation and simulated annealing. All three methods can be conditional or unconditional, although LU decomposition is more often used for unconditional simulation. A fourth method, the turning bands method, was once popular, but its disadvantages now outweigh its merits and it has fallen out of favour. We describe the first three in some detail.

### 12.3.1 Lower–upper decomposition

The LU decomposition technique is based on a standard result in matrix theory that any square symmetric positive definite matrix,  $\mathbf{C}$ , can be represented as a lower triangular form such that its upper triangular counterpart is its transpose:

$$\mathbf{C} = \mathbf{L}\mathbf{U} = \mathbf{L}\mathbf{L}^T, \quad (12.13)$$

where  $\mathbf{L}$  and  $\mathbf{U}$  are the lower and upper triangular matrices. The technique is due to Cholesky and is known also as the Cholesky decomposition. All covariance matrices, such as those in kriging systems, are amenable to LU decomposition.

To simulate a field of values unconditionally, we start with a variogram or covariance function of a standard normal variate. For this we compute the covariance matrix for the field,  $\mathbf{C}$ , with elements  $c_{ij}$  for all  $i$  and  $j$ . The matrix  $\mathbf{C}$  is decomposed to obtain  $\mathbf{L}$ . We then create a vector,  $\mathbf{g}$ , of random numbers drawn from a standard normal distribution,  $\mathcal{N}(0, 1)$ . Multiplying  $\mathbf{L}$  by  $\mathbf{g}$  gives the required vector of simulated values:

$$\mathbf{y} = \mathbf{L}\mathbf{g}, \quad (12.14)$$

and

$$E[\mathbf{y}\mathbf{y}^T] = \mathbf{C} = \mathbf{L}\mathbf{U}. \quad (12.15)$$

It is elegant.

For conditional simulation, let there be  $N$  data with which to simulate at  $T$  unsampled positions. Thus, we have to consider  $N + T$  points in total. The symmetric covariance matrix,  $\mathbf{C}$ , is also of dimension  $N + T$ , and it comprises four sub-matrices. It is decomposed as

$$\begin{bmatrix} \mathbf{C}_{NN} & \mathbf{C}_{NT} \\ \mathbf{C}_{TN} & \mathbf{C}_{TT} \end{bmatrix} = \begin{bmatrix} \mathbf{L}_{NN} & 0 \\ \mathbf{L}_{TN} & \mathbf{L}_{TT} \end{bmatrix} \begin{bmatrix} \mathbf{U}_{NN} & \mathbf{U}_{NT} \\ 0 & \mathbf{U}_{TT} \end{bmatrix}. \quad (12.16)$$

The values of  $T$  are drawn independently and at random from a standard normal distribution as above to give the vector  $\mathbf{g}$ . In addition, the  $N$  conditioning data are transformed as necessary to standard normal form and are denoted by the vector  $\mathbf{z}$ . The vector of conditionally simulated values,  $\mathbf{y}$ , of length  $N + T$  is

$$\mathbf{y} = \begin{bmatrix} \mathbf{z}_N \\ \mathbf{L}_{TN}\mathbf{L}_{NN}^{-1} + \mathbf{L}_{TT}\mathbf{g}_T \end{bmatrix}. \quad (12.17)$$

The resulting vector can be transformed back to the original scale.

The LU technique is neat and readily programmed to take advantage of efficient numerical library subroutines. Its major disadvantage is that it becomes computationally impracticable for many points because matrix  $\mathbf{C}$  must be held in memory and it becomes too large to decompose. This limits the number of sites to about 1000 ( $N + T$ ), but this could well increase as computer memory grows. However, if one wants many realizations for a small field of values it is very fast because matrix  $\mathbf{C}$  has to be decomposed only once. All that has to be done is to generate more vectors of random numbers drawn from a standard normal distribution.

### 12.3.2 Sequential Gaussian simulation

The sequential approach is the most straightforward method for simulating a multivariate Gaussian field. Each value is simulated sequentially according to its normal conditional cumulative distribution function, which must be determined at each location to be simulated. The conditioning data comprise all the original data and all previously simulated values within the neighbourhood of the point being simulated.

Sequential Gaussian simulation starts with the assumption that the kriging error is normally distributed with mean 0 and variance  $\sigma_K^2(\mathbf{x}_0)$ , i.e.  $\mathcal{N}(0, \sigma_K^2(\mathbf{x}_0))$ . In these circumstances the probability distribution for the true values is  $\mathcal{N}(\hat{Z}(\mathbf{x}_0), \sigma_K^2(\mathbf{x}_0))$ ; it is simply shifted by  $\hat{Z}(\mathbf{x}_0)$ .

To implement the technique the following are the steps needed.

1. Ensure that the data are approximately normal; transform to a standard normal distribution if necessary.
2. Compute and model the variogram.
3. Specify the coordinates of the points at which you want to simulate. These will usually be on a grid.
4. Determine the sequence in which the points,  $\mathbf{x}_j, j = 1, 2, \dots$ , will be visited for the simulation. Choosing the points at random will maximize the diversity of different realizations.

5. Simulate at each of these points as follows.
  - (a) Use simple kriging with the variogram model to obtain  $\hat{Z}(\mathbf{x}_i)$  and  $\sigma_K^2(\mathbf{x}_i)$ .
  - (b) Draw a value at random from a normal distribution  $\mathcal{N}(\hat{Z}(\mathbf{x}_i), \sigma_K^2(\mathbf{x}_i))$ .
  - (c) Insert this value into the grid at  $\mathbf{x}_i$ , and add it to the data.
  - (d) Proceed to the next node and simulate the value at this point in the grid.
  - (e) Repeat steps (a) to (c) until all of the nodes have been simulated.
6. Back-transform the simulated values if there is a need to.

### 12.3.3 Simulated annealing

Simulated annealing is a generic term for a set of algorithms that optimize rather than strictly simulate. The method is based on the general principle of stochastic relaxation described by Kirkpatrick *et al.* (1983), which Geman and Geman (1984) showed could be used to process and restore images. The concept derives from the way a molten metal cools. When the metal cools rapidly it solidifies to a more or less disordered state comprising many small crystals. If the solid is then heated for a long time and allowed to cool slowly the molecules in it rearrange themselves into larger crystals in which the free energy is less. This is the process of annealing. Deutsch and Journel (1992) introduced simulated annealing into geostatistics for creating random fields with specific characteristics. In geostatistics the values of a regionalized variable are equivalent to the molecules. These values can be moved around or replaced by the method so as to minimize some objective function that measures the deviation between the target and present characteristics of the realization at each  $i$ th perturbation of the data. The objective function embodied in Deutsch and Journel's (1992) GSLIB is to reproduce the variogram model,  $G$ ,

$$G_i = \sum_{m=1}^M \{\hat{\gamma}_i(\mathbf{h})_m - \gamma(\mathbf{h})_m\}^2. \quad (12.18)$$

In this equation the  $\gamma(\mathbf{h})_m, m = 1, 2, \dots, M$ , are the values of the empirical model, and the  $\hat{\gamma}_i(\mathbf{h})_m$  are the values computed for the current realization on the full grid and give the value  $G$ . The quantity  $M$  defines the limit within which  $G$  is to be computed. If the variation is isotropic then  $M$  is the maximum number of intervals on the grid.

The steps in simulated annealing are as follows.

1. The process starts with data, observed values of  $Z(\mathbf{x})$  for which we have an empirical model of the variogram  $\gamma(\mathbf{h})$ , and the simulation is conditioned on those data. The process generates additional values on a fine grid.

Ideally the data themselves should occupy nodes on this grid, but if any do not then in most algorithms they are moved to their nearest grid nodes. The unoccupied nodes of the grid are assigned values drawn at random from the same frequency distribution as the data.

2. Compute the initial value of the objective function from the initial realization

$$G(0) = \sum_{m=1}^M \{\hat{\gamma}_0(\mathbf{h})_m - \gamma(\mathbf{h})_m\}^2. \quad (12.19)$$

Typically  $\hat{\gamma}_0(\mathbf{h})_m$  for the initial realization will appear flat because the values were drawn independently of one another at the originally unoccupied nodes.

3. Perturb the realization by *swapping* pairs of values, as in the first version of GSLIB (Deutsch and Journel, 1992) or by *replacement*, as in the second version of the program (Deutsch and Journel, 1998).

*Swapping.* Values at two nodes, say  $z(\mathbf{x}_i)$  and  $z(\mathbf{x}_j)$ , are chosen at random and swapped, and  $G$  is recomputed. If  $G$  is diminished it means that the new experimental variogram,  $\hat{\gamma}_i(\mathbf{h})$ , is closer to  $\gamma(\mathbf{h})$  than the first, and so the swap is retained. This is analogous to a molecular rearrangement in annealing that results in a decrease in the Gibbs free energy. The process continues with a swap of two more values, the recalculation of  $G$ , and the retention of the swap if  $G$  is diminished.

*Replacement.* The value at a randomly chosen node is replaced by another value drawn at random from the same distribution as that of the original values, and  $G$  is recomputed. As in the swapping mode, the new value is retained if  $G$  is diminished.

4. The process ends when either  $G$  has become sufficiently small or the number of swaps has reached some preset limit.

Although there is little that is geostatistical in the process, the final outcome depends on (a) the initial random selection of values at the unoccupied nodes of the grid and (b) the random choice of pairs to be swapped or choice of grid node at which the value is to be replaced and the value that replaces it. There are therefore very many possible final outcomes with the desired variogram, and the one we obtain can therefore still be regarded as a realization of a random function.

The broad strategy is outlined above, but there are additional details to observe.

1. The observed values or conditioning data on the grid are never swapped or replaced.
2. Equation (12.18) gives equal weight to all the semivariances. Yet we know in general that those near the ordinate, i.e. those at short lag distances, are

more reliable than those further away. If we divide the squared differences by the squares of the model values then we shall give greater weight to the comparisons near the ordinate, thus:

$$G_i = \sum_{m=1}^M \frac{\{\hat{\gamma}_i(\mathbf{h})_m - \gamma(\mathbf{h})_m\}^2}{\gamma^2(\mathbf{h})_m}. \quad (12.20)$$

3. If we reject all swaps or replacements that fail to diminish  $G$  the result might be a local optimum far from the global one that we desire. So, some swaps that fail the above test are retained. Therefore, we have to refine our rule for accepting swaps, and we do so again with reference to physical annealing. The frequency with which apparently unfavourable swaps are retained is set to be proportional to

$$\exp\left(\frac{G_{\text{old}} - G_{\text{new}}}{t}\right).$$

The value of  $t$  is set large at the start and then decreased slowly, and in this way convergence to a local minimum is avoided. This quantity  $t$  mimics the temperature in the Boltzmann distribution, which decreases as a metal anneals. The precise steps by which  $t$  is diminished are found to some extent by trial and error. Goovaerts (1997) suggests that you start with a large  $t$  such that many apparently unfavourable swaps are retained. You then decrease  $t$  by some common factor  $\beta$ ,  $\beta < 1$ , when you have accepted enough swaps or too many have been tried (Farmer, 1991; Press *et al.*, 1986). The maximum number of accepted or attempted swaps is chosen as some multiple of the total number of grid nodes,  $T$ .

Finally, bear in mind that simulated annealing generates fields by making the experimental variograms converge to the input models. Therefore, this method should not be used to study fluctuations arising in the generating process.

### **12.3.4 Simulation by turning bands**

The method known as ‘Turning bands’, due to Matheron (1973) and Journel (1974), was the earliest for simulating autocorrelated random processes in three dimensions,  $\mathbb{R}^3$ . It involves first simulating independent one-dimensional realizations along lines radiating from a central point in the volume of interest. Then each point in the three-dimensional space for which a value is required is projected orthogonally on to every line, and the values at the nearest points to the projections are averaged.

Crucially, the one-dimensional covariance function must be known,  $C^1(h)$ , corresponding to that in three dimensions,  $C^3(\mathbf{h})$ . These are easily obtained for

the common three-dimensional functions such as the spherical and exponential models. Finding the correct one-dimensional functions corresponding to them in two dimensions,  $C^2(\mathbf{h})$ , turns out to be much more complex. This is presumably why most software packages do not include the turning bands method for  $\mathbb{R}^2$ . Even more worrying is that some of those that do include the method produce patterns of values in which the bands are plainly evident; see, for example, Figure V.8 on page 148 in Deutsch and Journel (1998). Such results are unacceptable.

Partly for the above reasons and partly because sequential Gaussian simulation and simulated annealing have proved so successful, the turning bands method has lost favour. We do not devote further attention to it therefore. Olea (1999) provides a detailed exposé of the method for those who are interested.

### 12.3.5 Algorithms

The above algorithms are expanded with proofs in Olea (1999). Goovaerts (1997) also describes them and illustrates their application with the data of Atteia *et al.* (1994) on soil of the Swiss Jura. The LU decomposition can be programmed readily in GenStat, MATLAB and S-Plus, for example, although it is also available in GSLIB (Deutsch and Journel, 1998), as are sequential Gaussian simulation and simulated annealing. The turning bands method was in the first edition only of GSLIB (Deutsch and Journel, 1992) and it is for three dimensions in that library.

## 12.4 USES OF SIMULATED FIELDS

As above, simulation is not a substitute for estimation; that is not its purpose. What it does do is give us pictures of the variation to expect between sampling points as distinct from the smoothed form provided by kriging. By conditioning the simulation on data we ensure that the fields generated do not stray far from reality.

