# SPATIAL STATISTICS:
## Past, Present, and Future

Proceedings from a Symposium, of the same name,
held on the campus of Syracuse University
from March to June, 1989.

Support for the Symposium, from the George B. Cressey/Preston E. James Geography Endowment Fund and from the Office of the Vice–President for Research and Graduate Studies (both of Syracuse University), is gratefully acknowledged. The site of the Symposium was the Department of Geography and the Maxwell School of Citizenship and Public Affairs of Syracuse University.


DANIEL A. GRIFFITH, Editor

L. Anselin
R. P. Haining
J. K. Ord
B. D. Ripley
D. Wartenberg

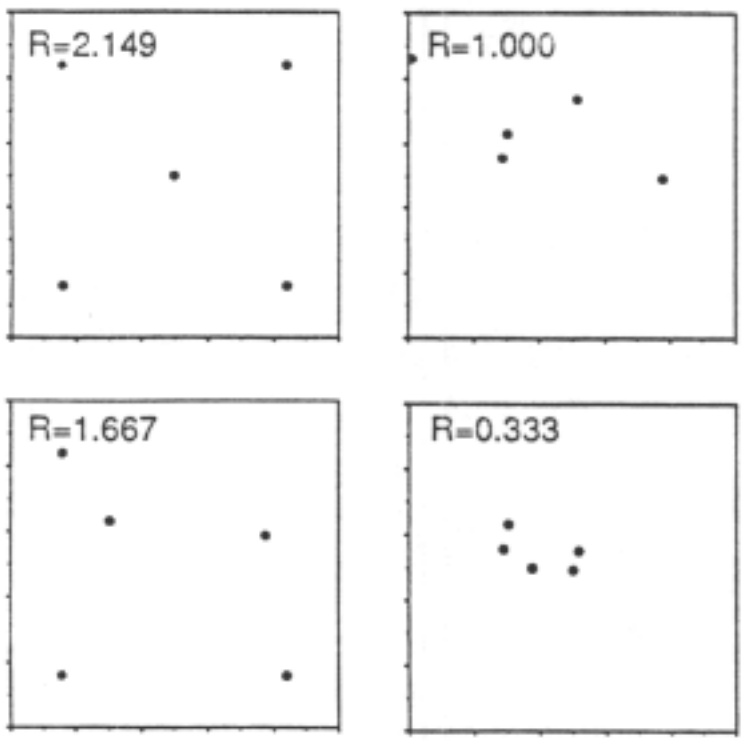
P. Doreian
K. V. Mardia
J. H. P. Paelinck
A. Sen

D. A. Griffith
R. J. Martin
S. Richardson
G. J. G. Upton

Institute of Mathematical Geography


? IMaGe

# PAST




point pattern analysis

# PRESENT

| | | | |
|---|---|---|---|
| A | E | I | M |
| B | F | J | N |
| C | G | K | O |
| D | H | L | P |

geographic configuration

| | | | |
|---|---|---|---|
| 19 | 83 | 84 | 13 |
| 38 | 55 | 58 | 26 |
| 50 | 41 | 38 | 75 |
| 16 | 78 | 23 | 27 |

no spatial autocorrelation

| | | | |
|---|---|---|---|
| 84 | 83 | 58 | 38 |
| 78 | 75 | 50 | 27 |
| 55 | 41 | 26 | 19 |
| 38 | 23 | 16 | 13 |

positive spatial autocorrelation

| | | | |
|---|---|---|---|
| 50 | 23 | 58 | 38 |
| 19 | 84 | 16 | 55 |
| 78 | 13 | 83 | 27 |
| 38 | 75 | 26 | 41 |

negative spatial autocorrelation

Gibb's states

# FUTURE

?

# TO LESLIE CURRY

WHO FIRST INTRODUCED THE EDITOR TO
THE PROBLEM AREA OF
SPATIAL AUTOCORRELATION

This book is published, on an on–demand basis, by the Insititute of Mathematical Geography (IMaGe), 2790 Briarcliff Street, Ann Arbor, Michigan 48105-1429 (Sandra L. Arlinghaus, Director). Printed and illustrative copy were provided by the Authors to the Editor of this monograph. Effort was made to achieve consistency between and within chapters without discarding individual or cultural preference. Thus, both American and British spellings of, for example, neighborhood/neighbourhood appear within the document but not within a single chapter. Order in the Index is determined, where significant, by American spelling. Typesetting was provided by IMaGe using Plain TEX, a trademark of the American Mathematical Society. Output for the master document, prepared by IMaGe, was generated on a laser–printing Xerox 9700, compatible with TEX, at The University of Michigan. Copies were printed from the master document by Michigan Document Services using a Xerox 5090, programmed to mass–produce books on recycled 20 pound paper with colored parchment inserts.



ISBNs: 1-877751-42-1; 1-877751-43-X.

# Preface

This reader is one product of a symposium held at Syracuse University, hosted by the Geography Department of the Maxwell School of Citizenship and Public Affairs, during the Spring semester of 1989. Each chapter is a contribution by a scholar affiliated with this symposium; most chapters essentially are written versions of presentations made at Syracuse University under the same or similar titles. Two exceptions are those chapters penned by Ord and by Griffith. Keith Ord was unable to deliver his lecture because he was on sabbatical leave from the Pennsylvania State University to the London School of Economics and Political Science during this time period. I have contributed a chapter based upon an invited paper presented to the Sixth European Colloquium on Theoretical and Quantitative Geography, held at the Centre Culturel "Les Fontaines" in Chantilly, France, during September of 1989.

Plans for this Symposium were stimulated by efforts associated with the 1987/88 competition for the NSF National Center for Geographic Information and Analysis (NCGIA), and the NSF Science and Technology Center (STC) program. Motivation for its organization and development arose from the NSF NCGIA solicitation calling for "improved methods of spatial analysis and advances in spatial statistics ..."; this is one reason why two of the lecturers had itineraries that included a presentation at the SUNY/Buffalo NCGIA site during their stays in Syracuse. A second impetus was supplied by an STC proposal for a "Center for Spatial Statistics and Spatial Econometrics." The organization of the Symposium was further inspired by a joint proposal with the Institute of Mathematical Geography, submitted to the Geography and Regional Science Program of NSF, requesting funds to support it.

The objectives of this Symposium were to evaluate the role that contributions in spatial statistics have played, are playing, and should play, in expanding the scope of spatial analysis in geography and the spatial sciences, to critically review state–of–the–art practices, and to establish a research agenda for the future. This edited collection of lectures should establish a timely foundation for building bridges to future applications of spatial statistics, and will disseminate findings that are at the frontiers of applied statistics to the international research community.

A general goal of this Symposium was to promote greater awareness of complications caused by the presence of spatial structure and spatial dependence that lie dormant in data sets, especially in terms of their effect on the validity of traditional statistical analysis. Much of the early work in this area was devoted to the development of indices to measure spatial dependence. Meanwhile, more recent advances, refinements, and applications in this field have been summarized in a number of books; but while these publications provide useful summaries of conceptual developments, numerical examples and issues, and case studies of particular problems, they fail to furnish an historical perspective, to assess meaningful progress to date, or to outline important problems for future research. The very recent spatial analysis literature is, however, lightly peppered with pieces invoking this broader viewpoint, suggesting the timeliness of the symposium theme. Given the emergence of this theme in the current literature, together with its attainment of increasing prominence, the topics and published results of this Symposium should find a receptive audience among researchers from many disciplines dependent on spatial analysis.

The Symposium lecturers visited Syracuse in accordance with the following schedule:

| *Lecturer* | *Date of visit* | *Date of lecture* |
|---|---|---|
| R. Haining | 3/18—4/05/89 | 3/31/89 |
| B. Ripley | 4/03—4/10/89 | 4/07/89 |
| R. Martin | 4/10—4/23/89 | 4/14/89 |
| S. Richardson | 4/16—4/22/89 | 4/21/89 |
| A. Sen | 4/22—4/29/89 | 4/27/89 |
| P. Doreian | 5/01—5/08/89 | 5/05/89 |
| K. Mardia | 5/06—5/14/89 | 5/12/89 |
| D. Wartenberg | 5/17—5/21/89 | 5/19/89 |
| J. Paelinck | 5/24—5/31/89 | 5/26/89 |
| G. Upton | 6/03—6/12/89 | 6/05/89 |
| L. Anselin | 6/09—6/14/89 | 6/09/89 |

The first four of these lectures also were part of the annual Geography Department Colloquium series. Ashish Sen and Graham Upton were the two lecturers who visited the NCGIA at SUNY/Buffalo.

The format of this compendium was shaped, in part, by several invaluable suggestions from lecturers. Robert Haining proposed that the context of papers as well as the audience would benefit from published discussions; the model he had in mind was that employed by the *Journal of the Royal Statistical Society*. Hence, I asked each Symposium participant to both referee one of the other papers, and write a discussion of one of the other papers. In retrospect, not only was Haining's suggestion a good idea, but I feel that the quality of both the papers and the book have been considerably enhanced by it. A second recommendation was made by Graham Upton, who mentioned that an editor's preamble to each paper would be a useful and integrative touch; the model he had in mind was that of a standard science fiction anthology. Again, I believe that this addition has greatly strengthened this publication. The quotations that I have cited were gleaned from *Best Quotations for All Occasions* and *The Pocket Book of Quotations*. Each quotation was judiciously selected in an attempt to capture the flavor of its accompanying paper as well as some special personality trait of the paper's author. As an aside, the topical mapping I found most suitable is as follows:

Anselin — statistics
Doreian — imagination
Griffith — perseverance
Haining — progress
Mardia — wisdom
Martin — criticism
Ord — time
Paelinck — thought
Richardson — inspiration
Ripley — knowledge
Sen — eloquence
Upton — teaching
Wartenberg — learning

At this time I would like to express my sincere appreciation to all of the Symposium participants for making these supplemental sections of the book a true success. The general organizational format used here is the one that I developed earlier for my three edited NATO Advanced Studies Institute volumes.

Each Symposium lecturer stayed in my home while visiting Syracuse, and took part in selected extracurricular activities during his/her stay. We had very enjoyable times hosting numerous receptions, attending special campus luncheons, sampling various restaurants of the city, and making site–seeing tours of the area. Most participants visited the Dinomania robotics dinosaur exhibit. A variety of other activities have made the time horizon of this Symposium memorable, too. For example, Haining accompanied us to an outstanding production by the Syracuse University experimental theater. We engaged in "Finger Lakes" wine tasting at the Plane's vineyard on Cayuga Lake with Ripley. Mardia and I scoured the rare book market of Syracuse. We visited Manas Chatterji, at SUNY/Binghamton, with Paelinck. Sen helped us become acquainted with the single East Indian restaurant of the city. Upton explored the Onondaga County park of Pratt's Falls with us. And, we hosted a backyard bar–b–que with Wartenberg. In addition, most of the European lecturers arrived by airplane in Toronto, facilitating visits to Niagara Falls during the trip between Syracuse and Toronto. All in all, my family helped me to plan several special events that accented each lecturer's stay.

Financial support from the George B. Cressey/Preston E. James Geography Endowment Fund of Syracuse University, and the Syracuse University Office of the Vice–president of Research and Graduate Studies is gratefully acknowledged. The patience, tolerance, and graciousness of my wife through a long parade of visitors to our home is most appreciated, too.

Daniel A. Griffith
Syracuse, New York
February 10, 1990

# Contents

*PREAMBLE*

*Knowledge is of two kinds. We know a subject ourselves, or we know where we can find information upon it.*

S. Johnson, **Boswell's Life of Johnson**

*17th century: discovery by Isaac Newton of the binomial theorem–18th century: Bernoulli's publication of the first book devoted to probability theory–19th century: development of the least squares principle by Gauss and Legendre, and emergence of the subject of statistics–20th century: Kendall, Fisher, Pearson, and the emergence of multivariate statistics–and, in our own day the appearance of spatial statistics. In many of his publications, Ripley has imparted knowledge to the spatial statistics audience about Gibbsian interaction models, one of the greatest successes of this new field. This paper furnishes a history of their development, with Ripley this time imparting knowledge to the spatial statistics audience concerning where in the literature different advances can be found. The purpose of this paper is twofold, namely, (1) to document the evolution of Gibbsian interaction models, and (2) to provide examples of their use. Thus, not only have the horizons of statistical science expanded, throughout the centuries, but they also continue to expand. Martin corroborates this latter contention by noting that much of the progress chronicled by Ripley is of very recent origin.*

The Editor

# Gibbsian Interaction Models

Brian D. Ripley *

*Department of Statistics, University of Strathclyde, 26 Richmond Street, Glasgow G1 1XH, UK.*

**Overview:** Gibbsian interaction models, often referred to as *Markov random fields*, have been one of the greatest successes of Spatial Statistics. They encompass conditional spatial autoregressions and a wide class of models for interacting point patterns. Originally borrowed from ideas in statistical physics, they have been applied to both regional measurements (lattice and non-lattice) and to point processes. The distinctive feature of their development in statistics has been the emphasis on statistical inference, in fitting models and estimating parameters. Over two decades a very satisfactory methodology has been developed, but this has never been documented in simple terms. This paper aims to document this progress and to provide some examples of the use of the methodology.

## 1. Introduction

It is common to classify the underlying forces producing a spatial pattern as either exogenous, producing *heterogeneity* or internal, producing *interaction*. The two effects tend to occur in opposite directions. For example, with point patterns (Figure 1), exogenous forces tend to produce 'patchy' patterns which are similar to those produced by clustering, whereas interaction tends to produce patterns which are more 'regular' than one would expect to happen by chance. This comment is not universal, and it *is* possible to produce clustered patterns by the methods described here. However, their *raison d' être* is to give us methods to describe interacting systems. These can be either systems of points such as Figure 1, in which case the interaction is reflected in their spatial positions, or systems of regions, where the interaction is reflected in their values. Geographical examples include the locations of market towns and supermarkets as point patterns (Glass & Tobler, 1971; Rogers, 1974; Ripley, 1979b) and Robert Haining's study of petrol pricing in Sheffield (Haining, 1983; Bennett & Haining, 1985).

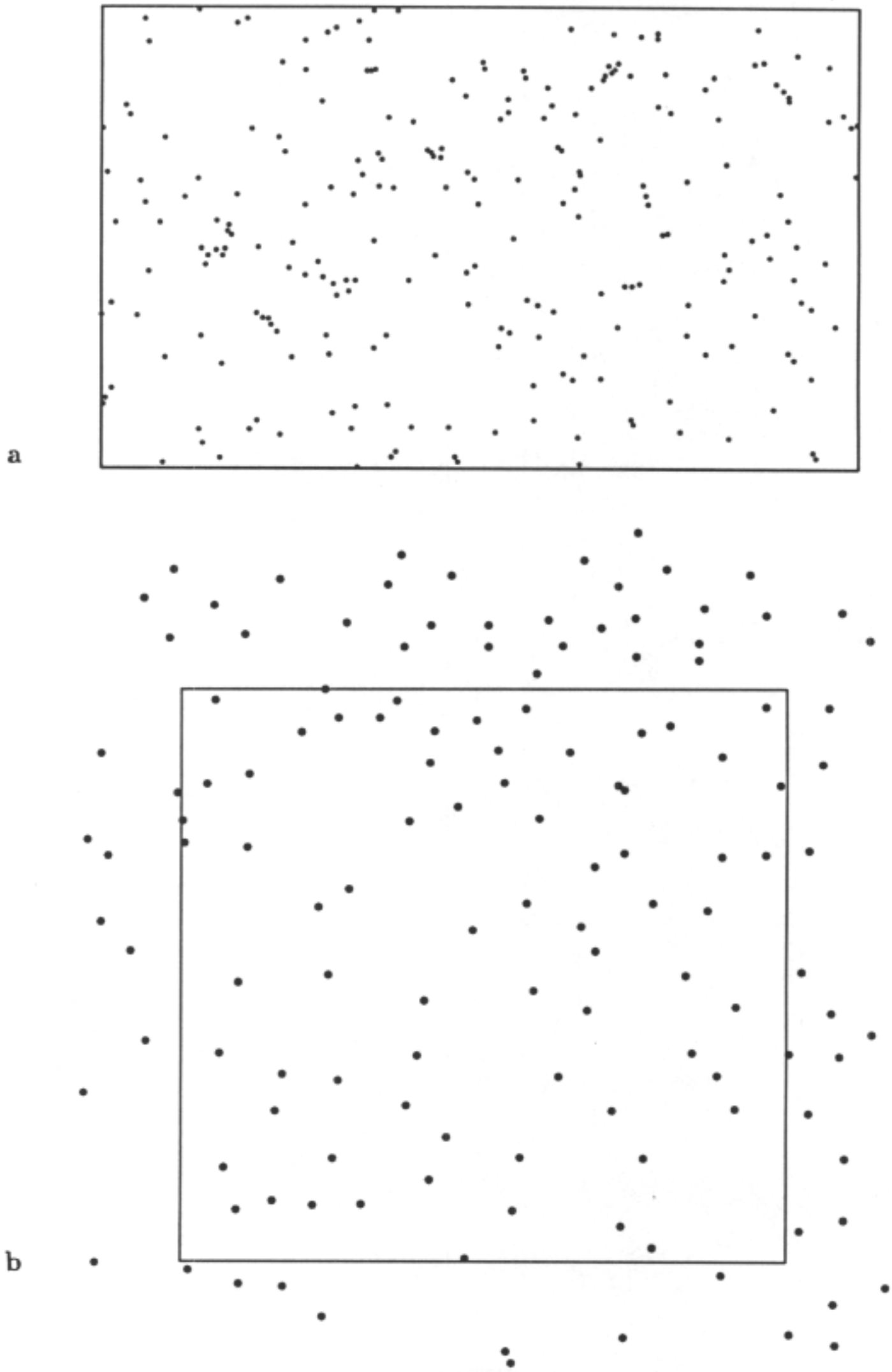
We will take for granted that studies of interaction are appropriate to at least *some* problems in geography; some of the issues here are discussed by other authors in this volume. It is worth noting that it is very difficult to consider exogenous forces and interaction simultaneously, especially for point patterns. One reason is that the interaction may be density–dependent. Suppose we studied the pattern of nests of small songbirds in a wood. Then we will expect territorial behaviour to lead to interaction and some regularity of spacing, *but* in parts of the wood where conditions are especially favourable the territories will be smaller and interaction stronger. Something similar might be expected to happen in the petrol–price competition study.

Gibbsian interaction models subsume what are often known as *Markov random fields*, in which sites or points interact if and only if they are neighbours. Of course the concept of 'neighbour' needs careful definition, but this has been extensively studied in a geographical

* After October 1, 1990, Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK.

**Figure 1.**

*Heterogeneity* (a) and *interaction* (b) in point patterns. (a) shows 232 meat stores in a 6.7744.158 km ward of Tokyo (Okabe & Miti, 1984). (b) shows towns on the Spanish plain, with a 40 mile square containing 70 of the 136 points (after Glass & Tobler, 1971).



a

b

context as part of the study of spatial autocorrelation. Formally, sites form a *graph*, with an edge joining every pair of sites which is a neighbour, giving a structure such as Figure 2. For definiteness, suppose we have $N$ sites as in Figure 2, and a random variable $X_s$ defined at each site. Then a (joint) probability distribution for the set of random variables, $P(X_1, \ldots, X_N)$, defines a Markov random field if for each site $s$, $P(X_s \mid X_t, t \neq s)$, the conditional distribution of the variable at $s$ given the values at all other sites, depends only on the values at sites which are neighbours of $s$. This is a minor restriction to local interactions. The fame of Markov random fields comes from the so–called Clifford–Hammersley theorem, which states that for a Markov random field, under a *positivity* condition,

$$P(X_1, \ldots, X_N) \propto \prod_{\text{cliques}} \phi(X_{i_1}, \ldots, X_{i_m}) \tag{1.1}$$

Here a 'clique' is a maximally connected component of the graph, that is a set of sites all of which are neighbours of all others, and which cannot be expanded (so the term is sociologically appropriate), and $\phi$ is an *interaction* function depending on the variables at all sites in the clique. This result was stated by Clifford and Hammersley in the early 1970's, and a long proof was in private circulation. They never published, and it became clear that the result had simpler proofs (*e. g.* Besag, 1974; Preston, 1974, 1976). The positivity condition essentially precludes processes for which certain combinations of values of $(X_1, \ldots, X_N)$ are excluded in a way inconsistent with (1.1).

The result (1.1) is often stated in an equivalent form. Let $V = -\log_e \phi$. Then

$$P(X_1, \ldots, X_N) \propto \exp[-\sum_{\text{cliques}} V(X_{i_1}, \ldots, X_{i_m})]$$

In this form $V$ is known as a *potential*, as this is the form in which it arises in statistical physics. The constant implied by the proportional sign is in most cases impossible to express in a closed form.

The importance of (1.1) is *not* the theorem, which guarantees its existence, but that it suggests a way of defining the probability distribution of interacting systems. In practice the interaction function $\phi$ is taken to be one for all but very small cliques. The simplest case is to consider cliques of only one site (a site being a neighbour of itself); we find

$$P(X_1, \ldots, X_N) \propto \prod_{\text{sites}} \phi(X_s)$$

so the random variables are independent. The first non–trivial case is to consider only pairs of sites which are neighbours. Expression (1.1) then becomes

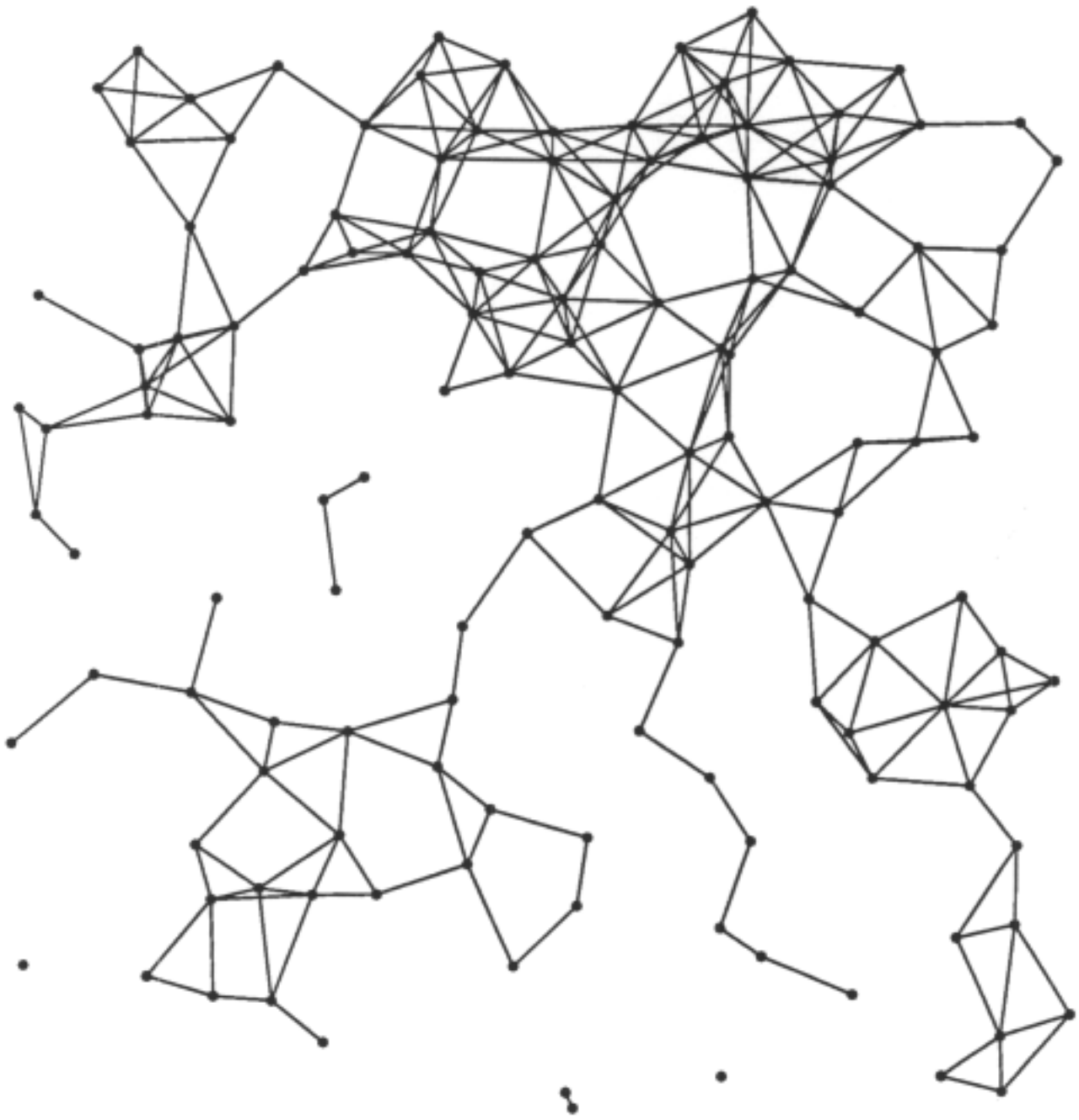
$$P(X_1, \ldots, X_N) \propto \prod_{s \text{ nhbr of } t} \phi(X_s, X_t) \tag{1.2}$$

and almost all Markov random fields ever used are of this form. (Rarely, three–way interactions are considered.) It seems unnecessary to insist that sites interact only if they are neighbours, but rather we should allow a gradual diminution of interaction with distance. We can easily relax (1.2) to

$$P(X_1, \ldots, X_N) \propto \prod_{s,t} \phi(X_s, X_t) \tag{1.3}$$

**Figure 2.**

An illustration of a graph joining the Spanish towns less than 6 miles apart.



called *pairwise interaction processes*, and these are the principal subject of this paper, as well as much recent work.

The term 'Gibbsian interaction' processes comes from statistical physics, where such models have been used for nearly a century to describe the behaviour of gases. Gravitational and electrostatic forces provide two physical examples in which interactions between objects are purely pairwise. The pairwise assumption makes sense in many other fields as well. Animals in defending territories have only pairwise fights, and this may be supposed to apply also to early human settlements. Nevertheless there are examples in which other

assumptions may be attractive. In competition studies it may be the minimum price in the town which is important, and the sizes of market towns could be governed by some notion of 'catchment area.' Thus we will define the general notion of a Gibbsian interaction process as

$$P(X_1, \ldots, X_N) \propto \prod_{\text{subsets}} \phi(X_{i_1}, \ldots, X_{i_m}) \tag{1.4}$$

with the presumption that the subsets of sites and the interaction function are chosen in some simple way.

An important alternative way to look at (1.4) is via the conditional distributions of the variable at each site given the values at all other sites. From (1.4) we have

$$P(X_s \mid X_t, t \neq s) \propto \prod \phi(X_{i_1}, \ldots, X_{i_m}) \tag{1.5}$$

the product being over subsets containing $s$, and (1.3) becomes

$$P(X_s \mid X_t, t \neq s) \propto \prod_{t \text{ nhbr } s} \phi(X_s, X_t) \tag{1.6}$$

The function $\phi$ must be symmetric, that is, the influence of site $s$ on site $t$ must be the same as that of $t$ on $s$.

Closely analogous ideas are used to define Gibbsian and pairwise interaction processes for point patterns. The variables $X_s$ are replaced by the locations $\mathbf{x}_s$ of the points, and probabilities must be re–interpreted as densities. Indeed, it is the technical details needed to make the simple formulae rigorously correct which makes the point process literature inaccessible to all but the highly mathematically trained. This is unfortunately necessary, but the treatment given in later sections is as non–mathematical as possible (and so should not be relied on as totally accurate, in that technical conditions on probability density functions have been omitted).

The distribution of market towns in a plain provides a nice example of an interacting point process. We could postulate a pairwise interaction, with strength decreasing with distance, or an interaction dependent on catchment area. Examples are shown in §4.

The Reference list at the end of the paper serves as a Bibliography to cover the papers of major methodological importance. Clearly from a biased view, I use my two monographs (Ripley, 1981, 1988) as reference sources. *'Spatial Statistics'* is meant for a general readership (but has been thought terse and mathematically tough) whereas *'Statistical Inference for Spatial Processes'* covers recent developments and their impact on statistical theory. It *is* intended to be mathematically tough, but is the only reference source for much of the material sketched here.

## 2. Regional processes

Interaction processes and Markov random fields were originally developed for lattice processes by Bartlett and Besag (then Bartlett's research assistant); Bartlett (1975) reports on those developments, which were motivated by looking at plausible space–time processes at just one time point. Much of the complication arises from the assumption of *stationary* lattice processes, that is processes defined throughout the two–dimensional lattice $\{(x, y) \mid x, y \text{ integer}\}$

in a translation–invariant way. My own belief is that processes on regular lattices are rarely useful except in connection with man–made phenomena; this includes their recent use in image analysis in which the pixels are man–made.

Besag (1975) gave an important insight that these processes do not need to be defined on a (regular) lattice, and it was in his form that they are given in our §1. If the random variables $(X_1, \ldots, X_n)$ have a normal distribution (jointly), this is defined by the vector of their means $(\mu_1, \ldots, \mu_N)$ and the covariance matrix $\Sigma = (\sigma_{ij})$ where $\sigma_{ij} = \mathrm{cov}(X_i, X_j)$. Then the conditional distribution of $X_i$ given the remaining sites is normal with mean $\mu_i + \sum_{j \neq i} b_j(X_j - \mu_j)$ and variance $\Sigma_{kk} - \mathbf{a}^T A^{-1}\mathbf{a}$, where $A$ is the covariance matrix of the remaining sites, $\mathbf{a}$ is the vector of covariances of site $i$ with the remaining sites, and $\mathbf{b} = A^{-1}\mathbf{a}$. Only the mean depends on the values at the remaining sites, so we have a Markov random field if and only if $b_j = 0$ unless $j$ is a neighbour of $i$. This has to be true for all $i$, and reduces to the assumption that the elements $\Sigma_{ij}^{-1}$ are non–zero only if $i$ and $j$ are neighbours. Note that since only covariances need to be considered, Gaussian processes are *always* pairwise interaction processes.

Markov random fields are normally specified by giving the conditional distributions of $X_s$ given the values at the remaining sites. In the Gaussian case we have found that the means are linear, so

$$E(X_s \mid X_t, t \neq s) = \mu_s + \sum_{t \neq s} B_{st}(X_t - \mu_t) \tag{2.1a}$$

$$\mathrm{var}(X_s \mid X_t, t \neq s) = \kappa_s \tag{2.1b}$$

say, where the matrix $B$ and numbers $(\kappa_s)$ are defined by the expressions in the previous paragraph and $B_{ss} = 0$. It is important to note that we can regard (2.1) as the *definition* of the process, which in this form is called a *conditional autoregression*. Elementary calculations (Ripley, 1981, p.89) give the mean and covariance matrix of $(X_s)$ as $(\mu_s)$ and $\Sigma = (I - B)^{-1}\mathrm{diag}(\kappa_s)$ respectively. A covariance matrix must be positive–definite and symmetric, which imposes an awkward condition on $B$. In particular, we must have $B_{st}\kappa_t = B_{ts}\kappa_s$ for all distinct pairs of sites $s, t$. The easiest case is when the conditional variances $(\kappa_s)$ are all equal, in which case the necessary and sufficient condition is that the matrix $(I - B)$ be positive–definite symmetric. I have given these conditions in some detail because they are often ignored in the literature. *Not every choice of connection matrix $B$ in (2.1a) defines a valid conditional autoregression.*

These conditional autoregressions should not be confused with another class of spatial processes used slightly earlier, now called *simultaneous* autoregressions. These are defined by

$$X_s - \mu_s - \sum_{t \neq s} B_{st}(X_t - \mu_t) = \epsilon_s \tag{2.2}$$

where the $\epsilon_s$ are *independent* normally distributed random variables.

The Gaussian case is by far the most important, but Besag (1974) showed that there were three other interesting cases of simple Gibbsian interaction processes, the auto–logistic, auto–binomial and auto–Poisson models, with the auto–logistic being a special case of the

auto–binomial. Their specifications are, for the auto–binomial

$$X_s \sim \text{binomial } (n, p_s)$$
$$\log(p_s/(1-p_s)) = \alpha_s + \sum_{t \text{ nhbr } s} B_{st} X_t$$

and for the auto–Poisson

$$X_s \sim \text{Poisson } (\mu_s)$$
$$\log(\mu_s) = \alpha_s + \sum_{t \text{ nhbr } s} B_{st} X_t$$

where for the auto–Poisson $B_{st} \leq 0$ so only competition between sites is allowed. The auto–logistic process is the special case $n = 1$ of the auto–binomial. They correspond to the special case of

$$\phi(X_s, X_t) = B_{st} X_s X_t$$

in (1.3) or (1.6), and Besag (1974) showed that these were the only discrete distributions for which this linear form is allowable. In all these processes the crucial parameter is the matrix $B$ governing the interactions. Knowledge of the subject is supposed to allow us to specify $B$ and $(\mu_i)$ [or $(\alpha_i)$] up to a few parameters, with the rest to be estimated from the data. For example, in studies of spatial autocorrelation (*e. g.* Cliff & Ord, 1981) it is common–place to specify a matrix $W$ of connection weights. If this is symmetric, we might take $B = \rho W$ for small positive $\rho$ to be estimated from the data. (In a weak sense, tests of spatial autocorrelation are tests of $\rho > 0$ *vs* $\rho = 0$ in this model.) In studies of plant competition, Mead (1971) based the weight function on the Voronoi polygons (also termed Dirichlet cells or Thiessen polygons) of the plants (representing 'catchment area'), and other suggestions are in Ord (1975) (although note that both these authors worked with simultaneous autoregressions).

### 2.1. Parameter estimation

In the early 1970's when these models were first proposed, people shied away from maximum likelihood estimation of the parameters in $B$ and in the mean vector. The computational difficulty comes from the normalizing constant which we have conveniently ignored up to now. For a conditional autoregression the joint probability density of $(X_1, \ldots, X_N)$ is (from its specification as a multivariate normal distribution)

$$\frac{|I-B|^{1/2}}{(2\pi\kappa)^{N/2}} \exp[-\frac{1}{2\kappa}(\mathbf{X}-\mu)^T(I-B)(\mathbf{X}-\mu)]$$

so minus twice the log likelihood is given by

$$N \ln 2\pi\kappa - \ln|I-B| + (\mathbf{X}-\mu)^T(I-B)(\mathbf{X}-\mu)/\kappa \tag{2.3}$$

This is easily maximized over $\kappa$ to give

$$\hat{\kappa} = N^{-1}(\mathbf{X}-\mu)^T(I-B)(\mathbf{X}-\mu)$$

so parameters in $B$ are chosen to minimize

$$N \ln \hat{\kappa} - \ln|I-B| \tag{2.4}$$

The perceived difficulty is the determinant $|I - B|$, but for realistically sized problems it is quite straightforward to minimize (2.4) numerically. In the special case $B = \rho W$ Ord (1975) pointed out that we can exploit

$$|I - \rho W| = \prod(1 - \rho\lambda_i)$$

the product being over the eigenvalues of $W$. For large systems on a regular lattice there are also asymptotic approximations to $|I - B|$ dating from Whittle (1954).

However, we should ask whether maximum likelihood is desirable *per se*. The classical justifications of maximum likelihood in statistics are asymptotic, for an infinite sequence of identically distributed *independent* observations. This is not a natural asymptotic regime in spatial statistics, although Mardia & Marshall (1984) have proved that for a related type of asymptotics some classical results hold. In general, we do not even know if (2.3) is a well–behaved likelihood function. Ripley (1988, Chapter 2) demonstrates that it can fail to be concave, but that with a known mean vector and $B = \rho W$ it *is* unimodal. (Much of the statistical theory of likelihood functions depends on concavity.) Thus although maximum likelihood estimation is almost always possible computationally, we must not assume that it is necessarily statistically desirable. The literature on this point is often misleading or plain wrong. [For example, Upton & Fingleton (1985, p. 284) quote classical results without comments on the lack of applicability of these results, and they are by no means alone.]

An alternative, *pseudo–likelihood* estimation, was introduced by Besag (1975). The pseudo–likelihood is defined as the product of the conditional densities $P(X_s \mid X_t, t \neq s)$, and is treated like a likelihood. There seems no simple explanation as to why this is a sensible idea, but it has proved to work well in practice. Besag (1975) sketched a proof that the pseudo–likelihood estimator would be consistent (converge to the true value) as the problem is increased, and an elegant general proof is given by Geman & Graffigne (1987). For a conditional autoregression we have

$$\ln PL = -\frac{N}{2}\ln(2\pi\kappa) - \frac{1}{2\kappa}\sum_s [(I - B)(\mathbf{X} - \mu)]_s^2$$

so pseudo–likelihood estimation amounts to least–squares fitting to the "residuals"

$$\eta_s = X_s - \mu_s \sum_{t \neq s} B_{st}(X_t - \mu_t)$$

Be warned that (2.1) cannot be treated like an ordinary regression, and that the residuals $(\eta_s)$ will themselves be spatially autocorrelated, unlike the $(\epsilon_s)$ in the simultaneous autoregression (2.2).

Pseudo–likelihood methods have been very successful, and have largely supplanted all others. Older methods such as 'coding' for regular lattices (Besag, 1974) should have disappeared by now!

## 2.2. Applications

These processes have been used less widely than their early promise suggested, and there are probably more papers on theory than applications. Part of the cause is that most attention has been to regular lattices, and as we suggested earlier, these are inevitably man–made. Another problem is defining well the set of sites to be considered. Interacting systems just do not occur in closed boxes, and it is always necessary to consider the effects of interaction with outside sites. Unless the number of sites is very large, there will be a high proportion of sites near the edge, and so it may be perilous to ignore the effect of the outside world. One reasonably successful application of spatial lattice processes has been to agricultural field trials, in which the system is closed and carefully controlled. Even there the impact has been more in emphasising the design of the trials to minimize spatial variation in soil fertility that in methods based on spatial autoregressions, except perhaps as part of rescue attempts on poorly designed experiments.

It seems that applications in image analysis will be successful. There the systems considered are vast by the standards of the 1970's, so edge effects may safely be neglected. We have been using Gibbsian interaction processes with astronomical images of $1024 \times 656$ pixels (Ripley, 1990; Molina & Ripley, 1989).

# 3. Point processes

Point patterns consist of $n$ point locations within a specified domain $D$, specified by their Cartesian coordinates $\{\mathbf{x}_i\} = \{(x_i, y_i)\}$. They differ from classic multivariate statistics in two ways. First, the number of points $n$ is thought of as variable, but more importantly, the coordinates of the points have very little rôle, the emphasis being on the *configuration* of the points. The domain $D$ has to be defined carefully, since edge–effects are important. There are two rather different cases. Either $D$ can be a natural region such as an island, or it can be thought of as a window into a much larger region, such as the examples of towns on a plain and supermarkets mentioned in §1.

Some further examples of studies of point patterns in geography are drumlins in Northern Ireland (Hill, 1973; Upton & Fingleton, 1985, pp. 68-9; Figure 7), schools in Southampton (Pinder & Witherick, 1972; Upton & Fingleton, 1985, pp. 76-7) and retail establishments in a Tokyo district (Okabe & Miki, 1984; Figure 1a).

The model for no interaction between points is the Poisson process. The number of points in $D$, $N(D)$, has a Poisson distribution with mean $\mu$, and the points are independently distributed over $D$ with density function $f$. Then $N(A)$, the number of points within a sub–region $A$ of $D$, has a Poisson distribution with mean $\mu \int_A f$. Further, the numbers of points in non–overlapping regions are independent. This means that we can consider a domain $D$ in isolation, as with a Poisson processes there is no interaction with points outside $D$. We will normally consider *homogeneous* processes in which case the density $f$ is uniform, and $\mu = \lambda \times \text{area}(D)$ where $\lambda$, the number of points per unit area, is an important parameter known as the *intensity*.

Gibbsian interaction processes are defined by stating how much more (or less) likely each configuration of points is than under a Poisson process. For example, pairwise interaction

processes have density (with respect to a Poisson process) of

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n) \propto \prod_i b(\mathbf{x}_i) \prod_{i<j} h(\mathbf{x}_i, \mathbf{x}_j) \tag{3.1}$$

The first product corresponds to a change of the underlying spatial intensity of points from $\mu \times f$ to proportional to $b \times f$, but the interaction is expressed by the second product. Since we almost always regard interaction as *stationary* throughout space, we can take $h(\mathbf{x}, \mathbf{y})$ as a function only of $d(\mathbf{x}, \mathbf{y})$, the distance between the two points. Then $h$ can be thought of as an *interaction function*, and examples are shown in Figure 3. More general interactions are possible, but have never been found necessary. (These processes were introduced by Ripley, 1977.) We will normally consider processes for which $b$ and $f$ are constant, so there is no heterogeneity in the model.

The simplest special cases are the so–called Strauss processes illustrated in Figures 3(a) and 3(b). The interaction function is defined by

$$h(d) = \begin{cases} c & \text{if } d = d(\mathbf{x}, \mathbf{y}) \leq R \\ 1 & \text{otherwise.} \end{cases} \tag{3.2}$$

introduced by Kelly & Ripley (1976) following earlier (incorrect) work of Strauss (1975). The case $c = 0$ is the 'hard–core' process in which points are prohibited from being closer than distance $R$ apart. It can be thought of as being produced by sampling from a Poisson process and throwing out all patterns which violate this condition. For $0 < c < 1$ there is a disincentive to close pairs, the density being proportional to $c^{y(R)}$ where $y(R)$ denotes the number of pairs of points closer than $R$. For $c = 1$ we have a Poisson process, and no interaction. For $c > 1$ the process only exists in a rather pathological way. Close pairs are encouraged, and most realizations for fixed $n$ will have a single 'clump' of points contained within a disk of diameter $R$. For variable $n$ the process cannot even be defined. A similar effect occurs whenever $h(d) > 1$ for some distance $d$, although this can be counter–balanced by $h(d) = 0$ for $d < R$, small.

There are a number of obvious extensions to the Strauss process. If we think of the points as having circular territories of diameter $R$, the interaction might depend on the area of the overlap (given by the term $[\cdots]$ as a proportion of the area of the territories),

$$\ln h(d) = \begin{cases} -\theta \left[ 1 - \frac{2}{\pi} \left\{ \frac{d}{R} \sqrt{\left(1 - \frac{d^2}{R^2}\right)} + \sin^{-1}\left(\frac{d}{R}\right) \right\} \right] & \text{if } d \leq R \\ 0 & \text{otherwise.} \end{cases} \tag{3.3}$$

(Pentinnen, 1984). Another idea is to allow more than one step in the interaction function, the 'multi–scale' process,

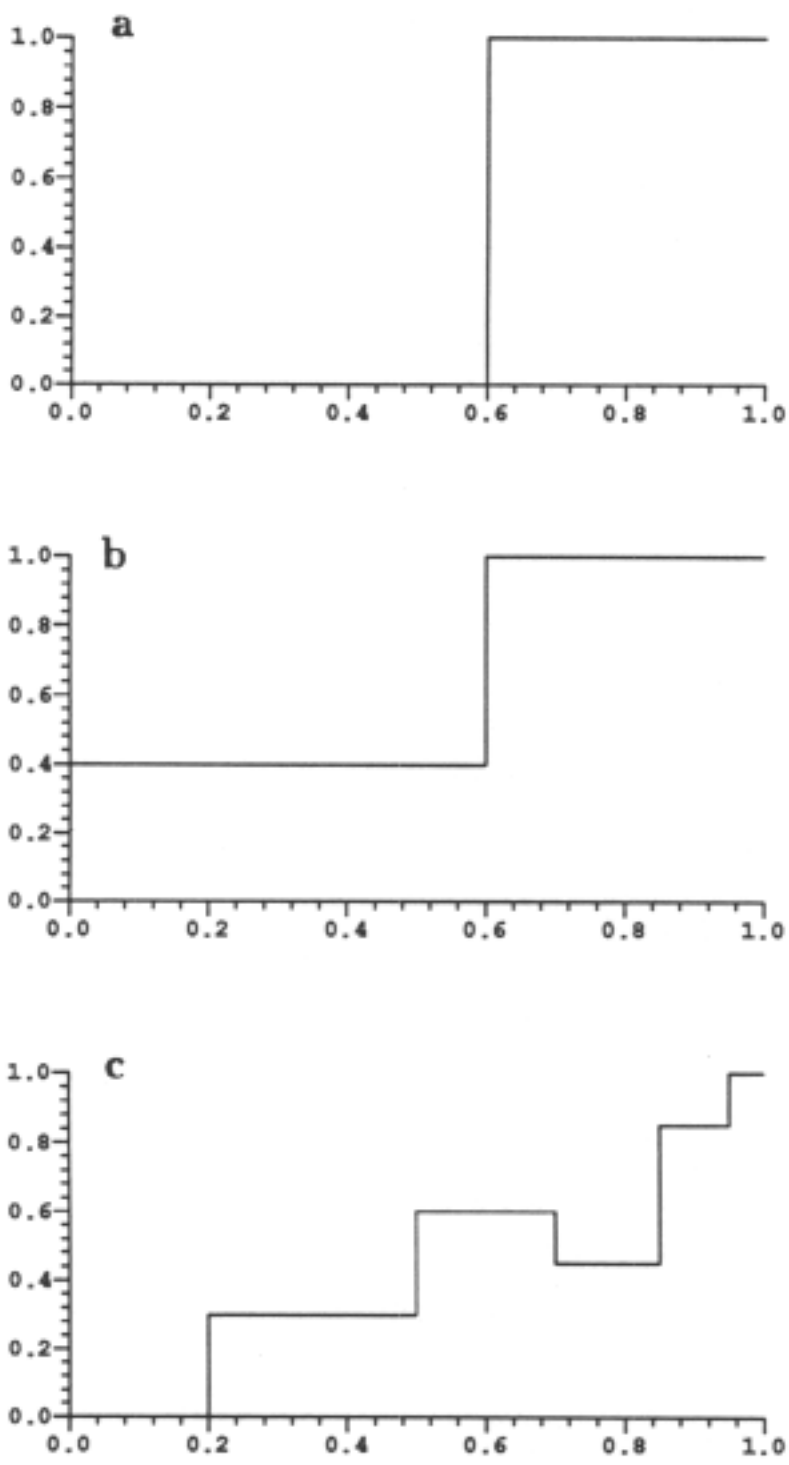
$$h(d) = \begin{cases} c_1 & \text{if } d \leq R_1 \\ c_2 & \text{if } R_1 < d \leq R_2 \\ c_3 & \text{if } R_2 < d \leq R_3 \\ \vdots & \vdots \\ 1 & \text{if } d > R_m \end{cases} \tag{3.4}$$

Pairwise interaction functions have been extensively considered in statistical physics, and other functional forms have been borrowed from that field, including the 'very–soft–core' potential (Ogata & Tanemura, 1984) illustrated in Figure 3(e),

$$h(d) = 1 - e^{-(d/\sigma)^2}$$

**Figure 3.**

Pairwise interaction functions $h$ as a function of distance. (a) 'Hard–core' interaction. (b) Strauss process. (c) Multiscale process. (d) Overlap area (Pentinnen, 1984) with $\theta = 4$. (e) 'Very–soft–core' (Ogata & Tanemura, 1984) with $\sigma = 0.5$.



Although some of these processes have been extensively discussed in statistical physics, the questions asked there are rather different. First, the size of the problem is very different, with of the order of $10^{23}$ points (molecules) so edge effects are indeed negligible. Further, the emphasis is the reverse of ours; the interaction function $h$ is assumed known and properties of the process are required, whereas we have a sample of the process from which to learn about $h$.

There is an analogue for point processes of Markov random fields, the Markov point processes of Ripley & Kelly (1977).

### 3.1. Simulation

One very effective way to understand the definition and parameters of a statistical distribution is draw some samples from it. Indeed, in my first stochastic processes course this was given by David Kendall as a criterion for having defined a process properly, that you could simulate from it. The problem with constructive definitions such as that given earlier for the hard–core process is that they are 'in principle' only. One will eventually obtain a realization of a Poisson process with no pair of points closer than $R$ (provided only that such a configuration can be packed into $D$), but the wait could be very long. Since simulation is such an effective way to understand these processes, it is important to know of reasonably efficient simulation methods.

All the known reasonably efficient simulation methods are iterative. At each stage a point is added or deleted, dependent on the configuration of the remaining points (Ripley, 1977, 1979a, 1987). The simplest case is when we prescribe the total number $n$ of points. Then each step of the simulation consists of deleting a point and replacing it. The point to delete is chosen at random (so each of the $n$ points is equally likely). Relabel the existing points as $\{\mathbf{x}_2, \ldots, \mathbf{x}_n\}$, so the replacement point will be $\mathbf{x}_1$. Then it has density proportional to

$$f_0(\mathbf{x}) = \prod_2^n h(\mathbf{x}, \mathbf{x}_i) \tag{3.5}$$

The unknown normalizing constant is once again an apparent problem, but this can be overcome by the use of *rejection sampling*, a well–known technique in simulation (Ripley, 1987). Consider the case of $h(d) \leq 1$ for all $d$. Then given $\{\mathbf{x}_2, \ldots, \mathbf{x}_n\}$, sample $\mathbf{x}$ from the underlying Poisson density $f$ on the domain $D$, and accept it with probability $f_0(\mathbf{x})$. (The assumption on $h$ ensures that $0 \leq f_0(\mathbf{x}) \leq 1$ and so it is indeed a probability. To accept with probability $p$, get your computer to generate a uniform random variable $U$ and accept if $U < p$.) If the sample is not accepted, try again until one is.

If $n$ is not fixed, a similar birth–and–death process is run, but additions and deletions no longer alternate. Details are given in Ripley (1977).

The theory says that this process produces samples whose distribution converges to that of the pairwise interaction process. The basic idea is due to Metropolis *et al.* (1953), the point process implementation to Ripley (1977, 1979a), the latter containing Fortran code. In practice the process has to be run for some time ($10n$—$100n$ steps) to settle down from a reasonable starting pattern, then it can be sampled every $2n$—$4n$ steps. For moderate $n$ the process runs fast enough on a personal computer to be fascinating viewing. For example, on the towns data shown in Figure 1(b) simulations of the Strauss process change about 10

points per second.

### 3.2. Parameter estimation

Parameter estimation in pairwise–interaction point processes proved to be difficult for a decade, and an obstacle to their wider use. One can of course use trial–and–error methods, matching some aspect of the simulated patterns to the data. This was illustrated for the Spanish towns data in Ripley (1977), using my K–function to measure fit, and there are more extensive comparisons in Ripley (1988, Figure 4.2). Diggle & Gratton (1984) raised this idea to the status of a theory! However, it is both non–intuitive and rather computer–intensive.

Even the simplest cases have shown difficulties. The maximum likelihood estimate of $R$ in a hard–core process is $d_{\min}$, the smallest distance between a pair of points. Since pairs closer than $R$ cannot occur, we know $d_{\min} > R$ and hence the estimator is biased. (This is not unexpected, since the discontinuity in $h$ at $R$ makes this a non–regular likelihood problem.) Ripley & Silverman (1978) showed how to correct this estimator for bias. Even if $R$ is known for the Strauss process, the maximum likelihood estimator of $c$ is not straightforward. The density is

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n) = a(b,c)\, b^n c^{y(R)}$$

and it is the unknown normalizing constant $a(b,c)$ which causes the trouble. To ease the notation, let us write $Y(R)$ for the random variable measuring the number of pairs of points within $D$ closer than $R$. Then after some calculation we find that the maximum likelihood estimators are given by the solution in $b, c$ to the equations

$$n = E_{b,c} N(D) \qquad y(R) = E_{b,c} Y(R) \tag{3.6}$$

if the number of points is allowed to vary, otherwise $b$ is irrelevant and $c$ solves

$$y(R) = E_c Y(R) \tag{3.7}$$

Here $y(R)$ is the fixed number of observed close pairs, and the right–hand side is the expected number of pairs. It is intuitively obvious that this increases from zero for $c = 0$ up to the value for a Poisson process for $c = 1$. The latter depends on the shape of the domain $D$, but is known for several common shapes (Ripley, 1988, pp. 28-9); for small $R$ it is approximately $n(n-1)\pi R^2/2$. It is perfectly possible for us to observe more close pairs than are the average for a Poisson process; in such a case we take $\hat{c} = 1$, although we ought to consider why we are fitting a process with interaction distance $R$!

Maximum likelihood estimation depends on our being able to calculate the means in (3.6). They are not known analytically, but as we shall see in §4, can be estimated by simulation. An alternative is to make approximations. When interactions are rare, we can derive the approximations (Ripley, 1988, pp. 56-7)

$$\hat{c} \approx y(R) \times \frac{2a}{n(n-1)\pi R^2} \tag{3.8}$$

for fixed $n$, and for variable $n$

$$\hat{b} \approx n/a, \qquad \hat{c} \approx y(R) \times \frac{2a}{n^2 \pi R^2}$$

where $a$ denotes the area of $D$. These approximations correspond to fitting a straight line in $c$ to $E_cY(R)$, although they are derived by assuming that any point enters into just one close pair. More sophisticated approximations have been considered, notably by Ogata & Tanemura (1981, 1984), who borrowed approximations from statistical physics for 'non-ideal' gases. Details are given in Ripley (1988, pp. 59-62). For the Strauss process, they give an approximation to the log-likelihood as

$$y(R)\ln c - \frac{n^2}{2a}(c-1)\pi R^2 - \frac{5.79n^3(c-1)^3R^4}{6a^2}$$

corresponding to

$$E_cY(R) \approx \frac{n^2\pi R^2}{2a}c\,[1+1.84(nR^2/a)(c-1)^2] \tag{3.9}$$

a cubic in $c$. (There is an error of a factor of 3 in the constants in Ripley, 1988.)

*3.2.1. Pseudo-likelihood*

We have seen that even in the simplest case maximum likelihood estimation causes difficulties, and just as in the regional data case, there is no reason to accord maximum likelihood methods special status from a theoretical viewpoint. This encourages us to consider pseudo-likelihood as it was so successful there. There are technical problems in the conditioning, but these can be overcome either by approximating by a lattice process (Besag, 1977; Besag, Milne & Zachary, 1982) or via the mathematically sophisticated theory of conditional intensities. The problem is that we want the probability of a point occurring at $\mathbf{x}$ given the locations of the remaining points. This probability is zero, but we can consider

$$\frac{P(\text{point in a region } \Delta \text{ containing } \mathbf{x} \mid \text{other points})}{\text{area}(\Delta)}$$

for a small region $\Delta$ around $\mathbf{x}$. This is the conditional intensity $\lambda(\mathbf{x};\mathbf{x}_1,\ldots,\mathbf{x}_n)$. If $\mathbf{x}$ is the location of one of the existing points it is omitted from the conditioning. Then the pseudo-likelihood is the product of the conditional intensity over all points in $D$, occupied or not, which leads to the log pseudo-likelihood as

$$\sum_1^n \ln\lambda(\mathbf{x}_i;\mathbf{x}_1,\ldots,\mathbf{x}_n) - \int_D \lambda(\mathbf{x};\mathbf{x}_1,\ldots,\mathbf{x}_n)d\mathbf{x}$$

For a pairwise interaction process the conditional intensity is proportional to

$$b(\mathbf{x}) \prod_{\text{pts other than } \mathbf{x}} h(\mathbf{x},\mathbf{x}_i)$$

so the log pseudo-likelihood is

$$\ln[\prod_i b(\mathbf{x}_i)\prod_{i,j} h(\mathbf{x}_i,\mathbf{x}_j)] - \int_D b(\mathbf{x})\prod_i h(\mathbf{x},\mathbf{x}_i)d\mathbf{x} \tag{3.10}$$

At first sight the first term is the likelihood (3.1) without the normalizing constant, but the second product is different in that each pair $\{i,j\}$ occurs twice. Nevertheless, pseudo-likelihood methods are very similar to maximum-likelihood ones, except that the normalizing

function is replaced by much simpler integrals. For the Strauss process the pseudo–likelihood estimators $\tilde{b}$ and $\tilde{c}$ solve

$$\begin{aligned} \tilde{b} \int_D \tilde{c}^{t(\mathbf{x})} d\mathbf{x} &= n \\ \tilde{b} \int_D t(\mathbf{x}) \tilde{c}^{t(\mathbf{x})} d\mathbf{x} &= 2y(R) \end{aligned} \tag{3.11}$$

for variable $n$, where $t(\mathbf{x})$ denotes the number of points of the pattern within distance $R$ of the test point $\mathbf{x}$. (Conditional intensities, hence pseudo–likelihood, make no sense if $n$ is fixed. The intensity will be zero or infinite depending on whether there are $n$ or $n-1$ points elsewhere.) The integrals in (3.11) can be estimated by sums over a grid of points within $D$, and sophisticated methods are available to find $t(\mathbf{x})$ rapidly using auxiliary data structures.

These pseudo–likelihood methods proved to be a special case of a family of moment measures derived by Takacs (1986) and Fiksel (1984). These compare the expected values of observations on the process with similar expectations conditional on a point of the pattern at $\mathbf{x}$. However, whereas the Takacs–Fiksel work depends on arbitrary choices of moments to compare (and is mathematically forbidding), pseudo–likelihood has some theoretical rationale.

### *3.2.2. Edge effects*

All the parameter estimation methods mentioned up to now have been for a process defined only on the domain $D$. This is frequently inadequate, in that we imagine the underlying process placing the points to occur within a much larger domain $D'$ but which is observable (or has only been recorded) within the window $D$. In such a case the number of points is necessarily variable. A rigorous treatment of this case is very difficult, as it would involve averaging over the positions of points outside $D$ which might interact with those within $D$. Rather, we choose to correct the estimators we have seen so far for edge–effects. This is particularly easy for the Strauss process, since the only statistic which occurs is $y(R)$, the number of $R$–close pairs. This occurs in the second–moment K–function which is often used to describe a spatial point pattern, and so much work has been done on correcting for edge–effects, going back at least as far as Glass & Tobler (1971). Full details of the corrections proposed and of their efficacy are given in Ripley (1988, Chapter 3).

## 3.3. Choosing an interaction function

How will we know what shape of interaction function to pick? Almost never is there any appropriate theory to suggest a particular functional form, and it is really the shape rather than the mathematical form which we seek. Two ideas have recently been suggested. The conceptually simplest is that of Ripley (1988, p. 73). Fit a multi–scale process (3.4) to obtain a step–function (histogram–like) estimate of the interaction function $h$, and choose a suitable functional form to fit parametrically. This is just like plotting a histogram in univariate statistics, and choosing a family of distributions (normal, gamma, ...) from its shape.

Fortunately, estimating the parameters in the multi–scale process by pseudo–likelihood is just as easy as for the Strauss process. The equations are similar to (3.11), with $c$ replaced by $c_i$ and $t_i(\mathbf{x})$ denoting the number of points of the pattern whose distance from $\mathbf{x}$ is

between $R_{i-1}$ and $R_i$. That is, $(\tilde{c}_i)$ solve

$$\tilde{b}\int_D t(\mathbf{x})\prod_j \tilde{c}_j^{t(\mathbf{x})}d\mathbf{x} = \sum_{\text{pts}} t_i(\mathbf{x}_j) = 2[y(R_i) - y(R_{i-1})]$$

for each $i$ together with

$$\tilde{b}\int_D \prod_j \tilde{c}_i^{t(\mathbf{x})}d\mathbf{x} = n$$

and the ratio is a function of $(c_1, \ldots, c_m)$ which is easily estimated numerically, and so the equations are solved numerically.

Care is needed in interpreting the shape of the function, as the estimates at small distances are very variable, unlike a histogram. For a Poisson process we would have that

$$\tilde{c}_i \approx [y(R_i) - y(R_{i-1})] \times \frac{2a}{n(n-1)\pi[R_i^2 - R_{i-1}^2]}$$

by an extension of (3.8), and the count of pairs $[y(R_i) - y(R_{i-1})]$ is approximately Poisson. Thus $\tilde{c}_i$ has standard deviation about the square root of the second term, and this will be large if the area of the annulus of points between $R_{i-1}$ and $R_i$ away from a fixed point is small. This will inevitably be the case for distances between 0 and $R_1$. For example, with the Spanish towns data [Figure 1(b)] we have $n = 70$ and the area $D$ is 40 miles square. If we take $R_1$ to be 1 mile, we find a standard deviation of $\tilde{c}_1$ of about $\sqrt{[2 \cdot 40^2/(70 \cdot 69\pi 1^2)]} \approx 0.5$ so any inference would be meaningless.

The other idea borrows from statistical physics a relationship between the interaction function $h$ and the second-moment function $K$ known as the Percus–Yevick formula. Whereas in physics this is used to calculate $K$ from $h$, Diggle, Gates & Stibbard (1987) had the idea to reverse the process to obtain a non-parametric estimator of $h$; full details are in their paper. The Percus–Yevick formula is approximate, and something of a mystery to me, but simulations have shown it to be reasonably accurate. It relates the interaction function $h$ to the pair-correlation function $g$. This is defined by

$$g(t) = (2\pi t)^{-1}\frac{dK(t)}{dt}$$

where $K(t)$ is my reduced second-moment function (Ripley, 1977, 1981). Then the formula is

$$h(t) \approx g(t)/[g(t) - c(t)]$$

where $c(\cdot)$ is the solution to

$$c(t) = g(t) - 1 - \lambda\int_0^{2\pi}\int_0^{\infty}[g(s) - 1]\,c(\sqrt{[t^2 + s^2 - 2ts\cos\theta]})\,s\,dsd\theta$$

Details of how $g(t)$ is estimated and how the equations are solved are given in the paper. The procedures are *not* straightforward, but it their authors' hands give good results for simulations of patterns of a few hundred points.

The cautionary note about the step–function method applies to all; one needs a very large number of points to have much success in fitting a non–parametric estimator of the shape of the interaction function $h$; about fifty points is not enough, rather a few hundred are needed. Examples occur in other fields, but rarely, I suspect, in geography. Okabe & Miki (1984) do have 393 greengrocer shops in their study area (28.2km$^2$ of Tokyo) but few other cities will have so many establishments within a homogeneous area.

## 4. Examples

In this section we consider two examples of fitting Gibbsian interaction models to point patterns drawn from geography. All have been mentioned earlier in the text and analysed previously by various statistical methods.

### 4.1. Spanish towns

The data come from Glass & Tobler (1971). They show the 'cities' (or 'towns', the terms are used interchangeably) in the 40 mile square centred at latitude 2° 30′ $W$ and longitude 39° 47′ $N$ which is claimed to be "especially homogeneous in climate, physiography, transport, population density and economy." An earlier version with 69 points has been considered from various points of view in Ripley (1977, 1979b, 1988). The data shown here have been re–digitized to show towns outside the study square, and 70 came within. The external points were used for edge–correction, for example in the integrals in (3.11). Glass & Tobler suggested fitting a hard–core model with interaction distance $R$ as 3.46 miles, this peculiar figure coming from a simple value of $\bar{\rho}$, the number of points divided by the maximum packing of discs into the same domain. This was fitted despite the presence of pairs of towns about 0.6 miles apart, and indeed there appear to be $y(R) = 30$ pairs closer than 3.5 miles, an appreciable part of the total of 2,346 pairs. Nevertheless, this is appreciably fewer pairs than we would expect for a Poisson process. Ignoring edge effects, we would expect $\lambda^2 a\pi R^2/2 = 57.3$ pairs. [The exact result is known from Borel (1925) as

$$\frac{\lambda^2 a^2}{2}\left[\frac{\pi R^2}{L^2} - \frac{8R^3}{3L^3} + \frac{R^4}{2L^4}\right]$$

for a square of side $L$, but at 53.1 this is only slightly less. (Further, to a fairly crude approximation the number of pairs will have a Poisson distribution, so the standard deviation of the number is about 7.5 and the shortfall is definitely significant. More precise and powerful tests given in Ripley (1979b) all reject the Poisson hypothesis at at least the 5% level, some at the 1% level. This suggests that a Gibbsian interaction process might be appropriate. The interaction function found by fitting a multi–scale process by pseudo–likelihood is shown in Figure 4. (Half–mile intervals were chosen, except at the smallest distances.) Remember that the values at short distances are unreliable. The Figure gives no reason to doubt that a Strauss process would be appropriate, although doubtless many other functional forms for $h$ would also suffice.

We can fit the parameter $c$ of the Strauss process by any of the methods discussed in §3. The exact value obtained depends on what form of edge correction (if any) is used and, of course, the method; but for $R = 3.5$ miles, values of $c$ between 0.4 and 0.5 are suggested (Ripley, 1988, §4.6). These correspond well to the trial–and–error value of 0.5 of Ripley (1977). This is a problem in which edge correction is *essential* as the study region does

**Figure 4.**

The interaction function $h$ fitted to the Spanish towns data by pseudo–likelihood, to suggest a parametric form.

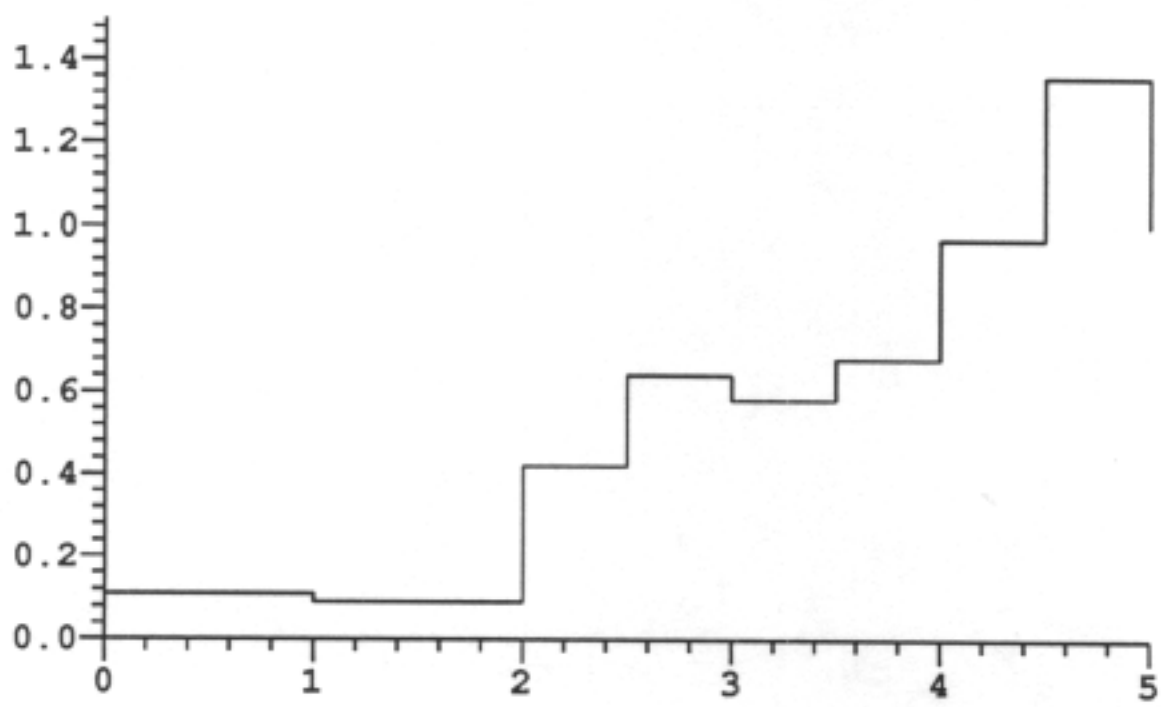
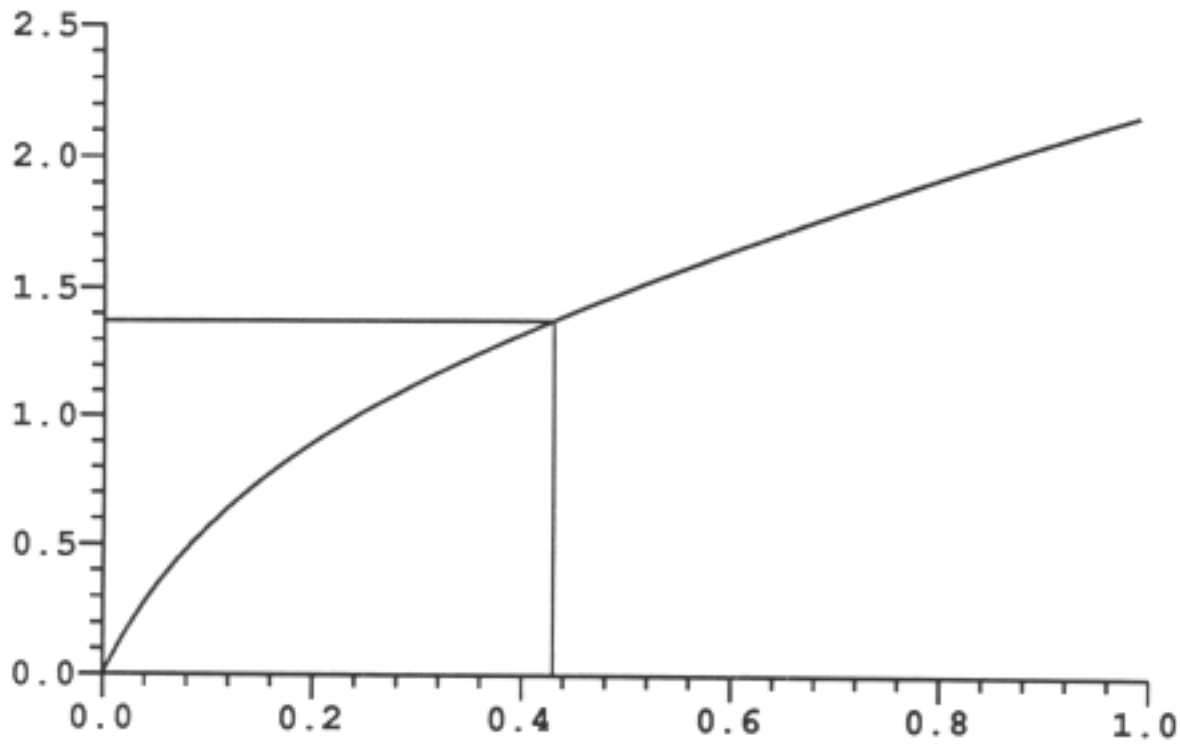



**Figure 5.**

Plot of the ratio of the equations (3.11) against $c$. The pseudo–likelihood estimator, $\tilde{c} = 0.43$ is obtained from the horizontal line at $2y(R)/n = 96/70 = 1.37$. This example is for the Spanish towns with $R = 4$ miles.



not exist in isolation but is part of a very much larger plain. Fortunately the locations of adjacent points are available, but even if they were not, 'internal' edge correction methods work well and were used in my earlier studies.

To illustrate the procedures let us consider $R = 4$ miles, as this is suggested by Figure 4. For all the estimators we need $y(R) = 48$, counting as one half pairs of point with only one of them inside the square. For the pseudo–likelihood estimator the ratio of the left–hand sides of (3.11) is evaluated for a range of values of $c$ by taking a grid of points within the square (I used $10 \times 10$) and averaging $t(\mathbf{x})c^{t(\mathbf{x})}$ and $c^{t(\mathbf{x})}$ at these points. [Remember that $t(\mathbf{x})$ is the number of towns closer that $R$ miles to the test point $\mathbf{x}$.] Figure 5 shows this ratio for all values of $c$ between zero and one. However, this range of values is not needed, as we can get a rough value of $c \approx 0.63$ from (3.8). (Note that this corresponds to fitting a straight line between the endpoints of the curve in Figure 5.) The value of the pseudo–likelihood estimator $\tilde{c}$ is 0.43.

This rough value is an approximation to the maximum likelihood estimator. We can refine it by (3.9) to be

$$c\,[1 + 1.29(c-1)^2] \approx 0.62 = \frac{2ay(R)}{n^2\pi R^2}$$

with solution $c \approx 0.45$. To find the exact maximum likelihood estimator we use simulation to calculate $E_c Y(R)$ and plot this against $c$ (Figure 6). On the other hand, the calculations for the solid line took 35 minutes on a Sun workstation (running the simulation 10,000 steps for each of twenty values of $c$), and still the curve is quite rough. From the observed value of $y(R) = 48$ we find $\hat{c} \approx 0.51$. As Figure 6 shows, the approximation (3.9) is good for large $c$, but fails for strong interactions ($c$ small). The simulations enable us to estimate the variance of $y(R)$ and hence of $\hat{c}$; this suggests a standard error of about 0.10.

Ogata & Tanemura (1984) fitted a very–soft–core model to the old dataset, using their approximation (but to higher order than is available for the Strauss process).

### 4.2. Drumlins in Northern Ireland

Figure 7 shows the pattern of 232 drumlins (glacial deposits) recorded by Hill (1973). Upton & Fingleton (1985) used sampling methods to analyse this pattern, and concluded "for this restricted region there is no evidence of other than a random distribution of 'plants'." More powerful methods show otherwise. For example, plots of my K–function show regularity up to 1 km, principally at distances of less than 600 m. This suggested fitting various multi–scale processes. With 5 equal steps of 200m each I obtained $\tilde{c}_i$ as 0, 0.27, 0.76, 1.30 and 1.06. This shows the interaction to be principally at shorter distances, and the most satisfactory result was obtained with an interaction function of 0.03 up to 150 m, 0.68 between 150 and 300 m, and one for larger distances.

Although there are many more points than in the previous example, the distances are smaller so the number of interacting pairs is only slightly greater [$y(300\text{m}) = 65$]. Thus the standard error of $\tilde{c}_2$ is around 0.08.

The pattern shows clear signs of heterogeneity, especially in the North–West, so it did not seem wise to take this analysis much further.

**Figure 6.**

Plot of $E_c Y(R)$ against $c$ for the Spanish towns example. The solid line is from simulations, the dashed line is equation (3.11).

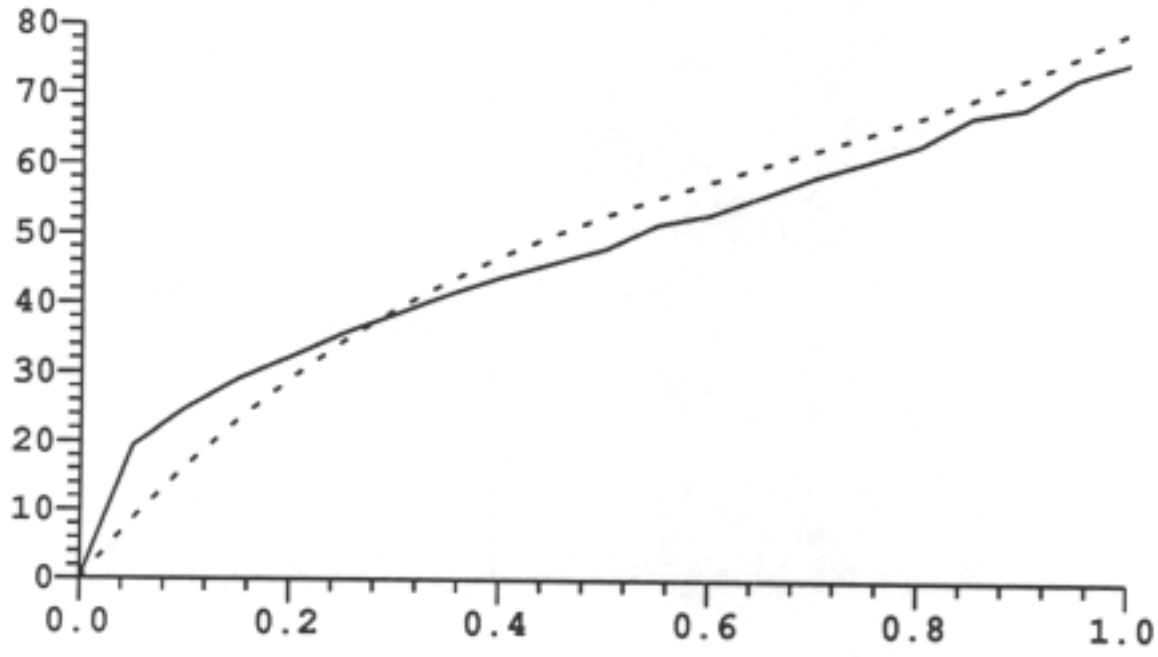
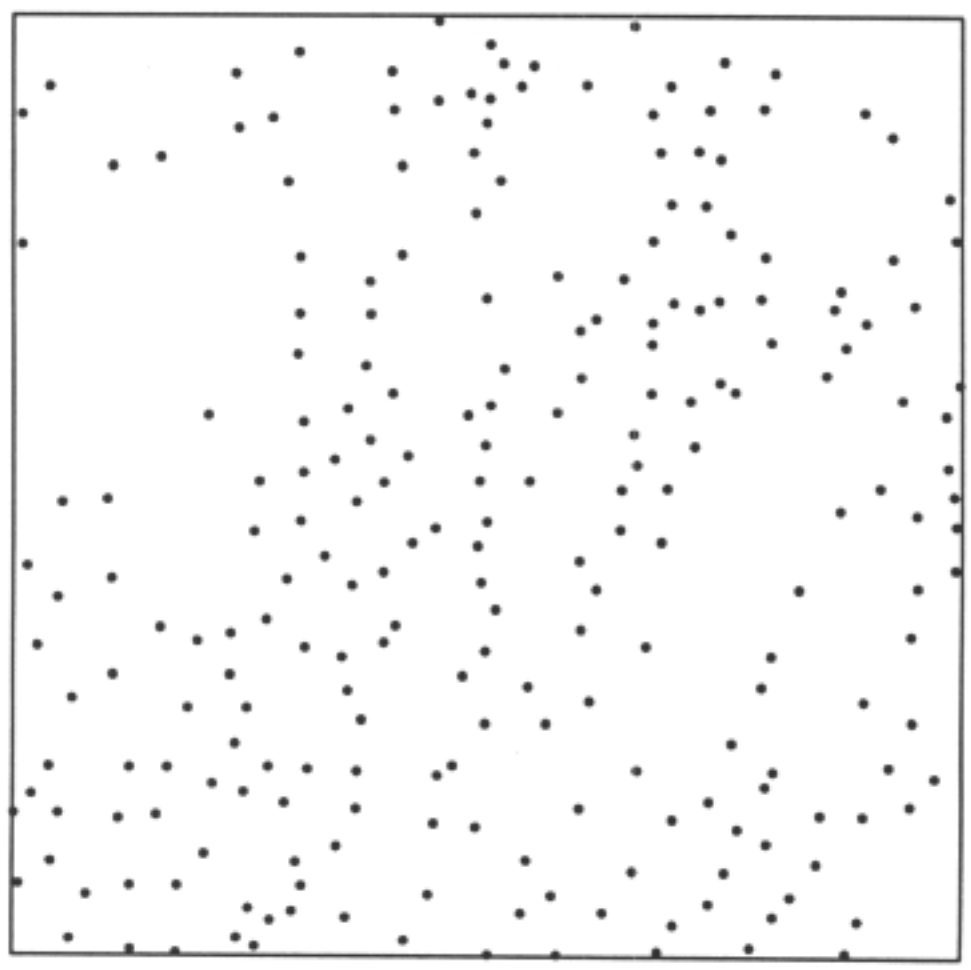



**Figure 7.**

Locations of 232 drumlins in an 8 km square of the Upper Ards peninsular, Co. Down. (After Hill, 1973).



### 4.3. Discussion

These two examples show that pairwise interaction models can be useful in geographical examples, and many more such models could have been tried. We do have to remember that parameter estimates are variable. It is difficult to quote accurate standard errors for the values quoted here, but since they are based on of the order of 50 pairs of points, and that number is approximately Poisson–distributed, we can expect standard errors of around 15% of the values quoted.

It is tempting to suggest smooth interaction functions and to interpret them. As Figure 3(d,e) shows, very different functional forms can be indistinguishable!

## 5. References


Bartlett, M. S. (1971a) Two–dimensional nearest–neighbour systems and their ecological applications. In *Statistical Ecology* eds G. P. Patil, E. C. Pielou & W.E. Waters, Penn. State Univ. Press, State College, PA, **I**, 179-194.

Bartlett, M. S. (1971b) Physical nearest–neighbour models and non–linear time series. *J. Appl. Prob.* **8**, 222-232.

Bartlett, M. S. (1975) *The Statistical Analysis of Spatial Pattern*. Chapman & Hall, London.

Bartlett, M. S. & Besag, J. (1969) Correlation properties of some nearest–neighbour systems. *Bull. Int. Statist. Inst.* **43(2)**, 191-193.

Bennett R. J. & Haining, R. J. (1985) Spatial structure and spatial interaction: Modelling approaches to the statistical analysis of geographical data. J. Roy. Statist. Soc. A **148**, 1-36.

Besag, J. E. (1972a) Nearest–neighbour systems: A lemma with application to Bartlett's global solutions. *J. Appl. Prob.* **9**, 418-421.

Besag, J. E. (1972b) Nearest–neighbour systems and the auto–logistic model for binary data. *J. Roy. Statist. Soc. B* **34**, 75-83.

Besag, J. E. (1972c) On the correlation structure of some two–dimensional stationary processes. *Biometrika* **59**, 43-48.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy. Statist. Soc. B* **36**, 192-236.

Besag, J. (1975) Statistical analysis of non–lattice data. *The Statistician* **24**, 179-195.

Besag, J. (1977) Some methods of statistical analysis of spatial data. *Bull. Int. Statist. Inst.* **47(2)**, 77-92.

Besag, J., Milne, R. & Zachary, S. (1982) Point process limits of lattice processes. *J. Appl. Prob.* **19**, 210-216.

Borel, E. (1925) *Principes et Formules Classiques du Calcul des Probabilités*. Gauthier–Villars, Paris.

Cliff, A. D. & Ord, J. K. (1981) *Spatial Processes*. Pion, London.

Diggle, P. J. (1983) *Statistical Analysis of Spatial Point Patterns*. Academic Press, London.

Diggle, P. J. & Gratton, R. J. (1984) Monte Carlo methods of inference for implicit statistical models (with discussion). *J. Roy. Statist. Soc. B* **46**, 193-227.

Diggle, P. J., Gates, D. J. & Stibbard, A. (1987) A nonparametric estimator for pairwise–interaction point processes. *Biometrika* **74**, 763-770.

Fiksel, T. (1984) Estimation of parameterized pair potentials of marked and non–marked Gibbsian point processes. *Elektron. Inform. Kybernet.* **20**, 270-278.

Geman, S. & Graffigne, C. (1987) Markov random fields and their applications to computer vision. *Proc. Inter. Cong. Math. 1986* ed. A.M. Gleason, Amer. Math. Soc., Providence, RI.

Glass, L. & Tobler, W. R. (1971) Uniform distribution of objects in a homogeneous field: Cities on a plain. *Nature* **233(5314)**, 67-68.

Haining, R. P. (1983) Modelling inter–urban price competition: An example of gasoline pricing. *J. Regional Sci.* **20**, 365-375.

Hill, A. R. (1973) The distribution of drumlins in County Down, Ireland. *Ann. Assoc. Amer. Geogr.* **63**, 229-241.

Kelly, F. P. & Ripley, B. D. (1976) On Strauss' model for clustering. *Biometrika* **63**, 357-360.

Kindermann, R. & Snell, J. L. (1980) *Markov Random Fields and Their Applications* Amer. Math. Soc., Providence, RI.

Mardia, K. V. & Marshall, R. J. (1984) Maximum likelihood estimation of models of residual covariance in spatial regression. *Biometrika* **71**, 135-146.

Mead R. (1971) Models for interplant competition in irregularly spaced populations. In *Statistical Ecology* eds G.P. Patil, E.C. Pielou & W.E. Waters, Penn. State Univ. Press, State College, PA, **2**,, 13-32

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953) Equations of state calculations by fast computing machines. *J. Chem Phys*, **21**, 1087-1092.

Molina, R. & Ripley, B. D. (1989) Using spatial models as priors in astronomical image analysis. *J. Appl. Statist.* **16**, 193-206.

Ogata, Y. & Tanemura, M. (1981) Estimation of interaction potentials of spatial point patterns through the maximum likelihood procedure. *Ann. Inst. Statist. Math.* **33B**, 315-328.

Ogata, Y. & Tanemura, M. (1984) Likelihood analysis of spatial point patterns. *J. Roy. Statist. Soc. B* **46**, 496-518.

Okabe, A. & Miki, F. (1984) A conditional nearest–neighbor spatial–association measure for the analysis of conditional locational interdependence. *Environment and Planning* **16**, 163-171.

Ord, K. (1975) Estimation methods for models of spatial interaction. *J. Amer. Statist. Assoc.* **70**, 120-126.

Pentinnen, A. (1984) Modelling interaction in spatial point patterns: parameter estimation by the maximum likelihood method. *Jyväskylä Studies in Computer Science, Economics and Statistics*, **7**.

Pinder, D. A. & Witherick, M. E. (1972) The principles, practice and pitfalls of nearest–neighbour analysis. *Geography* **57**, 277-288.

Preston. C. J. (1974) *Gibbs States on Countable Sets*. Cambridge University Press, London.

Preston, C. J. (1976) *Random Fields*. Lecture Notes in Mathematics **534**, Springer–Verlag.

Ripley, B. D. (1977) Modelling spatial patterns (with discussion). *J. Roy. Statist. Soc. B* **39**, 172-212.

Ripley, B. D. (1979a) Algorithm AS137. Simulating spatial patterns: Dependent samples from a multivariate density. *Appl. Statist.* **28**, 109-112.

Ripley, B. D. (1979b) The analysis of geographical maps. In *Exploratory and Explanatory Statistical Analysis of Spatial Data*, eds C.P.A. Bartels & R.H. Ketellapper, Martinus Nijhoff, Boston, 53-72.

Ripley, B. D. (1981) *Spatial Statistics*. Wiley, New York.

Ripley, B. D. (1987) *Stochastic Simulation*. Wiley, New York.

Ripley, B. D. (1988) *Statistical Inference for Spatial Processes*. Cambridge University Press, Cambridge.

Ripley, B. D. (1990) The use of spatial models as image priors. In *Spatial Statistics & Imaging* ed. A. Possolo, IMS Lecture Notes.

Ripley, B. D. & Kelly, F. P. (1977) Markov point processes. *J. Lond. Math. Soc.* **15**, 188-192.

Ripley B. D. & Silverman B. W. (1978) Quick tests for spatial interaction. *Biometrika* **65**, 641-642.

Rogers, A. (1974) *Statistical Analysis of Spatial Dispersion: The Quadrat Method*. Pion, London.

Strauss, D. G. (1975) A model for clustering. *Biometrika* **63**, 467-475.

Takacs, R. (1986) Estimator for the pair potential of a Gibbsian point process. *Statistics* **17**, 429-433.

Upton, G. J. G. & Fingleton, B. (1985) *Spatial Data Analysis by Example* volume I. Wiley, Chichester.

Whittle, P. (1954) On stationary processes in the plane. *Biometrika* **41**, 434-449.

# DISCUSSION

## "Gibbsian interaction models"

### by Brian D. Ripley

Gibbsian interaction models are usually better known as Markov random fields. These encompass many of the dependence models for regional or lattice data, and many models of point processes. The author aims to document the progress made over the past two decades in statistical inference and model fitting for these models.

About three–quarters of the paper is concerned with point processes. The Gibbsian interaction processes are defined, with particular emphasis on the Strauss processes, and the usual iterative method for simulation is given. Next there is a thorough discussion of maximum pseudolikelihood estimation. Approximate maximum likelihood is now feasible using theoretical approximations to the likelihood, or through simulation. Approximate maximum pseudolikelihood estimation is also feasible, and somewhat easier than approximate maximum likelihood. Two recent methods for choosing the interaction function are discussed. Many of these developments are very recent, and the discussion was welcome to me. Two examples are given. The first is a detailed illustration using the familiar 'Spanish towns' data, whilst the second comments on a previous analysis of some data on 'drumlins in Northern Ireland'. There are several figures that illustrate the text.