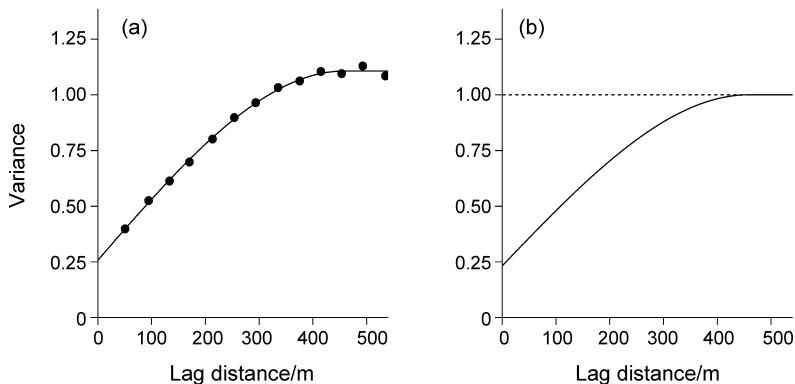
Unconditional and conditional simulation can give us dense fields of values on which we can study dispersion and sampling fluctuation, and from which we can construct confidence intervals on estimates of the variogram, as in the examples in Chapter 5.

We have already mentioned that repeated simulations enable us to judge the probability that a variable exceeds a threshold at two or more places in a neighbourhood. This is valuable for the delimitation of zones of pollution (for examples, see Goovaerts, 1997; Fabbri and Trevisani, 2005) and for estimating the travel time of water and solute through an aquifer which depends on the joint distribution of transmittivities (see Gotway, 1994).

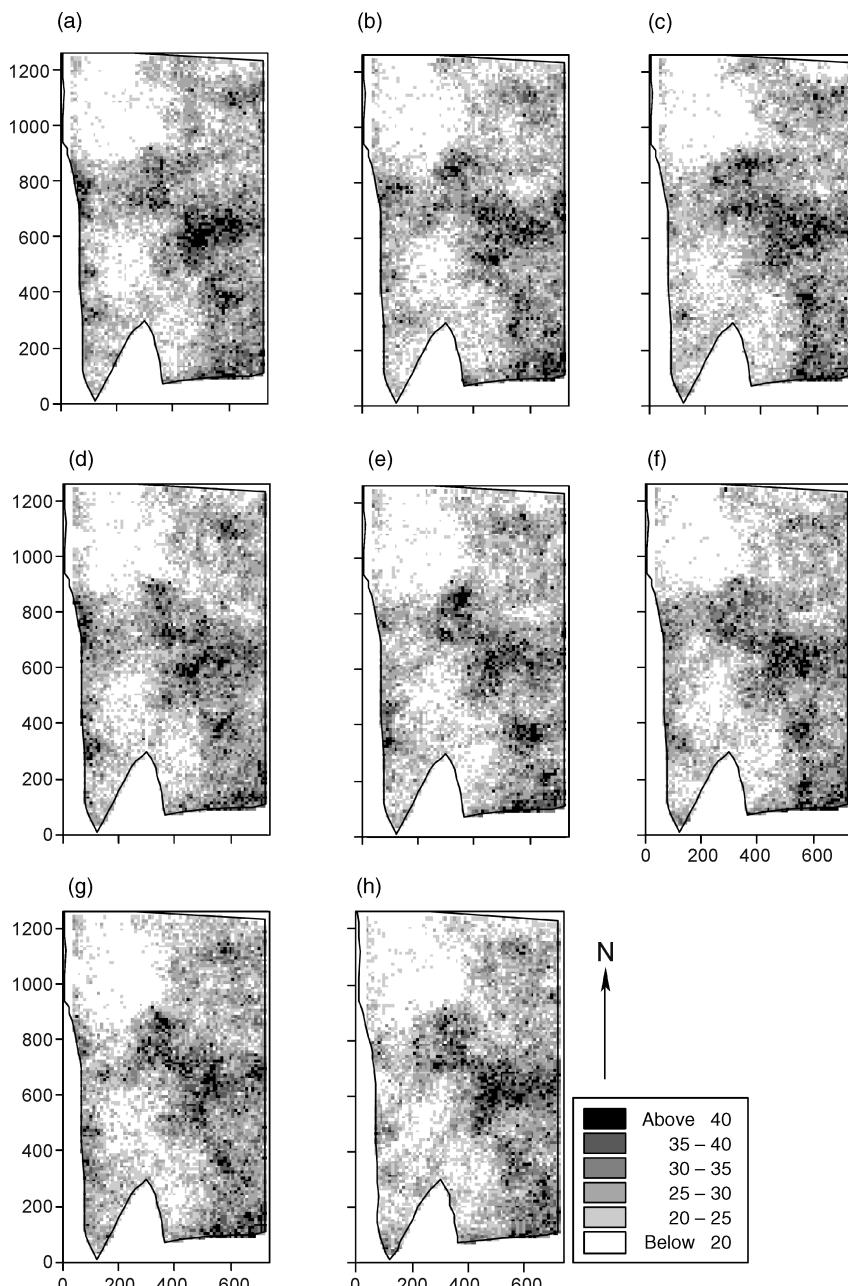
## 12.5 ILLUSTRATION

Figures 5.5 and 5.6 are examples of random fields produced by unconditional simulation. We illustrate the results of conditioning by returning to the case study at Broom's Barn Farm.

The potassium data for Broom's Barn Farm were transformed first to normal scores and the variogram computed on the transformed values. A spherical function fitted the experimental values with  $c_0 = 0.2536$ ,  $c = 0.8410$  and  $a = 458$  m; it is shown in Figure 12.1(a). The GSLIB (Deutsch and Journel, 1998) simulation program assumes that the sill variance of the normal score transform will be unity, and so the above model parameters were adjusted proportionately to  $c_0 = 0.2326$  and  $c = 0.7674$  to ensure this; see Figure 12.1(b). The latter was used to generate values by sequential Gaussian simulation on a 10 m  $\times$  10 m grid with the normal scores of K at 40-m intervals. Eight simulations were done with a unique random number seed each time, and the simulated normal scores were transformed back to the original scales. Their means and variances were calculated (Table 12.1), and are close to those of the data. Figure 12.2 shows the maps of the eight fields. There are differences in the local detail, but they reflect the general pattern of variation shown in Figure 8.22. In particular, there is more local variation in the maps of the simulated fields than in the kriged maps. Figure 12.3 shows the corresponding experimental variograms plotted as points at 40-m intervals. There are small differences among them for the individual fields, but all lie close to the generating function. We fitted spherical functions to each individually,



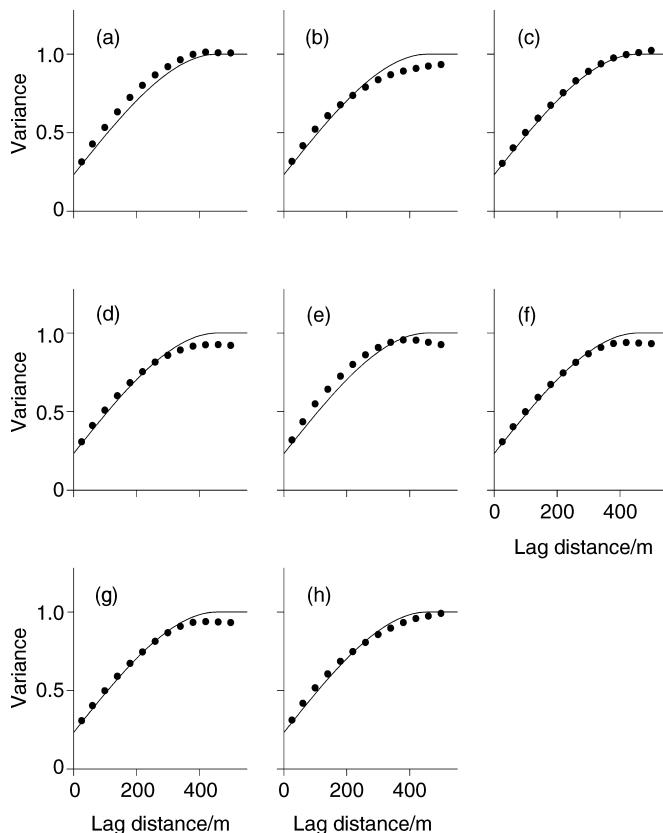
**Figure 12.1** (a) Experimental variogram of the normal scores of K at Broom's Barn and the fitted function; (b) the variogram rescaled to a sill of 1 and used for the sequential Gaussian simulation.



**Figure 12.2** Maps of eight fields of values produced by sequential Gaussian simulation for K at Broom's Barn.

**Table 12.1** Means, variances and standard deviations for eight fields simulated by sequential Gaussian simulation conditioned by the potassium data from Broom's Barn Farm.

Simulation	Mean	Variance	Standard deviation
1	25.56	70.76	8.412
2	26.25	62.13	7.882
3	26.27	68.81	8.295
4	26.43	59.34	7.703
5	25.91	63.79	7.987
6	26.33	61.96	7.872
7	26.50	59.85	7.736
8	25.74	58.94	7.677
Raw data	26.31	81.71	9.039



**Figure 12.3** Experimental variograms for the eight sequential Gaussian simulated fields plotted as points with the generating variogram shown in Figure 12.5(b) added to each as the solid line.

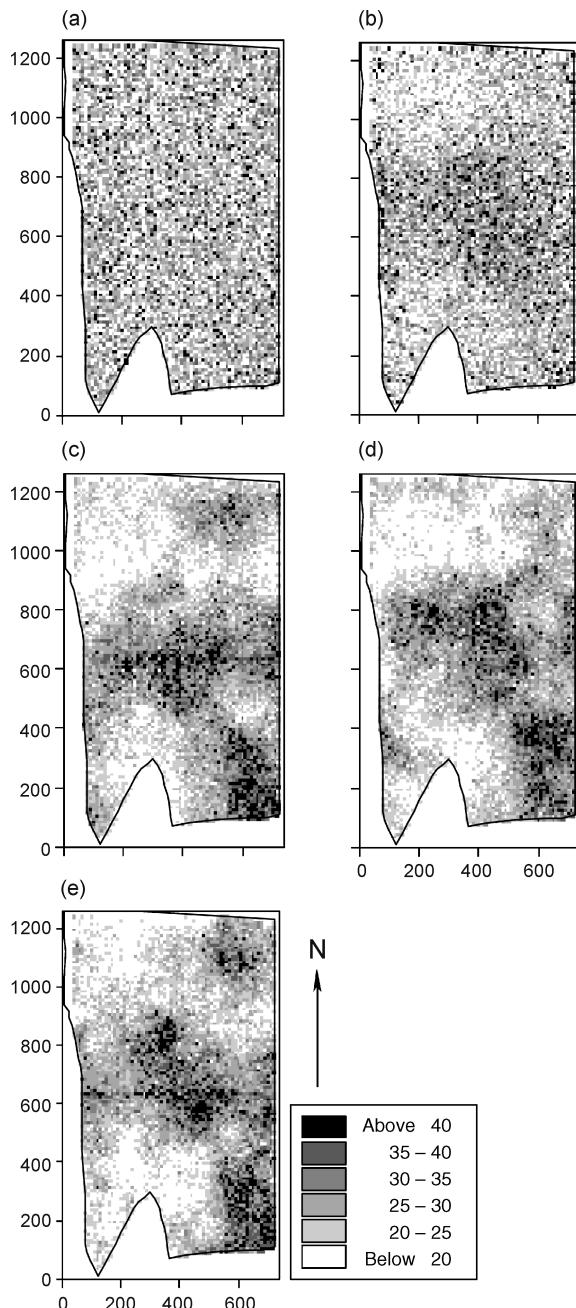
**Table 12.2** Model parameters for the spherical function fitted to experimental variograms from eight fields simulated by sequential Gaussian simulation conditioned by the potassium data standardized to mean 0 and variance 1 from Broom's Barn Farm.