The section on regional processes is much briefer, and, apart from reviewing the usual conditional and simultaneous autoregression models, mainly consists of contrasting maximum pseudolikelihood estimation with maximum likelihood estimation. For these dependence models, both estimation methods have been used for some years (the former was introduced in 1975), although the author rightly questions whether maximum likelihood estimation is always relevant. However, his claim that 'pseudolikelihood methods have been very successful' may reflect his interest in astronomically large data sets, and non–Gaussian distributions. The results of Besag (1977) show that pseudolikelihood estimators can be very inefficient when the dependence is not small, so that their use is questionable when other estimators, that may be more efficient, are readily available. Although the author refers to a recent rigorous proof of the consistency of the pseudolikelihood estimator, he does not refer to Guyon's (1987) theorem on asymptotic Normality. There are no examples given in this section, and the only applications referred to are in agricultural field trials and astronomical image analysis.

The author's presentation is, as usual, precise and comprehensive, though the mathematics may make demands on many. Perhaps the author could have considered more the questions of how appropriate these models and methods are for geographical research, and of how they are being, or have been, used in geographical research. Whilst it may be useful to give unreferenced publications in the bibliography, it would have been helpful if this policy had been noted, and a brief note had been given with each to explain its possible relevance.

## References

Besag, J. (1977) Efficiency of pseudolikelihood estimation for simple Gaussian fields. *Biometrika*, **64**, 616-618.

Guyon, X. (1987) Estimation d'un champ par pseudo–vraisemblance conditionnelle: étude asymptotique et application au cas Markovien, in *Spatial Processes and Spatial Time Series Analysis*, edited by F. Droesbeke, pp. 15-62. Brussells: Facultés Universitaires Saint–Louis.

*R. J. Martin*, University of Sheffield

## *PREAMBLE*

*There's a time for all things.*

W. Shakespeare, **Comedy of Errors** (Act II, Scene 2)

*Computers are immobile robots that compute, perform tasks in a logical manner, and collate information furnished to them by scientists. The dawn of a computer age is upon science, and computing devices will play an increasingly larger role in scientific research during the coming decades. Ord focuses attention on the increased computer power available today that enables progress to be made in the area of point pattern analysis, by rendering numerically intensive problems soluble, speculating that new computer advances will spur on subsequent developments, too. The purpose of this paper is to review the main stochastic models used in univariate point pattern analysis, their accompanying traditional statistical tests and inference problems, and their extensions to multivariate point pattern analysis. Ripley's commentary complements the contents of this paper by outlining other important developments in the areas of theory, point processes of objects, statistical inference, and mosaic models. He makes three interesting conjectures, namely that (1) point patterns per se are becoming less important, their few golden years having passed, with their future literature becoming less prolific, (2) fundamental problems of computational complexity do not exist for point pattern analysis, with computational intractability being a misnomer applied to situations of ignorance, and (3) point pattern analysis now is being eclipsed by statistical image analysis. Indeed, then, is now the time to begin finalizing the history of point pattern analysis? Ord thinks not; Ripley thinks so!*

The Editor

# Statistical Methods for Point Pattern Data

J. Keith Ord *

*Departments of Management Science and Statistics, The Pennsylvania State University, University Park, PA 16802.*

**Overview:** Recent developments in spatial point processes are reviewed. After presenting an outline of the major univariate stochastic models for such processes, a Poisson–Gaussian model is described which can be made operational for both areal and distance–based sampling. We then summarize work on tests of randomness using distance, quadrat and line transect methods. Robust estimation procedures for the intensity of the process are examined along with the descriptive tools provided by second order processes. Increased computer power has enabled some progress to be made in estimating the parameters of stochastic models from the likelihood function and further activity is likely in this area.

Finally, work on multitype processes is reviewed and future directions for research are outlined.

## 1. Introduction

The analysis of spatial data has become a major preoccupation of statisticians only rather recently and most of the advances in the study of spatial point patterns have happened within the past ten years. In part, this reflects a neglect of such problems by statisticians since heuristic statistical methods have been employed in other disciplines such as ecology and forestry for over fifty years. However, the other side of the coin is the level of computational difficulty facing statistical model builders. It is only with the computing power available in the 1980's that such difficulties are being overcome.

After outlining the different approaches to data collection in Section 2, we turn in Section 3 to the main stochastic models which have been developed for spatial point patterns. To avoid overworking the term "point" we shall reserve it for points in the study region and prefer the term "individual" for realized events.

After developing the basic models, Sections 4 and 5 cover the traditional areas of tests of "randomness" (*i. e.* whether the pattern is formed by a Poisson process) and statistical inference. In addition to the usual issues of intensity estimation, Section 5 covers the newer topics of second order methods and parametric model–building. Section 6 is devoted to multitype (or multivariate) point processes and the paper concludes with a brief section on the future directions of the subject.

### 1.1. A bibliography

Apart from the major pioneering effort by Matern (1960), the first theoretical text on spatial point patterns was that of Bartlett (1975). Since then the literature has expanded considerably. Chapter 4 of Cliff and Ord (1981) considers the analysis of spatial point patterns with an emphasis on geographical applications. Getis and Boots (1978) discuss spatial point, line and areal processes, again with an emphasis on problems in geography. Diggle (1983) gives

* I am grateful to several colleagues, especially Brian Ripley, for comments on an earlier version of this paper.

an excellent overview with many ecological examples. The major book of a more theoretical nature is that of Ripley (1981) which also includes a wide variety of applications. Kinderman and Snell (1980) provide a useful introduction to the theory of Markov random fields. Most recently, Upton and Fingleton (1985) provide a lively presentation of the methods of spatial analysis, drawing on examples from many different disciplines. The most recent and complete discussion of inference problems for spatial processes appears in Ripley (1988).

In addition to these texts, the review paper by Ripley (1984a) was a valuable source of information in the preparation of this paper. Further useful sources are the bibliography compiled by Naus (1979) and the volume of papers edited by Cormack and Ord (1979).

## 2. Data collection

The manner in which data on spatial point processes are collected and recorded has a major impact upon subsequent methods of analysis. When a single study area is selected and all the individuals within that study area are recorded, we say the data are *mapped*. Alternatively, when a series of sites is selected at random and individuals are recorded only in the neighborhood of those sites, we say that the process has been sparsely *sampled*. The definitions and the terminology follow Diggle (1983).

We may select sampling units (known as quadrats, whatever their shape!) of a prespecified size and record the number of individuals present in that sub–area: this is known as the method of quadrat counts. The other principal option is to select individuals, or points, at random and to measure either point–to–individual (PI) or individual–to–individual (II) distances. These distance methods are often known as *nearest–neighbor* methods, although this term should be taken to include all immediate neighbors, not just the nearest one. The more general label of *distance methods* will be used in this paper.

At one time, sparse sampling methods were widely used in ecology, although they are now less popular. Naturally, geographers prefer mapped data analyses. When complete mapping is undertaken, other approaches become feasible, such as "empty space" methods whereby we examine the probability that a region (disk) of given area is devoid of individuals. Subject matter and research objectives should and do play a major role in the types of analyses undertaken.

## 3. Modeling spatial point processes

The benchmark model for spatial patterns is the homogeneous Poisson process, often known simply as the Poisson process. In order to describe this we consider a planar region $\mathbf{C}$ (sample space) and (sub–)regions $\mathbf{A}, \mathbf{B} \subset \mathbf{C}$. The area of such regions is denoted by $|\mathbf{A}|$, $|\mathbf{B}|$ and so on. Let $\mathbf{N}(\mathbf{A})$ be the random variable denoting the number of individuals in $\mathbf{A}$. The expected number of individuals in $\mathbf{A}$ is then

$$E\{N(A)\} = \int_A \lambda(\mathbf{x})\, d\mathbf{x}, \tag{3.1}$$

where $\lambda(\mathbf{x})$ is the *intensity function* defined at all points $\mathbf{x} \epsilon \mathbf{C}$.

### 3.1. The Poisson Process

The assumptions underlying the Poisson process are

**PP1:** $\lambda(\mathbf{x}) = \lambda$ for all $\mathbf{x} \epsilon \mathbf{C}$.

**PP2:** $N(A)$ follows a Poisson distribution with mean $\lambda|A|$.

**PP3:** Given that $N(A) = n$, the $n$ events in $A$ are independent and form a random sample from the uniform distribution on $A$.

It then follows that

**PP4:** If $A$ and $B$ are disjoint regions, $N(A)$ and $N(B)$ are independent.

In fact, it may be shown that **PP1** and **PP2** imply **PP3** (Ripley, 1976b), although the proof is difficult.

Property **PP3** motivates the description of the Poisson process as being "purely random." Many tests of the Poisson model (see Section 4) are known as tests of "randomness." Another central property of the Poisson process relates to distance sampling:

**PP5:** If $P$ is a randomly selected point in the plane, $C$ and $I_1$ is the individual nearest to $P$, the distribution of $U_1 = (PI_1)^2$ is exponential with parameter $\pi\lambda$. Further if $I_k$ is the $k^{th}$ nearest individual to $P$,

$$U_k = (PI_k)^2 \text{ is gamma } (k, \pi\lambda). \tag{3.2}$$

It seems a rather minor comment to note that, if **P** denotes a randomly selected individual rather than a point in the plane, property **PP5** is unchanged. However, this additional result proves very valuable in estimation and testing problems.

### 3.2. Departures from "randomness"

Given the specialized nature of the assumptions underlying the Poisson process, a variety of departures from **P1** and **P2** may be considered. The two simplest schemes allow (a) heterogeneous intensity (**HI**) and (b) clusters of individuals respectively. Following Diggle (1983, pp 52-55) the heterogeneous intensity model may be described as follows:

**HIP1** $N(A)$ has a Poisson distribution with mean (3.1), $\lambda(\mathbf{x}) \neq \lambda$ for all $\mathbf{x}$

**HIP2** Given $N(A) = n$, the individuals in $A$ form a random sample from the distribution on $A$ with pdf proportional to $\lambda(\mathbf{x})$.

Kooijman (1979) discusses the analysis of mapped data when $\lambda(\mathbf{x})$ is represented by a low order polynomial.

An extension of the heterogeneous scheme is to assume that the intensity $\lambda(\mathbf{x})$ is itself determined by a random process, $\mathbf{\Lambda}(\mathbf{x})$. This yields the doubly stochastic process or Cox process, first developed in the time domain by Cox (1955); see also Bartlett (1964, 1975) and Matern (1971).

**CP1** $\{\Lambda(\mathbf{x}) : \mathbf{x} \epsilon \Re^2\}$ is a non–negative–valued stochastic process.

**CP2** Conditional on $\{\Lambda(\mathbf{x}) = \lambda(\mathbf{x}) : \mathbf{x} \epsilon \Re^2\}$ the number of individuals is described by a heterogeneous intensity Poisson process with intensity function $\lambda(\mathbf{x})$.

When the process in CP1 is stationary, it follows that the intensity is

$$\lambda = E[\Lambda(\mathbf{x})].$$

For two points $\mathbf{x}$ and $\mathbf{y}$ the conditional intensity given $\{\Lambda(\mathbf{x})\}$ is $\Lambda(\mathbf{x})\Lambda(\mathbf{y})$, and we define the second order intensity to be

$$\lambda_2(\mathbf{x},\mathbf{y}) = E\{\Lambda(\mathbf{x})\Lambda(\mathbf{y})\}.$$

When the process is both stationary and isotropic, this becomes

$$\lambda_2(\mathbf{x},\mathbf{y}) = \lambda^2 + \gamma(t), \tag{3.3}$$

where $\gamma(\mathbf{t}) = \text{cov}[\Lambda(\mathbf{x})\Lambda(\mathbf{y})]$ and $t = \|\mathbf{x}-\mathbf{y}\|^{1/2}$ is the distance between $\mathbf{x}$ and $\mathbf{y}$.

*3.2.1. Poisson cluster process*

**PCP1** The distribution of parent individuals in $A$ follows a Poisson distribution with intensity $\rho$.

**PCP2** Each parent produces a random number, $S$, of offspring, where $S$ is independent and identically distributed for each parent with probabilities $\{\pi_s, s = 0, 1, \ldots\}$.

**PCP3** The positions of the offspring relative to their parents are independently and identically distributed according to a bivariate distribution with pdf $f(\cdot)$.

When $f(\cdot)$ represents a degenerate distribution **PCP1** and **PCP2** yield the class of contagious Poisson distributions (cf. Ord, 1972, pp. 126-7).

When $S$ is described by a logarithmic series distribution, it is well–known that the resulting **PCP** is described by the negative binomial distribution. However, when $f(\cdot)$ is non–degenerate, Diggle and Milne (1983a) found that no plausible process exists which could produce counts following the negative binomial law.

Quite generally, the clustering processes may be described by their characteristic functionals (Bartlett, 1964). Bartlett went on to show that the Poisson cluster processes are Cox processes so that the generating mechanisms cannot be distinguished by any data analytic method.

Another existence result of interest is due to Kingman (1977), who showed that for reproducing populations formulated as Cox processes there is no process with independent displacements of individuals which leads to the Poisson process as an equilibrium. However, Kingman went on to demonstrate that when dependent displacements are allowed, a variety of processes lead to a Poisson process in the equilibrium state. These results help to justify the often implicit assumption that purely spatial models may be viewed as equilibrium processes for an (unobserved) spatio–temporal process.

### 3.3. Gibbs processes

Now consider a process defined on the planar region $C$ such that the joint pdf that there are exactly $n$ individuals in $C$ located at $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ is

$$a g_n(\mathbf{x}_1, \ldots, \mathbf{x}_n)\rho^n e^{-\rho|C|}/n!,$$

where $a$ is a normalizing constant, $g_n(\cdot)$ is symmetric in its $n$ arguments and $\rho$ represents intensity as before. It follows that

$$P\{N(C) = n\} = [a\rho^n e^{-\rho|C|}/n!] \int_{C^n} g_n(\mathbf{x}_1, \ldots, \mathbf{x}_n) d\mathbf{x}_1 \ldots d\mathbf{x}_n,$$

so that the joint pdf for the $\{x_i\}$ given $N(C) = n$, is

$$g(\mathbf{x}_1, \ldots, \mathbf{x}_n) / \int_{C^n} g(\mathbf{x}_1, \ldots, \mathbf{x}_n) d\mathbf{x}_1 \ldots d\mathbf{x}_n. \tag{3.4}$$

Writing $g_n = \exp(-\phi_n)$, we may consider $\phi_n$ to be a potential function and the joint pdf then defines a Gibbs process. In particular, the special case

$$\phi_n(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum_{i<j} \beta_2(\mathbf{x}_i - \mathbf{x}_j) \tag{3.5}$$

has provided the basis for most empirical work in statistics; $\beta_2 \equiv 0$ clearly reduces to the Poisson process. Ripley and Kelly (1977) gave a seminal development of Markov random fields and show that, for a general class of potential functions including (3.5), the Gibbs processes and Markov random fields are equivalent. Further, Ripley and Kelly (1977) demonstrate that the Gibbs processes provide a particularly useful framework for models of inhibitory processes. For example, setting

$$\beta_2(\mathbf{x}_i - \mathbf{x}_j) = 0, \quad \|\mathbf{x}_i - \mathbf{x}_j\| < d_0$$

provides a *hard–core* model whereby two individuals cannot co–exist within $d_0$ of each other. Following earlier work by Strauss (1975), Kelly and Ripley (1976) developed a clustering model using (3.5). Sanders *et al.* (1982) have extended this scheme to include local hard–core inhibition and clustering. Taking $\phi_n > 0$ ensures that the normalizing factor in (3.4) is finite; otherwise, the process may not be well defined. If $\phi_n < 0$ for some $\mathbf{x}$, extreme caution is advised.

### 3.4. Related processes

Several other processes have been developed and many of these are reviewed by Diggle (1983, Chapter 4). Most of the processes described thus far assume stationarity and isotropy. However, this is more for convenience than necessity. For example, the second order properties discussed in Section 5 below also hold for anisotropic processes.

Byth (1981, 1982a) developed a class of processes which are isotropic with respect to a particular "origin," but non–stationary. The intensity is a function of the distance from the "origin." Byth's motivating example was the pattern of fungi around a tree, but her process has potential value for many diffusion–type processes which emanate from a known origin.

A link between lattice processes and point processes has been provided by Besag *et al.* (1982) who show that the auto–Poisson process (of Besag, 1974) on a regular lattice approaches a limiting inhibitory (pairwise interaction) spatial point process.

Another class of models are the thinned processes discussed by Brown and Holgate (1974) and Brown (1979) among others. The method of thinning may be either random or position dependent.

### 3.5. Nearest–neighbor results

As suggested by property **PP5** in Section 3.1, distances from randomly selected points (or individuals) to first, second, ..., nearest individuals, often known as nearest–neighbor distances, are widely used in the analysis of spatial point patterns. Further recent developments in this area include Cox (1983) on the probability that $m^{th}$ nearest neighbors to individual $I_1$, are also $n^{th}$ nearest neighbors to $I_2$; Cox also develops higher order results of this type. Newman *et al.* (1983) consider the probability that an individual is the nearest neighbor of exactly $k$ other individuals; it is found that the probabilities are of Poisson form with $\lambda = 1$, for a variety of processes. Pickard (1982) gives a general treatment of the isolated nearest neighbor (or reflexive pairs) problem. Warren (1971) and Warren and Batcheler (1979) give nearest neighbor distributions for several specific non–Poisson schemes.

### 3.6. The heterogeneous Poisson–Gaussian process

The lack of any plausible spatial process underlying the negative *binomial* distribution for quadrat counts has been demonstrated by Diggle and Milne (1983a). This makes it difficult to interrelate analyses based on areal and distance sampling procedures yet, especially for mapped processes, some interconnection between these approaches is essential.

A review of the assumptions underlying Cox processes (Section 3.2) indicates that the intensity process must be non–negative but also infinitely divisible so that spatial aggregates may be defined for any region. A natural choice would be the Gaussian scheme except that, of course, it may take on negative values. We could condition upon $\Lambda(\mathbf{x}) \geq 0$ for all $\mathbf{x}$; done rigorously, however, this imposes considerable additional complexity upon the analysis. In practice, provided the probability of negative values is sufficiently small we need apply this conditioning only notionally and, algebraically at least, may consider the complete Gaussian process as a reasonable approximation. This approach has been employed in models for discrete data where the Poisson mixture with the normal gives rise to the Hermite distribution; Kemp and Papageorgiou (1982) describe the bivariate case and give references to earlier work. However, it should be noted that the correlation function must be non–negative for all $\mathbf{x} - \mathbf{y}$ to ensure a well–defined process (see Section 3.3). We now develop these ideas for a particular Cox process.

Following Bartlett (1975, p. 7) we may write the characteristic functional (c.fl.) for all points $\mathbf{u}$ in some region $Q$ as

$$C(\boldsymbol{\theta}) = E_{\Lambda}[\exp \int \Lambda(\mathbf{u})\{z(\mathbf{u}) - 1\}\, d\mathbf{u}], \quad z = e^{i\boldsymbol{\theta}},$$

where the integral is taken over all points $\mathbf{u} \in Q$. This expression is more readily understood if we first consider $Q$ to contain a finite number of points when the c.fl. becomes a multivariate probability generating function (p.g.f.), viewed as an argument in $\mathbf{z}$ rather than $\boldsymbol{\theta}$.

When $\Lambda(\mathbf{x})$ is a Gaussian process with mean function $\mu(\mathbf{x}) = \mu$, variance function $\omega(\mathbf{x}) = \omega$ and autocorrelation function

$$\rho(\mathbf{x}, \mathbf{y}) = \rho(\mathbf{x} - \mathbf{y}), \text{ for all } \mathbf{x} \text{ and } \mathbf{y},$$

with $\rho(\mathbf{0}) = 1$, it follows that

$$C(\boldsymbol{\theta}) = \exp[\mu \int \{z(\mathbf{u}) - 1\}\, d\mathbf{u} + 0.5\omega \iint \rho(\mathbf{u} - \mathbf{v})\{z(\mathbf{u}) - 1\}\{z(\mathbf{v}) - 1\}\, d\mathbf{u}\, d\mathbf{v}]. \quad (3.6)$$

If sampling is by quadrats, the p.g.f. for the quadrat count is obtained by setting $z(\mathbf{u}) = z$ for all $\mathbf{u} \epsilon Q$. If $Q$ has area $A$, (3.6) yields the p.g.f.

$$C_Q(z) = \exp[A\mu(z-1) + 0.5\omega(z-1)^2 \iint \rho(\mathbf{u}-\mathbf{v})\, d\mathbf{u}\, d\mathbf{v}], \tag{3.7}$$

where the double integral is taken over $(\mathbf{u}, \mathbf{v}) \epsilon Q$. Thus, (3.7) defines an Hermite distribution with p.g.f. of the general form

$$\exp[\lambda_1(z-1) + \lambda_2(z-1)^2], \tag{3.8}$$

reducing to the Poisson when $\lambda_2 = 0$. It follows from (3.8) that the quadrat count, $R$, has

$$E(R) = \lambda_1 < \text{var}(R) = \lambda_1 + 2\lambda_2, \text{ for } \lambda_2 > 0.$$

Expression (3.6) simplifies further in those cases where an explicit functional form is available for the correlation. In particular, suppose that the process is isotropic and

$$\rho(\mathbf{w}) = \exp\{-\alpha^2(w_1^2 + w_2^2)/2\}. \tag{3.9}$$

Further, we assume that the quadrats are rectangular, say $h_1$ by $h_2$, with $h_1 \cdot h_2 = A$. The integral in (3.7) partitions into two integrals of the form

$$\begin{aligned} G(h) &= \int_0^h \int_0^h \exp\{-\alpha^2(u-v)^2/2\}\, du\, dv \\ &= [4(2\pi)^{1/2}/\alpha^2][\alpha h F(\alpha h) - \alpha h/2 + f(\alpha h) - f(0)], \end{aligned}$$

where $f$ and $F$ denote the density and distribution functions, respectively, for the standard normal. Thus, the p.g.f. for a single quadrat becomes

$$C_Q(z) = \exp[A\mu(z-1) + \omega(z-1)^2 G(h_1)G(h_2)/2]. \tag{3.10}$$

By extending the argument to two (or more) quadrats, we obtain bivariate (multivariate) Hermite distributions. Comparison of (3.8) and (3.10) indicates that

$$\lambda_1 = A\mu \quad \text{and} \quad \lambda_2 = \lambda_2(\omega, \alpha, h_1, h_2),$$

so that there are three unknown parameters $(\mu, \omega, \alpha)$ relating to the mean, variance and correlation structures respectively. This is in a one–to–one correspondence with the bivariate Hermite with identical marginal distributions. The simplest procedure inferentially is probably to estimate the parameters of the bivariate Hermite using a pseudo–likelihood approach and then relate these values to the process parameters using the moments derived from (3.10) and its bivariate extension. Detailed development of the method is necessary, but it appears to have some potential.

Finally, we note that the process also corresponds to a **PCP** when the cluster size, $S$, is binomially distributed with index $n = 2$.

## 4. Tests of randomness

The Poisson process forms a natural null hypothesis for testing whether structure exists within a spatial data set. Such tests are often known as "tests of randomness" and may be designed with particular alternatives (clustering, inhibition) in mind or may use the completely general alternative of a non–Poisson process.

As has been remarked by Ripley (1981) and others, tests of randomness should represent only the first step in a spatial modeling paradigm, yet it has to be admitted that such tests have often been presented as the end result of an analysis. Either way, such tests clearly have value and we now summarize the main options for distance and area–based sampling respectively.

### 4.1. Distance methods

The majority of the tests proposed are based upon distance sampling. A summary list is presented in Table 1. When the region is sparsely sampled, the distributional properties listed hold. For mapped data, these results hold up less well (Byth and Ripley, 1980). An alternative procedure is to use Monte Carlo testing (cf. Cliff and Ord, 1981, pp 63-65; Diggle, 1979a). Diggle and Gratton (1984) present the first systematic development of inferential procedures based upon Monte Carlo methods.

Several comparisons of the relative power of these tests have been performed recently; their conclusions may be summarized as follows:

(1) In general, tests which include squared distances perform better than those which do not.

(2) The Hopkins ($D$) test is most powerful across a range of alternatives but is non–feasible for sparse data since randomly selected individuals cannot be selected (Besag and Gleaves, 1973). For mapped data the Hopkins test is often a good choice and the test may be performed by Monte Carlo methods (cf. Diggle, 1979a). Byth and Ripley (1980) have developed a semi–systematic sampling scheme for choosing individuals "at random," which allows use of the Hopkins test. When the semi–systematic sampling is feasible, this approach would appear to give the best power.

(3) The Besag–Gleaves ($G$,$H$) and Cox–Lewis ($L$) tests perform similarly across a variety of alternatives. Cox–Lewis has a slight edge but has a somewhat more difficult distribution theory. The Eberhardt ($J$) and Holgate ($E$,$F$) statistics are less powerful (Hines and Hines, 1979). The Hines–Hines ($K$) procedure is comparable in power to the Hopkins ($D$) test for clustered alternatives, but weaker for regular alternatives (Hines and Hines, 1979).

(4) Tests ($R$), ($S$), ($T$) and ($A$) perform better than tests ($P$) and ($Q$) and several other alternatives (Ripley, 1979a). Tests ($R$), ($S$) and ($T$) have the advantage that they may suggest an alternative model, but ($S$) and ($T$) are clearly dependent on the values of the constants $C1$ and $C2$.

Clayton (1984) suggests a procedure based upon the use of an exclusion angle, but its relative performance characteristics are as yet unknown.

Ripley and Silverman (1978) demonstrate that, when the alternative hypothesis is a Poisson hard–disk process, the uniformly most powerful test is based upon the smallest order

TABLE 1
DISTANCE TESTS OF RANDOMNESS

| Code Letter | Statistic | Distribution under $H_0$ | Reference |
|---|---|---|---|
| $A$ | $2\sqrt{\hat{\lambda}}\Sigma r_i/n$ | approx $N[1,(4-\pi)/(\pi n)]$ | Clark & Evans (1954) |
| $B$ | $\pi\hat{\lambda}\Sigma u_i/n$ | approx $N[1,(\lambda A+n+1)/(n\lambda A)]$ | Pielou (1959), Mountford (1961) |
| $C$ | $2\pi\lambda\Sigma u_i$ | approx $\chi^2_{2n}$ | Skellam (1951) |
| $D$ | $\Sigma u_i/\Sigma(u_i+u_i^*)$ | $B(n,n)$ | Hopkins (1954) |
| $E$ | $\Sigma u_i/\Sigma u_{i2}$ | $B(n,n)$ | Holgate (1965) |
| $F$ | $\Sigma(u_i/u_{i2})/n$ | approx $N(1/2,1/12n)$ | Holgate (1965) |
| $G$ | $\Sigma u_i/\Sigma(u_i+.5v_{iT})$ | $B(n,n)$ | Besag and Gleaves (1973) |
| $H$ | $\Sigma\{u_i/(u_i+.5v_{iT})\}/n$ | approx $N(1/2,1/12n)$ | Besag and Gleaves (1973) |
| $I$ | $-.5\Sigma\ell n\{u_i/(u_i+.5v_{iT})\}$ | $\chi^2_{2n}$ | Besag and Gleaves (1973) |
| $J$ | $n\Sigma u_i/(\Sigma r_i)^2$ | see Hines & Hines (1979) | Eberhardt (1967) |
| $K$ | $\Sigma(u_i+.5v_{iT})/\Sigma(r_i+.5\sqrt{v_{iT}})$ | see reference for tables | Hines and Hines (1979) |
| $L$ | $4/3\Sigma(1-\pi w_i)$ | approx $N(1/2,1/12n)$ | Cox and Lewis (1976), Cormack (1977) |
| $M$ | $\min_i(w_i)$ | beta $(1,n)$ | Cox and Lewis (1976) |
| $N$ | $n\ell n[\Sigma(u_i+.5v_{iT})/n]-\Sigma\ell n(u_i+v_{iT})$ | $\chi^2(n-1)$ | Diggle (1977b) |
| $P$ | $\Sigma(v_i-D)^2/D^2(n-1)$ | by simulation | Brown and Rothery (1978) |
| $Q$ | $(\pi v_i)^{1/n}/D$ $(D=\Sigma v_i/n)$ | by simulation | Brown and Rothery (1978) |
| $R$ | $\sup_{t\le t_0}\lvert L(t)-t\rvert$ | by simulation | Ripley (1979a), p. 369. |
| $S$ | $\Sigma\phi(\mathbf{x},\mathbf{y})$ | asymp. normal | Liebetrau (1977) |
| $T$ | $\Sigma\phi(\mathbf{x},\mathbf{y})$ | asymp. normal | Jolivet (1978) |

**Notation for Table 1**

$r_i$ = distance from random point to the nearest individual

$u_i = r_i^2$

$r_{ip}$ = distance from random point to p-th nearest individual

$u_{ip} = r_{ip}^2$

$r_i^*$ = distance from random individual to its nearest neighbor

$u_i^* = (r_i^*)^2$

$v_i$ = squared distance from nearest individual to a random point to its nearest neighbor

$v_{iT} = v_i$, but with nearest neighbor restricted to T–square sampling

$w_i^{-1} = 2\pi + \sin\theta_i - (\pi+\theta_i)\sin\theta_i$; $\sin(0.5\theta_i) = (v_i/u_i)^{1/2}$

$L(t) = [\hat{K}(t)/\pi]^{1/2}$, see (5.3) for definition of $\hat{K}(t)$

$\phi(\mathbf{x},\mathbf{y}) = [c_1 - |x_1-y_1|][c_2-|x_2-y_2|]$ if $|x_i-y_i| > c_i$, $i=1,2$; $\quad=0$, otherwise.

statistic, $r_{(1)}$, of a sample of first nearest–neighbor distances. Concerns about measurement and rounding errors may lead to the use of higher order statistics, but this would occasion loss of power unless the errors are substantial. Saunders and Funk (1977) demonstrate that

for a disk of radius $r_0$, $r_{(1)}^2 - r_0^2$ is approximately exponentially distributed; for further developments, see Silverman and Brown (1978).

As noted earlier, the Poisson cluster process may be represented as a Cox process so that clustering and heterogeneity cannot be distinguished. More pragmatically, it may be argued that clustering is essentially a local phenomenon whereas heterogeneity is manifested at a larger scale; that is, $\mathbf{\Lambda}(\mathbf{x})$ changes slowly with $\mathbf{x}$. From this perspective, Diggle (1977b) proposed test $(N)$ as a test of heterogeneity. Diggle recommends that one of the local distance tests be used to detect local departures from randomness and then test $(N)$ be used to detect larger scale patterns. As such it represents an alternative procedure to the transect methods described in Section 4.3. The Diggle procedure has the advantages of being somewhat simpler to apply and being applicable to both sparsely sampled and mapped data. Its power, relative to the methods in Section 4.3, is unknown.

### 4.2. Areal sampling methods

When data are collected from a Poisson process, the counts distribution is Poisson whether those data are mapped or sparsely sampled. In principle, any goodness–of–fit test could be used to test the null hypothesis that the distribution is Poisson. However, tests such as chi–square have been found to have relatively low power since it is the upper tail observations which serve to distinguish non–Poisson alternatives. The most popular test is the index of dispersion

$$D = \text{Sample variance/sample mean.}$$

Perry and Mead (1979) show that $D$ has good power properties in sparse sampling. Heltshe and Ritchey (1984) show that Stevens' test, defined as

$$Z = \text{number of empty quadrats} \quad | \; S \text{ individuals in } N \text{ quadrats}$$

has power comparable to $D$ for aggregated alternatives, but does not perform as well for regular patterns when large quadrats are used. This is understandable since

$$E(Z) \doteq Ne^{-\lambda}, \text{ where } S = N\lambda,$$

which becomes small for large $\lambda$ and fixed $N$.

For mapped data, quadrat methods necessarily lose power as they ignore the spatial dependence between quadrats; the distance–based methods of Section 5.2 are preferable. When only quadrat counts are available, the $D$ test can be complemented by a test for spatial autocorrelation between neighboring quadrats (Cliff and Ord, 1981, pp. 97-99).

### 4.3. Tests using transect data

There is a considerable literature on line transects which we shall not review in detail; see Gates (1979) and De Vries (1979). One topic relevant to our present discussion is the use of line transect data to test for different scales of spatial pattern. These methods are associated with the name of Greig–Smith (1952) who has used the approach extensively in his ecological work; see Greig–Smith (1979).

Briefly, the computational details of the method may be described as a hierarchical analysis of variance. As noted by Professor Bartlett and others, formal ANOVA tests are

unreliable since detection of one scale of spatial pattern may obscure higher levels. Mead (1974) developed a randomization procedure which avoids the difficulties noted by Bartlett.

In general, the Greig–Smith and Mead procedures have single starting points along the transect (or two–dimensional array). Hill's (1973) method gave improved power by considering all possible starting points and Upton's (1984) procedure does the same for Mead's test. Zahl (1977) developed a Scheffé–type procedure which considers all overlapping blocks and controls the overall probability of Type I error; simulation results indicate that this technique outperforms the Greig–Smith method, and also the random quadrat method of Goodall (1974). Moellering and Tobler (1972) and Cliff and Ord (1981, pp. 123-6) provide geographical applications of this approach.

A different approach is the use of spectral analysis, illustrated by Ripley (1978). This analysis appears to give at least as clear results as spatial domain methods and may indeed be a better vehicle for separating out different scales of pattern; see also Ripley (1981, pp. 112-129), Renshaw and Ford (1983) and Renshaw (1984).

## 5. Statistical inference

Inference for spatial point processes may focus upon either the first and second order properties required for weak stationarity or upon strict stationarity when the more ambitious goal of fitting a complete parametric model is attempted. We shall examine these in turn.

### 5.1. Intensity of a process

The simplest and most traditional problem is that of estimating process intensity, or the average number of individuals per unit area. When the process is stationary and isotropic, this is only a problem for sparsely sampled data; for non–stationary processes with mapped data the smoothing methods used in probability density estimation are particularly useful, see Cox (1979), Diggle (1981a) and, more generally, Silverman (1981).

For stationary and isotropic processes, the obvious real estimate

$$\widehat{\lambda} = \text{Total number of individuals}/(\text{Number of quadrats}) \cdot (\text{Quadrat area})$$

is unbiased for any spatial process. However, it is often the case that quadrat sampling is either too expensive or impractical which has led to an extensive search for distance–based estimators. Many of the early estimators were based upon precisely the same statistics used to test for randomness (see Table 1); this includes the maximum likelihood estimators for the Poisson process (cf. Pollard, 1971). To the extent that the tests are successful, the lack of robustness of the estimators is perhaps not surprising. Persson (1971) made this point forcibly and suggested several more robust alternatives; see Table 2. Robustness in this context is a somewhat elusive concept. It is very easy to consider a particular departure from a Poisson point process and to produce an estimator that is robust to that change, yet not robust to others. For example, if a pattern includes very tight clusters of variable size, any distance–based estimator is likely to fail for some configurations. Thus, robustness is a quality to be assessed in the eye of the beholder and the researcher should determine whether the method to be used will be robust for the spatial processes likely to be encountered. We now summarize the evidence on different estimators, bearing in mind that published studies have tended to focus upon a rather limited variety of non–Poisson processes.

TABLE 2
ROBUST ESTIMATORS FOR INTENSITY

| Code Letter | Estimator | Reference |
|---|---|---|
| $A$ | $c/(\Sigma r_{ip})^2$ | Persson (1971) |
| $B$ | $c/(\Sigma u_{ip})$ | Persson (1971) |
| $C$ | $c\Sigma u_{ip}^{-1}$ | Persson (1971), Cox (1976) |
| $D$ | $c/\text{median}(u_{ip})$ | Persson (1971) |
| $E$ | $c/(\Sigma u_i \Sigma u_i^*)^{1/2}$ | Diggle (1975) |
| $F$ | $c/(\Sigma u_i \Sigma v_{iT})^{1/2}$ | Diggle (1975) |
| $G$ | $c_1(\Sigma u_i)^{-1} + c_2(\Sigma u_{i2})^{-1}$ | Lewis (1975) |
| $H$ | $c_1\Sigma(u_i/w_i) + c_2(\Sigma u_{i2})^{-1}$ | Cox (1976) |
| $I$ | $c_1/(\Sigma u_i)c_2^{-K}$ | Batcheler and Hodder (1975), Warren and Batcheler (1979) |
| $J$ | $c\,k(n)u[k(n)]/n$ | Patil *et al.* (1979, 1982) |
| $K$ | $c/(\Sigma r_i)(\Sigma v_{iT}^{1/2})$ | Byth (1982b) |
| $L$ | $c/(\Sigma r_i \Sigma r_i^*)$ | Clayton and Cox (1986) |

**Notation for Table 2**

$c,\ c_1,\ c_2$ = constants

$K = c\Sigma(r_i - \bar{r})^2/(\Sigma r_i)(\Sigma v_i^{1/2})$

$k(n)$ = function of $n$ such that $k(n) \to \infty$ and $k(n)/n \to 0$ as $n \to \infty$,

[*e. g.*, $k(n) = n^{1/2}$]

$u[k(n)]$ = $[k(n)]$th order statistic from $u_1, \ldots, u_n$

Other notation as for Table 1.

Diggle (1977a) shows that the estimators $(E)$ and $(F)$ are more robust than the earlier suggestions $(A\text{-}D)$; $(E)$ performs better than $(F)$ when the clusters are rather diffuse but, of course, $(E)$ may not be usable in practice. Cox (1976) shows $(H)$ to be better than $(C)$. An extensive Monte Carlo study by Byth (1982b) indicates that $(E)$ and $(F)$ perform better than $(A\text{-}D)$ or $(I)$, but suggests that $(K)$ may be best of all unless there is a very high degree of clustering. The evidence on other methods is incomplete although method $(J)$ appears to have a high standard error despite its low degree of bias; the several possible estimators based upon $(L)$ seem to perform quite well.

Aherne and Diggle (1978) recommend the use of $(E)$ or $(F)$ in conjunction with a preliminary test of randomness. This opens up the possibility of using different estimators according to whether the spatial pattern is judged regular, random or clustered. For forestry applications, Ord (1978) gives an unbiased estimator based upon the angle–exclusion method.

## 5.2. Second moment methods

The second order intensity function may be defined more generally than in (3.3) as

$$\lambda_2(\mathbf{x},\mathbf{y}) = \lim_{\substack{|d\mathbf{x}|\to 0\\ |d\mathbf{y}|\to 0}} E[N(d\mathbf{x})N(d\mathbf{y})]/(|d\mathbf{x}|\cdot|d\mathbf{y}|). \tag{5.1}$$

For stationary processes $\lambda_2(\mathbf{x},\mathbf{y}) = \lambda_2(\mathbf{x}-\mathbf{y})$ and when the process is also isotropic

$$\lambda_2(\mathbf{x},\mathbf{y}) = \lambda_2(t),$$

where $t^2 = \|\mathbf{x}-\mathbf{y}\|$. For the Poisson process, $\lambda_2(\mathbf{x},\mathbf{y}) = \lambda^2$ for all $\mathbf{x}$ and $\mathbf{y}$.

An alternative representation of second order properties is by means of the function (Ripley, 1976a):

$$K(t) = \lambda^{-1}E[N(t)] \tag{5.2}$$

where $N(t)$ denotes the number of further individuals within distance $t$ of a given individual. For the Poisson process, $K(t) = \pi t^2$. The function $K(t)$ is a natural representation for mapped data since an unbiased estimator is then available as

$$\widehat{K}(t) = |A|\sum k(\mathbf{x},\mathbf{y})/N^2, \tag{5.3}$$

where

$|A|$ = area of the study region $A$,
$N$ = number of points in $A$, and
$[k(\mathbf{x},\mathbf{y})]^{-1}$ = proportion within $A$ of a disk of radius $t$ centered on $\mathbf{x}$ (and passing through $\mathbf{y}$),

and the summation is taken over all pairs $(\mathbf{x},\mathbf{y})\epsilon A$ less than distance $t$ apart. Other edge corrections are possible; see Ohser and Stoyan (1981). $\widehat{K}$ was introduced by Ripley (1976a); Ripley (1979) shows that the distribution of $\widehat{K}$ is approximately Poisson for distances small relative to the size of the study area. A normal approximation is reasonable for $N \geq 50$ provided the average number of points per unit area is not too small.

For non–Poisson processes, the distribution theory is intractable, but the sample $\widehat{K}(t)$ may be contrasted with the envelope of a set of random simulations to determine agreement with a specified model (cf. Ripley, 1977; Diggle, 1979a,b; Diggle and Gratton, 1984). Silverman (1978) and Besag (in the discussion on Ripley, 1977) note that $L(t) = [\widehat{K}(t)/\pi]^{1/2}$ has a linear plot against $t$ and also $\mathrm{var}[L(t)]$ is approximately constant. This provides the basis for test $(R)$ in Table 1.

As defined, $K(t)$ requires both stationarity and isotropy; however, isotropy is not necessary (Ripley, 1976a). Ohser and Stoyan (1981) suggest plotting a "rose" of directions with $\lambda K(t,\gamma)$ defined as the number of individuals within $t$ of the given individual, with an orientation angle $\theta \leq \gamma$. Hanisch (1983) considers higher order moments, both for isotropic and anisotropic processes.

Another form of analysis useful for mapped patterns is the estimation of

$$P(t) = P\,[\text{circle of radius } t \text{ contains no individuals}] \tag{5.4}$$

sometimes known as "empty space" methods. Again, edge corrections are necessary to produce an unbiased estimator (Ripley, 1977, 1984b,c). The information content of (5.2) and (5.4) is different, so both analyses should be performed (Ripley, 1977). Cox (1979) describes a simple approach for identifying "sparse" and "dense" regions which may then be incorporated into computer mapping of the study region. For further discussion of second order methods see Getis and Franklin (1987).

### 5.3. Model–based inference

The fitting of parametric process models to spatial point patterns has proved extremely difficult once the Poisson scheme is rejected as inadequate. Primarily this is due to the very heavy computational burdens which must be borne. These problems arise in evaluating normalization constants for Gibbs processes as in (3.3) or generally in evaluating joint likelihood functions. Ogata and Tanemura (1981, 1984) developed the likelihood for a Gibbs process based on a variety of local pair–potential functions satisfying (3.4). Even so, several heroic approximations were necessary before the computational effort proved feasible. Gates and Westcott (1986) showed that some of the Ogata–Tanemura simulated examples were unrealistic and that the potential functions under consideration violated a stability condition, at least for some parameter combinations. They conclude that data analysis based on models with unstable potential functions must be performed with great care and, in general, recommend the use of more traditional clustering methods. Alternative approximations for hard–core models are given by Westcott (1982).

Shapiro, Schein and De Monasterio (1985) give an interesting example of modeling a spatial point process. Their paper should be read in conjunction with the ensuing discussion by Diggle and Gates.

Ogata and Tanemura (1984) introduce a "very soft core" (VSC) model with potential function

$$\beta(r) = -\log[1 - \exp(-\alpha r^2)],$$

where $r^2 = \|\mathbf{x} - \mathbf{y}\|$. The VSC process represents a weaker form of local inhibition than the hard–core models where rounding error is present.