Simulation	$c_0$	$c$	$a/\text{m}$	Residual Mean Square
1	0.2660	0.7406	412.7	48.68
2	0.2942	0.6209	420.0	227.7
3	0.2546	0.7578	459.6	43.11
4	0.2618	0.6588	395.6	44.80
5	0.2691	0.6744	364.3	151.5
6	0.2635	0.6966	418.8	60.05
7	0.2523	0.6840	409.3	12.01
8	0.2899	0.6818	455.2	186.0
Generator	0.2536	0.8410	454.0	

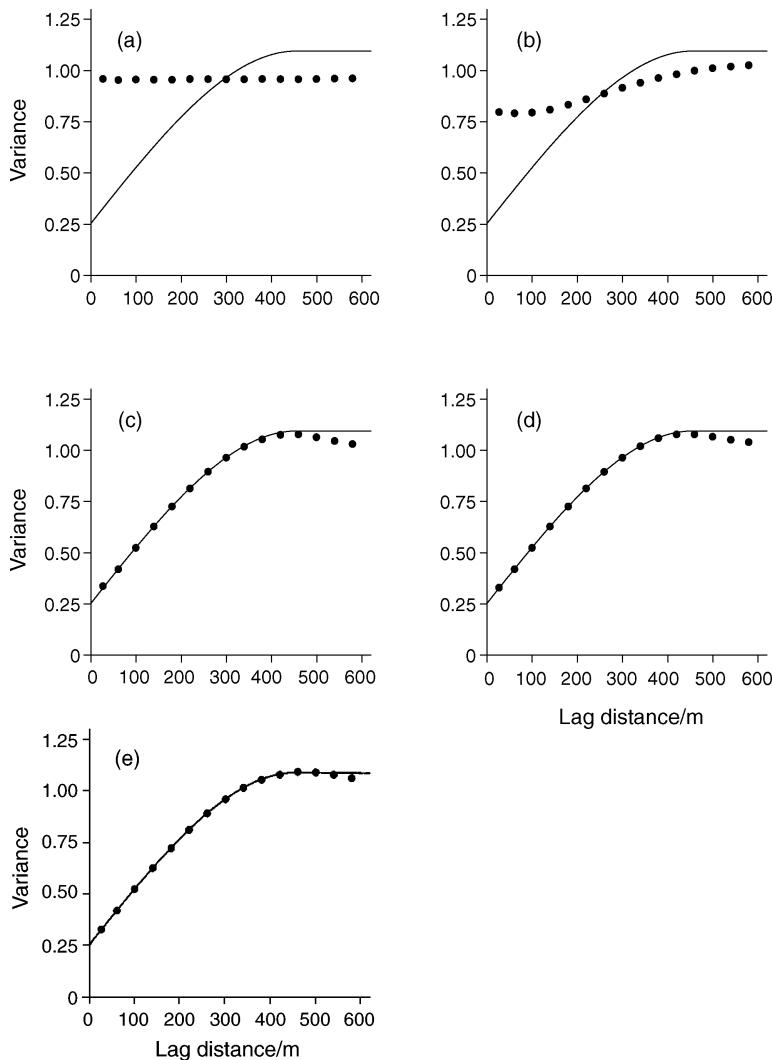
and Table 12.2 gives their model parameters. There are small differences in the nugget variances, but somewhat larger differences in the ranges from the generating function.

We also illustrate results using simulated annealing on the same grid and conditioned on the same normal scores of K. Figure 12.4 shows the stages in the annealing process. The first map, Figure 12.4(a), includes the data and additional values drawn from a normal distribution with the same mean and variance as the data. The field appears to lack any pattern, and its experimental variogram, shown in Figure 12.5(a), is effectively flat. The objective function at this stage, the beginning, is 1. Figure 12.4(b) shows the results after 65 000 swaps, at which stage the objective function had decreased from 1 to 0.5289. A broad pattern is beginning to develop and is just detectable. The variogram of the field, presented in Figure 12.5(b), shows only weak structure at this stage, which manifests itself as long-range autocorrelation. The experimental variogram is still fairly flat near the ordinate, however. After approximately 190 000 swaps the objective function reached 0.0004, and the experimental variogram is close to the theoretical curve, as is apparent in Figure 12.5(c). The main features of the spatial pattern are evident in Figure 12.4(c), and this map resembles those from sequential Gaussian simulation (Figure 12.2) with the same broad features as those from block kriging (Figure 8.22). Nevertheless, it is not as similar to the latter as are those in Figure 12.2. The maps in Figure 12.4(d) and (e) are for two more simulations that converged after some 195 000 swaps and an objective function of 0.0002. Figure 12.5(d) and (e) shows their variograms.

The maps in Figure 12(c)–(e) are clearly different from one another and seem to vary more than do those for sequential Gaussian simulation; there are more



**Figure 12.4** Stages in the annealing process: (a) the data plus additional values drawn from a normal distribution; (b) the result after 65 000 swaps; (c), (d) and (e) the results after some 190 000 swaps and convergence of the experimental variogram to the generator, Figure 12.1(a).



**Figure 12.5** Experimental variograms of the fields from the simulated annealing in Figure 12.4: (a) of the initial field; (b) after 65 000 swaps; (c), (d) and (e) after convergence with some 190 000 swaps. The solid line in each graph is the variogram, shown in Figure 12.1(a), of the function to which the annealing converges.

differences among them in the detail. Clearly, they all show more local detail than appears in the kriged map.

# ***Appendix A***

## ***Aide-mémoire for Spatial Analysis***

### **A.1 INTRODUCTION**

This appendix summarizes the steps that a scientist should take in a geostatistical analysis of survey data, beginning with error detection, summary statistics, exploratory data analysis, the variogram and its modelling, kriging, and mapping. In many instances data from remote imaging require the same treatment, and where that is so we mention it.

### **A.2 NOTATION**

The notation is the same as used in the main text, but we repeat it here for completeness. The geographic coordinates of the sampling points and target points for prediction are denoted  $x_1$  for eastings (or across the map from left to right) and  $x_2$  for northings (or from bottom to top on the map). The pair  $\{x_1, x_2\}$  are given the symbol  $\mathbf{x}$  in vector notation. The variates are denoted  $z_1, z_2, \dots$ , and the measured values are denoted  $z(\mathbf{x}_i)$  for  $i = 1, 2, \dots$ , for any one variate.

### **A.3 SCREENING**

Few large files of data are free of mistakes caused by instrumental malfunction and human error. When you receive data, whether from the field or laboratory or from remote scanners, check for such mistakes.

**Position.** Examine the positions of the data in relation to the bounds of the region. Plot them on a map, known as a ‘posting’.

- Do all the points lie within the region?
- Are there sampling points in the sea when they should be on the land? If so why?
- Have the coordinates been reversed inadvertently so that northings precede eastings, i.e. the  $x_2$  precede the  $x_1$ ?
- Do the points approximately fill the region?
- Are the coordinates properly scaled?

**Measurements.** Screen the measured values,  $z(x_1), z(x_2), \dots$ . Pass them through a program that compares each value in the file against the minimum and maximum possible of the scale and flags any that lie outside these bounds. Print out the minimum and maximum for each  $z$  and check that they are sensible.

## A.4 HISTOGRAM AND SUMMARY

For each variate compute a histogram and plot it, ensuring that all classes are of the same width. Examine it for outliers, i.e. individuals or small groups that are isolated from the main body of data. In addition, if you prefer box-plots, compute them for the  $z$ s, to show outliers.

**Outliers.** If there are outliers identify their positions on the map.

- Are they mistakes? If not, what do they represent?
- Are they part of the ‘target population’?

If not (e.g. water when you are interested only in land) then replace the recorded values by a symbol to indicate missing or not applicable.

The statistical treatment of outliers is a complex subject, and if you wish to retain outliers in your analysis then consult a statistician.

**Frequency distribution.** Study the shape of the histogram.

- Has it more than one peak?

If so, the scene or region almost certainly contains at least two distinct populations, e.g. land and water, farmland and forest.

- Is the distribution symmetric?
- If it is skewed is the longer tail towards the small values or towards the large?

**Summary.** Summarize the statistics for each variate by computing:

- the number of sampling points and the number of valid values;
- the minimum and maximum;
- the mean;

- the median;
- the variance;
- the standard deviation (the square root of the variance);
- the coefficient of variation (optional);
- the skewness (coefficient  $g_1$ ); and
- the kurtosis ( $g_2$ , optional).

## A.5 NORMALITY AND TRANSFORMATION

Geostatistical analysis is most efficient when done on variables that have normal, or Gaussian, distributions. Some analyses assume normality. You should therefore examine the form of the distribution of each  $z$ .

**Symmetric histogram.** If the frequency distribution appears symmetric, with a single central peak, try fitting a normal curve to it. If the fit ‘looks good’ then accept the variate as normal. If not, in what way does it depart from normal? For example, is it flat-topped, or light in the tails? These features may be matched with the coefficient of kurtosis,  $g_2$ . A flat-topped distribution suggests that you have more than one population in the image or region—see multiple peaks (Frequency distribution, Section A.4).

**Skewed histogram.** Asymmetry is the most common form of departure from normality, and in particular positive skewness (long upper tail, coefficient  $g_1 > 0$ ). In these circumstances the variance is likely to change from one part of the image or region to another, thereby violating one of the assumptions of stationarity on which analysis is usually based. Consider transforming the recorded  $z$  to stabilize the variance. Options are as follows.

- Skew positive,  $0 < g_1 < 0.5$ . Do not transform.
- Skew positive,  $0.5 < g_1 < 1$ . Consider transformation to square roots, i.e.  $y = \sqrt{z}$ .
- Skew positive,  $g_1 > 1$ . Transform. Try logarithmic transformation first, i.e.  $y = \ln z$  or  $y = \log_{10} z$ . Examine the resultant distribution. If it is approximately normal then accept it. If the result is still skewed then try subtracting a positional constant,  $a$ , so that  $y = \ln(z - a)$ .

You might be able to find a suitable value for  $a$  fitting the two- and three-parameter lognormal functions to the  $z$ .

Other transformations are available if these prove unsatisfactory.

Significance tests for normality are available. *Disregard them when analysing images!* With many pixels you will almost surely discover that the distributions are ‘significantly’ non-normal. They can be helpful if you have only 100 or so measurements from ground survey.

## A.6 SPATIAL DISTRIBUTION

Explore the spatial distribution of each  $z$ . Here you might need to treat image data differently from ground data.

**Ground data.** Make an isarithmic ('contour') map using a reputable program with a well-behaved algorithm for interpolation, such as inverse squared distance weighting or simple bilinear interpolation if the data are dense, and layer shading to indicate the magnitude of  $z$ , or  $y = f(z)$ .

If the data are on a grid then compute the row and column means.

Alternatively, find the medians of each row and column.

Is there any trend in them?

**Images.** If you are analysing images then map the distribution of pixel values using a computer program that will show the individual pixels coloured, or shaded grey, on a scale according to recorded values, or transformed to  $y = f(z)$ . Compute the row and column means or medians as for gridded ground data.

Examine either kind of map for trends and patches.

**Trend.** Is there any evident long-range trend over the scene or region?

- If so, what is its form and principal direction?

**Patches.** Are there patches?

- If so, how big are they on average?
- Are they isotropic?
- If not in which direction are they elongated?

Long-range trend is incompatible with the assumptions of stationarity on which most geostatistical analysis is based. If the trend is strong then consider removing it by some kind of filter, such as a global trend surface, before proceeding further.

Alternatively, adopt a model for  $z$  that incorporates the non-stationary trend. This will take you into more advanced technique, and you should consult a specialist about it.

## A.7 SPATIAL ANALYSIS: THE VARIOGRAM

The variogram summarizes the spatial distribution of  $z$  in the absence of trend. Three variograms are to be distinguished: the experimental variogram; the regional variogram; and the theoretical variogram.

**The experimental variogram.** This is the variogram that you compute from the data,  $z(\mathbf{x}_i)$ ,  $i = 1, 2, \dots$ . For ground data and images on regular

rectangular grids compute the semivariances separately along the rows and columns of the grid, incrementing the lag by one sampling interval or pixel at a time, and along the principal diagonals of the grid at intervals of  $\sqrt{2}$  sampling intervals.

For irregularly scattered data, and provided you have sufficient (several hundred), compute a variogram in four or more directions by discretizing the lag by both distance and direction. Compute also the variogram ignoring direction, i.e. with lag in distance only.

Plot the results as variance against lag distance with a unique symbol for each direction. Identify the main features, as follows.

**Anisotropy.** Does the variogram have approximately the same form and values in all directions? If so, then accept it as isotropic and compute an average experimental variogram over all the directions. If not, then in what way do the directions differ?

**Different spatial scale.** This indicates geometric anisotropy, which might be removed by a simple transformation of the spatial coordinates.

**Different semivariances.** These indicate ‘zonal’ anisotropy—there is simply more variance in some directions than in others.

**Different form.** Look especially for contrasts between convex (decreasing gradient with increasing lag distance) and concave (increasing gradient). This suggests trend in the direction of increasing gradient, and it should be compared with the evidence from the exploratory analysis, above.

**Bounds.** Does the variogram appear bounded, i.e. does the semivariance reach a maximum within the distance computed or appear as though it would reach a maximum if the lag distance was extended somewhat (bounded)? Alternatively, does it look as though it would increase without limit (unbounded)?

**Nugget.** Does an imaginary line drawn through the experimental values when projected cut the ordinate at a positive value (not 0)? If so, this intercept is known as the nugget variance.

**The regional variogram.** This is the variogram that you would compute if you had complete information in the region. It is approximated by the experimental variogram.

**The theoretical variogram.** This is the variogram of the process that you must imagine generated the field of which the measured data or pixels are a sample.

To proceed further you must fit a mathematical function to the experimental variogram as a model or approximation to the theoretical variogram (see below).

## A.8 MODELLING THE VARIOGRAM

1. Match the form of the experimental variogram with those of the common simple valid models for variance in two dimensions. Choose several that appear to have the right form.
2. Fit each of these models in turn using a numerically sound and well-tried program by minimizing a weighted least-squares criterion. Choose weights in proportion to the number of paired comparisons in the experimental values, and set approximate starting values for the non-linear parameters. Tabulate the residual sum of squares and residual mean square as criteria. You may use a more elaborate scheme of weights such as one of those mentioned in Chapter 6 if you wish to model the variogram better near the ordinate.
3. Select the function for which the criteria are least. Plot the fitted function on the same pair of axes as the experimental semivariances. Does the fit appear good on the graph? If not, inspect another. If none appear to fit well, then consider fitting a more complex model by combining two or more simple models from the standard repertoire, and repeat the process.

In principle, you can always improve the fit of a model by making it more complex, i.e. by increasing the number of parameters in it. To compare functions with different numbers of parameters, calculate the Akaike information criterion (AIC), and choose the model for which the AIC is least. This trades simplicity against goodness of fit. The AIC is defined as

$$\text{AIC} = -2 \ln(\text{maximized likelihood}) + 2 \times (\text{number of parameters}).$$

For any given experimental variogram it has a variable part:

$$\hat{A} = n \ln R + 2p,$$

where  $n$  is the number of experimental values,  $R$  is the mean squared residual, and  $p$  is the number of parameters.

Least-squares fitting minimizes  $R$ , but if  $R$  is diminished further only by an increase in  $p$  ( $n$  is constant) then there is a penalty, which might be too big.

4. Check that the models that appear to fit accord with prior knowledge. If they do not then investigate further. You might need to shorten the interval between successive lags, narrow the angular discretization, or extend the maximum distance over which you compute the experimental variogram. You might need to try fitting other models.
5. Tabulate the parameters of the final best model and any others that are almost equally good. You will need the parameters for kriging (below).

## A.9 SPATIAL ESTIMATION OR PREDICTION: KRIGING

The aim is to estimate or predict in a spatial sense the values of  $z$  at unsampled places, or ‘targets’, from the data. For images such targets are likely only where there are gaps in a scene. Ordinary kriging smoothes, however, and you might choose to use it to remove short-range noise in the image so that you can see a more general pattern. For ground surveys they are commonplace, and in this section ground survey is assumed. Further, ordinary kriging of  $z$  (or  $y = \ln z$  for lognormal kriging) is likely to serve in 90% of cases, and only this is covered.

You will need the original data and a legitimate model of the variogram. You now have several choices before you.

**Punctual or block kriging.** The targets may be points, say  $\mathbf{x}_0$ , in which case the technique is punctual kriging. Alternatively, they may be small blocks,  $B$ , which may be of any reasonable size and shape but are usually square; this is block kriging. The size of block should be determined by the application: what size of block does the user of the predictions want? It should not be determined by the data or the cosmetics of mapping (see below).

**Number of data points.** Ordinary kriging computes a weighted average of the data. The weights are determined by the configuration of the data in relation to the target in combination with the variogram model. They do not depend on the measured values, the  $z(\mathbf{x}_i)$ . Unless the model has a large proportion of nugget variance only the nearest few sampling points carry appreciable weight; more distant points have negligible weight. So kriging is local.

Take the nearest 20 points to the target. If the data points are exceptionally unevenly scattered then take the nearest two or three points in each octant around the target.

Form the kriging equations, and solve them to obtain the weights, the predicted values and the prediction variances (kriging variances).

If you are uncertain how many points to take then experiment with numbers between 4 and 40 and plot their positions in relation to the target and their weights. Do not be alarmed if some weights are negative, provided they are fairly close to 0.

**Transformation.** For lognormal kriging the data must be transformed to  $y = \ln z$  or  $y = \log_{10}z$ , and the variogram model must be of  $y$ . If you want estimates to be of  $z$  then you must transform the predicted  $y$  back to  $z$ .

**Kriging for mapping.** Krige at the nodes of a fine square grid. Write the kriged estimates and kriging variances to a file. For an isarithmic display the interval of the grid should be chosen such that it is no more than 2 mm on the final hard copy. The optimality of kriging will not then be noticeably degraded by non-optimal interpolation in the graphics program.

The grid interval need not be related to the block size if you block-krige. The blocks may overlap, or there may be gaps between them.

## **A.10 MAPPING**

Pass the file of kriged estimates and variances to a graphics program for the final display of the results as isarithms or small square cells. Choose colours or grey levels to represent the magnitude of the estimates and variances, as above.

Do not use graphics programs or geographic information systems for geostatistics unless you are in complete control and you know that they do exactly what you want.

# ***Appendix B***

## ***GenStat Instructions for Analysis***

The analyses summarized in Appendix A can be done in GenStat, the latest version of which is release 9 (Payne, 2006), for which we give commands below. The data used as the example are of the exchangeable potassium (K) in the soil of a 80-ha farm (Broom's Barn) in Eastern England. The farm was sampled at 40-m intervals on a square grid, and bulked cores of soil to 20 cm were taken and analysed in the laboratory to give 435 values for each variable.

The measured variable ( $z$ ) is here denoted by  $z$ , and the spatial coordinates ( $\{x_1, x_2\}$ ) by  $x$  and  $y$  in units of 40 m. Unless otherwise defined, variables are vectors, or in the GenStat language, *variates*.

### **B.1 SUMMARY STATISTICS**

GenStat enables you to obtain a statistical summary readily by means of standard functions:

```
calculate zbar=mean(z)
calculate zmed=median(z)
calculate zmax=maximum(z)
calculate zmin=minimum(z)
calculate zmed=median(z)
calculate zvar=var(z)
```

and

```
calculate zsdev=sqrt(zvar)
```

These and other summary statistics can be obtained alternatively with the GenStat procedure *describe*; thus

```
describe [selection=nobs, mean, median, min, max,\  
range, var, sd, skew, kurtosis] z
```

in which nobs is the number of non-missing observations and the other options are evident from their names.

## B.2 HISTOGRAM

The histogram of z is formed simply by the command

```
histogram [title=!t('Potassium')] z
```

for a device such as a line printer, and by

```
dhistogram [title=!t('Potassium')] z
```

for a high-quality graph. The title within the square brackets is an option.

You are likely to want to specify the limits to the classes, or 'bins' in statistical jargon. So define a variate containing them:

```
variate [values=10,15...100] binlims
```

Then write

```
dhistogram [limits=binlims; title=!t('Potassium')] z
```

If you want to see what the frequency distribution looks like on the logarithmic scale then z is readily transformed by

```
calculate lz=log10(z)
```

and you can then replace z by lz in the above commands.

## B.3 CUMULATIVE DISTRIBUTION

The cumulative distribution of z can be formed by the following set of commands

```
calculate az=sort(z)  
calculate cz=cum(az)  
calculate nz=nobservations(z)  
calculate pz=(!(1...nz)-0.5)/nz
```

in which sort assembles the values in z in order from smallest to largest, and cz contains the accumulated sum.

To draw a graph of the cumulative distribution you can write the command

```
dgraph x=az; y=pz
```

Further, to show it on a normal probability scale you can convert  $pz$  to ‘normal equivalent deviates’, as follows:

```
calculate nd=ned(pz)
dgraph x=az; y=nd
```

The normal equivalent deviate is the area beneath the standard normal curve of the probability density from  $-\infty$  to  $G(z)$ .

## B.4 POSTING

You can plot the data as a posting as follows. You should assemble the outline of the region of interest as pairs of coordinates in two variates, say  $ox$  and  $oy$ . Then you can write

```
pen 1; linestyle=0; method=point; symbols=4
pen 2; linestyle=1; method=line; symbols=0; join=given
dgraph y=y,ox; x=x,oy; pen=1,2
```

## B.5 THE VARIOGRAM

### B.5.1 Experimental variogram

You will first want to compute (form) the experimental or sample variogram from your data. You can do it using the command `fvariogram`. It is followed by options and parameters. Below is an example, in which the `fvariogram` command is preceded by the declarations of two variates to hold the directions in which you want to compute the variogram and the angles subtended by the segments:

```
variave [nvalues=4] angles; values=!(0,45,90,135)
variave [nvalues=4] segs; values=!(45,45,45,45)
fvariogram [y=y; x=x; step=1; xmax=13; \
directions=angles; segments=segs] z; \
variogram=zgam; counts=zcounts; distances=midpts
```

The lag is incremented in steps of 1, which is the interval on the grid, to a maximum of 13. The identifiers `zgam`, `zcounts` and `midpts` are matrices, and you will usually want the results as vectors (variates). These are readily obtained from the matrices by

```
variave vgram[#angles], lag [#angles], count [#angles]
calculate vgram[ ] = zgam$[*;1...4]
calculate lag[ ] = midpts$[*;1...4]
calculate counts[ ] = zcounts$[*;1...4]
```

which you can then print and graph.

An average or omnidirectional variogram can be computed with the fvariogram command, again preceded by the declarations of two variates to hold the directions and the segments:

```
variave [nvalues=1] angles; values=!(0)
variave [nvalues=1] segs; values=!(180)
fvariogram[y=y; x=x; step=1; xmax=13; \
directions=angles; segments=segs] data=z; \
variogram=zgam; counts=zcounts; distances=midpts
```

Vectors of the identifiers zgam, zcounts and midpts are obtained as follows:

```
variave vgram[#angles], lag[#angles], count[#angles]
calculate vgram[ ] =zgam$[*;1]
calculate lag[ ] =midpts$[*;1]
calculate counts[ ] =zcounts$[*;1]
```

## B.5.2 Fitting a model

GenStat has a procedure, mvariogram, for fitting several standard models to experimental variograms. You can call it by, for example,

```
mvariogram[model=spherical; print=model, summary,\ 
estimates; weighting=counts] zgam; counts=zcounts; \
distances=midpts
```

This will fit a spherical model to the experimental variogram in zgam and midpts with weights proportional to the counts in zcounts. The models available are unbounded linear, bounded linear, circular, spherical, pentaspherical, stable (including exponential), Gaussian, Whittle's (besselk1), cubic, cardinalsine and power.

The procedure makes use of the fitnonlinear command in GenStat, and you can write models of your own choosing with this command. For example, to fit a spherical model you can write:

```
expression spherical; \
value=!e(c=((1.5*lag/a-0.5*(lag/a)**3*(lag.le.a)\ 
+ (lag.gt.a))) )
model[weights=counts] vrgam
rcycle a; initial=10
fitnonlinear[calculation=spherical] c
```

The expression lag.le.a gives the value 1 if the lag is less than or equal to the range, a, and 0 otherwise. In like manner lag.gt.a returns 1 if the lag is greater than a and 0 otherwise. These two conditions ensure that the function remains constant once the lag exceeds the range. Notice that only

the non-linear part of the model has to be described in the `value` command. The `rcycle` command sets an initial value for `a`. This is to ensure that the search for a solution starts in roughly the right place. Otherwise the program might never converge.

## B.6 KRIGING

The kriging facility in GenStat creates a grid of estimates (predictions) for mapping using the command `krige`, as follows.

```
krige[x=x; y=y; youter=!(1,31); xouter=!(1,18); \
yinner=!(5,20); xinner=!(3,15); block=!(0,0); \
radius=4.5; minpoints=7; maxpoints=20; interval=0.5] \
z; isotropy=isotropic; model=spherical; nugget=0.00476; \
sill=0.01528; range=10.8; predictions=krigest; \
variances=krigvar
```

## B.7 COREGIONALIZATION

The commands for computing cross-variograms, for modelling the linear coregionalization and cokriging have the same general form as those for the autovariogram and kriging in Sections B.5 and B.6, but there are significant differences.

### B.7.1 Auto- and cross-variograms

The command for forming the experimental variograms is `fcovariogram`, and an example for Broom's Barn might be

```
fcovariogram[step=1; maxlag=13; \
directions=angles; segments=segs; covariogram=zcovgam] \
data=z1,z2; x1=x; x2=y
```

Note that there are now at least two variates of measurements, here denoted `z1` and `z2`. The identifier `zcovgam` is a pointer to store the auto- and cross-variograms and is called when the coregionalization is modelled.

As for the autovariograms alone, you can compute average or omnidirectional variograms with the the `fcovariogram` command preceded by the declarations of two variates to hold the directions and the segments:

```
variave[nvalues=1] angles; values=!(0)
variave[nvalues=1] segs; values=!(180)
```

### B.7.2 Fitting a model of coregionalization

The GenStat procedure for fitting a model to a set of variograms is `mcovariogram`, and the same widely used functions as those for single variograms are available. You can call the procedure by, for example,

```
mcovariogram [print=model, summary, estimates; \
weighting=counts; maxlag=13; covariogram=zcovgam] \
model=nugget, spherical; estimates=mcovparam
```

This will fit a spherical model plus nugget to the whole set of experimental variograms in `zcovgam` with weights proportional to the counts. Note that in `mcovariogram` you specify the combination of models you want in the `model` parameter. This gives you a wider range of combinations than `mvariogram` does. The input is provided from `fcovariogram` in the pointer `zcovgam`, and the output, the estimates of the model parameters, are stored in the pointer `mcovparam`. The procedure uses the algorithm of Goulard and Voltz (1992) but with the additional optimization of the distance parameters.

### B.7.3 Cokriging

The `cokrige` directive computes estimates using the model fitted by `mcovariogram`. Again, it is similar to that for `autokriging`:

```
cokrige [y=z1; x1outer=!(1,31); x2outer=!(1,18); \
x1inner=!(5,20); x2inner=!(3,15); \
blockdimensions=!(0,0); \
radii=4.5; minpoints=7; maxpoints=20; x1interval=0.5; \
x2interval=0.5; searchneighbourhood=local] \
data=z1,z2; x1=x; x2=y; estimates=mcovparam; \
predictions=cokrigest; variances=cokrigvar
```

The directive `cokrige` has many options, which you can find in Payne (2006). We point out here only two features in the above code. The target variable must be specified by the `y` option, and that variable must also appear in the parameter list of data. By default `cokrige` uses all the data for each prediction. This might not be what you desire or be sensible in the circumstances, and you can ensure that predictions are based on local data only by restricting the searches with the option `searchneighbourhood=local` as above. Note that the variogram parameters are transferred from `mcovariogram` in the pointer `mcovparam`.