Sager (1982) develops a non–parametric maximum likelihood estimation procedure using smoothing methods. Diggle and Gratton (1984) further develop the use of kernel methods for estimating the likelihood of an implicitly defined process. They give suitable numerical procedures for obtaining the maximum likelihood estimators from the estimated likelihood function. The central idea, developed from Diggle (1978), is that the sample realization, described by some function such as $\widehat{K}(t)$, is fitted as closely as possible to the theoretical function $\widehat{K}(t, \boldsymbol{\theta})$ estimated from $S$ simulated realizations of the process for different values of the parameters $\boldsymbol{\theta}$. The intractability of the likelihood functions led Diggle and Grattan to determine $\widehat{\boldsymbol{\theta}}$ by minimizing

$$d_s(\boldsymbol{\theta}) = \int_0^{t_0} [\{\widehat{K}(t)\}^{1/2} - \left\{\frac{1}{s}\sum_{j=1}^{s} \widehat{K}_j(t, \boldsymbol{\theta})\right\}^{1/2}]^2 \, dt$$

so that their estimates are "sensible" rather than "optimal." Diggle, Gates and Stibbard (1987) give an improved non–parametric estimator for interaction processes based on an

approximation drawn from statistical physics. This works well in simulations and probably represents the best approach to model estimation at the present time, although it is not immune from some numerical problems.

### 5.4. Simulation of processes

The simulation of spatial point processes represents another computer–intensive task. When the spatial model can be represented as the equilibrium form of a spatio–temporal process, the scheme can be run over sufficient time periods until an equilibrium may be assumed to have been achieved. Kelly and Ripley (1976) give a rejection method for hard–core processes; see also Lotwick (1981). The Dirichlet tessellation algorithm of Green and Sibson (1978) is used to identify empty areas and to speed up the assignment of individuals to available spaces. Lewis and Shedler (1979) simulate a non–homogeneous Poisson process by thinning. For further discussion, see Ripley (1987).

## 6. Multitype processes

Point processes involving several types of individual are variously known as *multivariate*, *marked* or *multitype* processes; we shall use the last of these terms. A useful discussion of the multitype process is given in Diggle (1983, Chapter 6).

### 6.1. Bivariate Cox processes

A convenient way to approach such schemes is to allow the individuals to be generated by a mechanism similar to the univariate scheme, and then to superimpose a "marking" scheme. The specification of the basic bivariate Cox process (Diggle, 1983, pp. 96-98) illustrates this:

**BCP1** $\Lambda(\mathbf{x}) = \{\Lambda_1(\mathbf{x}), \Lambda_2(\mathbf{x})\}$ is defined for all $\mathbf{x} \in \Re^2$ and represents a pair of non–negative valued stochastic processes.

**BCP2** Conditional on $\Lambda_j(\mathbf{x}) = \lambda_j(\mathbf{x}), \quad j = 1, 2$ and all $\mathbf{x} \in \Re^2$, type 1 and type 2 events form a pair of independent non–homogeneous Poisson processes with intensity functions $\lambda_j(\mathbf{x})$.

Dependence between the processes may then be achieved by modifying **BCP2**. For example,

$\mathbf{\Lambda_2(x) = c\Lambda_1(x)}$ represents extreme positive correlation

$\mathbf{\Lambda_1(x) + a\Lambda_2(x) = b}$ represents extreme negative correlation, and

$\mathbf{\Lambda_j(x) = \Lambda_0(x) + \Lambda_j^*(x)}$ represents a variety of processes with positive correlation.

The class of **BCP** schemes is studied in detail by Diggle and Milne (1983b), who also consider thinned bivariate processes. Brown *et al.* (1982) develop a class of multitype processes where each marginal process is Poisson but there is negative correlation between the processes.

Isham (1984) has developed a bivariate Markov point process extending the earlier work of Ripley and Kelly (1977). Ogata and Tanemura (1985) have extended their potential function models to the bivariate case.

### 6.2. Tests of dependence

Let $R_{ij}$ denote the distance from an arbitrary individual of type $i$ to the nearest individual of type $j$ and the subscript $i = 0$ denotes a randomly selected point in the study region. Let $F_{ij}(r) = P(R_{ij} \leq r)$.

If the two point processes are independent, stationary and isotropic it follows that

$$F_{01}(r) = F_{21}(r) \quad \text{and } F_{02}(r) = F_{12}(r). \tag{6.1}$$

Goodall (1965) tested the equality of the distributions in (6.1) by applying a two sample t–test to the square roots of the distances. Diggle and Cox (1981, 1983) show that Goodall's test is fairly robust against departures from the assumptions, but recommend the use of the corresponding Mann–Whitney test. A second pair of tests follows from testing the correlation between $R_{01}$ and $R_{02}$ for a pair of individuals near randomly selected points. The test based upon Kendall's tau appears to perform best but the form of the alternative hypothesis is critical. An advantage of the correlation test is that it does not require random sampling of individuals.

### 6.3. Second order properties

The appropriate second order function is

$$K_{ij}(t) = \lambda_j^{-1} E[N_{ij}(t)], \tag{6.2}$$

where $N_{ij}(t)$ denotes the number of type $j$ individuals within distance $t$ of an arbitrary type $i$ individual, which follows directly from the univariate case (Ripley, 1976). Two estimators, $\hat{K}_{12}(t)$ and $\hat{K}_{21}(t)$, are available and since $K_{12}(t) = K_{21}(t)$, a simple average of the estimates is usually taken for each value of $t$ (Lotwick and Silverman, 1982). Examples of data analyses are given in Diggle and Milne (1983b), Lotwick and Silverman (1982) and Harkness and Isham (1983). Lotwick (1984) shows that stationary ergodic processes exist for which the interactions cannot be detected using $K$, so that empty space methods should also be examined; the extension from (5.4) is relatively straightforward.

Byth (1981) develops methods for isotropic but non–stationary multitype processes.

## 7. Related topics

It is inevitable that any review must draw boundaries around the subject matter that is considered "most relevant" and omit discussion of other topics. Invariably, some topics are caught on the boundary. In this case, perhaps the two major omissions are mosaics and sampling. Comprehensive reviews of earlier work on mosaics is provided by Pielou (1977, pp. 181-199) and by Getis and Boots (1978, Chapters 6 and 7). Roach (1968) should also be consulted. Diggle (1981b) suggests a stochastic model for binary mosaics as a union of overlapping disks and uses the model to describe the spatial pattern of heather. Hall (1985) describes an alternative modeling approach; also see Ripley (1986) for discussion of this model in the context of pattern recognition methods.

A general view of recent developments in sampling is provided by the volume of papers edited by Cormack *et al.* (1979). With regard to sampling point patterns, the early work of Matern (1960) has been followed up by Diggle and Matern (1981) among others; Byth and Ripley (1980) also discuss sampling procedures for distance measurements.

For other recent developments in spatial processes, the review paper of Ripley (1984a) should be consulted. This paper has restricted attention to purely spatial, or static, models. For a recent survey of dynamic spatial models including birth–death and epidemic processes, see Renshaw (1986).

## 8. Future directions

The field of spatial point processes has undergone very major changes in the past fifteen years and it is a research area where the computer revolution will continue to have a strong impact on future model–building efforts.

It would seem that the major classes of stochastic model have been developed, at least for the univariate case, although further work is merited on interesting special cases, such as the Poisson–Gaussian. Much more research is needed to develop effective inferential methods. The combined use of likelihood approximations, Monte Carlo methods and second order methods seems to offer a fruitful road ahead for both univariate and multitype processes.

The development of tests for randomness would seem to have reached the point of diminishing returns, although some useful work remains to be done on power comparisons; perhaps similar comments are true for the robust estimation of intensity, although the use of preliminary tests in estimation seems worthy of further investigation.

An area with considerable potential for development is the use of more complex sampling schemes such as the semi–systematic scheme of Byth and Ripley (1980) or the divided quadrats scheme of Ord (1970). Close collaboration with researchers in the field will be necessary to ensure that such methods are cost–effective as well as statistically sound.

Finally, it should be noted that work on dependence among multitype processes has only just started and much remains to be done both to describe patterns of dependence and to develop the necessary inferential tools.

## 9. References

Aherne, W., and P. Diggle. (1978) The estimation of removal population density by a robust method. *Journal of Microscopy*, **114**, 285-293.

Bartlett, M. (1964) The spectral analysis of two–dimensional point processes. *Biometrika*, **51**, 299-311.

Bartlett, M. (1975) *The Statistical Analysis of Spatial Pattern*. London: Chapman and Hall.

Batcheler, C., and R. Hodder. (1975) Tests of a distance technique for inventory of pine populations. *New Zealand Journal of Forestry Science*, **5**, 3-17.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society*, **36B**, 192-236.

Besag, J., and J. Gleaves. (1973) On the detection of spatial pattern in plant communities. *Bulletin of the International Statistical Institute*, **45**, 153-158.

Besag, J., R. Milne, and S. Zachary. (1982) Point process limits of lattice processes. *Journal of Applied Probability*, **19**, 210-216.

Brown, D., and P. Rothery (1978) Randomness and local regularity of points in a plane. *Biometrika*, **65**, 115-122.

Brown, S., and P. Holgate. (1974) The thinned plantation. *Biometrika*, **61**, 253-262.

Brown, T. (1979) Position dependent and stochastic thinning of point processes. *Stochastic Processes and Their Applications*, **9**, 189-193.

Brown, T., B. Silverman, and R. Milne. (1981) A class of two–type point processes. *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete*, **58**, 299-308.

Byth, K. (1981) $\boldsymbol{\theta}$–stationary point processes and their second–order analysis. *Journal of Applied Probability*, **18**, 864-878.

Byth, K. (1982a) On kernel methods of estimating marginal, radial and angular probability density functions. *Biometrical Journal*, **24**, 49-58.

Byth, K. (1982b) On robust distance–based density estimators. *Biometrics*, **38**, 127-135.

Byth, K., and B. Ripley. (1980) On sampling spatial patterns by distance methods. *Biometrics*, **36**, 279-284.

Clark, P., and F. Evans. (1954) Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology*, **35**, 445-453.

Clayton, G. (1984) On the random–pairs method of density estimation. *Biometrics*, **40**, 199-202.

Clayton, G., and T. Cox. (1986) Some robust density estimators for spatial point processes. *Biometrics*, **42**, 753-67.

Cliff, A., and J. Ord. (1981) *Spatial Processes: Models and Applications*. London: Pion.

Cormack, R. (1977) The invariance of Cox and Lewis' statistic for the analysis of spatial patterns. *Biometrika*, **64**, 143-144.

Cormack, R. (1979) Spatial aspects of competition between individuals, in *Spatial and Temporal Analysis in Ecology*, edited by R. Cormack and J. Ord. Fairland, Md.: International Cooperative Publishing House, pp. 151-211.

Cormack, R., and J. Ord (eds.) (1979) *Spatial and Temporal Analysis in Ecology*. Fairland, Md.: International Cooperative Publishing House.

Cormack, R., G. Patil, and D. Robson (eds.) (1979) *Sampling Biological Populations*. Fairland, Md.: International Cooperative Publishing House.

Cox, D. (1955) Some statistical methods related with series of events. *Journal of the Royal Statistical Society*, **17B**, 129-164.

Cox, T. (1976) The robust estimation of the density of a forest stand using a new conditioned distance method. *Biometrika*, **63**, 493-500.

Cox, T. (1979) A method for mapping the dense and sparse regions of a forest stand. *Applied Statistics*, **28**, 14-19.

Cox, T. (1983) Nearest neighbors to nearest neighbors. *Statistics & Probability Letters*, **1**, 161-166.

Cox, T., and T. Lewis. (1976) A conditioned distance ratio method for analyzing spatial patterns. *Biometrika*, **63**, 483-492.

De Vries, P. (1979) Line intersect sampling–statistical theory, applications and suggestions for extended use in ecological inventory, in *Sampling Biological Populations*, edited by R. Cormack, G. Patil and D. Robson. Fairland, Md.: International Cooperative Publishing House, pp. 1-70.

Diggle, P. (1975) Robust density estimation using distance methods. *Biometrika*, **62**, 39-48.

Diggle, P. (1977a) A note on robust density estimation for spatial point patterns. *Biometrika*, **64**, 91-95.

Diggle, P. (1977b) The detection of random heterogeneity in plant populations. *Biometrics*, **33**, 390-394.

Diggle, P. (1978) On parameter estimation for spatial point processes. *Journal of the Royal Statistical Society*, **40B**, 178-181.

Diggle, P. (1979a) On parameter estimation and goodness–of–fit testing for spatial point patterns. *Biometrics*, **35**, 87-101.

Diggle, P. (1979b) Statistical methods for spatial point patterns in ecology, in *Spatial and Temporal Analysis in Ecology*, edited by R. Cormack and J. Ord. Fairland, Md.: International Cooperative Publishing House, pp. 95-150.

Diggle, P. (1981a) Some graphical methods in the analysis of spatial point patterns, in *Interpreting Multivariate Data*, edited by V. Barnett. Chichester: Wiley, pp. 55-73.

Diggle, P. (1981b) Binary mosaics and the spatial pattern of heather. *Biometrics*, **37**, 531-539.

Diggle, P. (1983) *Statistical Analysis of Spatial Pattern*. New York: Academic Press.

Diggle, P., and T. Cox. (1981) On sparse sampling methods and tests of independence for multivariate spatial point patterns. *Bulletin of the International Statistical Institute*, **49**, 211-229.

Diggle, P., and T. Cox. (1983) Some distance based tests of independence for sparsely–sampled multivariate spatial point patterns. *International Statistical Review*, **51**, 11-23.

Diggle, P., D. Gates, and A. Stibbard. (1987) A non–parametric estimator for pairwise interaction point processes. *Biometrika*, **74**, 763-770.

Diggle, P., and R. Gratton. (1984) Monte Carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society*, **46B**, 193-227.

Diggle, P., and B. Matern. (1981) On sampling designs for the estimation of point–event nearest neighbor distributions. *Scandinavian Journal of Statistics*, **7**, 80-84.

Diggle, P., and R. Milne. (1983a) Negative binomial quadrat counts and point processes. *Scandinavian Journal of Statistics*, **10**, 257-267.

Diggle, P., and R. Milne. (1983b) Bivariate Cox processes: some models for bivariate spatial point patterns. *Journal of the Royal Statistical Society*, **45B**, 11-21.

Eberhardt, L. (1967) Some developments in "distance" sampling. *Biometrics*, **23**, 207-216.

Gates, C. (1979) Line transect and related issues, in *Sampling Biological Populations*, edited by R. Cormack, G. Patil and D. Robson. Fairland, Md.: International Cooperative Publishing House, pp. 71-154.

Gates, D., and M. Westcott. (1986) Clustering estimates for spatial point distributions with unstable potentials. *Annals of the Institute of Statistical Mathematics*, **38**, 123-135.

Getis, A., and B. Boots. (1978) *Models of Spatial Processes*. Cambridge: Cambridge University Press.

Getis, A., and J. Franklin. (1987) Second order neighborhood analysis of mapped point

patterns. *Ecology*, **68**, 473-477.

Goodall, D. (1965) Plot–less tests of interspecific association. *Journal of Ecology*, **53**, 197-210.

Goodall, D. (1974) A new method for the analysis of spatial pattern by random pairing of quadrats. *Vegetatio*, **29**, 135-146.

Green, P., and R. Sibson. (1978) Computing Dirichlet tessellations in the plane. *Computing Journal*, **21**, 168-173.

Greig–Smith, P. (1952) The use of random and contiguous quadrats in the study of the structure of plant communities. *Annals of Botany*, **NS16**, 293-316.

Greig–Smith, P. (1979) Pattern in vegetation. *Journal of Ecology*, **67**, 755-779.

Hall, P. (1985) Counting methods for inference in binary mosaics. *Biometrics*, **41**, 1049-1052.

Hanisch, K. (1983) Reduction of the nth moment measure and the special case of the third moment measure of stationary and isotropic point–processes. *Mathematische Operationsforschung und Statistik*, **14**, 421-435.

Harkness, R., and V. Isham. (1983) A bivariate spatial point pattern of ants' nests. *Applied Statistics*, **32**, 293-303.

Heltshe, J., and T. Richey. (1984) Spatial pattern detection using quadrat samples. *Biometrics*, **40**, 877-885.

Hill, M. (1973) The intensity of spatial pattern in plant communities. *Journal of Ecology*, **61**, 225-235.

Hines, W., and R. Hines. (1979) The Eberhardt index and the detection of non–randomness of spatial point distributions. *Biometrika*, **66**, 73-79.

Holgate, P. (1965) Tests of randomness based on distance methods. *Biometrika*, **52**, 345-353.

Hopkins, B. (1954) A new method of determining the type of distribution of plant individuals. *Annals of Botany*, **18**, 213-226.

Isham, V. (1984) Bivariate Markov point processes–some approximations. *Proceedings of the Royal Society*, **391A**, 39-53.

Jolivet, E. (1978) Caractérisation et test du caractère agrégatif des processus ponctuels stationnaires sur R2, in *Lecture Notes in Mathematics*, #636. Berlin: Springer–Verlag, pp. 1-25.

Kelly, F., and B. Ripley. (1976) A note on Strauss' model for clustering. *Biometrika*, **63**, 357-360.

Kemp, C., and H. Papageorgiou. (1982) Bivariate Hermite distributions. *Sankhya*, **44A**, 269-280.

Kinderman, R., and J. Snell. (1980) *Markov Random Fields and Their Applications*. Providence: American Mathematical Society.

Kingman, J. (1977) Remarks on the spatial distribution of a reproducing population. *Journal of Applied Probability*, **14**, 577-583.

Kooijman, S. (1979) The description of point patterns, in *Spatial and Temporal Analysis in Ecology*, edited by R. Cormack and J. Ord. Fairland, Md.: International Cooperative Publishing House, pp. 305-332.

Lewis, P., and G. Shedler. (1979) Simulation of non–homogeneous Poisson processes by thinning. *Naval Research Logistics Quarterly*, **26**, 403-413.

Lewis, S. (1975) Robust estimation of density for a two–dimensional point process. *Biometrika*, **62**, 519-521.

Liebetrau, A. (1977) Tests of randomness in two dimensions. *Communications in Statistics: Theory and Methods*, **6A**, 1367-1383.

Lotwick, H. (1981) Simulation of some spatial hard core models, and the complete packing problem. *Journal of Statistical Computation and Simulation*, **15**, 293-314.

Lotwick, H. (1984) Some models for multitype spatial point processes, with remarks on analyzing multitype patterns. *Journal of Applied Probability*, **21**, 575-582.

Lotwick, H., and B. Silverman. (1982) Methods for analyzing spatial processes of several types of points. *Journal of the Royal Statistical Society*, **44B**, 406-413.

Matern, B. (1960) *Spatial Variation*. Stockholm: Meddelanden Statens fran Skogsforskningsinstitut, Vol. 49 pp. 1-144.

Matern, B. (1971) Doubly stochastic Poisson processes in the plane, in *Statistical Ecology*, Vol. 1, edited by G. Patil, E. Pielou and W. Waters. University Park, Pa.: The Pennsylvania State University Press, pp. 195-213.

Mead, R. (1974) A test for spatial pattern at several scales using data from a grid of contiguous quadrats. *Biometrics*, **30**, 295-307.

Moellering, H., and W. Tobler. (1972) Geographical variances. *Geographical Analysis*, **4**, 34-50.

Mountford, M. (1961) On E. C. Pielou's index of non–randomness. *Journal of Ecology*, **49**, 271-275.

Naus, J. (1979) An indexed bibliography of clusters, clumps and coincidences. *Review of the International Statistical Institute*, **47**, 47-78.

Newman, C., Y. Rinott, and A. Tversky. (1983) Nearest neighbors and Voronoi regions in certain point processes. *Advances in Applied Probability*, **15**, 726-751.

Ogata, Y., and M. Tanemura. (1981) Estimation of interaction potentials of spatial point patterns through the maximum likelihood procedure. *Annals* of the *Institute of Statistical Mathematics*, **33B**, 315-338.

Ogata, Y., and M. Tanemura. (1984) Likelihood analysis of spatial point patterns. *Journal of the Royal Statistical Society*, **46B**, 496-518.

Ogata, Y., and M. Tanemura. (1985) Estimation of interaction potentials of marked spatial patterns through the maximum likelihood method. *Biometrics*, **41**, 421-433.

Ohser, J., and D. Stoyan. (1981) On the second–order and orientation analysis of planar stationary processes. *Biometrical Journal*, **23**, 523-533.

Ord, J. (1970) The negative binomial model and quadrat sampling, in *Random Counts in Models and Structures*, edited by G. Patil. University Park, Pa.: The Pennsylvania State University Press, pp. 151-163.

Ord, J. (1972) *Families of Frequency Distributions*. London: Griffin.

Ord, J. (1978) How many trees in a forest? *Mathematical Scientist*, **3**, 23-33.

Patil, S., K. Burnham, and J. Kovner. (1979) Non–parametric estimation of plant density by the distance method. *Biometrics*, **35**, 597-604.

Patil, S., J. Kovner, and K. Burnham. (1982) Optimum nonparametric estimation of population density based on ordered distances. *Biometrics*, **38**, 243-248.

Perry, J., and R. Mead. (1979) On the power of the index of dispersion test to detect spatial pattern. *Biometrics*, **35**, 613-622.

Persson, O. (1971) The robustness of estimating density by distance measurements, in *Statistical Ecology*, Vol. 2, edited by G. Patil, E. Pielou and W. Waters. University Park, Pa.: The Pennsylvania State University Press, pp. 175-190.

Pickard, D. (1982) Isolated nearest neighbors. *Journal of Applied Probability*, **19**, 444-449.

Pielou, E. (1959) The use of point–to–plant distances in the study of the pattern of plant populations. *Journal of Ecology*, **47**, 607-613.

Pielou, E. (1977) *Mathematical Ecology*. New York: Wiley.

Pollard, J. (1971) On distance estimators of density in randomly distributed forests. *Biometrics*, **27**, 991-1002.

Renshaw, E. (1984) Competition experiments for light in a plant monoculture: an analysis based on two–dimensional spectra. *Biometrics*, **40**, 717-728.

Renshaw, E. (1986) A survey of stepping stone models in population dynamics. *Advances in Applied Probability*, **18**, 581-627.

Renshaw, E., and E. Ford. (1983) The interpretation of process from pattern using two dimensional spectral analysis: methods and problems of interpretation. *Applied Statistics*, **32**, 51-63.

Ripley, B. (1976a) The second order analysis of stationary point processes. *Journal of Applied Probability*, **13**, 255-266.

Ripley, B. (1976b) Locally finite random sets: Foundations for point process theory. *Annals of Probability*, **4**, 983-994.

Ripley, B. (1977) Modeling spatial patterns. *Journal of the Royal Statistical Society*, **39B**, 172-212.

Ripley, B. (1978) Spectral analysis and the analysis of pattern in plant communities. *Journal of Ecology*, **66**, 965-981.

Ripley, B. (1979) Tests of "randomness" for spatial point patterns. *Journal of the Royal Statistical Society*, **41B**, 368-374.

Ripley, B. (1981) *Spatial Statistics*. Chichester: Wiley.

Ripley, B. (1984a) Spatial Statistics: Developments 1980–3. *International Statistical Review*, **52**, 141-150.

Ripley, B. (1984b) Analyses of nest spacing, in *Statistics in Ornithology*, edited by B. Morgan and P. North. Berlin: Springer–Verlag, Lecture Notes in Statistics, pp. 151-158.

Ripley, B. (1984c) Edge corrections for spatial processes, in *Stochastic Geometry, Geometrical Statistics, Stereology*, edited by R. Ambartzumian and W. Weil. Leipzig: Teubner Texte, pp. 144-153.

Ripley, B. (1986) Statistics, images and pattern recognition. *Canadian Journal of Statistics*,

**14**, 83-102.

Ripley, B. (1987) *Stochastic Simulation.* New York: Wiley.

Ripley, B. (1988) *Statistical Inference for Spatial Processes.* New York: Cambridge University Press.

Ripley, B., and F. Kelly. (1977) Markov point processes. *Journal of the London Mathematical Society*, **15**, 188-192.

Ripley, B., and B. Silverman. (1978) Quick tests for spatial regularity. *Biometrika*, **65**, 641-642.

Roach, S. (1968) *The Theory of Random Clumping.* London: Methuen.

Sager, T. (1982) Nonparametric maximum likelihood estimation of spatial patterns. *Annals of Statistics*, **10**, 1125-1136.

Saunders, R., and G. Funk. (1977) Poisson limits for a clustering model of Strauss. *Journal of Applied Probability*, **14**, 776-84.

Saunders, R., R. Kryscio, and G. Funk. (1982) Poisson limits for a hardcore clustering model. *Stochastic Processes and Their Applications*, **12**, 97-106.

Shapiro, M., J. Schein, and F. De Monasterio. (1985) Regularity and structure of the spatial pattern of blue cones of macaque retina. *Journal of the American Statistical Association*, **80**, 803-812 (with discussion).

Silverman, B. (1978) Distances on circles, toruses and spheres. *Journal of Applied Probability*, **15**, 136-143.

Silverman, B. (1981) Density estimation for univariate and bivariate data, in *Interpreting Multivariate Data*, edited by V. Barnett. Chichester: Wiley, pp. 37-53.

Silverman, B., and T. Brown. (1978) Short distances, flat triangles and Poisson limits. *Journal of Applied Probability*, **15**, 815-825.

Skellam, J. (1951) Random dispersal in theoretical populations. *Biometrika*, **38**, 196-218.

Strauss, D. (1975) A model for clustering. *Biometrika*, **62**, 467-475.

Upton, G. (1984) On Mead's test for pattern. *Biometrics*, **40**, 759-766.

Upton, G., and B. Fingleton. (1985) *Spatial Data Analysis By Example*, Vol. 1. Chichester: Wiley.

Warren, W. (1971) The center–satellite concept as a basis for ecological sampling, in *Statistical Ecology*, Vol. 2, edited by G. Patil, E. Pielou and W. Waters. University Park, Pa: The Pennsylvania State University Press, pp. 87-116.

Warren, W., and C. Batcheler. (1979) The density of spatial patterns: robust estimation through distance methods, in *Spatial and Temporal Analysis in Ecology*, edited by R. Cormack and J. Ord. Fairland, Md.: International Cooperative Publishing House, pp. 247-270.

Westcott, M. (1982) Approximations to hard–core models and their application to statistical analysis. *Journal of Applied Probability*, **19**, 281-292.

Zahl, S. (1977) A comparison of three methods for the analysis of spatial pattern. *Biometrics*, **33**, 681-692.

## DISCUSSION

### "Statistical methods for point pattern data"

### by J. Keith Ord

The field of spatial point processes encompasses work in both probability and statistics as well as in several applied fields, and readers should note that Ord's review is much more restricted in scope than its title might suggest. Rather than comment on the details of the paper, I will attempt to complement it with other developments that I see as important. Almost all the issues addressed are covered in some detail in texts such as Diggle (1983), Ripley (1981) and Upton and Fingleton (1985), but many of the topics below are not.

**Theory.**

It is easy to overlook the importance of theory: major practical developments such as second-moment methods ($\widehat{\mathbf{K}}$) and Gibbs processes arose from theoretical ideas. Major theoretical contributions have been made by Kallenberg (1975, 1983), the East German school (Matthes *et al.*, 1978; Stoyan *et al.*, 1987), and Papangelou for conditional intensities (see Kallenberg, 1983). Apart from these monographs there are also expositions by Karr (1985) and Daley and Vere-Jones (1988). The area of Gibbs processes is on the interface with the axiomatization of statistical physics by Preston, Georgii and colleagues.

One theoretical contribution has been to find minimal conditions that define a Poisson process. Rather surprisingly, it suffices to know

$$P(N(A) = 0) \quad \text{for a sufficiently wide class of sets} \quad A \subset C$$
$$P(N(\{\mathbf{x}\}) > 1 \quad \text{for any } \mathbf{x}) = 0$$

to specify any point process that is *simple* (the second condition). In particular, a process with

$$N(A) \text{ Poisson mean } \Lambda(A), \qquad \Lambda(\{\mathbf{x}\}) = 0 \quad \text{for all } \mathbf{x}$$

is a Poisson process (a combination of results of Renyi, Kallenberg and myself in the period 1971-6).

**Point processes of objects.**

One of the major goals of spatial statistics has been the ability to handle more complicated image data than points. *Stochastic geometry* reduces objects to points in other spaces, and much work has been done on 'fibre processes' and their cousins (Stoyan *et al.*, 1987). These can represent lines and curves such as rivers, and Stoyan and Ohser (1982) have studied the interactions between points (trees), fibres (rivers) and areas (soil types) on Dresden Heath.

I would argue that point patterns *per se* are becoming less important. They are almost always approximations to the truth, representing objects with size and shape, and as we become able to cope with the image data we actually observe, the underlying point pattern descriptions will become important rather than direct inference from point patterns.

**Statistical inference.**

Much more has been done on inference from point patterns recently than appears in this review. Many of the issues are discussed in my own contribution to this volume and in Ripley (1988), but let me underline the importance of pseudo–likelihood and simulation methods, as well as Palm probabilities. The latter correspond to comparing the views of a point pattern from an arbitrary point with those from an 'individual' (in Ord's notation), and this has been exploited by Takacs (1986) and others subsequently.

I do not share Ord's pessimism (first paragraph) on computational intractability. The phrase seems to reflect our ignorance of what to do rather than fundamental problems of computational complexity. Over the last decade the picture has become much more promising due to all of

(a) developments in computational geometry (Preparata and Shamos, 1985),

(b) new ideas on approximations, and

(c) cheap desktop computing power, equal to that of a mainframe of a decade ago.

Unfortunately these subjects are not part of the traditional education of either statisticians or geographers. Indeed, the computing skills needed for modern methods may exclude many potential users, and one day we may see the `MINITAB` or `SAS` of spatial statistics.

**Mosaic models.**

Diggle (1981b) used distance methods to fit a point process model of objects to his heather data, and very successfully demonstrated how methods of low power allow one to fit inappropriate models! The inadequacies of his model are obvious to most people immediately on comparing images of the data with those of simulations of the model. Finding summary measures that are as good as 'eye–balling' took some time, and depended on measuring notions of shape as well as size (Ripley, 1986, 1988). Hall's (1985) methods suffer from similar difficulties.

Hall (1988) gives a comprehensive introduction to a family of *coverage processes* for mosaics. Other models are described by Ahuja and Schachter (1983).

Spatial point processes have had a golden few years, and the methods developed then are beginning to be widely used. It *is* the case that point patterns are now usually analysed as maps, and the K–function methods (specifically, plotting $\mathbf{L(t)} = \sqrt{\widehat{\mathbf{K}}(\mathbf{t})/\mathbf{t}}$ vs $\mathbf{t}$) have become standard in a number of biological fields. Nevertheless, currently point processes are being eclipsed by statistical image analysis, and I suspect developments in point processes in the near future will be less prolific.

**References**

Ahuja, N., and B. J. Schachter. (1983) *Pattern Models.* New York: Wiley.

Daley, D., and D. Vere-Jones. (1988) *An Introduction to the Theory of Point Processes.* New York: Springer–Verlag.

Hall, P. (1988) *An Introduction to the Theory of Coverage Processes.* New York: Wiley.

Kallenberg, O. (1975) *Random Measures.* Berlin: Akademie/Academic.

Kallenberg, O. (1983) *Random Measures*, 3rd ed. London: Academic Press.

Karr, A. (1985) *Point Processes and their Statistical Inference*. Marcel–Dekker, New York.

Matthes, K., J. Kerstan, and J. Mecke. (1978) *Infinitely Divisible Point Processes*. Chichester: Wiley.

Preparata, F. D. and M. L. Shamos. (1985) *Computational Geometry*. Springer, New York.

Ripley, B. D. (1986) Statistics, images and pattern recognition. *Canadian Journal of Statistics*, **14**, 83-112.

Ripley, B. D. (1988) *Statistical Inference for Spatial Processes*. Cambridge: Cambridge University Press.

Stoyan, D., and J. Ohser. (1982) Correlations between planar random structures, with an ecological application. *Biometrical Journal*, **24**, 631-647.

Stoyan, D., W. S. Kendall, and J. Mecke. (1987) *Stochastic Geometry and its Applications*. Chichester: Wiley.

Takacs, R. (1986) Estimator for the pair potential of a Gibbsian point process. *Statistics*, **17**, 429-433.

*Brian D. Ripley*, University of Strathclyde.

## A REJOINDER TO RIPLEY'S DISCUSSION

by J. Keith Ord

In some ways my reply must include a defense of Ripley–past from the comments of Ripley–present. But, overall, I do not believe our viewpoints are so very different.

It certainly was not my intention to underrate the role of theory. Rather, my review focuses more on methodology, as the title implies. I am grateful to Dr. Ripley for including the additional references to more theoretical literature.

I would agree entirely, for the case of image processing, that shapes are more important than points. However, there are many areas of application, especially in human geography, where point processes are, and will remain, of central interest.

My opening paragraph was not intended to be pessimistic. To the contrary, it is a comment on the history of spatial modeling. I agree wholeheartedly that the present need is for accessible software.

In terms of stochastic modeling, I suspect it is true that the golden age has passed for spatial point processes. However, the further development of inferential tools and diagnostic procedures remains a substantive challenge for the future. Only by such progress can innovative applications be assisted.

## *PREAMBLE*

*Statistics are like alienists–*
*they will testify for either side.*

F. H. LaGuardia, **Liberty**, (May, 1933)

*In a fashion somewhat similar to that found in the subsequent chapter by Haining, Anselin addresses a debate focusing on different perspectives regarding geo-referenced data analysis. In doing so, he promotes model validation and sensitivity analysis. But what conclusion should a researcher state when a slight, modest, or even radical change in underlying assumptions produces an opposite statistical decision? In recent years, scholars seem to be increasingly bombarded with contradictory statistical evidence extracted from data sets. Such pairs of findings could be amusing if consequences of their coexistence were not so unfortunate. The purpose of this paper is to review and evaluate various approaches to modelling and analyzing spatial data, as well as the role spatial errors play in these endeavors. By fulfilling this goal, Anselin helps to resolve these troublesome themes of conflicting statistical implications obtained from geo–referenced data. Haining goes on to emphasize three of the points raised by Anselin, stressing the importance of substance over method as a guiding light in data analysis.*

The Editor

# What Is Special About Spatial Data? Alternative Perspectives on Spatial Data Analysis

Luc Anselin *

*Department of Geography and Department of Economics, and National Center for Geographic Information and Analysis, University of California/ Santa Barbara, CA 93106, U.S.A.*

**Overview:** In this paper, some general ideas on fundamental issues are outlined, related to the distinctive characteristics of spatial data analysis, as opposed to data analysis in general. The emphasis is on the relevance for spatial data analysis of the ongoing debate about methodology in the disciplines of statistics and econometrics, and on the role of spatial errors in modeling and analysis. First, some general remarks are formulated on two opposing viewpoints regarding spatial analysis and spatial data: a data–driven approach versus a model–driven approach. This is followed by a review of a number of competing inferential frameworks that can be used as the basis for spatial data analysis. Next, the focus shifts to spatial errors and to the implications of various forms of spatial errors for spatial data analysis. Finally, some concluding remarks are formulated on future research directions in spatial statistics and spatial econometrics.

## 1. Introduction

The analysis of spatial data has always played a central role in the quantitative scientific tradition in geography. Recently, there have appeared a considerable number of publications devoted to presenting research results and to assessing the state of the art. For example, at an elementary level, Goodchild (1987a), Griffith (1987a), and Odland (1988) introduce the concept of spatial autocorrelation, and Boots and Getis (1988) review the analysis of point patterns. At more advanced levels, Anselin (1988a) and Griffith (1988) deal with a wider range of methodological issues in spatial econometrics and spatial statistics. Extensive reviews of the current state of the art for different aspects of spatial data analysis are presented in Anselin (1988b), Anselin and Griffith (1988), Getis (1988), Griffith (1987b), and Odland, Golledge and Rogerson (1989). In addition, spatial data analysis has received considerable attention as an essential element in the development of Geographic Information Systems (GIS), as outlined in Goodchild (1987b) and Openshaw (1987), and as an important factor in regional modeling, as argued in Anselin (1989a).

In this paper, I will take some distance from specific methods and techniques, and instead outline a few general ideas on fundamental issues related to the distinctive characteristics of spatial data analysis, as opposed to data analysis in general. I will focus on two issues that are often overlooked in technical treatments of the methods of spatial statistics and spatial

* Paper prepared for presentation at the Spring 1989 Symposium on Spatial Statistics, Past, Present and Future, Department of Geography, Syracuse University. The research reported on in this paper was supported in part by Grants SES 86-00465 and SES 87-21875 from the National Science Foundation, and by the National Center for Geographic Information and Analysis (NCGIA). An earlier version was presented at the NCGIA Specialist Meeting entitled "Accuracy of Spatial Databases," Montecito, CA, December 13-16, 1988.

econometrics. One is the relevance for spatial data analysis of the ongoing debate about methodology in the disciplines of statistics and econometrics. I will review and evaluate a number of different approaches towards modeling and analyzing spatial data, and put them in the context of the debate. Some recent examples of the opposing viewpoints that are taken in this debate can be found in Leamer (1978), Hendry (1980), Sims (1980, 1982), Lovell (1983), Swamy *et al.* (1985), Zellner (1985, 1988), Efron (1986), Pagan (1987), Kloek and Haitovsky (1988), and Durbin (1988). The second issue is much narrower and pertains to the role of spatial errors in modeling and analysis. This topic has recently received considerable attention in the context of GIS (*e. g.*, as evidenced in the 1988 Research Initiative of the National Center for Geographic Information and Analysis on "errors in spatial databases"), but many aspects of its relation to spatial data analysis remain to be explored.

The discussion in this paper is not intended to be comprehensive, but it is selective in the sense that I will focus on issues that seem to be most relevant to current modeling practice and most promising to lead to future research advances. Clearly, this selective treatment reflects my own biases and interests, and is focused on applications in regional science and analytical human geography.

The remainder of the paper consists of six sections. First, I formulate some general remarks on two opposing viewpoints regarding spatial analysis and spatial data: a data–driven approach versus a model–driven approach. This is followed by a review of a number of competing inferential frameworks that can be used as the basis for spatial data analysis. Next, I focus on spatial errors and on the implications of various forms of spatial errors for spatial data analysis. I close with some concluding remarks on future research directions in spatial statistics and spatial econometrics.

## 2. Spatial Analysis and Spatial Data

In general terms, spatial analysis can be considered to be the formal quantitative study of phenomena that manifest themselves in space. This implies a focus on location, area, distance and interaction, such as is expressed in Tobler's (1979) First Law of Geography, where "everything is related to everything else, but near things are more related than distant things." In order to interpret what "near" and "distant" mean in a particular context, observations on the phenomenon of interest need to be referenced in space (*e. g.*, in terms of points, lines or areal units). There are two opposite approaches towards dealing with spatially referenced data (Anselin, 1986b; Haining, 1986). In one, which I will call the data–driven approach, information is derived from the data without a strong prior notion of what the theoretical framework should be. In other words, one lets the "data speak for themselves" (Gould, 1981). In this largely inductive approach information on spatial pattern, spatial structure and spatial interaction is derived without the constraints of a pre–conceived theoretical notion.

In most respects, this approach falls under the category of "exploratory data analysis" (EDA) popularized by Tukey (1977) and Mosteller and Tukey (1977). It is also similar to the philosophy underlying time series analysis and forecasting of the Box–Jenkins (1976) type, and its extensions to vector autoregressive processes and the like (*e. g.*, Doan *et al.*, 1984; and the critique of Cooley and LeRoy, 1985).

The data–driven approach in spatial analysis is reflected in a wide range of different techniques, such as point pattern analysis (Getis and Boots, 1978; Diggle, 1983), indices of

spatial association (Hubert, 1985; Wartenberg, 1985), kriging (Clark, 1979), spatial adaptive filtering (Foster and Gorr, 1986), and spatial time series analysis (Bennett, 1979). All these techniques have two aspects in common. First, they compare the observed pattern in the data (*e. g.*, locations in point pattern analysis, values at locations in spatial autocorrelation) to one in which space is irrelevant. In point pattern analysis this is the familiar Poisson pattern, or "randomness," while in many of the indices of spatial association it is the assumption that an observed data value could occur equally likely at each location (*i. e.*, the null hypothesis for many tests for spatial autocorrelation, based on a normal or randomization approach).

The second common aspect is that the spatial pattern, spatial structure, or form for the spatial dependence are derived from the data only. For example, in spatial time series analysis, the specification of the autoregressive and moving average lag lengths is derived from autocorrelation indices or spatial spectra.

The data–driven approach is attractive in many respects, but its application is not always straightforward. Indeed, the characteristics of spatial data (dependence and heterogeneity) often void the attractive properties of standard statistical techniques. Since most EDA techniques are based on an assumption of independence, they cannot be implemented uncritically for spatial data. In this respect, it is also important to note that dependence in space is qualitatively more complex than dependence in the time dimension, due to its two–dimensional and two–directional nature. As a consequence, many results from the analysis of time series data do not apply to spatial data. As discussed in detail in Hooper and Hewings (1981), the extension of time series analysis into the spatial domain is limited, and only applies to highly regular processes. It goes without saying that most data in empirical spatial analysis for irregular areal units do not fit within this restrictive framework.

The second approach, which I will call model–driven, starts from a theoretical specification, which is subsequently confronted with the data. The theory in question may be spatial (*e. g.*, a spatial process or a spatial interaction model, as in Haining, 1978, 1984) or largely aspatial (*e. g.*, a multiregional economic model, as in Folmer, 1986), but the important characteristic is that its estimation or calibration is carried out with spatial data. The properties of this data, namely spatial dependence and spatial heterogeneity, necessitate the application of specialized statistical (or econometric) techniques, irrespective of the nature of the theory in the model.

Most of the methods that I would classify under this category deal with estimation and specification diagnostics in linear models in general, and regression models in particular (*e. g.*, Cliff and Ord, 1981; Anselin, 1980, 1988a). The main conceptual problem associated with this approach is how to formalize the role of "space." This is reflected in three major methodological problems, which are still largely unresolved to date: the choice of the spatial weights matrix (Stetzer, 1982a; Anselin, 1984, 1986a); the modifiable areal unit problem (Openshaw and Taylor, 1979, 1981); and the boundary value problem (Griffith, 1983, 1985; Griffith and Amrhein, 1983).

In order for the data–driven or the model–driven approaches to be operational, the various tests, diagnostics and estimators need to be incorporated in an inferential framework. More precisely, the uncertainty associated with a random variable, sampling error, or any other stochastic aspect of the data analysis needs to be assessed within a consistent framework that forms a logical basis for decisions. A number of competing frameworks have been suggested. They are discussed next.

## 3. Inferential Frameworks in Spatial Data Analysis

Spatial data analysis is not immune from the implications of the philosophical debates that go on in the broader disciplines of statistics and econometrics. Although the results of applied and empirical work are often presented as if only one particular view of statistics existed, there are in fact many competing perspectives (or even paradigms). Rather than repeating the various philosophical arguments, I will outline five dimensions of conflict or competition, and discuss some implications of the alternative viewpoints for spatial data analysis. Some of these dimensions are more fundamental than others, but all have direct applications to the practice of spatial statistics and spatial econometrics.

### 3.1. Classical versus Bayesian inference

The debate between the classical (Neyman–Pearson) and Bayesian approaches to statistical inference (or decision making) is undoubtedly the most fundamental one ongoing in the discipline. The arguments of both sides are well known and a compromise does not seem likely in the near future (*e. g.*, Efron, 1986; Durbin, 1988; Zellner, 1988). In a nutshell, the classical approach is "objective," and practical, but fraught with philosophical problems when applied in a strict sense: problems with multiple comparisons, the need to assume a "true" model, and the such. On the other hand, the Bayesian approach is generally considered to be superior in terms of overall consistency and as a perspective on "learning," but is "subjective" and difficult to apply to many practical problems, due to the need to construct complex prior distributions and to carry out numerical integration in multiple dimensions.