## B.8 CONTROL

Remember to terminate each GenStat job with  
`stop`

# References

- Ahmed, S. and de Marsily, G. (1987) Comparison of geostatistical methods for estimating transmissivity using data on transmissivity and specific capacity. *Water Resources Research*, **23**, 1727–1737.
- Aitchison, J. and Brown, J. A. C. (1957) *The Lognormal Distribution*. Cambridge University Press, Cambridge.
- Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In: *Second International Symposium on Information Theory* (eds B. N. Petrov and F. Csáki), pp. 267–281. Akadémiai Kiadó, Budapest.
- Armstrong, M. (1998) *Basic Linear Geostatistics*. Springer-Verlag, Berlin.
- Atteia, O., Webster, R. and Dubois, J.-P. (1994) Geostatistical analysis of soil contamination in the Swiss Jura. *Environmental Pollution*, **86**, 315–327.
- Badr, I., Oliver, M. A., Hendry, G. L. and Durrani, S. A. (1993) Determining the spatial scale of variation in soil radon values using nested sampling and analysis. *Radiation Protection Dosimetry*, **49**, 433–442.
- Barnes, R. J. (1989) Sample design for geologic site characterization. In: *Geostatistics*, Volume 2 (ed. M. Armstrong), pp. 809–822. Kluwer Academic Publishers, Dordrecht.
- Bartlett, M. S. (1966) *An Introduction to Stochastic Processes*, 2nd edition. Cambridge University Press, Cambridge.
- Baxter, S. J. and Oliver, M. A. (2005) The spatial prediction of soil mineral N using elevation. *Geoderma*, **128**, 325–339.
- Bierkens, M. F. P. and Burrough, P. A. (1993a) The indicator approach to categorical soil data. I. Theory. *Journal of Soil Science*, **44**, 361–368.
- Bierkens, M. F. P. and Burrough, P. A. (1993b) The indicator approach to categorical soil data. II. Application to mapping and land use suitability analysis. *Journal of Soil Science*, **44**, 369–381.
- Blackman, R. B. and Tukey, J. W. (1958) *The Measurement of Power Spectra*. Dover Publications, New York.
- Brigham, E.O. (1974) *The Fast Fourier Transform*. Prentice Hall, Englewood Cliffs, NJ.
- Brus, D. J. and de Gruijter, J. J. (1994) Estimation of non-ergodic variograms and their sampling variance by design-based sampling strategies. *Mathematical Geology*, **26**, 437–454.
- Burgess, T. M. and Webster, R. (1980a) Optimal interpolation and isarithmic mapping of soil properties. I. The semi-variogram and punctual kriging. *Journal of Soil Science*, **31**, 315–331.

- Burgess, T. M. and Webster, R. (1980b) Optimal interpolation and isarithmic mapping of soil properties. II. Block kriging. *Journal of Soil Science*, **31**, 333–341.
- Burgess, T. M. and Webster, R. (1984) Optimal sampling strategy for mapping soil types. I. Distribution of boundary spacings. *Journal of Soil Science*, **35**, 641–654.
- Burgess, T. M., Webster, R. and McBratney, A. B. (1981) Optimal interpolation and isarithmic mapping of soil properties. IV. Sampling strategy. *Journal of Soil Science*, **32**, 643–654.
- Burrough, P.A., Bregt, A.K., de Heus, M.J. and Kloosterman, E.G. (1985) Complementary use of thermal imagery and spectral analysis of soil properties and wheat yields to reveal cyclic patterns in the Flevopolders. *Journal of Soil Science*, **36**, 141–152.
- Carroll, Z. L. and Oliver, M. A. (2005) Exploring the spatial relations between soil physical properties and apparent electrical conductivity. *Geoderma*, **128**, 354–374.
- Chappell, A. and Oliver, M. A. (1997) Analysing soil redistribution in south-west Niger. In: *Geostatistics Wollongong '96*, Volume 2 (eds E. Y. Baafi and N. A. Schofield), pp. 961–972. Kluwer Academic Publishers, Dordrecht.
- Chauvet, P. (1982) The variogram cloud. In: *Proceedings of the 17th APCOM Symposium* (eds T. B. Johnson and R. J. Barnes), pp. 757–764. Society of Mining Engineers, New York.
- Chilès, J.-P. and Delfiner, P. (1999) *Geostatistics: Modeling Spatial Uncertainty*. John Wiley & Sons, Inc., New York.
- Clark, I., Basinger, K. L. and Harper, W. V. (1989) MUCK – a novel approach to cokriging. In: *Proceedings of the Conference on Geostatistical Sensitivity and Uncertainty Methods for Ground Water Flow and Radionuclide Transport Modeling* (ed. B. E. Buxton), pp. 473–493. Battelle Press, Columbus, OH.
- Cochran, W. G. (1946) Relative accuracy of systematic and stratified random samples for a certain class of populations. *Annals of Mathematical Statistics*, **17**, 164–177.
- Cochran, W. G. (1977) *Sampling Techniques*, 3rd edition. John Wiley & Sons, Inc., New York.
- Cooley, J. W. and Tukey, J. W. (1965) An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, **19**, 297–301.
- Cressie, N. (1985) Fitting variogram models by weighted least squares. *Journal of the International Association of Mathematical Geology*, **17**, 563–586.
- Cressie, N. (1990) The origins of kriging. *Mathematical Geology*, **22**, 239–252.
- Cressie, N. A. C. (1993) *Statistics for Spatial Data*, revised edition. John Wiley & Sons, Inc., New York.
- Cressie, N. (2006) Block kriging for lognormal spatial processes. *Mathematical Geology*, **38**, 413–443.
- Cressie, N. and Hawkins, D. M. (1980) Robust estimation of the variogram. *Journal of the International Association of Mathematical Geology*, **12**, 115–125.
- Dalenius, T., Hájek, J. and Zubrzycki, S. (1961) On plane sampling and related geometrical problems. In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (ed. J. Neyman), Volume 1, pp. 164–177. University of California Press, Berkeley.
- De Gruijter, J. J., Brus, D. J., Bierkens, M. F. P. and Knotters, M. (2006) *Sampling for Natural Resources Monitoring*. Springer-Verlag, Berlin.
- Delhomme, J.-P. (1978) Kriging in the hydrosciences. *Advances in Water Resources*, **1**, 251–266.

- Deutsch, C. V. and Journel, A. G. (1992) *GSLIB Geostatistical Software Library and User's Guide*. Oxford University Press, New York.
- Deutsch, C. V. and Journel, A. G. (1998) *GSLIB Geostatistical Software Library and User's Guide*, 2nd edition. Oxford University Press, New York.
- Dowd, P. A. (1984) The variogram and kriging: robust and resistant estimators. In: *Geostatistics for Natural Resources Characterization* (eds G. Verly, M. David, A. G. Journel and A. Marechal), pp. 91–106. D. Reidel, Dordrecht.
- Durbin, J. and Watson, G. S. (1950) Testing for serial correlation in least squares regression I. *Biometrika*, **37**, 409–428.
- Fabbri, P. and Trevisani, S. (2005) A geostatistical simulation approach to a pollution case in Northeastern Italy. *Mathematical Geology*, **37**, 569–586.
- Farmer, C. L. (1991) Numerical rocks. In: *The Mathematical Generation of Reservoir Geology* (eds J. Fayers and P. King), pp. 437–447. Oxford University Press, New York.
- Fisher, R. A. (1925) *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.
- Fisher, R. A. and Yates, F. (1963) *Statistical Tables for Biological, Agricultural and Medical Research*, 6th edition. Oliver and Boyd, Edinburgh.
- FOEFL (Swiss Federal Office of Environment, Forest and Landscape) (1987) *Commentary on the Ordinance relating to Pollutants in Soil* (VSBo of June 9, 1986). FOEFL, Bern.
- Frogbrook, Z. L. (1999) The effect of sampling intensity on the reliability of predictions and maps of soil properties. In: *Precision Agriculture '99, Part 1*, (ed. J. V. Stafford), pp. 71–80. Sheffield Academic Press, Sheffield.
- Frogbrook, Z. L., Oliver, M. A., Salahi, M. and Ellis, R. H. (1999) Comparing the relations in the spatial variation of soil and crop attributes. In: *Precision Agriculture '99, Part 1*, (ed. J. V. Stafford), pp. 397–406. Sheffield Academic Press, Sheffield.
- Galli, A., Gerdil-Neuillet, F. and Dadou, C. (1984) Factorial kriging analysis: a substitute to spectral analysis of magnetic data. In: *Geostatistics for Natural Resource Characterization* (eds G. Verly, M. David, A. Journel and A. Marechal), pp. 543–557. D. Reidel, Dordrecht.
- Gambolati, G. and Galeati, G. (1987) Analysis of nonintrinsic spatial variability by residual kriging with application to regional groundwater levels—comment. *Mathematical Geology*, **19**, 249–257.
- Gandin, L. S. (1965) *Objective Analysis of Meteorological Fields*. Israel Program for Scientific Translation, Jerusalem.
- Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.
- Genton, M. G. (1998) Highly robust variogram estimation. *Mathematical Geology*, **30**, 213–221.
- Gilmour, A. R., Gogel, B. J., Cullis, B. R., Welham, S. J. and Thompson, R. (2002) *ASReml User Guide, Release 1.0*. VSN International, Hemel Hempstead.
- Gleick, J. (1988) *Chaos*. William Heinemann, London.
- Golden Software (2002) *Surfer 8: Contouring and 3D Surface Mapping for Scientists and Engineers*. Golden Software, Inc., Golden, CO.
- Goovaerts, P. (1994) Performance of indicator algorithms for modelling conditional probability distribution functions. *Mathematical Geology*, **26**, 389–411.
- Goovaerts, P. (1997) *Geostatistics for Natural Resources Evaluation*. Oxford University Press, New York.

- Goovaerts, P. and Webster, R. (1994) Scale-dependent correlation between topsoil copper and cobalt concentrations in Scotland. *European Journal of Soil Science*, **45**, 77–95.
- Goovaerts, P., Webster, R. and Dubois, J.-P. (1997) Assessing the risk of soil contamination in the Swiss Jura using indicator geostatistics. *Environmental and Ecological Statistics*, **4**, 31–48.
- Gotway, C. A. (1994) The use of conditional simulation in nuclear-waste site performance assessment. *Technometrics*, **36**, 129–141.
- Goulard, M. and Voltz, M. (1992) Linear coregionalization model: tools for estimation and choice of cross-variogram matrix. *Mathematical Geology*, **24**, 269–286.
- Gower, J. C. (1962) Variance component estimation for unbalanced hierarchical classification. *Biometrics*, **18**, 168–182.
- Hallsworth, E. G., Robertson, G. K. and Gibbons, F. R. (1955) Studies in pedogenesis in New South Wales. VII. The 'gilgai' soils. *Journal of Soil Science*, **6**, 1–31.
- Hammond, L. C., Pritchett, W. L. and Chew, V. (1958) Soil sampling in relation to soil heterogeneity. *Soil Science Society of America Journal*, **22**, 548–552.
- Hengl, T., Heuvelink, G. B. M. and Stein, A. (2004) A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma*, **120**, 75–93.
- Hodge, C. A. H. and Seale, R. S. (1966) *The Soils of the District around Cambridge*. Memoirs of the Soil Survey of Great Britain. Agricultural Research Council, Harpenden.
- Hudson, G. and Wackernagel, H. (1994) Mapping temperature using kriging with external drift: theory and example from Scotland. *International Journal of Climatology*, **17**, 77–91.
- Isaaks, E. H. and Srivastava, R. M. (1989) *An Introduction to Applied Geostatistics*. Oxford University Press, New York.
- Jaquet, O. (1989) Factorial kriging analysis applied to geological data from petroleum exploration. *Mathematical Geology*, **21**, 683–691.
- Jenkins, G. M. and Watts, D. G. (1968) *Spectral Analysis and its Applications*. Holden-Day, San Francisco.
- Journel, A. G. (1974) Geostatistics for conditional simulation of ore bodies. *Economic Geology*, **69**, 673–680.
- Journel, A. G. (1983) Non-parametric estimation of spatial distributions. *Journal of the International Association of Mathematical Geology*, **15**, 445–468.
- Journel, A. G. (1988) Nonparametric geostatistics for risk and additional sampling assessment. In: *Principles of Environmental Sampling* (ed. L. H. Keith), pp. 45–72. American Chemical Society, Washington, DC.
- Journel, A. G. and Huijbregts, C. J. (1978) *Mining Geostatistics*. Academic Press, London.
- Jowett, G. H. (1955) Sampling properties of local statistics in stationary stochastic series. *Biometrika*, **42**, 160–169.
- Kantey, B. A. and Williams, A. A. B. (1962) The use of soil engineering maps for road projects. *Transactions of the South African Institution of Civil Engineers*, **4**, 149–159.
- Kerry, R. and Oliver, M. A. (2007a) Determining the effect of asymmetric data on the variogram. I. Underlying asymmetry. *Computers and Geosciences*, in press. doi: 10.1016/j.cageo.2007.05.008.
- Kerry, R. and Oliver, M. A. (2007b) Determining the effect of asymmetric data on the variogram. II. Outliers. *Computers and Geosciences*, in press. doi: 10.1016/j.cageo.2007.05.009.
- Kerry, R. and Oliver, M. A. (2007c) Comparing sampling needs for variograms of soil properties computed by the method of moments and residual maximum likelihood. *Geoderma*, in press. doi: 10.1016/j.geoderma.2007.04.019.

- Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P. (1983) Optimization by simulated annealing. *Science*, **220**, 671–680.
- Kitanidis, P. (1983) Statistical estimation of polynomial generalized covariance functions and hydrological applications. *Water Resources Research*, **19**, 909–921.
- Kitanidis, P. K. (1987) Parametric estimation of covariances of regionalized variables. *Water Resources Bulletin*, **23**, 557–567.
- Knotters, M., Brus, D. J. and Voshaar, J. H. O. (1995) A comparison of kriging, co-kriging and kriging combined with regression for spatial interpolation of horizon depth with censored observations. *Geoderma*, **67**, 227–246.
- Kolmogorov, A. N. (1939) Sur l'interpolation et l'extrapolation des suites stationnaires. *Comptes Rendus de l'Académie des Sciences de Paris*, **208**, 2043–2045.
- Kolmogorov, A. N. (1941) Interpolirovanie i ekstrapolirovaniye statisyonarnykh sluchainykh posledovatel'nostei (Interpolated and extrapolated stationary random sequences). *Isvestia AN SSSR, Seriya Matematicheskaya*, **5**(1)
- Krige, D. G. (1966) Two-dimensional weighted moving average trend surfaces for ore-evaluation. *Journal of the South African Institute of Mining and Metallurgy*, **66**, 13–38.
- Krumbein, W. C. and Slack, H. A. (1956) Statistical analysis of low-level radioactivity of Pennsylvanian black fissile shale in Illinois. *Bulletin of the Geological Society of America*, **67**, 739–762.
- Lake, J. V., Bock, G. R. and Goode, J. A. (1997) *Precision Agriculture: Spatial and Temporal Variability of Environmental Quality*. John Wiley & Sons, Ltd, Chichester.
- Langsaetter, A. (1926) Om beregning av middelfeilen ved regelmessige linjetakseringer. *Meddelanden fra det norske Skogsforsøksvesen*, **2 h 7**, 5–47.
- Lark, R. M. (2000) A comparison of some robust estimators of the variogram for use in soil survey. *European Journal of Soil Science*, **51**, 137–157.
- Lark, R. M. and Cullis, B. R. (2004) Model-based analysis using REML for inference from systematically sampled data on soil. *European Journal of Soil Science*, **55**, 799–813.
- Lark, R. M. and Webster, R. (2006) Geostatistical mapping of geomorphic variables in the presence of trend. *Earth Surface Processes and Landforms*, **31**, 862–874.
- Lark, R. M., Cullis, B. R. and Welham, S. J. (2006) On optimal prediction of soil properties in the presence of spatial trend: the empirical best linear unbiased predictor (E-BLUP) with REML. *European Journal of Soil Science*, **57**, 787–799.
- Laslett, G. M., McBratney, A. B., Pahl, P. J. and Hutchinson, M. F. (1987) Comparison of several spatial prediction methods for soil pH. *Journal of Soil Science*, **38**, 325–341.
- Leenhardt, D., Voltz, M., Bornand, M. and Webster, R. (1994) Evaluating soil maps for prediction of soil water properties. *European Journal of Soil Science*, **45**, 293–301.
- Lemmer, I. C. (1984) Estimating local recoverable reserves via indicator kriging. In: *Geostatistics for Natural Resources Characterization* (eds G. Verly, M. David, A. G. Journel and A. Marechal), pp. 349–364. D. Reidel, Dordrecht.
- Marcuse, S. (1949) Optimum allocation and variance components in nested sampling with application to chemical analysis. *Biometrics*, **5**, 189–206.
- Mardia, K. V. and Jupp, P. F. (2000) *Directional Statistics*. John Wiley & Sons, Ltd, Chichester.
- Marechal, A. (1976) The practice of transfer functions: Numerical methods and their application. In: *Advanced Geostatistics in the Mining Industry* (eds M. Guarascio, M. David and C. Huijbregts), pp. 253–276. D. Reidel, Dordrecht.

- Marquardt, D. W. (1963) An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society of Industrial and Applied Mathematics*, **11**, 431–441.
- Matérn, B. (1960) Spatial variation: Stochastic models and their applications to problems in forest surveys and other sampling investigations. *Meddelanden från Statens Skogsforskningsinstitut*, **49**, 1–144.
- Matheron, G. (1963) Principles of geostatistics. *Economic Geology*, **58**, 1246–1266.
- Matheron, G. (1965) *Les variables régionalisées et leur estimation*. Masson, Paris.
- Matheron, G. (1969) *Le krigage universel*. Cahiers du Centre de Morphologie Mathématique, No 1. Ecole des Mines de Paris, Fontainebleau.
- Matheron, G. (1973) The intrinsic random functions and their applications. *Advances in Applied Probability*, **5**, 439–468.
- Matheron, G. (1976) A simple substitute for conditional expectation: the disjunctive kriging. In: *Advanced Geostatistics in the Mining Industry* (eds M. Guarascio, M. David and C. Huijbregts), pp. 221–236, D. Reidel, Dordrecht.
- Matheron, G. (1979) *Recherche de simplification dans un problème de cokrigage*. Publication N-628, Centre de Géostatistique, Ecole des Mines de Paris, Fontainebleau.
- Matheron, G. (1982) *Pour une analyse krigeante de données régionalisées*. Note N-732 du Centre de Géostatistique. Ecole des Mines de Paris, Fontainebleau.
- Matheron, G. (1989) *Estimating and Choosing*. Springer-Verlag, Berlin.
- McBratney, A. B. and Webster, R. (1981) Detection of ridge and furrow patterns by spectral analysis of crop yield. *International Statistical Review*, **49**, 45–52.
- McBratney, A. B. and Webster, R. (1983) Optimal interpolation and isarithmic mapping of soil properties. V. Coregionalization and multiple sampling strategy. *Journal of Soil Science*, **34**, 137–162.
- McBratney, A. B. and Webster, R. (1986) Choosing functions for semivariograms of soil properties and fitting them to sampling estimates. *Journal of Soil Science*, **37**, 617–639.
- McBratney, A. B., Webster, R. and Burgess, T. M. (1981) The design of optimal sampling schemes for local estimation and mapping of regionalized variables. *Computers and Geosciences*, **7**, 331–334.
- McBratney, A. B., Webster, R., McLaren, R. G. and Spiers, R. B. (1982) Regional variation of extractable copper and cobalt in the topsoil of south-east Scotland. *Agronomie*, **2**, 969–982.
- McCullagh, M. J. (1976) Estimation by kriging of the reliability of the Trent telemetry network. *Computer Applications*, **2**, 357–374.
- McNeill, J. D. (1990) *Geonics EM38 Ground Conductivity Meter: EM38 Operating Manual*. Geonics Limited, Mississauga, Ontario.
- Mercer, W. B. and Hall, A. D. (1911) Experimental error of field trials. *Journal of Agricultural Science, Cambridge*, **4**, 107–132.
- Meul, M. and Van Meirvenne, M. (2003) Kriging soil texture under different types of nonstationarity. *Geoderma*, **112**, 217–233.
- Miesch, A. T. (1975) Variograms and variance components in geochemistry and ore evaluation. *Geological Society of America Memoir*, **142**, 333–340.
- Minasny, B. and McBratney, A. B. (2005) The Matérn function as a general model for soil variograms. *Geoderma*, **128**, 192–207.
- Ministry of Agriculture, Fisheries and Food (1986) *The Analysis of Agricultural Materials*, 3rd edition. MAFF Reference Book 427. Her Majesty's Stationery Office, London.

- Moffat, A. J., Catt, J. A., Webster, R. and Brown, E. H. (1986) A re-examination of the evidence for a Plio-Pleistocene marine transgression on the Chiltern Hills. I. Structures and surfaces. *Earth Surface Processes and Landforms*, **11**, 95–106.
- Morse, R. K. and Thornburn, T. H. (1961) Reliability of soil units. In: *Proceedings of the 5th International Conference on Soil Mechanics and Foundation Engineering, Volume 1*, pp. 259–262. Dunod, Paris.
- Mulla, D. J. (1997) Geostatistics, remote sensing and precision farming. In: *Precision Agriculture: Spatial and Temporal Variability of Environmental Quality* (eds J. V. Lake, G. R. Bock and J. A. Goode), pp. 100–115. John Wiley & Sons, Ltd, Chichester.
- Muñoz-Pardo, J. F. (1987) *Approche géostatistique de la variabilité spatiale des milieux géophysique*. Thèse de Docteur-Ingénieur, Université de Grenoble et l'Institut National Polytechnique de Grenoble.
- Myers, D. E. (1982) Matrix formulation of cokriging. *Journal of the International Association of Mathematical Geology*, **14**, 249–257.
- Myers, D. E. (1991) Pseudo-cross-variograms, positive definiteness, and cokriging. *Mathematical Geology*, **23**, 805–816.
- Nelder, J. A. and Mead, R. (1965) A simplex method for function minimization. *Computer Journal*, **7**, 308–313.
- Odeh, I. O. A., McBratney, A. B. and Chittleborough, D. J. (1994) Spatial prediction of soil properties from landform attributes derived from digital elevation models. *Geoderma*, **63**, 197–214.
- Odeh, I. O. A., McBratney, A. B. and Chittleborough, D. J. (1995) Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. *Geoderma*, **67**, 215–226.
- Olea, R. A. (1975) *Optimum Mapping Techniques using Regionalized Variable Theory*. Series on Spatial Analysis, no 2. Kansas Geological Survey, Lawrence.
- Olea, R. A. (1999) *Geostatistics for Engineers and Earth Scientists*. Kluwer Academic Publishers, Boston.
- Oliver, M. A. and Badr, I. (1995) Determining the spatial scale of variation in soil radon concentration. *Mathematical Geology*, **27**, 893–922.
- Oliver, M. A. and Carroll, Z. L. (2004) *Description of Spatial Variation in Soil to Optimize Cereal Management*. Project Report 330. Home-Grown Cereals Authority, London.
- Oliver, M. A. and Webster, R. (1986) Combining nested and linear sampling for determining the scale and form of spatial variation of regionalized variables. *Geographical Analysis*, **18**, 227–242.
- Oliver, M. A. and Webster, R. (1987) The elucidation of soil pattern in the Wyre Forest of the West Midlands, England. II. Spatial distribution. *Journal of Soil Science*, **38**, 293–307.
- Oliver, M. A., Webster, R. and McGrath, S. P. (1996) Disjunctive kriging for environmental management. *Environmetrics*, **7**, 333–358.
- Oliver, M. A., Webster, R., Edwards, K. J. and Whittington, G. (1997) Multivariate, autocorrelation and spectral analyses of a pollen profile from Scotland and evidence of periodicity. *Review of Palaeobotany and Palynology*, **96**, 121–141.
- Oliver, M. A., Webster, R. and Slocum, K. (2000) Filtering SPOT imagery by kriging analysis. *International Journal of Remote Sensing*, **21**, 735–752.
- Omre, H. (1987) Bayesian kriging—merging observations and qualified guesses in kriging. *Mathematical Geology*, **19**, 25–39.