In spatial data analysis, the Bayesian perspective is the exception, and it has found only limited application. Some Bayesian concepts are fairly familiar in image processing of remotely sensed data (Richards, 1986), but applications to spatial data analysis in human geography are fairly rare (some exceptions are provided in March and Batty, 1975; Odland, 1978; Hepple, 1979; and Anselin, 1982, 1988b). Although the classical approach reflected in the Neyman–Pearson inferential framework is by far the dominant one in geography, its uncritical application to spatial data analysis is inappropriate in a lot of respects. The many assumptions, judgements and multiple comparisons carried out in the practice of estimation and data analysis (both data–driven as well as model–driven) make a mockery out of the rigorous and elegant probabilistic calculus that underlies the classical approach (for more details, see Anselin, 1988b). It therefore would seem, at least from a conceptual viewpoint, that a number of spatial "problems" could be most fruitfully attacked from a Bayesian perspective. Examples are pattern recognition, or "learning" from data in general, the prior assumptions about a spatial weights matrix, spatial interpolation, and dealing with boundary effects. However, the practical implementation of a Bayesian analysis of these issues is not straightforward. Specifically, it has so far not been possible to develop useful prior distributions for the full range of patterns of spatial dependence (spatial weight matrices) that would be operational in spatial data analysis. Overall, dealing with the two–directional nature of spatial dependence in a Bayesian framework is still very much an unresolved research topic.

### 3.2. Parametrics versus non–parametrics and/or robustness

In applied spatial data analysis, the standard assumptions of normality and of perfect knowledge of the model specification are often rather crude abstractions of reality. Consequently, the relevance of a strict parametric approach has been increasingly questioned, and a nonparametric, qualitative or robust perspective has sometimes been suggested as an alternative, by, among others, Gould (1981), Costanzo (1983), Nijkamp, Leitner and Wrigley (1985) and Knudsen (1987). However, it is not as if nonparametric and robust procedures have not been introduced into spatial analysis. On the contrary, a number of well known indices for spatial association have been based on randomization, permutation and other nonparametric techniques. Examples range from a robust Moran index in Cliff and Ord (1973), and Sen and Soot (1977), to the general measures of spatial association in Hubert *et al.* (1981, 1985). Most of these methods would fall under the data–driven category of spatial data analysis. However, there have been some recent applications in the model–driven category as well, primarily based on the use of the Jackknife and bootstrap estimation techniques (Stetzer, 1982b; Folmer and Fischer, 1984; Anselin, 1989b).

In spite of the concerns about its appropriateness, the parametric approach remains the most common one in spatial data analysis. Most tests are based on an underlying distribution which is normal (for values) or Poisson (for point patterns) and the estimation method of choice is the maximum likelihood technique. As is well known, the parametric approach is optimal in a number of ways if the underlying assumptions are indeed satisfied. It is when this is not the case that problems occur. Since the robust and nonparametric techniques are not grounded in such a restrictive set of assumptions, they remain valid in a wider range of situations. However, this robustness comes at the cost of a loss in generality and precision. For example, the spatial association indices that are based on a permutation approach only pertain to the data at hand, and cannot be generalized to hold for a "population." Similarly, the variance estimates for parameters obtained by means of the bootstrap or Jackknife will tend to be larger than for the maximum likelihood (ML) approach, and thus will lead to a more conservative inference (*i. e.*, it will be "harder" to find significant coefficients). Clearly, when the assumptions underlying the ML approach do hold (primarily normality of the distribution), the larger variances of the robust approach will be inefficient and the parametric approach is superior. However, this is likely to be the exception rather than the rule in spatial data sets.

An important obstacle for the acceptance of robust or non–parametric techniques in spatial data analysis is that a great many of the techniques developed in mathematical statistics and econometrics (*e. g.*, as reviewed in Huber, 1981; Koenker, 1982; Efron, 1982; and Robinson, 1988) are not directly transferable, since they are based on an assumption of observational independence. An appropriate "spatial" theoretical framework for robust analysis remains to be developed.

### 3.3. Random sample versus stochastic process

The dependence that is inherent in many (if not most) spatial data runs directly counter to the postulate of a random sample of independent observations upon which most common statistical procedures are based. Nevertheless, much applied spatial data analysis still proceeds as if the standard assumptions hold (see Anselin and Griffith, 1988), and notions of sampling error, sampling variance, and the such, abound in the empirical literature. Clearly, this is

incorrect, and the loss of information that results from the dependence in the observations should be accounted for.

In most instances, the proper perspective is not to consider spatial data as a random sample with many observations, but instead as a single realization of a stochastic process. In contrast to the sampling approach, where each observation is taken to provide an independent piece of information, the dependence (and heterogeneity) embodied in a stochastic process implies that only one observation is available, which is the full spatial pattern (or space–time pattern) of values. Provided that the underlying stochastic process is sufficiently stable (stationary, isotropic, *etc.* ...) or that the structure of the instability (nonstationarity) is known, the observed pattern will yield information on the characteristics of that process. In contrast to the random sampling approach, where the notion of independence is exploited in order to derive exact statistical properties for estimates and hypothesis tests, an asymptotic reasoning is needed in the stochastic process approach. Specifically, the theory of mixing processes, which allows a degree of dependence as well as heterogeneity, forms a solid basis for the inference for spatial stochastic processes (for details see Anselin, 1988a, Chapter 5).

The consequence of spatial dependence, or, more precisely, positive spatial dependence is that the observations contain less information than if there had been independence. In other words, in order to obtain approximately the same degree of information as in an independent set of observations, a larger data set of (positively) dependent observations will be needed. Sometimes, the latter can be transformed into the former, by deleting observations that are contiguous or within a given distance of each other. For example, if only those observations are selected that are far enough apart so that no marked dependence can reasonably be expected (*i. e.*, dependence related to distance only) this new "sample" can be considered to be independent for most practical purposes. This "re–coding" lies at the basis of the so–called "conditional" approach to spatial modeling (Haining, 1986). Its advantage is that most standard statistical techniques can be applied unchanged to the re–coded data. However, the re–coding itself is not unique and somewhat arbitrary. Also, this is only a practical approach if the loss of information from discarding the "dependent" observations is not critical. Unfortunately, in many practical situations such a luxury does not exist, and the "simultaneous" (joint probability) stochastic process approach is the only feasible parametric framework.

A related issue is the extent to which spatial data constitute a sample, a realization of a stochastic process, or instead form the complete population of interest. It is sometimes argued that the latter is the only correct perspective, and not only that no inferential statistics are possible, but also that a descriptive approach is the only valid one (*e. g.*, Summerfield, 1983). Although this may be an acceptable viewpoint in the case of extreme heterogeneity (*i. e.*, each place is "unique" and no generalization is possible), it is more the exception than the rule. There are two crucial issues that need to be considered. The first pertains to the imperfect nature of measurement, and the inherent error (or noise). Since a mixture of signal and noise is observed in empirical practice, the stochastic nature of the data can be easily generated from the randomness in errors of measurement. As a consequence, the population in question pertains to the family of stochastic processes that may have generated a particular error pattern. Thus, a "statistical" approach is the only way in which conclusions can be formed about the underlying "signal" and an understanding of spatial errors is crucial.

The second issue pertains to the nature of space as a framework within which observations are ordered. In essence, the spatial unit of observation needs to be a representative unit for the phenomenon that is under study. Only then will it be possible to formulate and test general statements about "space." The real issue is whether the observations at hand are compatible with the complexity of the phenomenon of interest. If they are not, this does not mean that a statistical approach should be rejected, but rather that other types of data are needed. For example, this may necessitate the collection of micro–behavioral data to avoid problems of ecological fallacy, or may require the extension of a cross–section into the time dimension in order to formulate general conclusions about a specific region.

### 3.4. Finite sample versus asymptotics

The stochastic process approach to spatial data analysis is based on asymptotic properties for an "abstract" and infinitely large data set. This conceptual framework contrasts sharply with the reality of small data sets with a finite number of observations. Two issues merit some consideration. The first is practical and pertains to the extent to which the asymptotic properties are valid in finite samples. As is well known, this is not necessarily the case, and many properties of equivalence and optimality of asymptotic tests and estimators are not reflected in realistic data sets. Moreover, few analytic results are available and the properties of a number of approximations are questionable (see, for instance Taylor, 1983; and also Anselin, 1988b for spatial data). In other words, considerable caution (a conservative inference) is needed when interpreting the findings of spatial data analysis that are based on asymptotic properties.

The second issue related to asymptotics is more conceptual and pertains to the relevance of the notion of an infinitely large data set for spatial analysis. In essence, an asymptotic reasoning is only meaningful if an infinitely large number of replications of the observed spatial units can be conceived of. While this is fairly straightforward in the case of a continuous process that is observed on regularly spaced points or grids, it is not at all obvious for discontinuous processes or observations for irregular areal units (*e. g.*, a given set of counties in a state). There are two approaches to this conceptual problem. In one, the data for irregular spatial units are transformed (interpolated) to regular spatial units. Although this forms an elegant solution to the problem, it is only valid if the underlying process is sufficiently smooth and homogeneous. In the other approach, the dependence and heterogeneity in the data are recognized as a limiting factor, and the only way to obtain meaningful information from the observations is by adding an additional dimension (*i. e.*, the time dimension). In other words, by pooling time series data for a fixed set of cross–sectional units, the asymptotics in the time dimension provide the framework to carry out statistical inference about the spatial dimension. In either case, it is necessary to evaluate whether the complexity of the proposed hypotheses or models is compatible with the information available in the data. Unfortunately, in many situations encountered in applied empirical work this will not be the case. In those instances the stochastic framework for inference will be suspect, and give rise to legitimate concerns about the relevance of a "statistical" approach.

### 3.5. Analytics versus computing power

A final issue that has come to the fore as a result of the recent advances in computer technology is the choice between procedures based on rigorous analytics and those that replace the analytics by numerically intensive computation. The latter have led to the development

of combinatorial methods and resampling schemes in which the stochastic properties are derived from a large number of replications of pseudo–data (*e. g.*, Efron, 1979; Hubert, 1985; Knudsen, 1987). With the advent of large spatial data bases and geographic information systems, the distinction between description, analysis, modeling and simulation has become blurred. The technological possibilities are virtually unbounded, and have opened up new horizons for spatial data analysis. An example of a recent development in this respect is the creation of a so–called "geographical analysis machine" (GAM), as a combination of a GIS, a spatial statistical analysis and expert system that is designed to carry out an automated spatial data analysis (Openshaw *et al.*, 1987). This concept has many attractive features, but in its current form, the GAM is still rudimentary and limited to a specific application. Also, the statistical properties of the results obtained from a sequence of multiple comparisons (as in the GAM) are unclear. As is well known, a naïve impression of "significance" can always be obtained after a large number sequential tests. Consequently, important further developments are needed before this a–theoretical approach will be able to replace (or complement) the more traditional analytic approach for a wide range of spatial data analysis problems.

## 4. Spatial Errors

Basic to both the data–driven and the model–driven analysis of spatial data is an understanding of the stochastic properties of the data. The use of "space" as the organizing framework leads to a number of features that merit special attention, since they are different from what holds for aspatial or time series data. The most important concept in this respect is that of error, or, more precisely for data observed in space, spatial error. The distinguishing characteristics of spatial error have important implications for description, explanation, and prediction in spatial analysis. Some of these issues will be discussed in the next section. In this section, I present a simple taxonomy of the nature of spatial errors, and outline some alternative perspectives on how error can be taken into account.

### 4.1. The nature of spatial errors

Spatial errors can be due to a variety of sources. For spatial data analysis, the most relevant types of error are measurement error and specification error. Measurement error occurs when the location or the value of a variable are observed with imperfect accuracy. The former is an old cartographic problem and is still very relevant in modern geographic information systems (*e. g.*, errors due to a lack of precision in digitizing). The main problem is that the geometric and graphical representation of the location of points, lines or areal boundaries (*i. e.*, a map) gives an imperfect impression of the uncertainty associated with errors in the measurement of these features. Since these locational features are important elements in the evaluation of distance and relative position, and in the operations of areal aggregation and interpolation, the associated measurement error will affect many of the "values" generated in a spatial information system as well. Although similar errors occur in the time dimension, they are much simpler to take into account since they only propagate in one dimension and one direction. Moreover, spatial measurement errors, in contrast to the classical case, will tend not to balance out.

Other spatial errors of measurement have to do with the imperfect way in which data on socio–economic phenomena are recorded and grouped in spatial units of observation (*e. g.*, various types of administrative units). This interdependence of location and value in spatial data leads to distinctively spatial characteristics of the errors. These are the familiar spatial

dependence and spatial heterogeneity. Dependence is mostly due to the existence of spatial spill–overs, as a result of a mis–match between the scale of the spatial unit of observation and the phenomenon of interest (*e. g.*, continuous processes represented as points, or processes extending beyond the boundaries of administrative regions). Heterogeneity is due to structural differences between locations and leads to different error distributions (*e. g.*, differences in accuracy of census counts between low–income and high–income neighborhoods).

Specification error is particular to the model–driven approach in spatial data analysis. It pertains to the use of a wrong model (*e. g.*, recursive versus simultaneous), an inappropriate functional form (*e. g.*, linear as opposed to nonlinear), or a wrong set of variables. In essence, it is no different from misspecification in general, but it generates spatial patterns of error due to the use of spatial data. These spatial aspects can occur as a result of ignoring location–specific phenomena, spatial drift, regional effects or spatial interaction. When a false assumption of homogeneity is forced onto a model in those instances, spatial heterogeneous errors will result. Similarly, when the spatial scale or extent of a process does not correspond to the scale of observation, or when the nature of a process changes with different scales of observation, spatial dependent errors will be generated.

### 4.2. Perspectives on spatial errors

The treatment of spatial errors in data analysis is fundamentally different between the data–driven and the model–driven approaches. In the data–driven approach, errors are considered to provide information. The focus of attention is on how the spatial pattern of the errors relates to data generation processes. For example, in attempts to provide measures of uncertainty for spatial information in a GIS, the spatial pattern of errors is related to data collection and manipulation procedures. The spatial pattern of errors can often provide insight into the form of the underlying substantive spatial process, such as is exploited in the model identification stage of a spatial time series analysis. Important and still unresolved research questions deal with the formulation of useful spatial error distributions, in which error is related to location, distance to reference locations, area, and the such.

In the model–driven approach to spatial data analysis, error is considered to be a nuisance. The main focus is on how to identify the spatial distribution of the error process, and how to eliminate the effect of errors on statistical inference. In other words, once errors are identified, they are eliminated by means of transformations, corrections, or filters. Alternatively, robust estimation and test procedures can be applied that are no longer sensitive to the effect of errors. A major research question in this respect is how diagnostics can be developed that are powerful in detecting various types of errors, and are able to distinguish between them (*e. g.*, to distinguish between spatial dependence and spatial heterogeneity, or "real" versus "apparent" contagion).

## 5. Implications of Spatial Errors for Spatial Data Analysis

The presence of errors with a distinctive spatial pattern has obvious implications for the analysis of spatial data. These implications vary between the analysis of spatial pattern, the estimation and prediction of spatial models, and their validation.

### 5.1. Analysis of spatial pattern

Given the importance of distance and contiguity in the analysis of spatial pattern, errors of measurement in the location of points, lines, and areal units will greatly affect the distributional properties of tests and other indices. This aspect of spatial error is largely ignored in current statistical practice, but merits closer attention, particularly in light of the increased availability of large computerized spatial data bases with the explosion of the GIS field. Some indices of spatial pattern and spatial association that are routinely derived in a GIS (*e. g.*, based on nearest–neighbors) provide a misleading sense of precision, since they ignore the uncertainty associated with the location of spatial units themselves. Conceptually, the solution to this problem is straightforward, in that a spatial distribution needs to be specified for each location (and the associated values). However, the choice of the most appropriate distribution and its effect on the properties of the various spatial statistics are still largely unresolved topics of research.

### 5.2. Estimation and prediction

The effect of spatial errors on the estimation and prediction of standard linear models is probably the best understood aspect of spatial data analysis. In particular, for the linear regression model with normally distributed disturbance terms, many tests and estimators have been developed (see Anselin and Griffith, 1988, for a review). In those models, the error is taken to pertain to the dependent variable only and its effect is incorporated in the regression disturbance term. The more realistic situation where error is present in both dependent and independent variables has received much less attention, and is considerably more complex. The specification of interaction between the various spatial errors is largely unresolved, and so far only a robust estimation approach seems to hold promise.

Most methodological results obtained so far also are limited to the normal distribution case. Spatial effects in models with limited dependent variables, censored and truncated distributions, or in models for count data have been largely ignored. A major problem in this respect is that multivariate dependent distributions other than the normal are highly complex. Moreover, their application in an operational context is often hampered by limitations on carrying out numerical integration in multiple dimensions. Since the non–normal case is probably the rule rather than the exception in actual spatial data, a considerable agenda of research questions remains to be addressed.

### 5.3. Model validation

In model validation, the focus is on assessing the uncertainty associated with the output (interpretation) of alternative specifications. Clearly, this will be a function of the probabilistic model that has been adopted for the underlying (unobserved) spatial errors. A particular problem in spatial data analysis is how to provide a meaningful summary measure of spatial accuracy. If spatial heterogeneity is present, the accuracy is likely to vary systematically by location. On the other hand, if spatial dependence is present, the accuracy at one location will be affected by the accuracy associated with "neighboring" locations. A summary or holistic measure of accuracy will be an imperfect reflection of this partitive (observation by observation) accuracy. What is needed is a meaningful objective function (loss or risk function) that incorporates the relative importance of accuracy for particular locations or regions in space. It is unlikely that such an objective function can be developed with uni-

versal applicability, but instead, a flexible approach can be taken that is consistent with the use of spatial information systems as decision support systems.

## 6. Conclusion

The wide array of philosophical and methodological dilemmas that confront the analysis of spatial data necessitates an eclectic perspective. Many different ways of looking at a data set or at a model specification should be compared, and sensitivity analysis should play a central role. In other words, the extent to which the results are affected by changes in the underlying assumptions (as in fragility analysis) needs to be assessed. If different approaches yield the same qualitative conclusions, one can be more confident that meaningful insights have been gained. On the other hand, if the statistical findings turn out to be very sensitive to the approach taken, there is likely to be something wrong with the data and/or with the model, and not much faith should be put in the precise quantitive results.

The characteristics of errors that affect observations of spatial data clearly motivate the need for a specialized methodology of spatial statistics and spatial econometrics. However, much of the current state–of–the–art in these fields pertains to highly artificial and rather simplistic data structures. A major emphasis of future research should be to focus on realistic perspectives on spatial data. With the vast power of a user–friendly GIS increasingly in the hands of the non–specialist, the danger is great that the "wrong" kind of spatial statistics will become the accepted practice. Since the "easy" problems have more or less been solved, a formidable challenge lies ahead.

## 7. References

Anselin, Luc (1980) *Estimation Methods for Spatial Autoregressive Structures*. Ithaca, NY: Cornell University, Regional Science Dissertation and Monograph Series #8.

Anselin, Luc (1982) A note on small sample properties of estimators in a first–order spatial autoregressive model. *Environment and Planning A*, **14**: 1023-1030.

Anselin, Luc (1984) Specification tests on the structure of interaction in spatial econometric models. *Papers*, Regional Science Association, **54**: 165-182.

Anselin, Luc, (1986a) Non–nested tests on the weight structure in spatial autoregressive models: some Monte Carlo results. *Journal of Regional Science*, **26**: 267-284.

Anselin, Luc (1986b) Some further notes on spatial models and regional science. *Journal of Regional Science*, **26**: 799-802.

Anselin, Luc (1988a) *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer Academic.

Anselin, Luc (1988b) Model validation in spatial econometrics: a review and evaluation of alternative approaches. *International Regional Science Review*, **11**: 279-316.

Anselin, Luc (1989a) Quantitative methods in regional science: perspectives on research directions. Paper Presented at a Plenary Session of the Third World Congress of the Regional Science Association, April 2-7, Jerusalem, Israel.

Anselin, Luc (1989b) Some robust approaches to testing and estimation in spatial econometrics. *Regional Science and Urban Economics*, **19**: (forthcoming).

Anselin, Luc and Daniel A. Griffith (1988) Do spatial effects really matter in regression

analysis? *Papers*, Regional Science Association, **65**: 11-34.

Bennett, Robert (1979) *Spatial Time Series*. London: Pion.

Boots, Barry N. and Arthur Getis (1988) *Point Pattern Analysis*. Newbury Park, CA: Sage Publications.

Box, G. and G. Jenkins (1976) *Time Series Analysis, Forecasting and Control*. San Francisco: Holden Day.

Clark, I. (1979) *Practical Geostatistics*. London: Applied Science Publishers.

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

Cliff, A. and J. K. Ord (1981) *Spatial Processes, Models and Applications*. London: Pion.

Cooley, T. and S. LeRoy (1985) Atheoretical macro–econometrics: a critique. *Journal of Monetary Economics*, **16**: 283-308.

Costanzo, C. Michael (1983) Statistical inference in geography: modern approaches spell better times ahead. *The Professional Geographer*, **35**: 158-165.

Diggle, P. (1983) *Statistical Analysis of Spatial Point Patterns*. New York: Academic Press.

Doan, T., R. Litterman, and C. Sims (1984) Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews*, **3**: 1-100 (with discussion).

Durbin, James (1988) Is a philosophical consensus for statistics attainable? *Journal of Econometrics*, **37**: 51-61.

Efron, B. (1979) Computers and the theory of statistics: thinking the unthinkable. *SIAM Review*, **21**: 460-480.

Efron, B. (1982) *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM).

Efron, B. (1986) Why isn't everyone a Bayesian? *The American Statistician*, **40**: 1-11 (with discussion).

Folmer, Hendrik (1986) *Regional Economic Policy: Measurement of Its Effect*. Dordrecht: Martinus Nijhoff.

Folmer, H. and M. Fischer (1984) Bootstrapping in spatial analysis. Paper Presented at the Symposium of the IGU Working Group on Systems Analysis and Mathematical Models, Besancon, France.

Foster, S. and W. Gorr (1986) An adaptive filter for estimating spatially–varying parameters: application to modeling police hours spent in response to calls for service. *Management Science*, **32**: 878-889.

Getis, Arthur (1988) Second–order theory in spatial analysis. Paper Presented at the Symposium of the IGU Working Group on Mathematical Models, August 16-19, Canberra, Australia.

Getis, Arthur and Barry Boots (1978) *Models of Spatial Processes*. London: Cambridge University Press.

Goodchild, Michael (1987a) *Spatial Autocorrelation*. CATMOG.

Goodchild, Michael (1987b) A spatial analytical perspective on geographical information systems. *International Journal of Geographical Information Systems*, **1**: 327-334.

Gould, Peter (1981) Letting the data speak for themselves. *Annals of the Association of*

*American Geographers*, **71**: 166-176.

Griffith, Daniel A. (1983) The boundary value problem in spatial statistical analysis. *Journal of Regional Science*, **23**: 377-387.

Griffith, Daniel A. (1985) An evaluation of correction techniques for boundary effects in spatial statistical analysis: contemporary methods. *Geographical Analysis*, **17**: 81-88.

Griffith, Daniel A. (1987a) *Spatial Autocorrelation: A Primer*. Washington, D.C.: Association of American Geographers.

Griffith, Daniel A. (1987b) Toward a theory of spatial statistics: another step forward. *Geographical Analysis*, **19**: 69-82.

Griffith, Daniel A. (1988) *Advanced spatial statistics*. Dordrecht: Kluwer Academic.

Griffith, Daniel A. and Carl G. Amrhein (1982) An evaluation of correction techniques for boundary effects in spatial statistical analysis: contemporary methods. *Geographical Analysis*, **17**: 81-88.

Haining, Robert (1978) Estimating spatial interaction models. *Environment and Planning A*, **10**: 305-320.

Haining, Robert (1984) Testing a spatial interacting market hypothesis. *The Review of Economics and Statistics*, **66**: 576-583.

Haining, Robert (1986) Spatial models and regional science: a comment on Anselin's paper and research directions. *Journal of Regional Science*, **26**: 793-798.

Hendry, David F. (1980) Econometrics — alchemy or science? *Economica*, **47**: 387-406.

Hepple, L. (1979) Bayesian analysis of the linear model with spatial dependence. In C. Bartels and R. Ketellapper, *Exploratory and Explanatory Statistical Analysis of Spatial Data*, pp. 179-199. Boston: Martinus Nijhoff.

Hooper, P. and G. J. D. Hewings (1981) Some properties of space–time processes. *Geographical Analysis*, **13**: 203-223.

Huber, P. (1981) *Robust Statistics*. New York: Wiley.

Hubert, L. (1985) Combinatorial data analysis: association and partial association. *Psychometrika*, **50**: 449-467.

Hubert, L., R. Golledge and C. Costanzo (1981) Generalized procedures for evaluating spatial autocorrelation. *Geographical Analysis*, **13**: 224- 233.

Hubert, L., R. Golledge, C. Costanzo and N. Gale (1985) Measuring association between spatially defined variables: an alternative procedure. *Geographical Analysis*, **17**: 36-46.

Kloek, Teun and Yoel Haitovsky (1988) Competing statistical paradigms in econometrics. Special Issue, *Journal of Econometrics*, **37**: (1).

Knudsen, Daniel (1987) Computer–intensive significance–testing procedures. *The Professional Geographer*, **39**: 208-214.

Koenker, R. (1982) Robust methods in econometrics. *Econometric Reviews*, **1,**: 213-255.

Leamer, Edward (1978) *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: Wiley.

Lovell, M. C. (1983) Data mining. *The Review of Economics and Statistics*, **65**: 1-12.

March, L. and M. Batty (1975) Generalized measures of information, Bayes' likelihood ratio

and Jaynes' formalism. *Environment and Planning B*, **2**: 99-105.

Mosteller, F. and J. W. Tukey (1977) *Data Analysis and Regression*. Reading, Mass: Addison-Wesley.

Nijkamp, P., H. Leitner and N. Wrigley (1985) *Measuring the Unmeasurable*. Dordrecht: Martinus Nijhoff.

Odland, John (1978) Prior information in spatial analysis. *Environment and Planning A*, **10**: 51-70.

Odland, John (1988) *Spatial Autocorrelation*. Newbury Park, CA: Sage Publications.

Odland, John, Reginald G. Golledge, and Peter Rogerson (1989) Recent developments in mathematical and statistical analysis in human geography. In G. Gaile and C. Wilmott, *Geography in America*, pp. 719- 745. Columbus, OH: Merrill.

Openshaw, Stan (1987) An automated geographical analysis system. *Environment and Planning A*, **19**: 431-436.

Openshaw, Stan and Peter Taylor (1979) A million or so correlation coefficients: three experiments on the modifiable areal unit problem. In N. Wrigley, *Statistical Applications in the Spatial Sciences*, pp. 127-144. London: Pion.

Openshaw, Stan and Peter Taylor (1981) The modifiable areal unit problem. In N. Wrigley and R. Bennett, *Quantitative Geography, a British View*, pp. 60-69. London: Routledge and Kegan Paul.

Openshaw, S., M. Charlton, C. Wymer, and A. Craft (1987) A Mark 1 Geographical Analysis Machine for the automated analysis of point data sets. *International Journal of Geographical Information Systems*, **1**: 335-358.

Pagan, Adrian (1987) Three econometric methodologies: a critical appraisal. *Journal of Economic Surveys*, **1**: 2-24.

Richards, John A. (1986) *Remote Sensing Digital Image Analysis: An Introduction*. New York: Springer Verlag.

Robinson, P. (1988) Semiparametric econometrics: a survey. *Journal of Applied Econometrics*, **3**: 35-51.

Sen, A. and S. Soot (1977) Rank test for spatial correlation. *Environment and Planning A*, **9**: 897-903.

Sims, Christopher (1980) Macroeconomics and reality. *Econometrica*, **48**: 1-48.

Sims, Christopher (1982) Scientific standards in econometric modeling. In H. Hazewinkel and A.N.G. Rinnooy Kan, *Current Developments in the Interface: Economics, Econometrics, Mathematics*, pp. 317-337. Dordrecht: D. Reidel Publishing.

Stetzer, F. (1982a) Specifying weights in spatial forecasting models: the results of some experiments. *Environment and Planning A*, **14**: 571-584.

Stetzer, F. (1982b) The analysis of spatial parameter variation with jackknifed parameters. *Journal of Regional Science*, **22**: 177-188.

Summerfield, M. (1983) Populations, samples and statistical inference in geography. *The Professional Geographer*, **35**: 143-149.

Swamy, P., P. R. Conway, and P. von zur Muehlen (1985) The foundations of econometrics—are there any? *Econometric Reviews*, **4**: 1-61.

Taylor, W. E. (1983) On the relevance of finite sample distribution theory. *Econometric Reviews*, **1**: 213-255.

Tobler, Waldo (1979) Cellular geography. In S. Gale and G. Olsson, *Philosophy in Geography*, pp. 379-386. Dordrecht: Reidel.

Tukey, J. W. (1977) *Exploratory Data Analysis*. Reading: Addison–Wesley.

Wartenberg, Daniel (1985) Multivariate spatial correlation: a method for exploratory geographical analysis. *Geographical Analysis*, **17**: 263-283.

Zellner, Arnold (1985) Bayesian econometrics. *Econometrica*, **53**: 253-269.

Zellner, Arnold (1988) Bayesian analysis in econometrics. *Journal of Econometrics*, **37**: 27-50.

## DISCUSSION

### "What is special about spatial data? alternative perspectives on spatial data analysis"

### by Luc Anselin

Anselin has given an interesting review of issues underlying the analysis of spatial data. I will comment on three areas of his discussion, namely model–driven versus data–driven approaches to data analysis, problems raised by data accuracy, and the role of robust analyses in spatial data analysis.

Data analysis involves the stages of model specification, parameter estimation and model validation. The dangers inherent in the model–driven approach are first that data properties play a reduced role in the first and last stages, and second, as a consequence, there is a tendency to confirm or re–enforce existing theoretical 'prejudices'. In the case of the data–driven approach, analysis tends to emphasize current experience (the data) and disregard the results of previous analyses, and as a consequence there is a risk of reporting results that are mere artifacts of a particular data set. Data analysis in the social sciences tries to strike a balance between these two approaches, which lie at two ends of a continuum. The case against a purely model–driven approach in the social sciences is the dearth of good theory, while the case against a purely data–driven approach are first the conditions under which much data are collected (non–experimental, complex interactions), and second the accuracy of much social science data. These concerns make model specification, purely on the basis of data properties, an uncertain exercise.

Data accuracy is a major concern in all areas of inductive science. The rising tide of spatially referenced data (collected through both government and commercial agencies) offers both opportunities and pitfalls. The accuracy of these data, in terms of both spatial referencing and the reported values, should be a matter of concern. Although methods are being developed to reduce the influence of unusual or suspect values, in later stages of analysis this is hardly a substitute for the specification of minimal criteria for the procedure of data collection and the careful screening of data prior to and during computerized data storage. The question as to whether it is worthwhile analyzing large data sets, particularly those that may not satisfy such minimum criteria, is an important one. The increase in data allows us to be more discriminating in what we analyze, and should not necessarily lead to more analysis. There is a danger that the more data we have, the less we will know.

The following are two classes of problems that confront the data analyst: those that arise when statistical assumptions are not satisfied, and those that arise from the nature of the data. Robust and resistant estimation methods have been developed that provide improved estimates where the data follow some skewed distribution, or the data contain outliers or extreme values. Most of these estimation methods assume that observations are independent. Robust and resistant estimation methods are required for situations where observations are not independent in order to provide estimates of the parameters of spatial models (where there may be several different parameter sets associated with different aspects of the model). Not only the presence of extreme values but also their spatial distribution may affect parameter estimation when standard 'non–robust' methods are used on such spatial models.

Lastly, in addressing the issues raised in Anselin's paper and identifying directions for future research, it is important not to lose sight of the reason for developing these methods within any subject field such as geography or regional science. The importance of any methodological area of research ultimately depends upon the extent to which it better enables users to tackle substantive questions. The concern of statisticians is largely with the development of statistical theory and the methodology of data analysis. The concern of the applied scientist is with the development of the theory and methods in relation to important substantive issues within the specific field of study.

*Robert P. Haining*, University of Sheffield

## *PREAMBLE*

*What we call progress is the exchange of one Nuisance for another Nuisance.*

Havelock Ellis

*A brief compendium of spatial regression model types is presented by Haining. In his discussion of this topic, "firmer" models are exchanged for "soft" models, "relative" mathematical space is exchanged for "absolute" space, and geo–referenced data complications are exchanged for traditional independent data complications. The purpose of this paper is to outline a class of models that seems to successfully handle redundant information contained in data arising from the locational positions of observations. What new idiosyncrasies, nuances, or subtleties of data will become problematic during analysis with the utilization of these alternate perspectives? A transcending issue beginning to emerge through the confusion surrounding development of spatial regression models appears to be even more fundamental than its old counterpart issue of multicollinearity that troubles traditional regression models. Dorean echoes many of these same sentiments, while raising questions concerning spatial regression model implementation.*

The Editor

# Models in Human Geography: Problems in Specifying, Estimating, and Validating Models for Spatial Data

Robert P. Haining*

*Department of Geography, The University of Sheffield, Sheffield S10 2TN, England.*

**Overview:** The use of the linear regression model for analysing relationships between a set of predictor variables and a response variable when the data refer to areal units, raises a number of distinctive issues. These issues include: specification of the regression model to allow for possible "spillover" effects; how to get good estimates of the spatial parameters. Further, there may be a separate set of concerns that derive from the nature of the data and the spatial distribution of certain types of values. The paper examines these problems and discusses ways of handling them. We conclude with two short examples.

## 1. Introduction

This paper is concerned with the problem of accounting for variation in some attribute (a response or dependent variable), in terms of a set of other attributes (explanatory, predictor or independent variables) where measurements are taken at "locations" (point sites distributed across a map or areas that partition a map).

For all its many perceived conceptual shortcomings as a model for variable relationships, the regression model is still widely used to treat questions of this type. Denoting the response variable as $Y$ and the predictor variables as $X_1, \ldots, X_k$ the model is specified by a linear equation of the form

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \xi \tag{1.1}$$

where the $\beta$s are unknown parameters and the $\xi$s are statistical errors (or disturbances). The data to fit equation (1.1) form an $n$–by–$(k+1)$ array, where $n$ is the number of cases and $(y_i, x_{1,i}, \ldots, x_{k,i})$ is the vector of observations for case i. Given these data, equation (1.1) may be re–written as

$$y_i = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_k x_{k,i} + \xi_i \quad (i = 1, \ldots, n), \tag{1.2}$$

where the $\xi$s are assumed to be normally and independently distributed with mean zero and constant variance $\sigma^2$ $[\xi_i \sim NID(0, \sigma^2)]$.

If these assumptions are satisfied, least squares provides the best linear unbiased estimator for the unknown parameters. Given the data, let $\hat{\boldsymbol{\beta}}$ denote the least squares estimate for $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)^T$, where the superscript $T$ denotes the matrix operation of transpose. Then

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Y}), \tag{1.3}$$

* Part of the work for this paper was carried out while the author was in receipt of a Nuffield Social Science Research Fellowship.

where $\mathbf{X}$ is the $n$–by–$(k+1)$ matrix containing data on the predictor variables, with the $i^th$ row given by $(1, x_{1,i}, \ldots, x_{k,i})$, and $\mathbf{Y}^T = (y_1, \ldots, y_n)$. Further, letting $\hat{\boldsymbol{\xi}} = \mathbf{e}$

$$\hat{\sigma}^2 = \mathbf{e}^T\mathbf{e}/(n-k-1), \tag{1.4}$$

where $\mathbf{e} = (e_1, \ldots, e_n)^T$ is the vector of least squares residuals, with $e_i = y_i - (\hat{\beta}_0 + \cdots + \hat{\beta}_k x_{k,i})$.

The variation accounted for by the linear combination of predictor variables is referred to as the explained variation, and the unallocated portion (associated with the residuals) is the unexplained variation, in variate $Y$. The steps that are followed in fitting a model such as equation (1) are:

(a) identification of a model (selection of predictor variables and specification of relationships),

(b) estimation of the unknown parameters,

(c) assessment of goodness–of–fit (residual analysis), and

(d) perhaps modification of the current model (new assumptions, data transformation), followed by a return to step (b).

This procedure cycles until a satisfactory model emerges, meaning a model where the percentage of explained variation is as high as possible and the residuals are well behaved in the sense of satisfying model assumptions. The initial model [at Step (a)] is often referred to as a "soft" model, which is made "firmer" by the cycles of fitting and model assessment. In addition to ensuring that the least squares assumptions are met (and applying remedial action if they are not), further problems may arise depending upon the nature of the data.

In this paper we examine the use of the regression model for describing relationships between variables measured across a set of locations on a map. In Section 2 we consider generalizations of equation (1.1) that reflect the spatial ordering of the data. Then in Section 3 we examine the implications for least squares parameter estimation. In Section 4 attributes of the spatial system that influence regression analysis are described. In particular we discuss how boundaries should be handled, the treatment of order relations between the set of locations and the influence of the surface partitioning. We also consider problems that arise when trying to assess the influence of individual cases and extreme values (outliers) on model fit. The last section briefly discusses two data sets in order to exemplify some of the issues raised in this paper.

## 2. Spatial Regression Models

The regression model defined by equations (1.1) and (1.2) disregards the geographic location of the $n$ cases. Each case is treated as a distinct event. In specifying the regression model, the value of the response variable at any location is assumed only to be a function of values of the predictor variables at that *same* location (this accords with what sometimes is referred to as an "absolute" or "container" conceptualization of space). The location of each case only plays a significant role at the assessment stage of analysis. The errors in the regression model (1.2) are required, by assumption, to be independent. One aspect of model evaluation, therefore, consists of checking the residuals for evidence of pattern (spatial autocorrelation). Within an "absolutist" representation of space the presence of residual pattern is taken to

imply that important variables have been omitted, or that the functional relationship has been misspecified. In the former case new variables are sought that will eliminate the residual autocorrelation while in the latter case data transformations are used.

But space is not a series of separate, disconnected (independent) "boxes" or "containers," and the influence of events need not be restricted to the locations where they occur. The level of a response variable at a location may reflect the levels of predictor variables at other locations, and indeed a response variable at one location may act as a predictor variable for another location. Such considerations reflect the fact that events in space are not "parcelled-up." If these influences are present, then they may need to be taken into consideration when specifying a regression model. Several situations are presented next where such issues arise.

### 2.1. The spatially lagged response variable model.

In a lagged response variable model, the response variable at location $j$ may act as a predictor variable at other locations. For example,

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_k X_{k,i} + \rho \sum_{\substack{j=1 \\ j \neq i}}^{n} w_{ij} Y_j + \xi_i, \tag{2.1}$$

where $\rho$ is an unknown parameter and the set $\{w_{ij}\}$ denotes a prior weighting scheme that may reflect, for instance, the distance between locations $i$ and $j$. Often the influence of $Y_j$ on $Y_i$ is assumed to decrease as this distance increases.

Consider a set of retail sites scattered across a large urban area, with each site selling a more or less identical product. In such cases the market of each seller may be closely linked to neighboring markets, with the degree of interdependence lessening rather quickly with distance until this dependency becomes zero. The result of such local competitive interactions is a network of intricately interwoven markets—a "chain linking" (Chamberlin, 1956, p. 103) that is likely to be strongly influenced by the underlying movements of consumers within the city. If the retailers are gasoline retailers strung out along a road, for example, and if one retailer reduces his price, then it is probable that the nearest competitor also will drop his price. Such considerations lead to the specification of price models of the form given by equation (2.1), where $X_1, \ldots, X_k$ measure site effects (such as site quality, range of automotive services, brand type) and the set of weights $\{w_{ij}\}$ describes the structure of inter–site competition (Haining, 1986). The price charged at site $i$ (namely $Y_i$) is in part dependent on prices charged in surrounding local retail sites, since the level of demand at site $i$ is not only dependent on prices at $i$, but also on the level of prices at site $i$ relative to prices at other sites with which $i$ competes.

As an additional example, consider the problem of modeling variation in the total income accruing to the residents of a number of towns and cities distributed over a region. Let $\mathbf{Y}$ denote the vector of total income for the residents of the $n$ places. Then

$$\mathbf{Y} = \mathbf{X} + \mathbf{C},$$

where $\mathbf{X}$ denotes the vector of exogenous income (deriving from export earnings, investment, government outlays), and $\mathbf{C}$ denotes the vector of endogenous income (local consumption by community residents). Assuming

$$\mathbf{C} = c\mathbf{Y}, \tag{2.2}$$

where the scalar $c$ is the income creating local propensity to consume, then

$$\mathbf{Y} = (1 - c)^{-1}\mathbf{X}.$$

The vector $\mathbf{X}$ can be further decomposed into income earned from long distance (extra-regional) income transfers $(\mathbf{X}_1)$, and income earned from short distance (intra-regional) transfers $\mathbf{X}_2$). Thus

$$\mathbf{Y} = (1 - c)^{-1}(\mathbf{X}_1 + \mathbf{X}_2).$$

The vector $\mathbf{X}_2$ includes income accruing to each community arising from consumption expenditures by non-residents. Haining (1987) suggests that if equation (2.2) is reasonable in terms of intra-community consumption expenditure, then

$$\mathbf{X}_2 = \mathbf{\Omega Y},$$

where $\mathbf{\Omega}$ is an $n$-by-$n$ matrix with diagonal values equal to zero, and non-negative off-diagonal entries that reflect the structure of inter-community movements for the purposes of purchasing consumer goods. Given a hierarchical ordering of the urban places, the non-zero elements of matrix $\mathbf{\Omega}$ can be specified using central place arguments. Accordingly the model becomes

$$\mathbf{Y} = (1 - c)^{-1}(\mathbf{X}_1 + \mathbf{\Omega Y}),$$

where again the response variable at location $i$ may appear as a predictor variable at other locations. Unlike the price model, where interactions are reciprocal ($Y_i$ is a function of $Y_j$ **and** vice versa, so that matrix $\mathbf{W} = \{w_{ij}\}$ contains non-zero values above and below the main diagonal), central place principles suggest that interaction will be directional (consumers in low-order centers spending in high-order centers, but not vice versa), so that matrix $\mathbf{\Omega}$ will be an upper- (or lower-) triangular matrix.

### 2.2. The spatially lagged predictor variable model.

The simplest form of this model may be stated as

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_k X_{k,i} + \tau \sum_{j=1}^{n} w_{ij} X_{r,j} + \xi_i, \tag{2.3}$$

where the set of weights $\{w_{ij}\}$ are as before, and $\tau$ is an unknown parameter. Variable $X_r$ is usually a member of the set $\{X_1, \ldots, X_k\}$.

Models of this type have arisen in the study of the housing market, and in particular the modeling of spatial variation in house prices. House price is a function of structural characteristics of the house, the location of the house with respect to the city center, and characteristics of the area in which the house is situated (including environmental and demographic characteristics). Furthermore, depending upon the scale of the areal units, the characteristics of neighboring areas also may be significant. Hence, a house located in a desirable residential area that is adjoined by other desirable residential areas will tend to have a higher price than an equivalent house located in an equally desirable residential area, but where some of the adjoining residential areas are of lower status. Anas and Eum (1984, p. 105) remark that "the spillovers among neighboring and otherwise substitutable sub-markets can be taken into account to specify models in which market information from other submarkets becomes capitalized into housing prices." The estimate of each coefficient implicitly measures the price of that attribute.

### 2.3. The spatially correlated error model.

Equation (1.1) is modified here, yielding

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_k X_{k,i} + \nu_i \tag{2.4}$$

where now $\nu_i$ are the statistical errors, with non–zero covariances $E(\nu_i, \nu_j) \neq 0$ for some $i$ and $j$ $(i \neq j)$. Therefore $E[\boldsymbol{\nu}\boldsymbol{\nu}^T] = \sigma^2\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is a matrix having some non–zero entries in its upper– and lower–triangles. Traditionally, in geography, matrix $\boldsymbol{\Sigma}$ has been modelled as a first–order simultaneous autoregressive scheme, such that

$$\boldsymbol{\Sigma} = [(\mathbf{I} - \rho\mathbf{W})^T(\mathbf{I} - \rho\mathbf{W})]^{-1},$$

where $\mathbf{I}$ is the $n$–by–$n$ identity matrix, $\mathbf{W}$ is an $n$–by–$n$ matrix of given weights $\{w_{ij}\}$ reflecting order relations on the map, and $\rho$ is the (unknown) autoregressive parameter. However, many other models could be used.

In experimental situations (*e. g.*, agricultural uniformity trials), where the set of predictor variables and their levels are determined by the experiment, a correlated errors model is a natural choice when residuals are found to be correlated. In this context attention focuses on the choice of error model. In non–experimental situations the justification for this model is less clear cut, since the set of relevant predictor variables is not defined. In such cases it is usual to consider fitting a model such as equation (2.4), if the residuals are found to be correlated and if

(a) no further variables can be identified, or

(b) data are not available on other variables that might be significant, or

(c) adding further variables to the model does not remove this property of the residuals.

Residual correlation may be present because of the influence of a large number of variables that are difficult to specify, but that together display spatial persistence or continuity. Omission of variables representing these influences is responsible for the correlation detected in the residuals. Rather than attempt to model such influences, which might prove very difficult, a model such as equation (2.4) enables "safer" inference to be made with respect to those variables that can be included in the model. If such effects display smooth variation (such as trend), then replacing the error model by some order of polynomial trend surface may be preferable.

Loftin and Ward (1983) use a correlated errors model in examining the effects of population density on fertility rates in areas of Chicago. Such a modification apparently enables better estimates and safer inferences to be made on the influence of the included predictor variable. Further details are discussed in, for example, Cliff and Ord (1981), and Upton and Fingleton (1985).

## 3. Parameter Estimation

Estimation procedures for each of the three models presented in Section 2 will be discussed now, in turn.

### 3.1. The spatially lagged response variable model.

Using matrix notation, equation (2.1) may be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \rho\mathbf{W}\mathbf{Y} + \boldsymbol{\xi},$$

where matrix $\mathbf{W} = \{w_{ij}\}$. By re–arranging terms this expression becomes

$$(\mathbf{I} - \rho\mathbf{W})\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi}, \tag{3.1}$$

and letting matrix $\mathbf{A} = (\mathbf{I} - \rho\mathbf{W})$ and matrix $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, substitutions into equations (1.3) and (1.4) yield

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\hat{\mathbf{A}}\mathbf{Y}), \tag{3.2}$$

$$\hat{\sigma}^2 = \mathbf{Y}^T\hat{\mathbf{A}}^T\mathbf{M}\hat{\mathbf{A}}\mathbf{Y}/(n-k-2), \tag{3.3}$$

where matrix $\hat{\mathbf{A}}$ denotes that the parameter $\rho$ has been estimated (the problem of estimating $\rho$ will be considered later), so that another degree of freedom is lost.

### 3.2. The spatially lagged predictor variable model.

Using matrix notation, equation (2.3) becomes

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \tau\mathbf{W}\mathbf{X}_r + \boldsymbol{\xi},$$

where vector $\mathbf{X}_r$ denotes one of the columns of matrix $\mathbf{X}$ (but not the first column). Again by re–arranging terms this equation may be expressed as

$$[\hat{\boldsymbol{\beta}}^T : \hat{\tau}]^T = (\mathbf{Z}^T\mathbf{Z})^{-1}(\mathbf{Z}^T\mathbf{Y}),$$

where the vertical dots symbol, :, denotes matrix partitioning, matrix $\mathbf{Z} = [\mathbf{X} : \mathbf{W}\mathbf{X}_r]$, and

$$\sigma^2 = \mathbf{e}^T\mathbf{e}/(n-k-2),$$

where $\mathbf{e} = \mathbf{Y} - (\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\tau}\mathbf{W}\mathbf{X}_r)$.

The regression parameters $\boldsymbol{\beta}$ and $\tau$ are estimated simultaneously. If variable $\mathbf{X}_r$ is spatially correlated, then $(\mathbf{X}^T\mathbf{X})^{-1}$ may be unstable (since in extreme cases vectors $\mathbf{X}_r$ and $\mathbf{W}\mathbf{X}_r$ may be linearly dependent, causing matrix $\mathbf{X}^T\mathbf{X}$ to be singular). This numerical problem produces inflated estimates of the parameter estimator variances, giving misleading or erroneous inferences.

### 3.3. The spatially correlated error model.

As noted in Section 2, a commonly used regression model with correlated errors is given by the pair of equations

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\nu} \\ \nu &= \rho\mathbf{W}\nu + \boldsymbol{\xi} \end{aligned} \tag{3.4}$$

so that $\boldsymbol{\nu} \sim MVN(0, \sigma^2\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = (\mathbf{A}^T\mathbf{A})^{-1}$. By substituting the second of these equations into the first one, and algebraically manipulating the result,

$$(\mathbf{I} - \rho\mathbf{W})\mathbf{Y} = (\mathbf{I} - \rho\mathbf{W})\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi} \tag{3.5}$$

Then by substitution into equations (1.3) and (1.4),

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-1}(\mathbf{X}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Y}), \tag{3.6}$$

$$\hat{\sigma}^2 = \mathbf{u}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{u}/(n-k-2), \tag{3.7}$$

where vector $\mathbf{u} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}}^{-1} = (\hat{\mathbf{A}}^T\hat{\mathbf{A}}) = (\mathbf{I} - \hat{\rho}\mathbf{W})^T(\mathbf{I} - \hat{\rho}\mathbf{W})$.

Intuitively speaking, it appears that the general effect of including $\hat{\boldsymbol{\Sigma}}^{-1}$ in the estimate of $\hat{\boldsymbol{\beta}}$ is to downweight those observations with high spatial correlation. The presence of spatial correlation means that the information content of an observation is partially duplicated by those other observations (usually nearby) with which it is strongly correlated. Therefore a natural approach to this problem is to reduce the influence of such data duplication in the model fit. Those observations that are to be most strongly downweighted depend upon how matrix $\mathbf{W}$ is specified; often (because of the way matrix $\mathbf{W}$ is specified in applications) they tend to be observations associated with the highly connected interior sites of a spatial partition, particularly if these interior areas also are small relative to areal units closer to the boundary of the region.

### 3.4. Properties of the spatial parameter.

In the case of the lagged response and correlated error models, there is the additional spatial parameter, $\rho$, to be estimated. Estimation of this parameter could be avoided by evaluating the regression equation for different values of $\rho$ contained in the permissible range which is, in fact, very restricted since $1/\eta_{\min} < \rho < 1/\eta\max$, where $\eta\max$ and $\eta_{\min}$ respectively are the largest and smallest eigenvalues of matrix $\mathbf{W}$. This identifies the sensitivity of $\hat{\boldsymbol{\beta}}$ to values of $\hat{\rho}$, and it is usually $\hat{\boldsymbol{\beta}}$ we are most interested in. As an extension of this strategy, a grid search can be conducted in order to find the minimum residual sum of squares for either equation (3.1) or equation (3.5).

Estimation of $\rho$ in the case of the lagged response variable model is described in, amongst other sources, Upton and Fingleton (1985). The maximum likelihood estimate of $\rho$ is obtained by minimizing (with respect to $\rho$)

$$(n/2) \star LN(\hat{\sigma}^2|\mathbf{A}|^{-2/n}), \tag{3.8}$$

where $\hat{\sigma}^2$ is given by equation (3.3), replacing the degrees of freedom value $n-k-2$ by the sample size $n$, and the pair of parallel lines, $|\bullet|$, denotes the determinant of a matrix. One

should note that, by substituting equation (3.3) into equation (3.8), the estimate of $\rho$ does not depend upon $\boldsymbol{\beta}$, so that once $\hat{\rho}$ is obtained $\hat{\boldsymbol{\beta}}$ can be estimated. The estimation of $\rho$ in the spatially correlated error model (with autoregressive errors) is more complicated. Again expression (3.8) is minimized, but $\hat{\sigma}^2$ is now given by equation (3.7), replacing $n-k-2$ by $n$, which depends upon $\hat{\boldsymbol{\beta}}$. A recommended estimation procedure here is to start with an initial estimate of $\boldsymbol{\beta}$ [*e. g.*, equation (1.3)], estimate from expression (3.8), then evaluate equation (3.6), and iteratively repeat these last two steps until convergence occurs. Mardia and Marshall (1984) discuss other possibilities, including computation of the standard error of $\hat{\rho}$. It is usually necessary to write a separate routine to estimate $\rho$, which includes evaluation of a matrix determinant. These problems currently limit the size of data sets that can be handled easily.

## 4. Regression Problems Associated with Spatial Data

In this section we examine a variety of problems encountered in the fitting of regression models that arise in one form or another from the spatial context of the data. The problems raised here fall into three generic groups, namely

(a) problems associated with representing the spatial distribution of observations—these usually call for modeling assumptions that cannot be directly tested;

(b) problems associated with the surface partitioning (in the case where observations are areal aggregates or densities) and which give rise to problems for least squares estimation; and,

(c) data problems that arise either as a consequence of the areal units system or independently of it.

### 4.1. The spatial distribution of observations.

The fitting of each of the models described in Sections 2 and 3 require an explicit, and largely *a priori*, representation of the areal units system to which the observations refer. This representation has two aspects to it: first defining order relationships between the sites or areas; and, second treating the boundary of the study area. For the most part any representation constitutes a set of modeling assumptions that are largely untestable (for instance, there is rarely sufficient information to estimate the elements of matrix $\mathbf{W}$). Since they are untestable, the sensitivity of results to different assumptions should be examined both as part of parameter estimation, and as assessment of fit (*e. g.*, inspecting whether or not the residuals are better behaved under one set of assumptions than under another).

#### *4.1.1. Order relationships.*

Order relationships across the map are specified by the matrix $\mathbf{W}$. This specification involves two separate decisions: (i) which sites or areas should be considered joined, and (ii) what weights should be attached to the joins. The first decision identifies which entries in matrix $\mathbf{W}$ are non–zero, whereas the second decision enables a particular value to be attached to each element $w_{ij}$. These issues are discussed at length in Cliff and Ord (1981). Table 1(a) identifies some of the main criteria used to select a join structure. The use of proximity criteria seems most appropriate where inter–site connections are not limited to special transport networks, whereas the use of interaction criteria seems most appropriate

where such a network does exist. Some level of flow existing between all pairs of sites or areas may necessitate introducing a cut off level, or alternatively an analysis of the system of flows, in order to identify the key linkages in the system (Holmes and Haggett, 1977). In addition, order relations might be hierarchical and directional [as on a central place lattice (Haining, 1987)] or discontinuous (for example, neighbors might be areas with stations on a railway track, if the analyst is looking at the spread of a rumor or an infectious disease). Table 1(b) identifies possible weighting schemes. Again one should note that these weighting schemes can be standardized by setting row sums to 1. This aspect of model specification clearly involves many *ad hoc* decisions, which implies that a need exists for assessing the sensitivity of results to *plausible* alternative definitions. The specification of matrix $\mathbf{W}$ has a direct effect on the fit of the model, since it enters into the estimation of $\boldsymbol{\beta}$ and $\sigma^2$ in all the models outlined in Sections 2 and 3, and in the case of the spatially correlated errors model it influences the relative downweighting of observations.

*4.1.2. Boundary effects.*

Boundary effects are likely to be more serious for the analysis of map data than for the analysis of time series data. In the time series case border effects are of order $1/n$, where $n$ is the length of the observed series. In the case of an $n = N \star M$ rectangular lattice there are usually at least $2N + 2M - 4$ border sites, although the number of border sites depends on the specification of order relationships on the map.

Suppose the study area is not naturally bounded (for example it is a subarea of a much larger region). Here the analyst must consider how to model boundary effects. Consider the problem of modeling county death rates in Pennsylvania due to the spread of an infectious disease. The border counties of Pennsylvania will be influenced by death rates in immediately adjacent counties in New York, Ohio, West Virginia, Maryland, Delaware, and New Jersey. If these influences are simply ignored, the overlooked effects may distort the fit of the model. Border county residuals may be inflated (relative to non–border county residuals) because these external influences will be felt most strongly close to the border of the study area.

In the case of a model such as equation (2.1), if death rates are available in non Pennsylvanian counties, such values at the boundary could be included as additional exogenous predictor variables. Where such boundary information is not available, options include shrinking the study area (which would seem wasteful of data), or assuming values for the (non–observed) boundary counties. Fixed values could be assigned that preserved gradients at the boundary in some sense. Any selection would tend to be arbitrary, and the sensitivity of results to the choices made probably should be assessed.

Regardless of how boundary values are treated, the problem remains of estimating the spatial parameters in models such as equations (2.1) and (2.4). The issues are far from straightforward [see Ord (1981), Künsch (1983), Martin (1987), Griffith (1988)]. However, if the primary interest is in estimating the regression parameter vector $\boldsymbol{\beta}$ rather than the spatial parameter $(\rho)$, then these problems may be rather less serious than they appear from the arguments put forward in the literature cited above. If some adjustment is thought necessary, since residuals might be larger for those cases in the study region close to the boundary, might not robust or resistant regression be considered? Alternatively, an *a priori* weighting might be considered in which observations close to the boundary are downweighted in the fit of the regression model. Different levels of downweighting could be tried and the

TABLE 1
SPECIFICATION OF **W**: DEFINING JOINS AND WEIGHTS

**(a) Joins**

*Proximity*:

i. Distance: each site/area is linked to all other sites/areas within a specified distance.

ii. Nearest neighbors: each site is linked to its $k$ ($k = 1, 2, 3, \ldots$) nearest neighbor(s).

iii. Gabriel graphs: "any two sites $A$ and $B$ are said to be contiguous if and only if all other sites are outside the $A - B$ circle, that is the circle on whose circumference $A$ and $B$ are on opposite points" (Matula and Sokal, 1980).

iv. Delaunay triangulation: all sites that share a common border in a Dirichlet partitioning of the area are joined. Where the sites refer to areas that already partition the map, then the joins may be based upon whether the areas have a boundary in common.

*Interaction*:

All sites/areas between which there is a flow (measured, for example, by traffic movement, telephone calls, or person to person contact).

**(b) Weights**

*Binary*:

$w_{ij} = 1$ if areas $i$ and $j$ are joined; $w_{ij} = 0$ otherwise

*Inverse distance*:

$w_{ij} = d_{ij}^{-\gamma}$ $(\gamma > 0)$, where $d_{ij}$ is the distance separating areas $i$ to $j$

*Exponential*:

$w_{ij} = \exp(-d_{ij}^{\gamma})$

*Boundary length*:

$w_{ij} = (l_{ij}/l_i)^{\tau}$, where $l_{ij}$ is the length of the common boundary between areas $i$ and $j$, $l_i$ is the perimeter of the border of area $i$, and $\tau$ is a constant.

*Boundary and distance*:

$w_{ij} = (l_{ij}/l_i)^{\tau} d_{ij}^{\gamma}$

results compared. As posed here, however, perhaps the most fundamental question concerns the sensitivity of $\hat{\boldsymbol{\beta}}$ to $\hat{\rho}$, and the influence of boundary assumptions on the estimation of $\rho$.

### 4.2. The surface partitioning.

Spatial data often refer to aggregate or density attributes of subareas into which the study area has been partitioned. The results of fitting a regression model to such data will be sensitive to the particular partition involved. Partitions that lump together individual micro–level units (*e. g.*, households), which are alike with respect to the important predictor variables, are generally considered better than those that lump together individual units that are unalike, simply because the level of the predictor variable then will be more representative of the area to which it refers. However such ideal partitions do not usually arise in practice, and there may be several "plausible" alternative partitions; ideally the sensitivity of results to these alternatives should be assessed.

Often the size of areal units within any surface partitioning varies with respect to either geometric size or total number of objects captured. If the response variable is a density measure (*e. g.*, with respect to the population size), then it is often the case that some density measures are taken with respect to subareas with large populations while others relate to subareas with small populations. If $Y$ is a density variable derived from equally variable units, then we might expect

$$\mathrm{Var}(Y_i) = \sigma^2/n_i, \quad n_i > 0,$$

where $n_i$ is a measure of the size (*e. g.*, total population) of areal unit $i$. This result is attributable to the law of large numbers, since the average is taken over more individual units. It follows that the errors are heteroscedastic:

$$\mathrm{Var}(\boldsymbol{\xi}_i) = \sigma^2/n_i,$$

and thus

$$E[\boldsymbol{\xi}\boldsymbol{\xi}^T] = \sigma^2\mathbf{P},$$

where $\mathbf{P}$ is a diagonal matrix with element $p_{ii} = 1/n_i$. The weighted least squares estimator for $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_w = (\mathbf{X}^T\mathbf{P}^{-1}\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{P}^{-1}\mathbf{Y}), \text{ and}$$
$$\hat{\sigma}^2_w = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_w)^T\mathbf{P}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_w)/(n-k-1).$$

Table 2 shows equivalent weighted estimators for the three models of Sections 2 and 3. The parameter $\rho$ may be estimated as before. The determinant $|\mathbf{A}|$ is unchanged if $\mathbf{P}$ does not depend upon $\rho$, but $\hat{\sigma}^2$ is now given by $\hat{\sigma}^2_w$.

### 4.3. Data problems.

Certain data problems arise that often have nothing to do with the spatial nature of the data *per se*. These include such problems as multicollinearity (which renders parameter estimates unreliable), excessive numbers of predictor variables (which makes efficient analysis difficult), missing and unreliable values, and outliers. The latter group of problems relate to specific data values. However, even if these latter problems arise independently of the areal units system, how they are dealt with (*e. g.*, estimation of missing values, adjustments of the fit to the presence of extreme values) might be affected by where the problem observations are located on the map. If a missing value is near the boundary, for example, interpolation of its

TABLE 2
WEIGHTED LEAST SQUARES ESTIMATION RESULTS
FOR THE THREE REGRESSION MODELS

1. The regression model with spatially correlated errors:

$$\hat{\boldsymbol{\beta}}_w = (\mathbf{X}^T\hat{\mathbf{A}}^T\mathbf{P}^{-1}\hat{\mathbf{A}}\mathbf{X})^{-1}(\mathbf{X}^T\hat{\mathbf{A}}^T\mathbf{P}^{-1}\hat{\mathbf{A}}\mathbf{Y})$$
$$\hat{\sigma}_w^2 = \mathbf{u}^T\hat{\mathbf{A}}^T\mathbf{P}^{-1}\hat{\mathbf{A}}\mathbf{u}/(n-k-2)$$

2. The regression model with lagged response variable:

$$\hat{\boldsymbol{\beta}}_w = (\mathbf{X}^T\mathbf{P}^{-1}\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{P}^{-1}\hat{\mathbf{A}}\mathbf{Y})$$
$$\hat{\sigma}_w^2 = (\hat{\mathbf{A}}\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_w)^T\mathbf{P}^{-1}(\hat{\mathbf{A}}\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_w)/(n-k-2)$$

3. The regression model with lagged predictor variables:

$$\hat{\boldsymbol{\beta}}_w = (\mathbf{Z}^T\mathbf{P}^{-1}\mathbf{Z})^{-1}(\mathbf{Z}^T\mathbf{P}^{-1}\mathbf{Y})$$
$$\hat{\sigma}_w^2 = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_w - \hat{\tau}\mathbf{W}\mathbf{X}_\tau)^T\mathbf{P}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_w - \hat{\tau}\mathbf{W}\mathbf{X}_\tau)/(n-k-2)$$

value might be more difficult than if it is near the center. Similarly, several missing values may be more difficult to estimate if they are clustered rather than scattered.

In regression analysis it is often of particular interest to assess the sensitivity of the model fit to individual cases, particularly cases with extreme values either in the response variable or in one or more of the predictor variables. In the standard regression model, individual cases can be deleted, one at a time, and the model refit. But with the models of Section 2, deletion of individual cases alters the order relationships between the sites, and creates internal boundaries within the study area if the cases refer to areal units. So procedures that are exact for the standard regression model no longer apply. Martin (1984) gives the estimator for vector $\boldsymbol{\beta}$, for the case of a general Gaussian correlated errors model, when one or more cases are (treated as) missing, and although implementation of the procedure taking each of the $n$ cases in turn might be lengthy, his results can be used to develop an appropriate check on the sensitivity of $\hat{\boldsymbol{\beta}}$ to individual observations. The results also can be used to assess the sensitivity of estimates of $\hat{\boldsymbol{\beta}}$ to individual observations in the case of a general matrix $\boldsymbol{\Sigma}$.

The spatial distribution of extreme values may need to be considered, particularly whether they are scattered or clustered. Suppose the distribution of extreme regression residuals in equation (3.4) (defined by vector $\mathbf{u}$) is clustered. This might have a greater impact on the estimate of $\rho$ than if the extreme values are scattered. A plot of the elements of vector $\mathbf{u}$ against the corresponding elements in vector $\mathbf{Wu}$ may highlight this problem. The potentially important consideration is the influence of the *distribution* of extreme values on the estimation of $\rho$, and thence on the estimation of $\boldsymbol{\beta}$ (since vector $\hat{\boldsymbol{\beta}}$ is a function of $\hat{\rho}$). Methods for resistant estimation of regression parameters by iterative downweighting of observations based on the frequency distribution of residuals are available [see Hoaglin,

Mosteller and Tukey (1983, 1985); see Besag (1981) for a spatial application]. Resistant estimators for $\boldsymbol{\beta}$ are of the general form

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{Q}\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Q}\mathbf{Y}),$$

where matrix $\mathbf{Q}$ downweights observations with large residuals. The estimation is often performed by iterated weighted least squares, where the residuals at one iteration step are used to specify the elements of matrix $\mathbf{Q}$ at the next iteration step through a selected "weighting function." This procedure is distinguished from the weighting scheme discussed in Section 4 ("surface partitioning"), where the weights are specified *a priori*, as a function of areal unit attributes, and are not subsequently re–estimated. However, in the case of the regression models discussed in Sections 2 and 3, there is the further problem of acquiring resistant estimates of the spatial parameter $\rho$, a problem that has not as yet received much attention in the literature. Indeed many of the problems raised in this section have not yet received detailed consideration.

## 5. Two Case Studies

We conclude this discussion by briefly examining two applications that exemplify some of the points made in earlier sections of this paper.

### 5.1. Standardized Mortality Rates (SMRs) for areas of Glasgow.

Cancer data are available on SMR's for 87 community medicine areas (CMAs) in Glasgow (1981/82). An SMR is obtained for any area by dividing the observed number of deaths ($O_i$) by the expected number ($E_i$), given the age and sex composition of the area, and multiplying by 100. Data also are available on 15 relevant social and economic variables for the areas that are to act as predictors for the SMR data.

Scatterplots of the SMR values against the predictor variables suggest that the relationships are more linear and have better spread properties if the SMR data are subjected to a logarithmic transformation. Table 3 summarizes the fit of the best fitting model selected by a stepwise regression procedure.

TABLE 3
SUMMARY OF THE FIT OF THE REGRESSION MODEL
TO THE SMR DATA FOR GLASGOW

$$\mathrm{LN}(\widehat{\mathrm{SMR}}) = 4.23 + 0.014X_1 + 0.017X_2$$

$$R^2 = 67.9\%, \qquad \hat{\sigma} = 0.100$$

Values of t–statistics corresponding respectively to terms in the regression model containing variables $X_1$, $X_2$:

3.95, 13.07

$X_1$ = % of population that are pensioners living alone.

$X_2$ = % of population in Social Classes 4 and 5.

The residuals from the model show evidence of pattern with generally higher values near the center of the city, declining out towards the suburbs. A Moran test for residual autocorrelation, using a binary connectivity matrix **W** in which CMAs that share a common boundary are defined as being joined, leads to a rejection of the null hypothesis of no pattern, at the 10% level of significance. However, a correlated errors model has proven to be an unsatisfactory data descriptor in this example. The regression model was augmented with a second–order trend surface. The fit of this model is summarized in Table 4. Although the $R^2$ does not improve substantially, its increase is significant, and the residuals are better behaved (no residual autocorrelation is detected). The trend surface model peaks at the city center and declines towards the suburbs.

TABLE 4
SUMMARY OF THE FIT OF THE REGRESSION MODEL
WITH SECOND ORDER TREND COEFFICIENTS
TO THE SMR DATA FOR GLASGOW

$$\mathrm{LN}(\widehat{\mathrm{SMR}}) = 4.27 + 0.010X_1 + 0.017X_2 - 0.165X_E + 0.202X_N$$
$$-0.610X_E^2 - 0.710X_N^2 + 1.296X_EX_N$$
$$R^2 = 73.7\%, \qquad \hat{\sigma} = 0.093$$

Values of t–statistics corresponding respectively to terms in the regression model containing variables $X_1$, $X_2$, $X_E$, $X_N$, $X_E^2$, $X_N^2$, $X_EX_N$:

2.55, 12.27, −0.38, 0.65, −1.65, −2.35, 3.24

$X_E$ and $X_N$ are trend surface co–ordinates.

$X_1$ and $X_2$ are defined as in Table 3.

There is evidence in the residuals of an inverse relationship between residual variance and the observed number of deaths. Pocock *et al.* (1981) have argued that this should be expected, and have shown that observations should be weighted, with areas having a small number of observed deaths being downweighted. In the notation of Section 4, they suggest that

$$p_{ii} = 1 + 1/(\sigma^2 O_i). \tag{5.1}$$

In addition, high leverage values have been noted for three of the suburban CMAs, which are of large areal extent and hence, when represented by their centroids for purposes of fitting the trend surface component of the model, isolated from the rest of the map. Leverage effects in fitting trend surface models have been discussed in Unwin and Wrigley (1987). The best fit model derived from reanalyzing the data and downweighting observations using equation (5.1) are reported in Table 5. The effect of deleting the three CMA's with high leverages resulted in variable $X_1$ ceasing to be significant in the fit (failed to reject $H_0 : \beta_1 = 0$). Finally, Table 6 reports the results of using a resistant R–estimator (Li, 1985, p. 331) to fit the regression model. This uses a rank–based criterion. The table provides evidence of the resistance of the fit to the few large residuals.

TABLE 5
SUMMARY OF THE FIT OF THE REGRESSION MODEL
USING MATRIX **P** TO DOWNWEIGHT THE OBSERVATIONS

$$\mathrm{LN}(\widehat{\mathrm{SMR}}) = 4.26 + 0.009X_1 + 0.016X_2 - 0.142X_E + 0.242X_N$$
$$-0.588X_E^2 - 0.704X_N^2 + 1.210X_E \star X_N$$
$$\hat{\sigma} = 0.093$$

Values of t–statistics corresponding respectively to terms in the regression model containing variables $X_1$, $X_2$, $X_E$, $X_N$, $X_E^2$, $X_N^2$, $X_E \star X_N$:

$$2.49,\ \ 12.27,\ \ -0.32,\ \ 0.71,\ \ -1.57,\ \ -2.19,\ \ 3.03$$

$X_E$ and $X_N$ are defined in Table 4.

$X_1$ and $X_2$ are defined as in Table 3.

TABLE 6
R–RESISTANT REGRESSION

$$\mathrm{LN}(\widehat{\mathrm{SMR}}) = 4.34 + 0.009X_1 + 0.017X_2 - 0.375X_E + 0.135X_N$$
$$-0.461X_E^2 - 0.682X_N^2 + 1.373X_E \star X_N$$
$$R^2 = 46.0\%, \quad \hat{\sigma} = 0.087$$

Values of standard errors corresponding respectively to terms in the regression model containing variables $X_1$, $X_2$, $X_E$, $X_N$, $X_E^2$, $X_N^2$, $X_E \star X_N$:

$$0.003,\ \ 0.001,\ \ 0.412,\ \ 0.294,\ \ 0.347,\ \ 0.284,\ \ 0.375$$

$X_E$ and $X_N$ are defined in Table 4.

$X_1$ and $X_2$ are defined as in Table 3.

An interesting feature of this analysis is the presence of an "inner city" factor as an added risk factor that is in addition to the usual class and age variables. Such a factor might be associated either with the environmental characteristics of the inner cities or the characteristics of the inner city population (related to, for example, diet, exercise, higher levels of stress and overcrowding).

### 5.2. Agricultural consumption and accessibility.

Cliff and Ord (1981, pp. 209 and 237) report the results of an analysis of spatial variation in the percentage, in value terms, of the gross agricultural output of each county in Ireland consumed by itself ($Y$) as a function of a measure of county accessibility in terms of the arterial road network ($X$). Table 7 reports the results of the least squares regression fit. The residuals show evidence of spatial autocorrelation under three different order definitions

of matrix **W** (the binary matrix is constructed by setting $w_{ij} = 1$ if counties $i$ and $j$ share a common boundary, zero otherwise; the matrices are standardized by setting row sums to 1).

TABLE 7
ORDINARY LEAST SQUARES ANALYSIS OF IRISH DATA

$$\hat{Y} = -8.44 + 0.0053X_1$$

Values of t–statistics corresponding in order to the two terms in the regression model:

$$-2.65, \quad 7.42$$

$$R^2 = 69.7\%, \quad R^2(\text{adjusted}) = 68.4\%$$

$$n = 26, \quad \text{residual variance} = 13.58$$

Lag correlations computed for the residuals:

| lag | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| correlation | 1.000 | 0.387 | −0.089 | −0.218 |
| number of pairs | | 58 | 93 | 94 |

Autocorrelation tests on the residuals:

| Matrix type: | GMC | E[GMC] | standard deviation [GMC] | standard normal deviate |
|---|---|---|---|---|
| binary | 1.726 | −0.238 | 0.535 | 3.67 |
| standardized binary | 0.315 | −0.057 | 0.126 | 2.95 |
| weighted | 0.429 | −0.057 | 0.146 | 3.32 |

Source: Cliff and Ord, 1981, p. 230.

We consider the effects of adding spatially lagged forms of the original variables in order to deal with the problem of residual spatial autocorrelation. Added variable plots have been used in order to determine whether vector **WX** or vector **WY**, or both, should be added to the model, as well as what form of matrix **W** provides the best fit (Haining, 1990). The evidence of these plots, irrespective of how matrix **W** is constructed, is that adding vector **WY** is preferable to adding vector **WX**; if vector **WX** is added then vector **WY** also should be added, whereas if **WY** is added then **WX** is not needed.

Table 8 reports the results of regression model fitting with vector **WX** and then with **WY**. The evidence here confirms the superiority of including vector **WY** rather than **WX**. One should note that vectors **X** and **WX** are correlated, raising the problem of multicollinearity in the fit of the lagged predictor variable model. Table 9 reports the results of fitting a regression model with a spatially correlated errors model. Three error models have been tried: a simultaneous autoregressive (SAR) scheme, a moving average (MA) scheme (parameter $\theta$), and a conditional autoregressive (CAR) scheme (parameter $\delta$). Details of these models are summarized in Cliff and Ord (1981), Upton and Fingleton (1985) and

Ripley (1981). The first error model provides the best fit, although it is not quite as good as that obtained by the lagged response variable model. On the other hand, the spatial error model is probably more substantively justifiable than the lagged response variable model for this data set.

TABLE 8
FITTING DIFFERENT SPATIAL REGRESSION MODELS TO THE IRISH DATA

**(a) Lagged predictor variable:**

$$\mathbf{Y} = \beta_0\mathbf{1} + \beta_1\mathbf{X} + \tau\mathbf{WX} + \boldsymbol{\xi}$$

| | Binary Matrix **W** | Standardized Binary Matrix **W** | Weighted Matrix **W** (Cliff & Ord, 1981, p. 230) |
|---|---|---|---|
| $\hat{\beta}_0$ | −14.13 | −23.97 | −20.59 |
| | (−3.55) | (−5.24) | (−4.22) |
| $\hat{\beta}_1$ | 0.0056 | 0.0026 | 0.0030 |
| | ( 8.25) | ( 3.13) | ( 3.25) |
| $\hat{\tau}$ | 0.0002 | 0.0063 | 0.0050 |
| | ( 2.15) | ( 4.05) | ( 3.02) |
| $R^2$ | 74.7% | 82.3% | 78.3% |
| adjusted $R^2$ | 72.5% | 80.7% | 76.4% |
| $r_{x,wx}$ | 0.06 | 0.57 | 0.58 |
| error variance | 11.80 | 8.28 | 10.16 |

**(b) Lagged response variable:**

$$\mathbf{Y} = \beta_0\mathbf{1} + \beta_1\mathbf{X} + \rho\mathbf{WY} + \boldsymbol{\xi}$$

| | Standardized Binary Matrix **W** | Weighted Matrix **W** (Cliff & Ord, 1981, p. 230) |
|---|---|---|
| $\hat{\beta}_0$ | −6.24 | −6.71 |
| | (−3.10) | (−3.21) |
| $\hat{\beta}_1$ | 0.0024 | 0.0028 |
| | ( 4.38) | ( 5.11) |
| $\hat{\rho}$ | 0.731 | 0.646 |
| | ( 6.38) | ( 5.13) |
| $R^2$ | 87.2% | 86.2% |
| error variance | 5.25 | 5.67 |

Figures in parentheses under the coefficient estimates are t–statistic values; $r_{x,wx}$ is the Pearson correlation between variable vectors **X** and **WX**; these results agree with those given in Anselin (1988) and Bivand (1984).

Finally we report the fitting of a robust form of the regression model with SAR errors (Table 10) using Tukey's biweight function to downweight the influence of large residuals (**e**) in the estimation of $\boldsymbol{\beta}$. Outliers also will affect the estimation of $\rho$, but after computing vector **u** at the first iteration and inspecting the plot of (**u**, **Wu**) a non–robust estimator

TABLE 9
FITTING A REGRESSION MODEL WITH SPATIALLY CORRELATED ERRORS TO THE IRISH DATA

**(a) SAR errors model:**

$$E[\mathbf{u}\mathbf{u}^{\mathbf{T}}] = \sigma^{\mathbf{2}}[(\mathbf{I} - \rho\mathbf{W})^{T}(\mathbf{I} - \rho\mathbf{W})]^{-1}$$

| | Binary Matrix **W** | Standardized Binary Matrix **W** | Weighted Matrix **W** (Cliff & Ord, 1981, p. 230) |
|---|---|---|---|
| $\hat{\beta}_0$ | 1.155 | 4.670 | 1.359 |
| | ( 0.33) | ( 1.04) | ( 0.36) |
| $\hat{\beta}_1$ | 0.0032 | 0.0024 | 0.0030 |
| | ( 5.84) | (37.79) | ( 4.74) |
| $\hat{\rho}$ | 0.177 * | 0.843 + | 0.780 + |
| | (14.29) | ( 9.44) | ( 7.06) |
| $R^2$ | 87.0% | 85.7% | 86.1% |
| error variance | 5.36 | 5.89 | 5.73 |

**(b) Other spatial error models:**

| | Moving Average | | Conditional |
|---|---|---|---|
| | Invertible Range | Unrestricted | Autoregressive |
| $\hat{\beta}_0$ | −1.290 | 1.766 | −3.725 |
| $\hat{\beta}_1$ | 0.0037 | 0.0034 | 0.0041 |
| $\hat{\theta}$ | 0.193 | 0.944 | *** |
| $\hat{\delta}$ | *** | *** | 0.184 |
| $R^2$ | 78.3% | 80.4% | 81.9% |
| error variance | 8.92 | 8.07 | 7.47 |

* denotes that the maximum value is 0.194.

+ denotes that the maximum value is 1.00.

Figures in parentheses under the coefficient estimates are t–statistic values.

for $\rho$ was chosen that employed the usual expression (16) (indeed an R–estimator fit of **u** on **Wu** gives an estimate for $\rho$ very close to the maximum likelihood estimate). The results reported here are for $B = cS$ where $c = 6$ and $S$ is the median absolute deviation of the residuals (Li, 1985, p. 293). The downweighting is strongest for western counties. It happens that these are the counties for which $X$ values are most unreliable because of the way the index was constructed (Cliff and Ord, 1981, p. 207).

TABLE 10
ROBUST ESTIMATION OF EQUATION (3.4) USING TUKEY'S BI-WEIGHT

$$\hat{Y} = 1.9941 + 0.0028X$$

$$\hat{\rho} = 0.794, \text{ error variance } (\hat{\sigma}^2) = 5.68, \; R^2 = 86.2\%$$

Final set of Tukey weights (in alphabetical order across rows)

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 0.989 | 0.962 | 0.918 | 0.866 | 0.971 | 0.980 | 0.959 | 0.908 | 0.997 |
| 0.972 | 0.999 | 0.821 | 0.991 | 0.970 | 0.999 | 0.829 | 0.847 | 0.999 |
| 0.990 | 0.873 | 0.893 | 0.928 | 0.990 | 0.973 | 0.970 | 0.999 | |

MAD = 1.715, B = 10.293

## 6. Conclusions

Social scientists are often accused of selecting models that derive too much from theory and too little from data properties. A "firm" model is specified on the basis of some theoretical argument (less kindly put, some "preconceived" idea) and attention then focuses on fitting the models and, at best, making comparisons with a small range of alternative models.

In developing "spatialized" forms of equation (1.1) in Section 1, the aim is to broaden the range of possible models that may be considered "soft" or "firm," depending upon the substantive context and the stage of data analysis reached. In discussing data related problems in Section 4, the aim was to draw attention to the sorts of fitting and assessment issues that often prove endemic to regression modeling with spatial data and to suggest some possible lines of treatment.

The development of interactive statistical packages with a range of graphical data inspection options should encourage closer inspection of data properties. Unfortunately, however, some of the procedures needed to fit the "extended" range of spatial regression models described here are yet to be made widely available in easy–to–use packages.

## 7. References

Anas, A., and S. Eum. (1984) Hedonic analysis of a housing market in disequilibrium. *Journal Urban Economics*, **15**: 87-106.

Anselin, L. (1988) Lagrange multiplier test diagnostics for spatial dependence and spatial heterogeneity. *Geographical Analysis*, **88**: 1-17.

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

Bivand, R. (1984) Regression modeling with spatial dependence: an application of some class selection and estimation methods. *Geographical Analysis*, **16**: 25-38.

Chamberlin, E. (1956) *The Theory of Monopolistic Competition*. London: Oxford University Press.

Cliff, A., and J. Ord. (1981) *Spatial Processes*. London: Pion.

Griffith, D. (1988) A reply to R. Martin's 'Some comments on correction techniques for

boundary effects and missing value techniques'. *Geographical Analysis*, **20**: 70-75.

Haining, R. (1986) Intra urban retail price competition: corporate and neighbourhood aspects of spatial price variation. In *Spatial Pricing and Differentiated Markets*, edited by G. Norman, pp. 144-164. London: Pion, London Papers in Regional Science #16.

Haining, R. (1987) Small area aggregate income models; theory and methods with an application to urban and rural income data for Pennsylvania. *Regional Studies*, **21**: 519-530.

Haining, R. (1990) The use of added variable plots in regression modelling with spatial data. *The Professional Geographer*, (to appear).

Hoaglin, D., F. Mosteller, and J. W. Tukey (1983) *Understanding Robust and Exploratory Data Analysis*. New York: Wiley.

Hoaglin, D., F. Mosteller, and J. Tukey (eds.) (1985) *Exploring Data Tables, Trends and Shapes*. New York: Wiley.

Holmes, J., and P. Haggett (1977) Graph theory interpretation of flow matrices: a note on maximization procedures for identifying significant links. *Geographical Analysis*, **9**: 388-399.

Künsch, H. (1983) Approximations to the maximum likelihood equations for some Gaussian random fields. *Scandinavian Journal of Statistics*, **10**: 239-246.

Li, G. (1985) Robust regression. In *Exploring Data Tables, Trends and Shapes*, edited by Hoaglin, D., F. Mosteller, and J. Tukey, pp. 281-343. New York: Wiley.

Loftin, C., and S. Ward (1983) A spatial autocorrelation model of the effects of population density on fertility. *American Sociological Review*, **48**: 121-8.

Mardia, K., and R. Marshall (1984) Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, **71**: 135-146.

Martin, R. (1984) Exact maximum likelihood for incomplete data from a correlated Gaussian process. *Communications in Statistics A—Theory and Methods*, **13**: 1275-1288.

Martin, R. (1987) Some comments on correction techniques for boundary effects and missing value techniques. *Geographical Analysis*, **19**: 273-282.

Matula, D., and R. Sokal (1980) Properties of Gabriel graphs relevant to geographic variation and the clustering of points in the plane. *Geographical Analysis*, **12**: 205-222.

Ord, J. (1981) Towards a theory of spatial statistics: a comment. *Geographical Analysis*, **13**: 86-91.

Pocock, J., D. Cook, and A. Beresford (1981) Regression of area mortality rates on explanatory variables: what weighting is appropriate? *Applied Statistics*, **30**: 286-296.

Ripley, B. (1981) *Spatial Statistics*. New York: Wiley.

Unwin, D., and N. Wrigley (1987) Control point distribution in trend surface modelling revisited: an application of the concept of leverage. *Transactions of the Institute of British Geographers*, **12**: 147-160.

Upton, G., and B. Fingleton (1985) *Spatial Data Analysis by Example*, Volume 1: Point pattern and quantitative data. New York: Wiley.

# DISCUSSION

## "Models in human geography: Problems in specifying, estimating and validating models for spatial data"

### by Robert P. Haining

Haining considers the following three kinds of model: (i) those with a spatially lagged response variable; (ii) those with a spatially lagged predictor variable; and, (iii) those with spatially correlated disturbance terms. Using his notion, with $\mathbf{A} = \mathbf{I} - \rho\mathbf{W}$, the presence of $\mathbf{ln}|\mathbf{A}|$ in the log–likelihood function complicates the estimation of the parameters for the first and third classes of models. With an emphasis on maximum likelihood methods, only a relatively small set of areas (making up a region) can be considered, as Haining notes. In turn, the smaller the number of cases (areas), the greater the vulnerability of the estimated regression parameters to problems stemming from the presence of high leverage data points and outliers. This vulnerability can be tackled on two fronts, namely (i) modifying software and data structures in order to analyze larger data sets, and (ii) paying close attention to diagnostic signals (of the presence of problems). Both are crucial. As the first strategy does little for small systems, it is necessary to take regression diagnostics very seriously and to look closely at robust regression methods. Haining's discussion of these issues for spatially distributed variables is particularly welcome.

### Theories, models and statistical methods.

He notes that "social scientists are often accused of selecting models that derive too much from theory and too little from data properties." Excluding theorists who disavow any systematic examination of empirical evidence, the problem for social scientists is not that they choose models based on theories, but that they choose models from a superficial examination of data properties. This "examination" usually excludes consideration of the diagnostic procedures discussed by Haining. There is, however, another basis for the superficial consideration of data that is rooted in the modeling cycle outlined by him. As described, the process of moving from a "soft" model to a "firmer" model capitalizes on chance. It is not clear why making "the percentage of explained variation as high as possible" is of any real value in assessing the utility of an estimated model. Certainly, we must have well behaved residuals, but this criterion can be invoked without slavish adherence to the maxim of maximizing $\mathbf{R^2}$. A satisfactory model may "emerge" but, at most, it is a specification of a model that can be assessed with a different data set. Of course, such a model has a better chance of serving further tests if the problems discussed by Haining are addressed in a (modified–only modest attention given to $\mathbf{R^2}$) modeling cycle.

### Specifying interdependence.

The specification of the weights matrix $\mathbf{W}$ is crucial. Leaving it out is problematic when linear models are specified and it is known that the data points are interdependent. Haining's Table 1 is helpful in laying out *some* of the possible specifications of $\mathbf{W}$. Many specifications of $\mathbf{W}$ in terms of joins and weights remain little more than guesses about the processes generating interdependencies. Until we know more about these processes, specifications of $\mathbf{W}$ will remain individual guesses or customary behavior. Even so, doing something with regard

to $\mathbf{W}$ for models where the disturbance terms are autocorrelated may be of some value for "safer" inference. For this class of models, the interdependence is a technical problem. However, for models with a spatially lagged response variable, the specification of $\mathbf{W}$ must be substantive, as it is an explicit part of the theoretical statement directly of interest.