- Pannatier, Y. (1995) *Variowin. Software for Spatial Analysis in 2D*. Springer-Verlag, New York.
- Papritz, A. and Webster, R. (1995a) Estimating temporal change in soil monitoring: I. Statistical theory. *European Journal of Soil Science*, **46**, 1–12.
- Papritz, A. and Webster, R. (1995b) Estimating temporal change in soil monitoring: II. Sampling from simulated fields. *European Journal of Soil Science*, **46**, 13–27.
- Papritz, A., Künsch, H. R. and Webster, R. (1993) On the pseudo cross-variogram. *Mathematical Geology*, **25**, 1015–1026.
- Pardo-Igúzquiza, E. (1997) MLREML: a computer program for the inference of spatial covariance parameters by maximum likelihood and restricted maximum likelihood. *Computers and Geosciences*, **23**, 153–162.
- Pardo-Igúzquiza, E. (1998) Inference of spatial indicator covariance parameters by maximum likelihood using REML. *Computers and Geosciences*, **24**, 453–464.
- Parzen, E. (1961) Mathematical considerations in the estimation of spectra. *Technometrics*, **3**, 167–190.
- Patterson, H. D. and Thompson, R. (1971) Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545–554.
- Payne, R. W. (ed.) (2006) *The Guide to GenStat Release 9 – Part 2: Statistics*. VSN International, Hemel Hempstead.
- Pettitt, A. N. and McBratney, A. B. (1993) Sampling designs for estimating spatial variance components. *Applied Statistics*, **42**, 185–209.
- Press, W. H., Flannery, B. P., Teukolsky, S. A. and Vetterling, W. T. (1992) *Numerical Recipes in Fortran*, 2nd edition. Cambridge University Press, Cambridge.
- Priestley, M. B. (1981) *Spectral Analysis and Time Series*. Academic Press, London.
- Quenouille, M. H. (1949) Problems in plane sampling. *Annals of Mathematical Statistics*, **20**, 355–375.
- Ratkowsky, D. A. (1983) *Nonlinear Regression Modeling*. Marcel Dekker, New York.
- Rendu, J.-M. (1980) Disjunctive kriging: comparison of theory with actual results. *Journal of the International Association of Mathematical Geology*, **12**, 305–320.
- Rivoirard, J. (1994) *Introduction to Disjunctive Kriging and Non-linear Geostatistics*. Oxford University Press, Oxford.
- SAS Institute (1999) *SAS/STAT User's Guide: Version 8*. SAS Institute Inc., Cary, NC.
- Scott, R. M., Webster, R. and Lawrence, C. J. (1971) *A Land System Atlas of Western Kenya*. Military Vehicles and Engineering Establishment, Christchurch, Dorset.
- Searle, S. R., Casella, G. and McCulloch, C. E. (1992) *Variance Components*. John Wiley & Sons, Inc., New York.
- Serra, J. (1968) Les structures gigognes: morphologie mathématique et interprétation métallogénique. *Mineralium Deposita*, **3**, 135–154.
- Shafer, J. M. and Varljen, M. D. (1990) Approximation of confidence limits on sample semivariograms from single realizations of spatially correlated random fields. *Water Resources Research*, **26**, 1787–1802.
- Shepard, D. (1968) A two-dimensional interpolation function for irregularly-spaced data. *Proceedings of the Association for Computing Machinery* (1968), 517–523.
- Sibson, R. (1981) A brief description of natural neighbour interpolation. In: *Interpreting Multivariate Data* (ed. V. Barnett), pp. 21–36. John Wiley & Sons, Ltd, Chichester.
- Snedecor, G. W. and Cochran, W. G. (1967) *Statistical Methods*, 6th edition. Iowa State University Press, Ames.

- Stein, M. L. (1999) *Interpolation of Spatial Data: Some Theory for Kriging*. Springer-Verlag, New York.
- Sullivan, J. (1984) Conditional recovery estimation through probability kriging: Theory and practice. In: *Geostatistics for Natural Resources Characterization* (eds G. Verly, M. David, A. G. Journel and A. Marechal), pp. 365–384. D. Reidel, Dordrecht.
- Taylor, C. C. and Burrough, P. A. (1986) Multiscale sources of spatial variation in soil. III. Improved methods for fitting the nested model to one-dimensional semi-variograms. *Mathematical Geology*, **18**, 811–821.
- Tukey, J. W. (1977) *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.
- Von Neumann, J. (1941) Distribution of the ratio of the mean square difference to the variance. *Annals of Mathematical Statistics*, **12**, 367–395.
- Von Steiger, B., Webster, R., Schulin, R. and Lehmann, R. (1996) Mapping heavy metals in polluted soil by disjunctive kriging. *Environmental Pollution*, **94**, 205–215.
- Wackernagel, H. (1994) Cokriging versus kriging in regionalized multivariate data analysis. *Geoderma*, **62**, 83–92.
- Wackernagel, H. (2003) *Multivariate Geostatistics*, 3rd edition. Springer-Verlag, Berlin.
- Webster, R. (1977) Spectral analysis of gilgai soil. *Australian Journal of Soil Research*, **15**, 191–204.
- Webster, R. (1991) Local disjunctive kriging of soil properties with change of support. *Journal of Soil Science*, **42**, 301–318.
- Webster, R. (1994) Estimating trace elements in soil: a case study in cobalt deficiency. In: *Introduction to Disjunctive Kriging and Non-linear Geostatistics* (J. Rivoirard), pp. 128–145. Oxford University Press, Oxford.
- Webster, R. and Beckett, P. H. T. (1968) Quality and usefulness of soil maps. *Nature, (London)*, **219**, 680–682.
- Webster, R. and Beckett, P. H. T. (1970) Terrain classification and evaluation using air photography: a review of recent work at Oxford. *Photogrammetria*, **26**, 51–75.
- Webster, R. and Boag, B. (1992) A geostatistical analysis of cyst nematodes in soil. *Journal of Soil Science*, **43**, 583–595.
- Webster, R. and Butler, B. E. (1976) Soil survey and classification studies at Ginninderra. *Australian Journal of Soil Research*, **14**, 1–26.
- Webster, R. and McBratney, A. B. (1987) Mapping soil fertility at Broom's Barn by simple kriging. *Journal of the Science of Food and Agriculture*, **38**, 97–115.
- Webster, R. and McBratney, A. B. (1989) On the Akaike Information Criterion for choosing models for variograms of soil properties. *Journal of Soil Science*, **40**, 493–496.
- Webster, R. and Oliver, M. A. (1989) Optimal interpolation and isarithmic mapping of soil properties. VI. Disjunctive kriging and mapping the conditional probability. *Journal of Soil Science*, **40**, 497–512.
- Webster, R. and Oliver, M. A. (1990) *Statistical Methods in Soil and Land Resource Survey*. Oxford University Press, Oxford.
- Webster, R. and Oliver, M. A. (1992) Sample adequately to estimate variograms of soil properties. *Journal of Soil Science*, **43**, 177–192.
- Webster, R. and Oliver, M. A. (1997) Software review. *European Journal of Soil Science*, **48**, 173–175.
- Webster, R. and Oliver, M. A. (2006) Modeling spatial variation of soil as random functions. In: *Environmental Soil-Landscape Modeling: Geographic Information Technologies*

- and Pedometrics (ed. S. Grunwald), pp. 241–287. CRC Taylor and Francis, Boca Raton, FL.
- Webster, R. and Rivoirard, J. (1991) Copper and cobalt deficiency in soil: a study using disjunctive kriging. In: *Cahiers de Géostatistique, Compte-Rendu des Journées de Géostatistique*, Volume 1, pp. 205–223. Ecole des Mines de Paris, Fontainebleau.
- Webster, R., Atteia, O. and Dubois, J.-P. (1994) Coregionalization of trace metals in the soil in the Swiss Jura. *European Journal of Soil Science*, **45**, 205–218.
- Webster, R., Welham, S. J., Potts, J. M. and Oliver, M. A. (2006) Estimating the spatial scale of regionalized variables by nested sampling, hierarchical analysis of variance and residual maximum likelihood. *Computers and Geosciences*, **32**, 1320–1333.
- Whittle, P. (1954) On stationary processes in the plane. *Biometrika*, **41**, 434–449.
- Wiener, N. (1949) *Extrapolation, Interpolation and Smoothing of Stationary Time Series*. MIT Press, Cambridge, MA.
- Wold, H. (1938) *A Study in the Analysis of Stationary Time Series*. Almqvist and Wiksell, Uppsala.
- Wood, G., Oliver, M. A. and Webster, R. (1990) Estimating soil salinity by disjunctive kriging. *Soil Use and Management*, **6**, 97–104.
- Yaglom, A. M. (1987) *Correlation Theory of Stationary and Related Random Functions. Volume 1: Basic Results*. Springer-Verlag, New York.
- Yates, F. (1948) Systematic sampling. *Philosophical Transactions of the Royal Society of London A*, **241**, 345–377.
- Yates, F. (1981) *Sampling Methods for Censuses and Surveys*, 4th edition. Griffin, London.
- Yates, S. R. and Yates, M. V. (1988) Disjunctive kriging as an approach to management decision making. *Soil Science Society of America Journal*, **52**, 1554–1558.
- Yates, S. R., Warrick, A. W. and Myers, D. E. (1986a) Disjunctive kriging. I. Overview of estimation and conditional probability. *Water Resources Research*, **22**, 615–621.
- Yates, S. R., Warrick, A. W. and Myers, D. E. (1986b) Disjunctive kriging. II. Examples. *Water Resources Research*, **22**, 623–630.
- Yfantis, E. A., Flatman, G. T. and Behar, J. V. (1987) Efficiency of kriging estimation for square, triangular and hexagonal grids. *Mathematical Geology*, **19**, 183–205.
- Youden, W. J. and Mehlich, A. (1937) Selection of efficient methods for soil sampling. *Contributions of the Boyce Thompson Institute for Plant Research*, **9**, 59–70.
- Yule, G. U. and Kendall, M. G. (1950) *An Introduction to the Theory of Statistics*, 14th edition. Griffin, London.
- Zimmermann, D. L. and Zimmermann, M. B. (1991) A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors. *Technometrics*, **33**, 77–91.

# **Index**

- Akaike information criterion (AIC) 105, 290  
analysis of variance 35, 44, 127–132  
angular transformation 22  
anisotropy 59, 99  
    affine or geometric 59, 100–101  
    anisotropy ratio 101  
    exploring and displaying 70, 73  
    transects 70  
    zonal 59  
ASReml 203  
*a priori* variance 52, 58, 84, 92, 98, 102  
asymmetry 17, 110–112  
    long-tail, outliers 110  
asymmetric covariance 221  
authorized models 80, 82–95  
autocorrelation 3, 53  
    autocorrelation coefficients 55, 66, 76  
    autocorrelation function 55, 57  
autocovariance 53, 56  
    autocovariance function 57  
  
balanced differences 33–34  
Bessel functions 92, 98, 99  
bias 43, 144, 186, 199  
binary variables 11, 246  
block kriging 159, 167–171, 175–180, 188–189  
Borders Region of Scotland 24, 105–107  
factorial kriging analysis 216–217  
histogram 24  
variogram 216  
  
bounded variation 84  
    bounded models, bounded variograms 84–94  
bounded linear variogram model 85  
box plots, box-and-whisker plots 14, 25  
Broom’s Barn Farm 257–263, 278–283  
    maps 262, 264, 279, 282  
pH 160  
posting 25, 27  
potassium 13, 23, 25, 67, 278, 280, 283  
variograms 102–103, 175–179, 278, 280, 283  
Brownian motion 83  
  
capacity variables 15  
Caragabal transect 140–142, 150–151  
    variogram 142,  
        spectral analysis 150–151  
CEDAR Farm 219–220, 226–228  
    coregionalization 226–228  
central limit theorem 32  
chi-square distribution 31  
circular variogram model 87  
circular scales 12  
classical sampling theory 28–33, 43, 124, 127  
CNSD 80, 224  
codispersion coefficient 222  
coefficient of variation 17, 287  
cokriging 228–234  
    benefits 231–234  
    fully sampled case 231  
    undersampled case 231  
variance 229, 231, 234

- combining models 95–97  
 conditional negative semidefinite,  
*see* CNSD  
 conditional probability 256  
 confidence intervals of variograms  
   119–125  
 confidence limits 29–31  
 continuity 57  
   continuous function 57  
   continuous lag 57  
 continuous variables 12  
   continuous scales 12  
 coregionalization 219 *et seq.*  
   CEDAR Farm 219–220,  
   226–228  
   linear model 222–224  
   matrices 235  
 correlation 19–20  
   correlation coefficient 20, 111  
 correlation range, *see* range  
 correlograms 74  
 cross-correlogram 222  
 covariance 19–20, 47–60  
   covariance function 53–55  
   covariance matrix 57  
   equivalence with variogram 54  
   estimation 74  
 cross-correlation 219  
   cross-correlation coefficient 221  
   estimating 222  
   modelling 222–224  
 cross-covariance 220–222  
   cross-covariance function  
   220–221  
 cross-indicator variograms 248–249  
 cross-validation 191–193  
 cross-variograms 220  
   cross-semivariance 220  
 cubic trend surface 40  
 cubic variogram function 93  
 cumulative distribution 15, 19, 23, 24,  
   26, 247, 250–260  
 cumulative distribution function 250,  
   256, 258, 260
- degrees of freedom 128, 132  
 design-based estimation 28  
 Dirac function 58
- Dirichlet tessellation, tiles, *see* Thiessen polygons  
 discontinuity 81, 177  
 disjunctive kriging 243 *et seq.*  
   Gaussian 251  
   Hermite polynomial 252–255  
   variance 256  
 dispersion 16–17  
   dispersion variance 60–64, 102,  
   120  
 distance parameters 89, 91–96, 224, 237,  
   298  
 double spherical variogram model 97  
   double spherical examples 96, 107,  
   216, 237–238  
 drift 59, 195–205  
   external drift 203–211
- E-BLUP 202  
 efficiency 32–33  
 environmental data, notation 12  
 environmental variables 11–12  
   binary 11  
   continuous 12  
 ergodicity 53  
 estimation 26–30, 32  
   classical, design-based 26–30, 33  
   estimation variance 29, 33  
   local 153–181  
   regional 181–183  
   simple random samples 28–32  
   stratification 32–33  
   systematic sampling 33–35  
 exhaustive variogram 122  
 experimental covariance function 73–74  
 experimental spectrum, *see* spectral analysis  
 experimental variogram 60, 68–73, 288,  
   295  
   experimental semivariances 60, 68–73,  
   288, 295  
 exploratory data analysis and display  
   22–25, 285–288  
 exponential variogram model 88, 91, 92,  
   94, 95
- F ratio 130  
 factorial kriging analysis 212–217  
 first moment 52