Another important issue concerns the match, if any, between the level of aggregation of the data and the spatial scale of the phenomenon under study. Intuitively, it is unlikely that a social process with a spatial scale defined in terms of local neighborhoods can be captured in data aggregated to the ward, postal ZIP Code, or city levels. The larger units are likely to contain many diverse and distinct neighborhoods that have been grouped together in the (usually implicit) aggregation. At the other extreme, a process with a spatial scale at the county level will be modeled, at best, inefficiently with data assembled at the local neighborhood level.

**Spatially lagged predictor variable models.**

The simplest form, as stated by Haining, of such a model is as follows:

$$\mathbf{y}_i = \beta_0 + \beta_1 \mathbf{X}_{1,i} + \cdots + \beta_R \mathbf{X}_{R,i} + \tau \sum_{j=1}^{n} \mathbf{W}_{ij} \mathbf{X}_{\tau,j} + \epsilon_i.$$

Attention can be focused on the estimation of $[\beta : \tau]$. Haining writes, "if variable $\mathbf{X}_\tau$ is spatially correlated, the matrix $(\mathbf{X}'\mathbf{X})^{-1}$ may be unstable (since in extreme cases vectors $\mathbf{X}_\tau$ and $\mathbf{W}\mathbf{X}_\tau$ may be linearly dependent, causing $\mathbf{X}'\mathbf{X}$ to be singular)." A *worse* situation would be an undiagnosed near singularity that would lead to inflated standard errors and compromised inference. (For an *exact* singularity, estimation would breakdown and diagnosis would be straightforward.) Why include both $\mathbf{X}_\tau$ and $\mathbf{W}\mathbf{X}_\tau$ in the specification of the model? If $\mathbf{X}_\tau$ is spatially autocorrelated, then $\mathbf{X}_{\tau,i}$, would be given by the weighted sum of the values of $\mathbf{X}_\tau$ for the areas, $j$, having non–zero $\mathbf{W}_{ij}$. The data value for $\mathbf{X}_{\tau,i}$ would include little new information beyond that contained in

$$\sum_{j=1}^{n} \mathbf{W}_{ij} \mathbf{X}_{\tau,j}$$

Use of $\mathbf{X}_\tau$ and $\mathbf{W}\mathbf{X}_\tau$ seems certain to generate collinearities and, as it is a specification problem, recourse to reduced rank methods seems premature. (Of course if $\mathbf{X}$ is not of full rank for a set of measured variables, then techniques like ridge regression *may* be of some value.)

The possibility of $\tau$ being a vector may merit further consideration. If one $\mathbf{X}_\tau$ is autocorrelated, then it is possible that other $\mathbf{X}$'s are autocorrelated, too. Further, each $\mathbf{X}_\tau$ may have its own $\mathbf{W}_\tau$ regime. This may be introducing an identification and specification nightmare, but there is no reason (other than simplicity) for assuming only one $\mathbf{X}_\tau$ is spatially autocorrelated. Coupled spatial processes with distinct autocorrelation regimes seem quite reasonable.

**Data problems.**

Haining is correct in directing our attention towards data problems. Missing and unreliable data values are a major problem, as are outliers. Influential data points can be included here also. In addition to the statistical problems discussed by Haining, there are many database management issues. Techniques for re–estimating a specified equation, when data points are dropped one at a time, must rest on an adequate database management system. Not only are $\mathbf{Y}$ and $\mathbf{X}$ changed when one (or more) observation(s) is(are) removed, but $\mathbf{W}$ also is changed. As Haining notes, deletion of a data point does create internal boundaries and it changes geometric relationships between areas. These are serious technical and substantive issues that can only be addressed if there is in place a sophisticated and flexible database management system that can handle dropped cases and the accompanying implied changes for matrix $\mathbf{W}$.

Haining's discussion uses a variety of alternative estimation procedures (*e. g.*, resistant regression methods) given a specific problem has been identified. This is useful, but many researchers will be left unsatisfied if they can neither implement nor have an adequate database management system to support such statistical procedures. It took a decade before regression diagnostics were widely available in the standard statistical software. As even fewer analysts grapple with interdependent data points, it may take even longer to have widely available software to support generalized autocorrelation modeling. Of course, if the importance of grappling with interdependent data points is recognized more widely, the (badly needed) software will become available for general use sooner. Haining has helped to push us in that direction.

*Patrick Doreian*, University of Pittsburgh

## *PREAMBLE*

*I had rather be hissed for a good verse than applauded for a bad one.*

Victor Hugo

*Some scholars practicing what is called pure science are convinced that their ways of doing science are theoretical, and hence superior to that done in what is called applied science. On the other hand, many scholars in the applied sciences stress that superiority of theory over practice is a myth, and that theory and practice cannot be separated. They continually point out the numerous possibilities of doing science that mix pure theoretical research goals and applied research goals, each worthy of equal respect and dignity. The relationship between statisticians and geographers in the realm of spatial statistics is a point in question here. The purpose of this paper is for Martin to give his personal view of the application of spatial statistical analysis in geographic research, mostly noting shortcoming of its use by selected geographers. Martin argues that if geographers believe they have theoretical research findings that contribute to statistics, then statisticians should be allowed to scrutinize this research. Throughout his discussion he hints that geographers do not have the expertise necessary for making such contributions, and that geographers should restrict themselves to applications while enticing statisticians into undertaking the requisite theoretical developments. A number of publications concerning spatial statistics have made clear that it is both an oversimplification and even an error to view geography as solely application–oriented, and statistics as theory–oriented, for scholars in both areas hold a variety of talents and viewpoints. Richardson softens Martin's message, noting that all scholars have an interest in avoiding abusive uses of statistics, and echoes Martin's belief that theoretical developments in spatial statistics need to be linked to relevant examples and realistic geographical problems. In many ways, this paper effectively highlights the contrasts between statistical and geographical approaches to spatial statistics.*

The Editor

# The Role of Spatial Statistical Processes in Geographic Modelling

R. J. Martin

*Department of Probability and Statistics, University of Sheffield, Sheffield S3 7RH, England*

**Overview:** In this paper I give a personal view of the role of spatial statistical processes in geographic modelling. I consider models used by geographers, and comment on the statistical shortcomings of their use. I discuss the role of the geographer in statistical research, and the role of the statistician in geographic research. I also expand on the discussion of two particular topics of interest to geographers—boundary effects and missing values.

## 1. Introduction

There has been a considerable amount of published research in geography in which spatial statistical models have been used or investigated—see for example the review papers of Cliff and Ord (1975) and Bennett and Haining (1985), and the references therein. I will discuss in this paper one particular part of this research—that part in which spatial stochastic processes are used to model the dependence between observations on the same variable at different spatial sites (or on different regions). This is the topic in Section 6 of Cliff and Ord (1975), and Sections 3.1 and 4.1 of Bennett and Haining (1985). Even in this restricted area I am only going to discuss aspects of which I have some statistical knowledge. My viewpoint is that of a theoretical statistician, and I claim no geographical knowledge or understanding.

In the discussion to Bennett and Haining (1985), I expressed my reservations (Martin, 1985) about the published research in geography that I had seen. Some stronger views were given by Besag (1985). My reservations concerned two main aspects. Firstly, that the models used by geographers did not appear to receive the validation from data that has become standard in current statistical analyses, and there was no indication that geographers felt that such validation was necessary. Secondly, that research by geographers that purported to advance statistical theory and methodology was being published in non–statistical journals, and was clearly receiving inadequate refereeing and not receiving the attention of statisticians.

In Section 2 I shall discuss in detail the role of statistical models in geography, whilst in Section 3 I shall discuss the role of the geographer in statistical research and of the statistician in geographical research.

Two particular topics that have received much attention in geographical publications on spatial statistics are boundary effects and missing values. I discussed boundary effects in Martin (1987), and will reconsider some of that discussion in Section 4. In response to a question from a geographer, I wrote up some research of mine on missing values in Martin (1984). Some further comments are in Martin (1987). A subsequent publication (Haining, Griffith and Bennett, 1989) has considered numerically one aspect of this—the information loss. As a result I derived some theoretical results covering this aspect, which are in Martin (1989a). I discuss and extend some of these results in Section 5.

## 2. Statistical Models in Geography

### 2.1. Justification of models

I am not a geographer, and I have no basis for discussing geographic models unless these models are statistical. Unfortunately, I have been unable to understand those statistical models that I have seen used in geography. I mentioned in the introduction my concerns about these models expressed in Martin (1985). In their reply, the authors (Bennett and Haining, 1985) confirmed that the 'model is paramount', and justified this by stating that 'It must be remembered that in human geography and planning we are dealing with individuals and social groups and this results in a problem of legitimizing models, and often in planning applications, it requires the participation of individuals who often will be non–numerate.', and that data analytic methods are 'not appropriate for planning a city'.

I will reply to this in two ways. Firstly, if the data are of no relevance to the model, I do not see the point of presenting the models to statisticians and hoping that 'a research agenda ... may have stimulated the Fellows of the Society to help in their solution'. However, whatever the context and whether or not individuals are involved, I would still be concerned that models are not validated through an examination of relevant data. I also cannot see the relevance of the possible non–numeracy of the participating individuals. My concern is with the non–numeracy of the geographical researchers. Secondly, my comments were not actually aimed at planning models, but at the spatial dependence models discussed below. To concentrate the reply on one area, which did not appear to be represented in the paper, is misleading.

I will now elaborate on my misgivings about models for spatial dependence. When it is possible to envisage a development over time, in which present events depend in some way on previous events, it may be reasonable to attempt to model this development using a 'generative' model. For purely spatial data it is not possible to imagine such a development. Besag (1974) says of spatial models that '... our models will not be mechanistic and must be seen as merely attempts at describing the "here and now" of a wider process'. Some of the early discussions in statistics over simultaneous and conditional models appeared to depend on the belief that such generative models had some meaning outside describing the data. This attitude still appears to pervade geographic research.

Thus, without consulting the data, it is forthrightly assumed that the covariance structure is modelled by, for instance, a one–parameter first–order conditional process or a one–parameter first–order simultaneous process. For example, Haining, Griffith and Bennett (1989) state that 'a first–order conditional autoregressive model ... has ... a monotonically decaying correlation function which seems appropriate for social and environmental spatial data'. They then use the non–stationary edge–corrected version of this model on some remotely sensed data, with only a cursory check for suitability, although they do allow the possibility that the model represents the deviations from a second–order trend surface. I know of no physical or geographic reason why the dependence should of necessity be adequately fitted by this model. The data collection may require the participation of non–numerate individuals (and instruments), but I would not find the argument convincing. There are of course many other correlation functions that monotonically decay with lag.

## 2.2. First–order models

Assume henceforth that for a given set of n sites or regions there is an n–vector of observations $\mathbf{y}$ with mean $\mu = E(\mathbf{y})$ and covariance (or dispersion) matrix $V\sigma^2$. The one–parameter first–order conditional process is usually taken in geographic modelling as specifying the inverse covariance matrix $V^{-1}$ in the form $I - \beta W$, where $W$ is a symmetric matrix of non–negative weights that are assumed known, and are usually taken as zero down the main diagonal. In this form it is a very simple and convenient model. Parameter estimation is particularly simple. Note that there is no need for the row sums of $W$ to be constant, nor any great advantage when they are; and that elements could be negative. Also, the diagonal elements do not need to be zero, although the conditional means are not then linear in $\beta$, as noted below.

Gaussian maximum likelihood requires (see, for instance, Martin, 1984) the minimization with respect to $\beta$ of $|V^{-1}|^{-1/n}\mathbf{e}'V^{-1}\mathbf{e}$, where $\mathbf{e} = \mathbf{y} - \hat{\mu}$, and $\hat{\mu}$ is an estimate of $\mu$. This involves the calculation of the quadratic form $\mathbf{e}'V^{-1}\mathbf{e}$ and the determinant $|V^{-1}|$. Both of these are very easy for a given $\beta$, since $\mathbf{e}'V^{-1}\mathbf{e} = \mathbf{e}'\mathbf{e} - \beta\mathbf{e}'W\mathbf{e}$ and so is linear in $\beta$, whilst $|V^{-1}| = \prod(1 - \beta\lambda_i)$, where the $\lambda_i$ are the eigenvalues of $W$.

Thus exact Gaussian maximum likelihood is easily performed using a one–dimensional search over the admissable range of $\beta$ (to ensure that $V^{-1}$ is positive definite), which is in general, provided $\lambda_{\min} < 0$ and $\lambda_{\max} > 0$, $(\lambda_{\min}^{-1}, \lambda_{\max}^{-1})$, where $\lambda_{\min} = \min_i\{\lambda_i\}$ and $\lambda_{\max} = \max_i\{\lambda_i\}$. This is the appropriate range when the diagonal elements of $W$ are zero. Note that this range is more general than that given in many geographical publications—see, for example, Haining (1987, 1988), and Haining, Griffith and Bennett (1989). For example, if $n = 3$ and the off–diagonal elements of $W$ are all $1/2$, as used by Brandsma and Ketellapper (1979), then the admissable range of $\beta$ is $(-2, 1)$. Note also that that if $W$ consists of non–negative elements and does have constant row sums $c$, then $\lambda_{\max} = c$, so that we require $\beta < c^{-1}$. Also, by the Perron–Frobenius theorem, $\lambda_{\min} \leq -c^{-1}$. Thus $\beta$ has a simple upper bound, rather than $|\beta|$ having a simple bound, as was stated in Martin (1987).

The differential and second differential of the likelihood can also be easily found, so that maximisation routines that use the differential can be used. For example, the Newton–Raphson procedure given by Ord (1975) for the one–parameter first–order simultaneous scheme can easily be adapted. In this case, using

$$f(\beta) = -n^{-1}\sum \ln(1 - \beta\lambda_i) - \ln(\mathbf{e}'V^{-1}\mathbf{e}),$$

where $\mathbf{e}'V^{-1}\mathbf{e} = \mathbf{e}'\mathbf{e} - \beta\mathbf{e}'W\mathbf{e}$, the iteration for $\hat{\beta}$ becomes

$$\hat{\beta}_{r+1} = \hat{\beta}_r - f_\beta(\hat{\beta}_r)/f_{\beta\beta}(\hat{\beta}_r),$$

where

$$f_\beta(\beta) = n^{-1}\sum\{\lambda_i/(1 - \beta\lambda_i)\} - (\mathbf{e}'W\mathbf{e})/(\mathbf{e}'V^{-1}\mathbf{e})$$

and

$$f_{\beta\beta}(\beta) = n^{-1}\sum\{\lambda_i/(1 - \beta\lambda_i)\}^2 - \{(\mathbf{e}'W\mathbf{e})/(\mathbf{e}'V^{-1}\mathbf{e})\}^2.$$

However, care should be taken to ensure numerical accuracy and to monitor convergence, as $\hat{\beta}$ is often close to the upper limit of its admissable range. Ripley (1988, Section 2.1) notes that the Newton–Raphson procedure may fail for some data sets.

For some configurations of the sites, the eigenvalues are known theoretically. For example, for an $n_1$ by $n_2$ rectangular lattice with $W$ containing ones for immediate horizontal or vertical neighbours, the $n_1 n_2$ eigenvalues of $W$ are given by

$$\lambda_{ij} = 2\cos\{\pi i/(n_1+1)\} + 2\cos\{\pi j/(n_2+1)\} \text{ for } i = 1,\ldots,n_1 \text{ and } j = 1,\ldots,n_2.$$

Gasim (1988) has obtained eigenvalues of $W$ when further neighbours are included, although it is difficult to see the practical use of such $W$ (he actually obtains his results for a one–parameter simultaneous process, but they hold equally for the conditional). In other situations the eigenvalues of $W$ need only be calculated numerically once.

Results on the (Fisher) information under Normality can also be easily obtained for this model. Formulæ for the information, defined as the expected value of the second differential of the log likelihood, are given by Mardia and Marshall (1984)—see also Martin (1984). Now, when $V^{-1} = I - \beta W$ we get $\frac{\partial^2 V^{-1}}{\partial \beta^2} = 0$, so that the most convenient form to take for twice the information on $\beta$ is $2I_{\beta\beta} = -\frac{\partial^2 \ln|V^{-1}|}{\partial \beta^2}$, which is $-\frac{\partial^2 \{\sum \ln(1-\beta\lambda_i)\}}{\partial \beta^2}$. Therefore, $2I_{\beta\beta}$ is $\sum\{\lambda_i/(1-\beta\lambda_i)\}^2$. Since $\frac{\partial V^{-1}}{\partial \beta} = -W$, another convenient form for $2I_{\beta\beta}$ is

$$\text{trace}(V\frac{\partial V^{-1}}{\partial \beta}\ V\frac{\partial V^{-1}}{\partial \beta}) = \text{trace}\{(V\frac{\partial V^{-1}}{\partial \beta})^2\}.$$

For small $n$, this is easiest found as the sum of squares of the elements of $V\frac{\partial V^{-1}}{\partial \beta} = -VW$, using $\text{trace}(A^2) = \Sigma\Sigma a_{ij}^2$ when $A$ is symmetric. Otherwise, we can use the fact that the trace of a matrix is the sum of its eigenvalues, and that the eigenvalues of $-VW = (I-\beta W)^{-1}W$ are $\lambda_i/(1-\beta\lambda_i)$, $i = 1,\ldots,n$, so that those of $(-VW)^2$ are $\{\lambda_i/(1-\beta\lambda_i)\}^2$.

If we want to get the asymptotic variance of $\hat{\beta}$, the maximum likelihood estimator of $\beta$, then we also need $I_{\beta\sigma^2}$. The simplest form for $2\sigma^2$ times this is $\text{trace}(-V\frac{\partial V^{-1}}{\partial \beta})$, which is therefore $\sum\{\lambda_i/(1-\beta\lambda_i)\}$. Although the previous result on $I_{\beta\beta}$ has been used by geographers, this result on $I_{\beta\sigma^2}$ has not—see Haining, Griffith and Bennett (1989), and Martin (1989a). These results give a simple formula for the asymptotic variance of $\hat{\beta}$ as 2 over the corrected sum of squares of the $\lambda_i/(1-\beta\lambda_i)$, $i = 1,\ldots,n$, (Martin, 1989a), although asymptotically equivalent forms are easier to use. The latter are considered in Besag and Moran (1975) and Besag (1977b).

It is important to realise that the form $V^{-1} = I - \beta W$ is not the inverse variance matrix of a second–order stationary one–parameter first–order conditional process when the sites or regions form a finite regular lattice. That is, $V$ is not proportional to a correlation matrix. There are several ways of seeing this. A simple way is to invert numerically a given $V^{-1}$, and note that, for instance, the diagonal elements are not constant. Although geographers are becoming more aware of this fact—see, for instance, Haining (1987)—there does still appear to be some confusion. For example, Haining, Griffith and Bennett (1989) use $V^{-1} = I - \beta W$, but also appear to assume $V$ proportional to a correlation matrix—see my comments in Martin (1989a).

For a given $W$ with zeroes on the main diagonal, the one–parameter first–order conditional process can be written in the form

$$E(y_i \mid \cdot) = \mu_i + \beta \sum w_{ij}(y_j - \mu_j), \quad \text{with } \text{var}(y_i \mid \cdot) = {\sigma_\eta}^2,$$

where the conditioning is on all other values, $y_j$, $j \neq i$. Note that the assumption that all sites have the same conditional variance is often not reasonable for a finite set of sites—in particular, it is usually preferable that the conditional variance is smaller for the interior points. It is possible to postulate unequal conditional variances for the one–parameter first–order conditional process, but $W$ must then be asymmetric with $w_{ij}\sigma_j^2 = w_{ji}\sigma_i^2$, where $\mathrm{var}(y_i \mid \cdot) = \sigma_i^2$ (Besag, 1975). Another possibility would be to use a symmetric $W$, but for $W$ to have non–zero diagonal elements. Then, provided $1 - \beta w_{ii} > 0 \ \forall i$, $\mathrm{var}(y_i \mid \cdot) = \sigma_\eta^2/(1 - \beta w_{ii})$. However, the conditional mean would now have the form

$$E(y_i \mid \cdot) = \mu_i + \frac{\beta}{1 - \beta w_{ii}} \sum_{j \neq i} w_{ij}(\mathrm{y}_j - \mu_j)$$

which is non–linear in $\beta$.

On an infinite lattice, the second–order stationary process is such that $w_{ij}$ only depends on the lag $i - j$. For a finite lattice, define an interior site as a site $i$ for which all the sites appearing in the expression for $E(y_i \mid \cdot)$ are observed. Then, part of the confusion about stationary is due to the fact that provided either site $i$ or site $j$ is an interior site, the $(i, j)$ element of $V_S^{-1}$, where $S$ denotes the stationary form, is precisely that element of $I - \beta W$. This is easily seen by direct multiplication of $V_S$ and $V_S^{-1}$, and using known relationships between the correlations—see equation (5.12) of Besag (1974).

Therefore, many results obtained for $V^{-1} = I - \beta W$ do hold for interior sites of the stationary process without modification. Nevertheless, there are many results that do not simply carry over from one form to the other. Of particular importance are the results of Guyon (1982), who showed that using Gaussian maximum likelihood for one form may lead to estimators with undesirable properties for a different form. Thus, great care should be taken to precisely define which form is being postulated. This care is not yet sufficiently in evidence.

Similarly, the one–parameter first–order simultaneous process has $V^{-1}$ of the form

$$(I - \beta W)'(I - \beta W),$$

where in this case $W$ does not need to be symmetric. Note that the diagonal elements of $W'W$ usually differ, and are usually greater for interior sites, so that the conditional variance at these sites is reduced. In fact, assuming $w_{ii} = 0$, $\mathrm{var}(y_i \mid \cdot) = \sigma_\eta^2/(1 + \beta^2 \sum_j w_{ji}^2)$. Whilst it may be desirable that the conditional variance is smaller for interior sites, the precise variances arise from the modelled neighbours, rather than being directly specified.

This model has also been used without question—for instance, see Haining (1987). The model is also simple and convenient, although not as simple as the conditional process. Since $|A'A| = |A|^2$, exact Gaussian maximum likelihood can easily be performed using the eigenvalues of $W$ (Ord, 1975), although these eigenvalues may be complex if $W$ is not symmetric. In many cases, the postulated $W$ is the same as a possible $W$ for the conditional process, in which case $W$ is symmetric and has the same eigenvalues as before.

In this common case that $W$ is symmetric, it is possible to get simple results for the information. Then $V^{-1} = (I - \beta W)^2$, so that its eigenvalues are $\{(1 - \beta\lambda_i)^2\}$, and $V$ and $W$ commute. Therefore $-V\frac{\partial V^{-1}}{\partial \beta} = 2W(I - \beta W)^{-1}$, with trace $2\sum\{\lambda_i/(1 - \beta\lambda_i)\}$. Since $W$

and $W'$ have the same eigenvalues, this formula also holds even when $W$ is not symmetric. Also

$$V\frac{\partial V^{-1}}{\partial\beta}\ V\frac{\partial V^{-1}}{\partial\beta} = 4W^2(I-\beta W)^{-2},$$

with trace $4\Sigma\{\lambda_i/(1-\beta\lambda_i)\}^2$. These are just multiples (2 and 4) of the values for the conditional process. The asymptotic variance of $\hat{\beta}$ is therefore exactly one quarter of the value for the conditional process, which was discussed above. Since Haining (1987) uses a symmetric $W$, the general formulæ misquoted from Ord (1975), and the approximation given, are quite unnecessary. Note that Ord's (1975) $\alpha$ should be $-\Sigma\{\lambda_i/(1-\beta\lambda_i)\}^2$.

### 2.3. More general models

Extensions to the one–parameter conditional or simultaneous forms have been suggested. For instance it is easy to extend the conditional form for $V^{-1}$ to the two–parameter form $I-\beta_1W_1-\beta_2W_2$, which either extends the range of dependence or can be used for the same range of dependence as before, but with $W$ split into two parts to allow different degrees of dependence in different directions. This extension, at least in the simultaneous form, is usually attributed in the geographic literature to Brandsma and Ketellapper (1979), although the idea was hardly new to statistics. Even in spatial statistics the use of more than one parameter is well established—see Whittle (1954). The simultaneous process can itself, at least on an infinite lattice, be represented as a special case of a conditional form with separate parameters for each of the immediate horizontal or vertical neighbours, the immediate diagonal neighbours, and the lag–two horizontal or vertical neighbours—see Besag (1974).

However, if it is wished to keep some of the simplicity of the one–parameter conditional or simultaneous forms, there are not many extensions available. The ability to obtain eigenvalues of $V^{-1}$ that are linear in the parameters $\beta_i$ is hampered by the requirement that the $W_i$ matrices need to commute. Using powers of $W$ is possible, but is not always satisfactory. Unless $W$ is triangular, $W^2$ has some diagonal elements positive, so that the conditional variance for those sites is reduced. For a rectangular lattice the most general conditional form with known eigenvalues that does not use powers has $V^{-1}=I-\beta_1W_1-\beta_2W_2-\beta_3W_3$, where $W_1$ is for horizontal neighbours, $W_2$ for vertical neighbours, and $W_3$ for the four diagonal neighbours. Squaring this gives the most general simultaneous form.

Note that whenever $V$ (or $V^{-1}$) has a simple eigenvalue/eigenvector decomposition, $V=P\Lambda P'$ where the columns of $P$ are the standardized (or normalized) eigenvectors of $V$ and $\Lambda$ is a diagonal matrix of the corresponding eigenvalues, then a simulation is easily obtained using $\mathbf{y}=\mu+P\Lambda^{1/2}P'\varepsilon$, where $\Lambda^{1/2}$ is a diagonal matrix of the square roots of the eigenvalues, and $\varepsilon$ is a vector of simulated independent random variables with mean zero and variance $\sigma^2$. There is no need to numerically find the Cholesky square root of $V$, as suggested by Haining, Griffith and Bennett (1983). Similarly, if $V$ has the simultaneous form $B'B$ where $B$ has a simple eigenvalue/eigenvector decomposition, $B=P\Lambda P^{-1}$, we can use $B^{-1}=P\Lambda^{-1}P^{-1}$ where $\Lambda^{-1}$ is a diagonal matrix of the inverses of the eigenvalues, so that $\mathbf{y}=\mu+B^{-1}\varepsilon$. This should be preferable to numerically inverting $B$, as suggested by Haining, Griffith and Bennett (1983), and reasonable for even moderately large $n$.

Another simple extension is to use the above forms for $V^{-1}$ as forms for $V$—finite dependence or moving–average models. Note that the finite dependence models are per-

fectly reasonable models of the covariance, although the attempt to derive them through a 'generating' mechanism in Cliff and Ord (1981, p. 150) is incorrect. The eigenvalues of $V$ will again be linear in the parameters. The quadratic form $\mathbf{e}'V^{-1}\mathbf{e}$ can be quickly computed as $\mathbf{f}'\mathbf{\Lambda}^{-1}\mathbf{f}$, where $\mathbf{f} = P\mathbf{e}$ and $V$ has the eigenvalue/eigenvector decomposition $P\mathbf{\Lambda}P'$. Simple results can be obtained for the information using, this time, trace$(V\frac{\partial V^{-1}}{\partial\beta})$ and trace$(V\frac{\partial V^{-1}}{\partial\beta}\ V\frac{\partial V^{-1}}{\partial\beta})$, so that essentially the same results are obtained as before. The rapid decay to zero of the covariances makes this form less attractive in many practical situations. It is also possible to combine the two forms, and still keep the same eigenvectors provided the $W_i$ matrices commute.

One other possible extension that does preserve some simplicity in the likelihood is the errors–in–variables formulation (Besag, 1977a). Essentially, this approach uses one of the above $V$ matrices, and adds to it $\alpha I$, so that var$(\mathbf{y}) = (V + \alpha I)\sigma^2$. This is useful when the sample correlation function does not appear to tend to 1 as the lag tends to 0, as when there is an extra independent error, such as measurement error, at each site. The quadratic form $\mathbf{e}'V^{-1}\mathbf{e}$ is found in the same way as when $V$ is specified.

Another extension for data on a rectangular lattice is to the separable processes, which can often be very easily fitted (Martin, 1990). These processes are somewhat restrictive in the range of possible covariance structures—in particular correlations must be reflection symmetric and decay exponentially—but the ease of specification and fitting makes them attractive whenever the assumption is reasonable. Simulation of a separable process is relatively easy provided the sites can be represented as a subset of a rectangular lattice, since on an $n_1$ by $n_2$ rectangular lattice $V$ is a Kronecker product of dispersion matrices of orders $n_1$ and $n_2$, and square roots of these matrices are usually easily found (Martin, 1990).

### 2.4. Comparison of models

From a data analytic point of view, it is important to be able to fit different models, and compare their fits. If the models are hierarchical, each more general than the previous, then standard statistical tests often can be used as a guide, although the theoretical justification is frequently lacking. Note that if $n$ is not too large, say less than 100, then provided care is taken to ensure numerical accuracy, any model can be fitted, whether or not $V$ has simple eigenvalues. If computing problems are ignored, it is easy, for regularly arranged sites or regions, to postulate a series of models, each more general than the previous one, with natural orderings of the neighbours (with, in general, different parameters for the two directions, but the possibility that the two parameters are the same). The next extension is to include the four diagonal neighbours (again, in general there would be different parameters for the two directions). For the subclass of separable processes, this procedure can actually be easily accomplished, because of the ease of fitting most processes (Martin, 1990).

It is much harder to say what should be done with irregular sites or regions. Note that despite the attempts at developing theory on a rectangular lattice, it is the irregular sites or regions that are most common for natural geographical data. Irregular sites can still be modelled with a particular dependence structure, and the form for $V$ deduced from it. Irregular regions cause the most problems. Modelling for data on irregular regions has tended to be extremely arbitrary. A particular form is postulated, such as the one–parameter conditional form of $V^{-1} = I - \beta W$, and then the weights $w_{ij}$ also are arbitrarily

chosen from a wide range of possibilities. These include functions of the distance between arbitrary centroids, and functions of the common boundary length (see Besag, 1975). In most applications the weights are more like parameters than known constants.

It is easy to criticise, but less easy to make constructive suggestions. My own belief is that for a given set of neighbours the elements of $W_i$ should be parameterized in terms of a small number of parameters, so that different choices of $W_i$ can be compared using standard statistical theory. This approach also has been suggested by Brandsma and Ketellapper (1979). Unfortunately, separability does not appear to have any relevance for data on irregular regions.

So far, I have discussed the modelling of the covariance structure. Of course there are other considerations. The mean structure can also be specified, and may be dogmatically specified, or chosen after examination of the data. The use of a trend surface to represent the large scale variation, as in Haining (1987), is fraught with difficulties when the dependence is also modelled through the covariance. Even a second–order stationary process can exhibit trend–like behaviour, so that the partition into a fixed trend and a random component is not clear. In addition, a parameterized trend surface is usually far too inflexible over a large region, and may require many parameters. Unless there is a clear planar trend over the sites, many statisticians would prefer to model the trend–like behaviour through differencing, or the use of the intrinsic processes introduced in geostatistics (see, for example, Journel and Huijbregts, 1978; and, the extension to intrinsic autoregressions of Künsch, 1987).

Another consideration is the distribution. Normality is almost always assumed, often implicitly. There is usually no check on normality; and, where there is, it usually consists of a univariate histogram of the original data. Apart from the necessity of correcting for the mean function, I have remarked before (Martin, 1983) on how misleading the marginal histogram can be for correlated data. The histogram of normal correlated data can often be multimodal and skewed. The need for correcting the significance values of a goodness–of–fit test for two–dimensional data was shown by Patankar (1954). My view (see Martin, 1990) is that some attempt should be made to obtain approximately uncorrelated residuals, on which standard tests of normality (for example, using as a test statistic the sample correlation coefficient associated with a normal probability plot) can be approximately used.

Note that, whilst it is important to check the distributional properties when simulating data, there is no point in checking the derived data, as suggested in Haining, Griffith and Bennett (1983). It is better and simpler to check that the original simulated data for $\varepsilon$ satisfy the assumptions of normality, constant variance, and independence.

## 3. Statisticians and geographers

There have been calls for statisticians to become involved in geographic research (see, for instance, Cliff and Ord, 1975; Bennett and Haining, 1985). My own involvement has been by a somewhat strange route. The published research in geography, which uses spatial statistics, held, and still holds, no particular interest to me. As I already have mentioned, I cannot see the point of most of it, and much is riddled with statistical errors. If that published research were concerned with applications in geography, I would probably have remained uninvolved. However, somewhat to my surprise I found that the vast majority of the publications that I had seen were not about geography at all. Although published in geographical journals, they were claiming to contain advances in statistical theory (by statistical theory I mean any

non–trivial mathematical manipulations performed within a statistical context).

I take care to seek opinions and comments, before submission of a paper, from others who may have more knowledge of the topics in the paper. I do not find it easy to get my papers published in statistical journals. I rigorously check any paper many times between the first draft and the final proof version for errors and misprints. I therefore was disturbed by the apparent ease with which these papers were able to appear in geographical journals, the kudos the authors received, the statistical errors the papers contained, and the lack of generality of reported results. Of course, there are poor papers in statistical journals, and there is good statistical work done by members of disciplines other than statistics. There also are many papers abusing statistics of which I am thankfully unaware. It seems unfortunate to me that non–statisticians who publish statistics, however flawed, are held in high esteem in their professions, whilst statisticians who publish statistics, however good, are seen as doing no more than trade. I can only say that I became aware of these "geographical" papers, and responded to them.

My first response (Martin, 1984) was a purely statistical work, but resulted from seeing some tentative beginnings by geographers. In this case I had already looked at the theory, but had not had the time (or the stimulus) to prepare it for publication (see my comments in Section 5). My second response (Martin, 1987) was a direct result of seeing published work by geographers. In that paper I attempted to straighten out what I saw as some muddled thought (see my comments in Section 4), and to correct some of the errors I had noted. To the credit of the geographic community, this response was published. However, this credit is somewhat diminished by the fact that I was not made aware of the reply (Griffith, 1988), nor given an opportunity to comment on it before or after publication. My third response (Martin, 1989a) also was a purely statistical work; it attempted to give the theory behind some numerical results obtained by geographers. This work was interesting in that it would never have occurred to me to obtain the results if it had not been for the other paper. Indeed, I still doubt that the results have any practical significance (see my discussion in Section 5). This present paper constitutes an invited fourth response, which I have used to develop my previous arguments.

I thus have four papers that may be of interest to geographers, but I still do not see myself as being, or wanting to be, a statistician interested in geographical research. I have reacted to geographical research, and may continue to do so. However, I have long thought that I would have nowhere near enough time, even if it were my only aim, to keep up with, and correct, statistical publications by geographers in the area of my research interest.

Perhaps fortunately, I am not kept in touch with current geographical research in this area, so that I only find out about it on the rare occasions that I am asked to referee a paper, or a paper appears in a journal that I notice. In connection with this, Bennett and Haining (1985) note, on geographic modelling, that I appeared unaware of "an extensive methodological discussion of these points ... in the social sciences and geography." I am happy to acknowledge my unawareness of this discussion. I feel that if geographers wish statisticians to become involved in their research, then the onus is on them to help make their research accessible to statisticians. It is quite unreasonable to expect a statistician to keep abreast of all the research in all the areas that use or abuse statistics.

My second point in Martin (1985) was that research purporting to contain statistical advances should be submitted to the scrutiny of statisticians. My hope was that better

refereeing would result in better papers, and that more statisticians would become aware of the research. I am pleased to see that there are now submissions to statistical journals; for example, Haining (1987, 1988), and Haining, Griffith and Bennett (1989). The drawback is that the statistical community must now share the blame for any criticisms of these papers. That I do have criticisms can be seen from my comments in Martin (1989a), and in this paper. As an additional example, I will mention that the reference to Matérn's lower bound of $-0.403$ (Haining, 1987, p. 464) is incorrect. This bound is for an isotropic process in continuous space, and has no relevance in discrete space. Even the concept of "isotropy" has little meaning or relevance for regional data, such as pixel measurements.

However, worse things are still occurring in quantitative geographical journals. For instance, Griffith (1987) gives (p. 72) an 11–line derivation of the simple result $E(\mathbf{x}) = (1-\rho)^{-1}\mu_\xi \mathbf{1}$ when $\mathbf{x} = (1-\rho C)^{-1}\xi$, $E(\xi) = \mu_\xi \mathbf{1}$, and $C$ has row sums equal to one (and does not achieve this result). The paper contains several errors, admittedly relatively minor once the necessary assumptions have been deduced. Also, the torus limit of $\rho_{\mathrm{g,h}}$ must equal the planar value (see Martin, 1986).

Poor published research does not represent a step forward, but several steps backwards. It sets a standard for subsequent publications, and deters those who might have worthwhile contributions to make. I wonder what the reaction of the geographic community would be if a statistician published, in a statistical journal, articles suggesting the present state of, and future research necessary in, geography. I wonder why people wishing to do research in statistics do not liase with statisticians who are expert in that area of research. I wonder why geographers do not concentrate on the many interesting geographic problems that are amenable to a sensible use of statistics.

As examples, I would like to see geographers investigating, by examining many relevant data sets, what models are reasonable for the sorts of data that arise in geographical applications. I would like to see investigations of different methods of predicting "missing values" in geographical situations in which an answer is actually required.

I also would like to see geographers who meet a statistical problem actively seeking the views and help of statisticians well before the stage of seeking publication. I am confident that many statisticians would be interested in such investigations and ready to help, when asked, with any necessary theoretical developments.

## 4. Boundary effects

There have been several papers discussing the "problem" of boundary values, and possible solutions to this "problem" (see Griffith, 1980, 1983, 1985, 1987; Griffith and Amrhein, 1983; Ord, 1981). In Martin (1987), written before I had seen Griffith (1987), I examined the "problem" and came to the conclusion that the published research was unsuccessful because the problem had not been sufficiently well defined, and the research had not considered problems that might be of interest. Some of the discussion is worth elaborating on here.

Much of the previous discussion on the topic of the "boundary value problem" appeared to assume that the problem was a well–defined one, and that a statistical solution to the problem was possible. I suspect that some of the confusion was due to the use of the term "boundary value," which has certain connotations in Applied Mathematics. In solutions to differential or difference equations, a general solution is found that depends on certain

initial conditions. Once these initial conditions are specified, the solution is unique. Also, a geographical boundary between sites is different from the boundary sites at which "boundary values" may occur.

However, the spatial statistical models used by geographers are of a different kind. I already have stated my view that these models are only descriptive of the covariance of the data, and have no meaning as generative models of the data. This still applies even when the model can be expressed in a "generative" or "causal" form. Thus the fact that these "generative" models include, for some "boundary" sites, dependence on unobserved sites, is irrelevant. Also irrelevant is any attempt to predict these boundary values in order to estimate parameters (Martin, 1987). Even the definition of what are boundary sites is unclear. One definition was used here in Section 2, but many others are possible.

A simple example is given by the first–order autoregression in one dimension, namely $x_i = \alpha x_{i-1} + \varepsilon_i$, where $\{\varepsilon_i\}$ is a sequence of uncorrelated random variables with zero mean and constant variance. With finite data $\{x_i\}$, $i = 1, 2, \ldots, n$, it appears that an assumption about $x_0$ needs to be made. However, for spatial data we could equally well "assume" that the data were "generated" from the right, with the model $x_i = \alpha x_{i+1} + \varepsilon_i$, so that now it appears that we need an assumption on $x_{n+1}$. If we write the model in conditional form, then $E(x_i|\cdot) = \beta(x_{i-1} + x_{i+1})$, where $E(x_i|\cdot)$ denotes the mean of $x_i$ given all other $x$s and $\beta = \alpha/(1 + \alpha^2)$, so that in this formulation we need assumptions about both $x_0$ and $x_{n+1}$. All these requirements concern what I term "exterior boundary" values; but it also is possible to formulate the problem in such a way that we need assumptions about $x_1$ and/or $x_n$—what I term "interior boundary" values.

A preferable way of considering the problem is through the covariance structure of $x_1, x_2, \ldots, x_n$. If the covariance matrix is $V$, then the only elements that change under the different assumptions on the first–order autoregression are the $(1, 1)$ and/or the $(n, n)$ elements of $V^{-1}$. We therefore essentially require assumptions about these. It also is possible to define the covariance structure of a larger set of $x$s, for instance $x_0, x_1, \ldots, x_{n+1}$, and then derive $V$ from this. This is discussed in Martin (1987), and eight different forms that have been suggested for $V^{-1}$ are given in Kunert and Martin (1987). Note that it is quite unnecessary to believe that the data were generated temporarily from an infinite past, or spatially from an infinite space, in order for $V$ to have the stationary form. Thus the use of a finite geographic region does not, of itself, rule out the use of the stationary $V$. However, doubtlessly it is true that when the region considered has natural boundaries, it may be reasonable to expect those sites on the geographic boundary to have different properties from those sites in the geographic interior.

Note that we can produce identical effects by including different assumptions on the variances of some of the "innovations." For instance, the assumptions that $x_0 = 0$ and $\text{var}(\varepsilon_1) = \sigma_\varepsilon^2/(1 - \alpha^2)$ lead to the stationary form for $V^{-1}$.

Thus the first possible boundary effect is that for a given model different "boundary" assumptions lead to different dispersion matrices. Since it is unlikely that even a large data set would allow statistical differentiation between mildly different "boundary" assumptions, the choice is largely a matter of convenience, unless there are strong prior arguments for one form over all other forms. The specification of a reasonable model for the "interior" sites usually will be more important than the specification of the precise form of the model to be used, although attempts should always be made to incorporate good prior information.

In more that one dimension there is another problem. The stationarity assumption for many statistical models does not lead to a $V$ or a $V^{-1}$ that can easily be numerically calculated. This means that exact likelihood is not feasible. Since it is, at least for the finite conditional or simultaneous schemes, the "boundary" sites that cause problems for $V^{-1}$, we have a second possible boundary effect, which is that some stationary models cannot, at present, sensibly be fitted by exact likelihood. This, together with results reported by Guyon (1982) on approximate likelihood not necessarily being $\sqrt{n}$–consistent, suggest that we should not attempt to fit the stationary form, but one of the other forms that is associated with different "boundary" assumptions.

The third possible boundary effect is that estimators may be biased, and that different "boundary" assumptions may reduce this bias. Whilst this is undoubtedly true, it is not at all clear in what way some boundary assumptions reduce the bias for estimators of parameters of $V$ for the same, or other, forms; nor is it clear whether changing the "boundary" assumptions is a good way to reduce estimator bias, or even whether the bias is large enough to cause concern.

The geographical discussion of "boundary effects" is greatly complicated by the lack of clear definitions of what are the effects that are causing concern, and what are the attempted solutions to them. It is difficult to comment on ambiguous or unclear work, since there is always the possibility that there are hidden assumptions that make the analysis correct. A step forward in research would be for all assumptions and aims to be clearly stated.

## 5. Missing Values

Although I had looked at the theory for estimation of the parameters of a spatial model when observations at some sites are not available as early as 1978, it was not until 1983, when I was told that geographers were working on a special case of the problem and were encountering difficulties, that I completed and wrote up the work (Martin, 1984). The results were circulated earlier, and mentioned in Martin (1983). This work covered in full generality the estimation of parameters using exact maximum likelihood for a Gaussian process. The work was referred to in several subsequent publications by geographers—see for example Bennett, Haining and Griffith (1984), Griffith, Haining, and Bennett (1985), and Haining, Griffith and Bennett (1984, 1989), although not always correctly, as I pointed out in Martin (1987).

There are two aspects to 'missing values'. One is the ability to use, with possibly minor modifications, known estimation methods on a given configuration of sites, usually a regular rectangular lattice. The results are of most usefulness when the covariance matrix, or its inverse, is of a known simple form on a given configuration, and m, the number of unobserved sites, is small. The other aspect is the prediction of the unobserved values.