- fitting models 101–107, 290, 296, 298  
complexity 105  
computer programs 103  
difficulties 101–102  
GenStat 290, 296  
recommended procedure 102  
Fourier transform, *see* spectral analysis  
frequency distribution 13–15, 286  
frequency domain, *see* spectral analysis
- gamma function 31  
Gaussian diffusion process 251  
Gaussian disjunctive kriging 251  
Gaussian distribution 18  
Gaussian variogram model 93  
Gaussian simulation 273–274, 278–281  
GenStat 293–298  
geometric or affine anisotropy 100–101  
geometric mean 21  
geostatistics, general 1–6  
    history 6–8  
    overview 1–6  
    roles 2  
goodness-of-fit criterion 104  
GSLIB 274, 275, 277
- heavy metals 235–241  
Hermite transformation 252–257  
    Hermite polynomials 252–257  
hierarchical analysis of variance 127–132  
nested sampling and analysis 127–128, 131–132  
histogram 13, 286, 294  
hole effect 56, 58  
hole effect models 98–99
- inclined plane 40, 206  
indictor variables (indicators) 246  
    indicator coding 246  
    indicator covariance function 249  
    indicator kriging 249–251  
    indicator variograms 247  
intensity variables 15  
interpolation 37–42  
intraclass correlation 44  
intrinsic corgionalization 224–225  
intrinsic hypothesis 54
- intrinsic random function of order  $k$  59  
intrinsic variation 54  
inverse functions of distance 40  
isarithmic chart 73, 75  
isotropic variation 70, 82, 124, 160, 187, 289, 297  
isotropic variograms 70, 73
- joint cdf 52  
joint distribution 20, 52, 66  
joint pdf 20
- Krige's relation 60–61, 63  
kriging 153 *et seq.*, 291, 297  
    Bayesian 155  
    block kriging 156–159  
    cokriging 228–234  
    disjunctive kriging 243 *et seq.*  
    factorial kriging 212–217  
    general characteristics 154  
    general theory 155–159  
    indicator kriging 249  
    kriging with external drift 203–205  
    kriging equations 172–173  
    kriging neighbourhood 172–173  
    kriging variance 158, 159, 163,  
        178–180, 182, 184, 185,  
        188–189, 198, 209, 211,  
        256  
    kriging weights 159–160  
    kriging with trend 195–211  
        E-BLUP 202  
        kriging with external drift 203–205  
        universal kriging 196–203  
lognormal kriging 184–185  
mapping 173–174, 181–191  
ordinary kriging 155–160  
    ordinary kriging equations  
probability kriging 155  
regression kriging 199  
simple kriging 183–184  
universal kriging 196–203  
    universal kriging equations 197–198  
Kronecker delta 57, 95  
kurtosis 18
- lag 53, 57  
increments, interval 69–73

- Lagrange multiplier 157  
 least-squares methods 40, 102–103, 105,  
   199, 290  
 Levenberg–Marquardt method 103  
 linear drift 197  
 linear mixed model 134, 200  
 linear sequences 139–140  
 local estimation 153 *et seq.*  
 logarithmic transformation 21, 259,  
   287  
 logit transformation 22  
 log-likelihood 202  
 lognormal kriging 184–185  
 long-range trend 82, 198, 215
- mapping 291–292  
   interpolation 37–42  
   kriging 173–174  
   optimal sampling 185–191  
   posting 27  
 Marcuse model II 127  
 Matérn variogram function 94  
 MATLAB 277  
 mean 15  
 mean error (ME) 191  
 mean squared deviation ratio (MSDR)  
   192  
 mean squared difference (MSE)  
   191  
 mean squared error, prediction (MSE)  
   45–46  
 mean squared residual (MSR) 107  
 measurement error in kriging 180  
 median 16  
 missing values 68, 70, 286  
 mode 16  
 model fitting, *see* fitting models  
 Monte Carlo methods 121, 270  
 multiple regression 40  
 multi-stage sampling 127
- natural neighbours 39  
   interpolation 39  
 negative exponential variogram model, *see*  
   exponential variogram model  
 nested sampling and analysis  
   127–138  
   balanced designs 127, 128
- components of variance 127, 128  
 REML estimation 132–138  
 unbalanced designs 131–132  
 Wyre Forest 132–138
- nested spherical variogram model, *see*  
   double spherical variogram  
   model
- non-ergodic variogram 60, 120  
 non-linear regression 103  
 normal distribution 18–20, 252, 287  
   random variables 49  
 normalized difference vegetation index  
   (NDVI) 47  
 notation 12  
 nugget, nugget variance 56–58, 79–84,  
   131  
   nugget:sill ratio 110, 161–163  
 nugget variogram 95
- Occam’s razor 77  
 ordinary kriging 155–160  
 outliers 22, 65, 113–118
- Pearson product-moment correlation  
   coefficient 20  
 pentaspherical variogram model 84,  
   88  
 periodic variation 97–99, 139–152  
   amplitude 97–98  
   periodic variogram model 97–99  
   phase 97–98  
 point samples 3  
 Poisson process 87, 90  
 positive intercept 79, 81  
 positive semidefiniteness 57, 79  
 posting 5, 25, 27, 285, 295  
 power function variogram 83  
 power spectrum, *see* spectral analysis  
 prediction 37, 153–194  
   general formula 37  
   kriging 153 *et seq.*  
   prediction error 43  
   prediction variance 43  
   purposively chosen sample 45  
   random sample 44  
   probability density 18, 49  
   probability density function 49  
 process control 5

- projection matrix 201  
pseudo-cross-variogram 241–242  
punctual kriging 155 *et seq.*  
pure nugget, pure nugget variogram 56, 95
- quadratic trend surface 40, 41  
  quadratic drift 197, 207  
quasi-stationarity 55
- random effects model 127, 200  
random sample, prediction 28  
random variables 49 *et seq.*  
  random functions, random process 49  
  random variation 49, 59, 79  
random walk model 83  
range 84  
  effective range 89  
realization 49  
regional estimation 181–183  
regional variogram 49, 60, 121  
regionalized variables, theory 48–60  
regression 40–44  
  regression kriging 199  
regression surfaces, *see* trend surfaces  
regular sampling for variogram  
  in one dimension 68–69  
  in two dimensions 70–71  
regularization 63–65  
  regularized variogram 64  
relative precision 33  
residual maximum likelihood (REML)  
  132–134, 200–202  
components of variance 133–134  
variogram estimation 202
- sample correlogram 74  
sample mean 15, 29  
sample variogram, *see* experimental variogram  
sampling 26 *et seq.*  
  design, plan 28, 186  
  intensity, density, spacing 164, 186 *et seq.*  
  sample size for variogram estimation 119–126  
  theory 28 *et seq.*  
SAS 103
- scatter diagram 22, 66, 193, 210, 263  
Schwarz's inequality 223  
screening 285  
second moments 17, 52  
second-order polynomial, *see* quadratic  
second-order stationarity 52  
semivariance 54 *et seq.*  
  estimation 65 *et seq.*  
short-range drift 59, 81  
sill 56, 79  
  sill variance 84  
simple kriging 183–184  
simple random sampling 28–30  
  estimation 28  
  estimation variance 29  
  standard error 29  
simulation, stochastic 267–283  
case study, illustration 278–283  
Cholesky (LU) decomposition 272–273  
conditional 270–271  
  purpose 271  
sequential Gaussian 273–274  
simulated annealing 274–276  
turning bands 276  
  unconditional 270  
sinusoidal function 97–98  
skewness 17, 287  
  skewed histogram 24, 287  
smoothing function, *see* spectral analysis  
soil classification 42–44  
soil maps 42–44  
spatial analysis, *aide mémoire* 285–292  
spatial classification 42–44  
spatial correlation 55 *et seq.*  
  spatial correlation functions, characteristics 55–60  
  spatial covariance 50  
  spatial dependence 58  
spatial distribution 288  
spatial domain, *see* spectral analysis  
spatial estimation, *see* kriging  
spatial interpolation 37–40  
spatial prediction 37–46  
spatial processes 47 *et seq.*

- spectral analysis 139–152  
 Bartlett windows 145  
 Caragabal transect 140–142, 147,  
 150–151  
 confidence limits and intervals 149  
 estimation 144  
 Fourier transformation 143  
 frequency domain 143  
 Parzen windows 146  
 power spectrum 143–144  
 smoothing 145–146  
 spatial domain 140, 142  
 spectral density 140  
 theory 140–143  
 spectrum *see* spectral analysis  
 spherical variogram, spherical model 84,  
 87–88, 100, 164, 166  
 splines 42  
 SPOT 50  
 S-Plus 277  
 square root transformation 21  
 stable variogram models 91, 93  
 standard deviation 13  
 standard error 29  
 standard normal deviate 30  
 standard normal distribution 31  
 stationarity 52 *et seq.*  
 statistical fitting 102  
 statistics, basic 11 *et seq.*  
 stochastic process 49  
 stratified sampling 32  
   estimates 32  
   precision 32  
   stratification 32  
   full stationarity 53  
 structural variance–covariance matrices,  
   *see* coregionalization matrices  
 Student's *t* 30  
 sum of squares 31, 130  
 summary statistics 13 *et seq.*, 293  
 support 61  
 Swiss Jura 235–241  
   coregionalization 237–240  
   principal component analysis 236  
   trace metals 235–241  
   variograms 238–239  
 symmetric distributions 16  
 systematic sampling 33–35  
 target population 28  
 theoretical variogram 60, 288  
 Thiessen polygons 38  
 transformations 20–22, 99  
   back-transformation 185  
   Fourier transformation 143  
     *et seq.*  
   Hermite transformation  
     252–254  
 trend 40, 59, 81, 195 *et seq.*  
   trend surfaces 40–42  
 triangulation 38  
 two-dimensional variogram, *see*  
   variogram  
 unaligned sampling 34  
 unbalanced sampling design  
   131–132  
 unbounded random variation 83  
   unbounded variogram 58  
 units, *see* sampling  
 universal kriging 196–203  
 variance 16, 29  
 variance ratio 130  
 variogram 54–76, 288, 295  
   behaviour near the origin  
     80–82  
   behaviour towards infinity 82  
   block-to-block integration 64  
   definition 54  
   equivalence with covariance 54  
   estimation 65–76, 295  
   linear approach to origin 81  
   modelling 77–107, 296  
   parabolic approach to  
     origin 81  
   regularized variogram 63–65  
   reliability 109 *et seq.*  
 variogram cloud 65–66  
 variogram functions, limitations on  
   79–80  
 Voronoi polygons, *see* Thiessen  
   polygons  
 weak stationarity 52  
 weighted average 37  
 weighted least squares 102, 104

weighting function 252  
weights  
  interpolation weights 38–40  
  kriging weights 159–172  
white noise 58, 83  
Whittle elementary correlation 92  
  Whittle variogram model 92  
within-class variance 44

Wyre Forest survey 134–138  
  nested sampling and analysis 134–138  
Yattendon 205–211  
  kriging with drift 207–211  
  REML estimation 208–211  
  variograms 207  
zonal anisotropy 59

## ***Statistics in Practice***

### *Human and Biological Sciences*

- Berger – Selection Bias and Covariate Imbalances in Randomized Clinical Trials  
Brown and Prescott - Applied Mixed Models in Medicine, Second Edition  
Chevret (Ed) – Statistical Methods for Dose Finding Experiments  
Ellenberg, Fleming and DeMets – Data Monitoring Committees in Clinical Trials:  
A Practical Perspective  
Hauschke, Steinijans and Pigeot – Bioequivalence Studies in Drug Development: Methods and Applications  
Lawson, Browne and Vidal Rodeiro – Disease Mapping with WinBUGS and MLwiN  
Lui – Statistical Estimation of Epidemiological Risk  
\*Marubini and Valsecchi - Analysing Survival Data from Clinical Trials and Observation Studies  
Molenberghs and Kenward – Missing Data in Clinical Studies  
O'Hagan – Uncertain Judgements: Eliciting Experts' Probabilities  
Parmigiani – Modeling in Medical Decision Making: A Bayesian Approach  
Pintilie – Competing Risks: A Practical Perspective  
Senn – Cross-over Trials in Clinical Research, Second Edition  
Senn – Statistical Issues in Drug Development  
Spiegelhalter, Abrams and Myles – Bayesian Approaches to Clinical Trials and Health-Care Evaluation  
Whitehead - Design and Analysis of Sequential Clinical Trials, Revised Second Edition  
Whitehead – Meta-Analysis of Controlled Clinical Trials  
Willan – Statistical Analysis of Cost-effectiveness Data  
Winkel and Zhang – Statistical Development of Quality in Medicine

### *Earth and Environmental Sciences*

- Buck, Cavanagh and Litton – Bayesian Approach to Interpreting Archaeological Data  
Glasbey and Horgan – Image Analysis in the Biological Sciences  
Helsel – Nondetects and Data Analysis: Statistics for Censored Environmental Data

McBride – Using Statistical Methods for Water Quality Management  
Webster and Oliver – Geostatistics for Environmental Scientists, Second Edition  
*Industry, Commerce and Finance*

Aitken - Statistics and the Evaluation of Evidence for Forensic Scientists, Second Edition

Balding - Weight-of-evidence for Forensic DNA Profiles

Lehtonen and Pahkinen - Practical Methods for Design and Analysis of Complex Surveys, Second Edition

Ohser and Mücklich - Statistical Analysis of Microstructures in Materials Science

Taroni, Aitken, Garbolino and Biedermann - Bayesian Networks and Probabilistic Inference in Forensic Science

\*Now available in paperback