Much of the original impetus for the interest of geographers appears to have been as a possible 'solution' to the 'boundary problem'. As I discussed in Martin (1987), and have commented again above, missing value techniques are quite irrelevant to the 'boundary problem'. Since then, more realistic problems have been proposed in which missing value techniques may be valuable. One is in the area of analysis of remotely–sensed data. For such data, it is possible to have unobserved sites for several reasons. Two possibilities are cloud cover when a passive sensor is used, and instrument malfunction. The former will result in unobserved data in all bands over a region on the ground, whilst the latter may result in the

loss of data on individual pixels or lines of pixels, and may only affect one band. Another situation in which 'missing data' arises is when the data contain possible outliers, that is observations that appear unusual amongst the others, or influential observations, that is data which have a large effect on the analysis. It may then be desirable to perform any analysis with such observations omitted. Some general theory on influence and residuals for known $V$ is in Martin (1989b). It may even be sensible to routinely calculate such 'leave–$k$–out statistics' as a diagnostic procedure—see the time series case in Bruce and Martin (1989). Procedures for dealing with an unknown mean and an unknown dispersion matrix require further investigation (Martin, 1989c).

Although the application to remotely–sensed data has been mentioned recently (Haining, Griffith and Bennett, 1989), the example given is unsatisfactory, in that there is little indication that the chosen covariance structure, the one–parameter first–order conditional process, is an adequate representation.

The main purpose of that paper appears to be to advance statistical theory on the loss in information (here meaning the Fisher information) when some sites are unobserved. The interest appeared to be on how the loss varies over different spatial configurations of the spatial sites. Results were obtained numerically for the special case of the one–parameter conditional process on a rectangular lattice.

The paper does not explain what the purpose of obtaining these results is. Since in any application the unobserved sites are given, and are not in the control of the investigator, it is difficult to see what the point is in comparing different configurations and different numbers of unobserved sites. However, if we assume that the results are of interest, it is easy to obtain theoretically much more powerful results. I have given the appropriate theoretical results in Martin (1989a). Special cases can easily be found—all the cases considered by Haining, Griffith and Bennett (1989) are also given in Martin (1989a). Many other special cases can also be considered, although it is only for the one–parameter conditional process that the formulæ are at all simple. Mrs. T. Krug at Sheffield has obtained formulæ for more sites and for the one–parameter first–order simultaneous model.

Assuming that there is an interest in these results, I shall outline some of them, elaborate on some of the details omitted in Martin (1989a), and include some new results. Assume that the $n$–vector of observations (strictly the random variable) is $\mathbf{u}$, with dispersion matrix $\mathrm{var}(\mathbf{u}) = V\sigma^2$, where $\sigma^2$ is a scale parameter and $V$ depends on $\beta$. Although it is possible to allow the mean to include trend and other fixed effects, I shall just discuss here the case of a constant mean, so that $E(\mathbf{u}) = \mu\mathbf{1}_n$, where $\mathbf{1}_n$ is an $n$–vector of ones. Also, it is easy to generalize to the case that $V$ is a function of the $q$–vector $\lambda$.

Assume also that data are unavailable at $m$ of the sites, and that $\mathbf{u}$ is permuted into $\mathbf{x}$, where the first $n-m$ elements of $\mathbf{x}$ are $\mathbf{y}$ and correspond to the observed sites, while the last $m$ elements are $\mathbf{z}$ and correspond to the unobserved sites. Similarly, let $\mathrm{var}(\mathbf{x})/\sigma^2 = V_{xx}$ be partitioned as

$$\begin{pmatrix} V_{yy} & V_{yz} \\ V_{zy} & V_{zz} \end{pmatrix}$$

and $V_{xx}^{-1}$ as

$$\begin{pmatrix} V^{yy} & V^{yz} \\ V^{zy} & V^{zz} \end{pmatrix} = \begin{pmatrix} V^{yx} \\ V^{zx} \end{pmatrix}$$

Then, in general, the loss in information on $\mu$ when $m$ sites are not observed is $\mathbf{c}'(V^{zz})^{-1}\mathbf{c}\sigma^2$, where $\mathbf{c} = V^{zx}\mathbf{1}_m$. For the one–parameter conditional process with $V^{-1} = I - \beta W$, we find that $V^{zz} = I - \beta W_{zz}$, where $W$ is partitioned similarly to $V$. Also, for any interior site of the first–order process on a rectangular lattice, $\mathbf{c} = \alpha\mathbf{1}_m$, where $\alpha = 1 - 4\beta$. Thus the information loss then becomes $\mathbf{1}_m'(I - \beta W_{zz})^{-1}\mathbf{1}_m$ times $\alpha^2\sigma^2$.

Exact formulæ can be obtained for this situation. Some general results, plus particular formulæ for the different configurations when $m = 1, 2, 3, 4$ are given in Martin (1989a). Note that when $m = 4$, one of the configurations was omitted by Haining, Griffith and Bennett (1989). For this case one configuration is

```
 xxx
  x
```

and the loss is $\frac{4+6\beta}{1-3\beta^2}$ times $\alpha^2\sigma^2$. This is intermediate in its loss between cases 3(d) and 3(e) of Haining, Griffith and Bennett (1989). For the values of $\beta$ they consider, 0.075, 0.150, and 0.225, the loss is 2.218, 0.840, and 0.063 respectively.

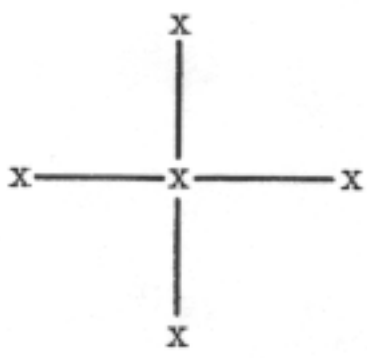
Exact results can be obtained for greater values of $m$, although the number of essentially different configurations increases rapidly with $m$, as does the difficulty in general in obtaining the formulæ for the elements of $(V^{zz})^{-1}$. The recursion given below is often useful. However, good approximations are also possible. Provided that $|\beta|$ is not too large, the information loss on $\beta$ for $m$ missing sites is approximately $\{m + 2m_1\beta + 2(m_1 + m_2)\beta^2\}$ times $\alpha^2\sigma^2$, where $m_1$ is the number of 'links' of length one among the missing sites, and $m_2$ is the number of 'links' of length two. These links are found using the usual city–block metric.

As an example, consider the case $m = 5$. There are several cases in which not all the sites are joined, but I will only consider those four cases in which all sites are connected.

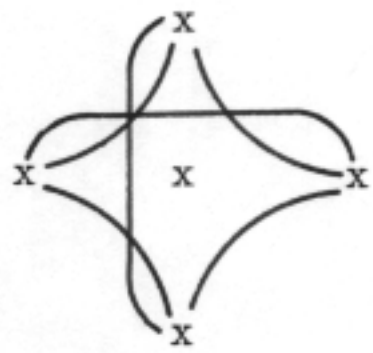
| *Case 1* | *Case 2* | *Case 3* | *Case 4* |
|---|---|---|---|
| x<br>xxx<br>x | xx<br>xxx | x<br>xxxx | xxxxx |
| (4,6) | (5,6) | (4,4) | (4,3) |

The pair of numbers associated with each configuration are $(m_1, m_2)$. The figure below shows how these are obtained for Case 1.

The 4 links of length 1



The 6 links of length 2



In general, the exact result requires the inversion of $V^{zz}$. However, using results on partitioned matrices, it is possible to obtain recursively formulæ for the information loss.

Partition the $m$ sites into $m-1$ and 1, so that

$$V^{zz} = \begin{pmatrix} A & \mathbf{b} \\ \mathbf{b}' & d \end{pmatrix}$$

where $A$ is an $m-1$ square matrix, $\mathbf{b}$ is an $m-1$ vector and $d$ is a scalar. Then

$$(V^{zz})^{-1} = \begin{pmatrix} A^{-1} & \mathbf{0} \\ \mathbf{0}' & 0 \end{pmatrix} + \frac{1}{d - \mathbf{b}'A^{-1}\mathbf{b}} \begin{pmatrix} -A^{-1}\mathbf{b} \\ 1 \end{pmatrix} \cdot \begin{pmatrix} -A^{-1}\mathbf{b} \\ 1 \end{pmatrix}'.$$

Thus, if $\mathbf{c} = \alpha \mathbf{1}_m$, which is so for interior points for conditional and simultaneous schemes, then the loss in information on $\mu$ is the loss for the $m-1$ sites,

$$(\mathbf{1}'_{m-1} A^{-1} \mathbf{1}_{m-1}) \alpha^2 \sigma^2,$$

plus $\alpha^2\sigma^2$ times $\frac{(1-\mathbf{b}'A^{-1}\mathbf{1}_{m-1})^2}{d-\mathbf{b}'A^{-1}\mathbf{b}}$. This latter additional term can be very easy to calculate if the extra point is carefully selected. For instance, for the one–parameter first–order conditional process, if in Case 1 above the centre point is chosen, then $A = I_4$ and $\mathbf{b} = -\beta \mathbf{1}_4$. Thus the loss is $\alpha^2\sigma^2$ times $4 + \frac{(1+4\beta)^2}{1-4\beta^2}$.

Note that the above result can easily be extended to $m$ sites being partitioned into $m - m'$ and $m'$.

Results for the one–parameter first–order simultaneous model can be obtained, but are nowhere near as simple. There are several reasons for this. Firstly, $(V^{zz})^{-1}$ does not have the form $(I - \beta W_{zz})'(I - \beta W_{zz})$. Secondly, because $V^{-1}$ has non–zero terms for $(1,1)$ and $(2,0)$ lags, there are more cases to consider. For example, when $m = 2$ there are 3 different configurations, and when $m = 3$ there are 12. The numbers for the conditional process are 2 and 3 respectively.

Thus, when $m = 2$ the four cases are:

| *Case 1* | *Case 2* | *Case 3* | *Case 4* |
|---|---|---|---|
| immediate neighbours | lag 2 neighbours | diagonal neighbours | all other configurations |
| xx | x·x | x·<br>·x | |

For the conditional process Cases 2 and 3 are included with Case 4. An interesting point with the simultaneous process is that when $\beta > 0$, the smallest loss is not associated with Case 4, but with Case 3. This follows from the element of $V^{-1}$ associated with diagonal neighbours being $2\beta^2$, which is positive. The next smallest is for Case 2, as the element of $V^{-1}$ associated with lag 2 neighbours is $\beta^2$, which is also positive.

So far I have considered the easier case of the loss of information on $\mu$. The loss of information on $\beta$ was also considered in Haining, Griffith and Bennett (1989), and Martin (1989a). This is more complicated for several reasons. Firstly, there are more configurations to consider, and secondly, the formulæ involve both $V^{zz}$ and $(V^{zz})^{-1}$. Because of the second point, the formulæ depend not just on the configuration of sites, but also on the actual positions of the sites. However, provided attention is restricted to interior points of

the *stationary* process, then the result only depends on the configuration. Although Haining, Griffith and Bennett (1989) do evaluate their results for the stationary process, it appears that they are also assuming $V^{-1} = I - \beta W$.

Note that if the loss of information on $\beta$ is being considered because of an interest in $\text{var}(\hat{\beta})$, then the information required is that for $\beta$ conditional on $\sigma^2$, which was considered in Section 2.

The formulæ for the loss of information on $\beta$ are most easily obtained by using the missing information principle of Orchard and Woodbury (1972). I take their principle to be their equations (2.13) and (2.15); that is, the use of the expectation with respect to $x$ of the conditional likelihood of $z|y$. Setting the mean $\mu$ to 0, since its value does not affect the information on $\beta$, the distribution of $z|y$ is Normal with mean $-(V^{zz})^{-1}V^{zy}\mathbf{y}$ and dispersion matrix $(V^{zz})^{-1}\sigma^2$.

Since the second differential with respect to $\beta$ of both $V^{zy} = -\beta W_{zy}$ and $V^{zz} = I - \beta W_{zz}$ is 0, the second differential of the conditional log likelihood becomes

$$-\frac{1}{2}\frac{\partial^2 \ln|V^{zz}|}{\partial\beta^2} + \frac{1}{2\sigma^2}\frac{\partial^2\{\mathbf{y}'V^{yz}(V^{zz})^{-1}V^{zy}\mathbf{y}\}}{\partial\beta^2}.$$

The first term can be evaluated as before. Taking the expectation over $\mathbf{y}$ of the second term gives

$$\frac{1}{2}\text{trace}\left[V_{yy}\frac{\partial^2\{V^{yz}(V^{zz})^{-1}V^{zy}\}}{\partial\beta^2}\right].$$

Now, $V^{yz}(V^{zz})^{-1}V^{zy} = \beta^2 W_{yz}(I - \beta W_{zz})^{-1}W_{zy}$, and so its second differential with respect to $\beta$ is $2W_{yz}(I - \beta W_{zz})^{-3}W_{zy}$ {compare this with the second differential with respect to $x$ of $x^2/(1-ax)$, which is $2/(1-ax)^3$ }. Then using $(V^{zz})^{-1}V^{zy}V_{yy} = -V_{zy}$ [see Martin (1984)] and $V_{zy}V^{yz} + V_{zz}V^{zz} = I$, it follows that this expectation becomes

$$\beta^{-1}\,\text{trace}\{(V^{zz})^{-2}V_{zy}W_{yz}\} = \beta^{-2}\,\text{trace}\{(V^{zz})^{-1}V_{zz} - (V^{zz})^{-2}\}.$$

The second term here can be evaluated as before, using the sum of squares of the elements of $(V^{zz})^{-1}$ for small $m$. The first term involves $V_{zz}$, as stated above. For small $m$, exact formulæ can be found (Martin, 1989a). Again, approximate formulæ can be derived—see Martin (1989a).

Also, these formulæ can be extended to larger $m$, and to other processes. Although the mathematics is interesting, I feel that further theory should be justified by practical needs. Which models are reasonable for a given application needs to be discovered, as well as why it is of interest to know the loss in information.

## 6. Conclusion

I have given a personal view of some of the spatial statistical models used in geography, and of some of the publications concerning these models. I hope that the papers in this volume will lead to an improvement in modelling, and in published research. If geographers stimulate statisticians by presenting problems of practical interest, then valuable joint research should result.

If my comments have been unduly negative, I should say that I have been heartened by the apparent willingness with which geographers accept criticism of their mistakes, although I would prefer that the mistakes were not made. I should also emphasize that similar comments could be made about workers in other disciplines, or even within the statistical community. I have tried to ensure there are no mathematical or statistical errors in this paper, and will endeavour to correct any that I notice subsequently or that are bought to my attention.

## 7. References


Bennett, R. J. and Haining, R. P. (1985) Spatial structure and spatial interaction: modelling approaches to the statistical analysis of geographical data (with discussion). *Journal of the Royal Statistical Society A* **148**: 1-36.

Bennett, R. J., Haining, R. P. and Griffith, D. A. (1984) The problem of missing data on spatial surfaces. *Annals of the Association of American Geographers* **74**: 138-56.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society B* **36**: 192-236.

Besag, J. (1975) Statistical analysis of non–lattice data. *The Statistician* **24**: 179-95.

Besag, J. (1977a) Errors–in–variables estimation for Gaussian lattice schemes. *Journal of the Royal Statistical Society B* **39**: 73-8.

Besag, J. (1977b) Efficiency of pseudolikelihood estimation for simple Gaussian fields. *Biometrika* **64**: 616-8.

Besag, J. (1985) Contribution to the discussion of a paper by R. J. Bennett and R. P. Haining. *Journal of the Royal Statistical Society A* **148**: 31.

Besag, J. and Moran, P. A. P. (1975) On the estimation and testing of spatial interaction in Gaussian lattice processes. *Biometrika* **62**: 555-62.

Brandsma, A. S. and Ketellapper, R. H. (1979) A biparametric approach to spatial autocorrelation. *Environment and Planning A* **11**: 51-8.

Bruce, A. G. and Martin, R. D. (1989) Leave–k–out diagnostics for time series (with discussion). *Journal of the Royal Statistical Society B* **51**: 363-424.

Cliff A. D. and Ord, J. K. (1975) Model building and the analysis of spatial pattern in human geography (with discussion). *Journal of the Royal Statistical Society B* **37**: 297-348.

Cliff A. D. and Ord, J. K. (1981) *Spatial Processes*. Pion: London.

Gasim, A. A. (1988) First–order autoregressive models: a method for obtaining eigenvalues for weighting matrices. *Journal of Statistical Planning and Inference* **18**: 391-8.

Griffith, D. A. (1980) Towards a theory of spatial statistics. *Geographical Analysis* **12**: 325-39.

Griffith, D. A. (1983) The boundary value problem in spatial statistical analysis. *Journal of Regional Science* **23**: 377-87.

Griffith, D. A. (1985) An evaluation of correction techniques for boundary effects in spatial statistical analysis: contemporary methods. *Geographical Analysis* **17**: 81-8.

Griffith, D. A. (1987) Towards a theory of spatial statistics: another step forward. *Geographical Analysis* **19**: 69-82.

Griffith, D. A. (1988) A reply to R. Martin's "Some comments on correction techniques for boundary effects and missing value techniques". *Geographical Analysis* **20**: 70-5. Correction (1989) **21**: 359.

Griffith, D. A. and Amrhein, C. G. (1983) An evaluation of correction techniques for boundary effects in spatial statistical analysis: traditional methods. *Geographical Analysis* **15**: 352-60.

Griffith, D. A., Haining, R. P., and Bennett R. J. (1985) Estimating missing values in space–time data series. In *Time Series Analysis: Theory and Practice 6*, Eds. O. D. Anderson, J. K. Ord and E. A. Robinson, pp. 273-82. North Holland : Amsterdam.

Guyon, X. (1982) Parameter estimation for a stationary process on a d–dimensional lattice. *Biometrika* **69**: 95-105.

Haining, R. (1987) Trend–surface models with regional and local scales of variation with an application to aerial survey data. *Technometrics* **29**: 461-9.

Haining, R. (1988) Estimating spatial means with an application to remotely sensed data. *Communications in Statistics—Theory and Methods* **17**: 573-97.

Haining, R., Griffith, D. A. and Bennett R. J. (1983) Simulating two–dimensional autocorrelated surfaces. *Geographical Analysis* **15**: 247-55.

Haining, R., Griffith, D. A. and Bennett R. J. (1984) A statistical approach to the problem of missing spatial data using a first–order Markov model. *Professional Geographer* **36**: 338-45.

Haining, R., Griffith, D. A. and Bennett R. J. (1989) Maximum likelihood estimation with missing spatial data and with an application to remotely sensed data. *Communications in Statistics—Theory and Methods* **18**: 1875-94.

Journel, A. G. and Huijbregts, Ch. (1978) *Mining Geostatistics*. Academic Press: London.

Kunert, J. and Martin, R. J. (1987) Some results on optimal design under a first–order autoregression and on finite Williams type II designs. *Communications in Statistics—Theory and Methods* **16**: 1901-22.

Künsch, H. (1987) Intrinsic autoregressions and related models on the lattice $Z^2$. *Biometrika* **74**: 517-24.

Mardia, K. V. and Marshall, R. J. (1984) Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71**: 135-46.

Martin, R. J. (1983) A review of continuous valued spatial processes. Paper presented at the Institute of British Geographers Quantitative Methods Study Group Conference on Advances in Applied Spatial Analysis, Sheffield.

Martin, R. J. (1984) Exact maximum likelihood for incomplete data from a correlated Gaussian process. *Communications in Statistics—Theory and Methods* **13**: 1275-88.

Martin, R. J. (1985) Contribution to the discussion of a paper by R. J. Bennett and R. P. Haining. *Journal of the Royal Statistical Society A* **148**: 29-30.

Martin, R. J. (1986) A note on the asymptotic eigenvalues and eigenvectors of the dispersion matrix of a second–order stationary process on a d–dimensional lattice. *Journal of Applied Probability* **23**: 529-35.

Martin, R. J. (1987) Some comments on correction techniques for boundary effects and

missing value techniques. *Geographical Analysis* **19**: 273-82.

Martin, R. J. (1989a) Information loss due to incomplete data from a spatial Gaussian one–parameter first–order conditional process. *Communications in Statistics—Theory and Methods* **18**: 4631-45.

Martin, R. J. (1989b) Leverage, influence and residuals when errors are correlated. Research report no. 326/89, Department of Probability and Statistics, University of Sheffield.

Martin, R. J. (1989c) Contribution to the discussion of a paper by A. G. Bruce and R. D. Martin. *Journal of the Royal Statistical Society B* **51**: 414.

Martin, R. J. (1990) The use of time–series models in the analysis of agricultural field trials. *Communications in Statistics—Theory and Methods* **19**: 55-81.

Orchard, T. and Woodbury, M. A. (1972) A missing information principle: theory and applications. In *Proceedings of the 6th Berkeley Symposium 1*, Eds. L. M. Le Cam, J. Neyman and E. L. Scott, pp. 697-715. University of California Press: Berkeley.

Ord, J. K. (1975) Estimation methods for models of spatial interaction. *Journal of the American Statistical Association* **70**: 120-6.

Ord, J. K. (1981) Towards a theory of spatial statistics: a comment. *Geographical Analysis* **13**: 86-93.

Patankar, V. N. (1954) The goodness of fit of frequency distributions obtained from stochastic processes. *Biometrika* **41**: 450-62.

Ripley, B. D. (1988) *Statistical inference for spatial processes*. Cambridge University Press: Cambridge.

Whittle, P. (1954) On stationary processes in the plane. *Biometrika* **41**: 434-49.

# DISCUSSION

## "The role of spatial statistical processes in geographic modelling"

## by R. J. Martin

Statistical models of spatial dependence have been used quite commonly in geographic research. In his presentation, the author both reviews and comments on their use. He further takes up the topics of boundary effects and missing values, attempting to clarify the former and giving some new results on the latter.

The paper starts with a substantial section (Section 2) on models that includes mathematical details on their fitting as well as the author's view on how a modelling exercise should be justified. Computations for fitting first–order models are thoroughly discussed, and the author gives convenient, simple forms for the Fisher information matrix of the parameters, both for the conditional and the simultaneous versions. Then restrictions of first–order models are developed, leading to a review of selected extensions, still using contiguity matrices, which would allow some form of non–isotropy for the dependence or an increase of its range.

The author omits from his review a class of models where the covariance between sites $\mathbf{i}$ and $\mathbf{j}$ is not modelled through arbitrarily defined contiguity matrices, but rather has a parametrised functional form. This class of models rarely has been used in geographical studies, although it has received attention in the statistical, epidemiological and geostatistical literature (Ripley, 1988; Cook and Pocock, 1983; Mardia and Marshall, 1984; Vecchia, 1988); it would be interesting to see applications of this model in geography.

Section 2 starts and ends with some methodological considerations about justification and comparison of models. This is certainly an important area that, until now, has not received enough attention, and the author's emphasis and suggestions are most welcome. I would add that the strategy used to justify or compare different models depends upon whether the aim of the modelling exercise is explanatory, for forecasting purposes, or to be used in a generalised regression framework.

Section 3 recounts some of the arguments that have arisen between the author and geographers concerning the application of statistics. Although part of this section may be difficult to follow for a reader who does not have all of the quoted papers on hand, the author develops a convincing case on the desirability of constructive discussions between statisticians and geographers that should benefit both professions. It is in everyone's interest to avoid incorrect uses of statistics. Discussions of this kind often stimulate new research.

Section 4 is of a general nature and argues for a precise definition of what is called the boundary value problem, whether it influences the dispersion matrices or the bias in estimators. In contrast, the final section gives some results on the loss of information due to missing values on the mean $\mu$ and the parameter $\beta$ of the first–order conditional or simultaneous process. Since the author wanted to expand on some new results, this section is the least self–contained. Useful approximations for the loss of information are given when the number of missing sites becomes large.

In this paper the author presents original and thoughtful considerations on the use of spatial statistics in geography, emphasising throughout the need to link theoretical developments (like those arising for missing values) to relevant examples, and to relate models to geographical problems.

## References

Cook, D., and S. Pocock. (1983) Multiple regression in geographic mortality studies with allowance for spatially correlated errors. *Biometrics*, **39**: 361-371.

Mardia, K., and R. Marshall. (1984) Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, **71**: 135-146.

Ripley, B. (1988) *Statistical inference for spatial processes*. Cambridge: Cambridge University Press.

Vecchia, A. (1988) Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society*, **50B**: 297-312.

*Sylvia Richardson*, INSERM

## *PREAMBLE*

*The three foundations of learning:*
*Seeing much, suffering much, and studying much.*

Catherall

*The terms "traditional" and "classical" refer to statistics governed by very restrictive assumptions, and cover much of statistical theory and practice prior to the widespread advent of numerically intensive computing capabilities. It was increased computing power that enabled statisticians to gradually develope abilities and skills that can distinguish between tenable and untenable assumptions. Hence to its benefit spatial statistics has seen much that has gone before it. Outliers—leverage points—influence functions—these and other diagnostics have been devised in order to better assess, deal with and understand statistical assumption violations. But what do these diagnostics reveal about geo-referenced data? Nothing perhaps; everything perhaps. Wartenberg suggests that geo-referenced data analyses may have suffered much from a lack of comprehending what such diagnostic tests tell about spatial data. The purpose of this paper is to help determine which aspects of spatial patterns and individual geo-referenced observations contribute most to spatial autocorrelation, based upon these standard diagnostic statistics. Upton, while questioning some of the specifics, agrees with the thrust of Wartenberg's work, supporting the contention that there is a need for methods to conduct exploratory spatial data analysis. In Upton's opinion, this work helps to address a new and fruitful research area in spatial statistics, and accordingly he views it as one step in developing exploratory spatial data analysis. Indeed, much studying remains!*

The Editor

# Exploratory Spatial Analyses: Outliers, Leverage Points, and Influence Functions

Daniel Wartenberg *

*Department of Environmental and Community Medicine, Robert Wood Johnson Medical School, 675 Hoes Lane, Piscataway, NJ 08854, U. S. A.*

**Overview**: Exploratory data analysis provides quick, easy to calculate summaries of data that convey much of the information relevant to interpretation of a sample. While development of exploratory methods in traditional applications has been extensive over the past decade, development of analogous methods that exploit the spatial relationships among observations have lagged. This paper presents three approaches for exploratory tools for use with spatial autocorrelation analysis that emphasize spatial aspects of the data.

The first approach proposes a method for detecting outliers called local trend surface residuals (LTSR). For each observation, a trend surface is fit to neighbors of a point and the difference between the observed value and the prediction based on the trend surface is evaluated. Highly deviant values are termed outliers. The method can detect spatial outliers, points that are outliers with respect to their neighbors while not being outside the overall range of observations. The second approach evaluates the location of observations relative to random placement of observations. Isolated points should not be considered in the same context as clustered points. The third approach develops influence functions for spatial autocorrelation analysis. This approach evaluates the importance of each observation in the determination of the value of the spatial autocorrelation coefficient.

These methods are applied to simulated data and to one real data set, the rate of population growth in Ireland 1926–1961. Results demonstrate the utility of these approaches for identifying unusual values and for characterizing the basic structure in a data set.

## 1. Introduction

The need for quick, informative and easy to perform descriptive methods for the analysis of data have catapulted exploratory data analysis (EDA) methods into the forefront of statistical development over the past 10 to 20 years. Development of analogous methods for the description of spatial data have lagged considerably, although recent efforts in this direction hold promise. This paper considers some of these developments, proposes a context for these approaches, and presents some recent suggestions for additional consideration.

The motivation and direction for work in exploratory, descriptive analysis owes much of its development, insight and widespread acceptance to the pioneering and innovative work of John Tukey (Olmstead and Tukey, 1947; Tukey, 1949; Tukey, 1951; Tukey, 1977; Mosteller

* I thank Daniel Griffith for his time, patience and useful suggestions that led to improvements in the original manuscript. This research was supported, in part, by funds from the Comprehensive Environmental Response, Compensation, and Liability Act trust fund through Cooperative Agreement No. U50/CCU101045-04, from the Centers for Disease Control and the Agency for Toxic Substances and Disease Registry, U. S. Public Health Service.

and Tukey, 1977). By showing that quick did not mean inaccurate, and that approximate did not mean without statistical foundation, Tukey was able to use EDA methods to lead the development of a field of statistical investigation that ran counter to many statisticians' ways of thinking; it was easy to do, and results were instantaneously apparent and heuristically pleasing. One did not need years of statistical training to appreciate the importance of numerical results. And yet, the methods are founded in rich statistical traditions. Tukey, among others, provided much of the rigor and statistical underpinning necessary for the discipline to gain acceptance.

Following Tukey's development of "quick and dirty" methods for evaluation of data, others began looking at observations that were inconsistent with a data set (outliers—see Hawkins, 1980; Barnett and Lewis, 1984), observations that had a disproportionate effect on a summary statistic or result (leverage points—see Belsey, Kuh and Welsch, 1980; Cook and Weisberg, 1982; Atkinson, 1985) and observations whose omission would result in a vastly different summary statistic or result (influential points—see Belsey, Kuh and Welsch, 1980; Cook and Weisberg, 1982; Atkinson, 1985). These methods allow for characterizations of a data set that go beyond simple statistical summaries. They relate information about consistency of observations, stability of parameter estimates and homogeneity of observations. After 20 years of development, the field of EDA is an accepted branch of statistics, and results of EDA analyses are reported routinely along side more traditional statistical summaries.

Development of EDA methods specifically designed for geographical data have lagged behind developments of more general EDA approaches. As the EDA methods have become popular, some geographers have incorporated EDA evaluations into geographical studies, but only as formulated in the aspatial context (*e. g.*, Unwin and Wrigley, 1987a; 1987b). That is, geographers, like other data analysts, look for outlying observations in a data set using methods that ignore geographic information and consider only the aspatial variate values. Or, geographers conducting regression analyses use leverage and influence curves to evaluate regression results without consideration of each observation's neighbors. While this approach is useful, it neglects the additional information available to geographers, namely, the spatial location of each observation and the variate values at each observation's neighbors.

To exploit this additional information, I will present a few methods that emphasize the geographic components of a data set in the EDA spirit. Most previous work in this area has taken place in the field of geostatistics, principally directed at enhancing the robustness of geostatistical predictions or kriging (*e. g.*, Cressie and Hawkins, 1980; Diamond and Armstrong, 1984; Hawkins and Cressie, 1984; Dowd, 1984; Omre, 1984; Brooker, 1986; Bardossy, 1988). Considerably less attention has been devoted to descriptive analyses of the data, process–oriented interpretations of correlograms (variograms), and identification and characterization of contamination or error variance (however, see Cressie, 1984; 1986; Cressie and Chan 1989; and Griffith, 1988 for work in this direction).

The goal of exploratory spatial analysis is to provide a quick and meaningful summary of both the spatial and aspatial characteristics of a data set. That is, we want ways to describe unusual observations, trends, patches, clusters, and systematic pattern in our data. In such descriptions, we must consider location, neighbors, observed values and the covariance of these characteristics. I present a few ways of decomposing the spatial structure of a given data set. The goal is to determine which aspects of the data contribute most to the

spatial autocorrelation structure of the data, and to seek some substantive interpretation of this structure. To begin, I undertake a preliminary outlier assessment to find individual observations that are unlike all others. I consider these in an aspatial as well as a spatial context, and evaluate whether or not any localities are extremely isolated. Upon detecting outliers, I remove them from the data set. Then, I evaluate their spatial autocorrelation structure and the contribution each observation makes to the overall spatial pattern.

## 2. The Problem: Detecting Unusual Observations

Unusual observations, or outliers, are troublesome data points for most statistical analyses. By definition, outlier observations are data points that stand apart from the rest, those that are extremely large or extremely small when compared to the distribution of all other observations, the definition of the term "extremely" taking on different meanings for different investigators and purposes; generally speaking, it refers to that which appears to be inconsistent with the rest of the data (Barnett and Lewis, 1984). Outliers are troublesome in that they may unduly influence summary statistics or other characterizations of a data set. Since they are uncharacteristic of the rest of the data set, by definition, summary statistics that reflect the few outliers rather than values of the other observations can be misleading.

Outliers can be defined operationally in terms of a variety of properties (Welsch, 1985). Three are used most frequently. First, we can define an outlier as an observation that is substantially greater or less than all other observations, as noted above. Second, we can define an outlier as an observation that contributes disproportionately to a summary statistic, a leverage point. Third, one can define an outlier as an observation whose deletion would effect a disproportionate change in the statistic being estimated or evaluated, an influential point. In discussing leverage points in regression analysis, Hoaglin and Welsch (1978) suggest that individual elements of the "hat matrix" (which is based entirely on the independent variables) should not deviate too far from a balanced design (each point having an equal influence) or else a few observations could disproportionately dominate the calculation of the regression coefficients. That is, in concert with the dependent variable, they could control the value of the regression coefficients to the near exclusion of all other observations. Influence points in regression, while similar in concept to the leverage point, reflect the covariance between independent and dependent variables in addition to each observation's consistency with the rest of the observations. (That is, an outlier for the independent variables will not have a large effect on the regression coefficients if the dependent variable is near the mean. Such a point would have high leverage but low influence. But to be influential, a point must have moderate to large leverage.)

Investigators are interested in influence and leverage because traditional analyses of data sets with high leverage or influence points may lead to misinterpretation if these properties are not noticed. Generally speaking, one assumes implicitly that a summary statistic reflects properties of an entire data set. For data sets with high influence points, the summary statistic may disproportionately reflect this one data point in preference to all others. While this information is important, it must be put into context. Investigators not only want to know about this data point, but also want to know about structure in the rest of the data set. Once identified, influential points can be removed, summary statistics calculated and both sets of results (with and without the influential point) should be reported.

Investigators have proposed a variety of ways of analyzing data sets with outliers. Most simply, one can detect outliers by *a priori* evaluation without consideration of which statistical analyses will be undertaken later. This would identify points that are unusual in a distributional sense, and give one an indication of the variability and consistency of one's data. The outliers could be removed from the data set and then more traditional analyses could be undertaken with the reduced data set. The results without the outlier can be interpreted on their own as well as in comparison to similar analyses with all the data points. One must, however, be wary of interpreting the results for the analysis with the outlier included in terms of underlying pattern or process, as the results reflect that status of the outlier in disproportion to the other data points.

A more rigorous approach to accommodating outliers is to develop statistical methods that are insensitive to or diagnostic of the influence of individual outlier observations. Such methods are called robust statistics (*e. g.*, Huber, 1981; Hampel *et al.*, 1986). Robust statistical methods either identify data points that are unlike the others (outliers) or have undue influence or leverage on a summary statistic of interest, or provide results and summaries that are insensitive to individual observations that may be aberrant. These methods are different from the *a priori* methods in that they tend to evaluate the effect of each data point on some statistical summary that is of interest to the investigator, identify unusual values and provide results that do not allow individual data points to dominate the analysis.

To clarify these concepts, consider the following example. Given a data set for a regression analysis, one can look for outlying values for each of the independent variables, in turn, and also separately for the dependent variable. This would correspond to the *a priori* considerations I described first. Then one could evaluate the independent variables as a set of variables separate from the dependent variable, detecting observations that could have disproportionate effects on the regression coefficients due to individual observation distances from the mean values of these variables. These are called leverage points. Then, one could determine which observations have a large impact or influence on the regression itself in two ways. One could evaluate quantitatively how much each point contributes to each regression coefficient. And, one could calculate the regression coefficients for the entire data set except one point. One could do this calculation repeatedly, omitting each data point one at a time. The scaled difference between the regression coefficient for all data points and that with a given point removed is called the influence of that point.

Usually, points identified as unusual in terms of leverage also will be identified as unusual in terms of influence. However, an outlying dependent variable observation might not affect a regression greatly if the corresponding values of the independent variables were inconsequential. In terms of the analysis of the overall data set, after unusual points are identified, such observations can be culled from the data set to remove their influence entirely, and one can compare analytic results with and without the outliers.

When analyzing geographic data (or any data set in which the observations are not independent of each other), one may encounter even more types of outlier observations. First, one may find an observation whose value would be considered unusual in any data set, that which is substantially different from all other observations, which I call an *aspatial* or *global outlier*. These are the same as the outliers discussed above and identified *a priori*. Or, one may find an observation that is not larger or smaller than all other observations, lying well within the range of variation of other values. This observation, however, may be

very different from all those observations nearby it, and this is what I call a *spatial* or *local outlier*. The concept of near or local is critical to this definition. For regional or quadrat data, near may mean contiguous. For point data, it may mean that the distance between the point in question and a neighbor is within a specified threshold. Or, neighborliness may be defined by the connections of a Delaunay tessellation. In all of these, the definition of neighborliness is in the hands of the investigator.

Time series analysts also undertake evaluations of the similarity of neighbors, although temporal data has a natural ordering defining neighbors. That is, in temporal analyses one studies sequential observations with neighbors being defined as the previous and successive observations. Temporal data also can have local outliers. These would be observations that are different from values near them in time, but not outside the range of observed values. Because of the possible dependencies of neighboring values, detecting outliers in time series is more complicated than for the case of independent observations (although less complicated than for spatial data). Fox (1972) defines two types of outliers in time series: (1) observation errors that affect single data points only, and (2) innovation errors that affect many nearby points. Denby and Martin (1979), Abraham and Box (1979) and Muirhead (1986) further emphasize this distinction and argue that different types of outliers warrant different types of adjustments. Others have focused on model fitting, filtering (Kleiner *et al.*, 1979) and influence functions (Künsch, 1984). Putterman (1988) considers data with autocorrelated errors by modifying diagnostic indices for data of independent observations, and shows the importance of considering outliers in the evaluation of data with first–order dependencies.

Two–dimensional dependencies add further complications. Some aspects are considered explicitly in geostatistical and geographical analyses, but others are omitted. Even when conducting analyses of geographic data, most investigators who have subjected their data to outlier tests have done so without consideration of where the values occur, even though locally extreme values (those substantially different from a group of neighboring observations) may be as problematic as globally extreme values. Investigators have discussed the robustness of geostatistical methods and the influence that an individual observation can have on the geostatistical prediction methods produce (*e. g.*, Cressie and Hawkins, 1980; Hawkins and Cressie, 1984; Dowd, 1984; Omre, 1984; Diamond and Armstrong, 1984; Brooker, 1986; Bardossy, 1988). However, even in these contexts little attention has been paid to identifying or characterizing outliers.

Cressie (1984; 1986) develops some methods for detecting outliers when kriging or calculating variograms. Cressie is particularly concerned with the position of unusual values and their neighbors, and develops a number of tools to detect troublesome observations. Most of his methods are designed for regularly spaced or gridded data. When confronted with irregularly spaced data (*e. g.*, Cressie and Read, 1989), he superimposes a grid and assigns observations to the nearest node. While a practical solution, this procedure may distort small scale spatial structure. Further, it deflates the relative importance of individual observations that are geographically clustered.

Unwin and Wrigley (1987a; 1987b) and Griffith (1988) investigate leverage of geographic data. Unwin and Wrigley consider the case of trend surface analysis, which is the regression of an independent variable on powers and cross products of geographic coordinates (Chorley and Haggett, 1965). Using traditional indices of aspatial leverage (*e. g.*, Belsey, Kuh and Welsch, 1980; Cook and Weisberg, 1982) on geographic data, they show that observations

near the edges of a study area or isolated observations often have disproportionate effects of trend surface regressions.

Griffith (1988) extends the traditional diagnostic indices for an aspatial framework to a spatial framework by modifying the indices to incorporate the non–independence of observations in a generalized least squares (rather than ordinary least squares) regression. He shows that evaluation of unusual values is of considerable importance in regressions using spatial data. Failure to do so can lead to erroneous conclusions about the presence or absence of outliers.

Additionally, it is worth noting that various investigators have considered the impact of outliers on non–geographic data in which observations are not independent. This particular work has focused on generalized least squares regression in which the dependence is modeled by factors other than geographic location (*e. g.*, Pierce and Schafer, 1986; DeGruttola, Ware and Louis, 1987; Lee, 1988).

Given the general importance of considering outliers in statistical analyses, the additional complexity of geographic data, and the paucity of methods directed towards geographic outliers, it is the purpose of this paper to suggest some methods of detecting and describing such unusual observations. In biological and medical applications the identification and description of these sorts of observations may be as informative in their own right, as well as important for the reduction of influence of these outliers on one's eventual statistical goals.

## 3. Methods: Characterizing Spatial Outliers

In conducting spatial analyses there are at least three ways one can characterize outliers. First, outliers can be identified as traditional, aspatial outliers; these are observations that are simply numerically different from all others in a data set. Second, outliers can be identified as outlying locations, referring to locations for variate observations that are far from all others, locations that have no nearby neighbors (*i. e.*, isolated points). These correspond to points with high leverage. Third, outliers can be identified as spatial outliers that are defined as observations whose variate values are unlike the values of their neighbors (although these values may be well within the range observed for the entire data set). These observations can be very influential. I consider each outlier type in turn. I note that the second type of outlier described here is analogous to leverage. But, since the focus of this paper is on the methods of spatial autocorrelation in which there are no dependent variables, I consider the methods as slightly different.

### 3.1. Aspatial Outliers

Outliers are observations that are unlike all others in a data set. They may be due to observation, recording or transcription errors. They may represent heterogeneous populations. They may be the result of data contamination. Or, they may even be the result of random fluctuations. Various authors have presented reviews of methods and theories for detecting and identifying outliers (Barnett and Lewis, 1984; Hawkins, 1980; Atkinson, 1985). These methods can be applied to spatial data to detect the most flagrantly different observations (*i. e.*, global outliers). I consider just a few of these methods as others have written extensively on their use. For simplicity, we will use only a few methods based on normally distributed data, namely N2, N8, N14 and N15, as described by Barnett and Lewis (1984). N2 assesses the significance of the most extreme standardized normal deviate, high or low.

N8 compares the gap between the two largest values with the overall data range and the gap between the two smallest values with the overall data range, picking the maximum of these two differences. N14 is the sample skewness and N15 is the sample kurtosis. These latter two methods test a data set for normality, which is a useful exercise since an outlier tends to distort the observed frequency distribution of an otherwise normally distributed set of data points. With these four measures in mind, it is useful to provide sample estimates of the first four moments of the distribution of observations (mean, variance, skewness, kurtosis), nonparametric data summaries (minimum, maximum, median, hinges) and a stem–and–leaf plot of the data. The significance cutoffs for the aspatial tests are derived from the tables provided in Barnett and Lewis (1984).

### 3.2. Outlying Locations and Leverage

The second type of outlier I consider is an outlying location. That is, in some data sets, one (or a few) observations are situated far away from all other observations in geographic space. This positional anomaly will affect their influence on spatially weighted statistics. For instance, if one is using an inverse squared distance weighting in spatial autocorrelation analysis (Cliff and Ord, 1981), influence of an outlying observation would be limited. Similarly, if one is calculating a correlogram, the influence of outlying observations would be relegated to the far distance classes.

In essence, this is a consideration of the spatial point pattern of the data locations. Various investigators have developed extensive reviews on the analysis of spatial point patterns (*e. g.*, Ripley, 1981; Diggle, 1983; Upton and Fingleton, 1985; Ord, 1990). I note that while most evaluations of spatial point patterns seek to identify clustering or shorter than expected interpoint distances, outlier detection is based upon finding longer than expected interpoint distance. Thus, while many tests exist to find non–random distributions of observations, most are not well suited to the task of finding outliers.

A complementary problem, that of point clustering, has been discussed by Journel (1983, Appendix A). He argues that if too many points occur in close proximity, then averaging functions will overemphasize these points. He proposes a method of evening out the density of observations which he calls declustering. To decluster a data set, one superimposes a grid on a data set and then replaces the observations within a grid cell by their mean. This has the effect of removing undue weight of clustered points, but also obscures any variation that might exist within the grid. One also must worry about small–scale anisotropy that might cause the placement of the grid to affect the value of the declustered data. A similar averaging method has been proposed by Robinson and Mathias (1972).

As noted above, Cressie (1984; 1986) and Cressie and Read (1989) also consider the non–uniform distribution of sampling locations. For their study of Sudden Infant Death Syndrome, Cressie and Read wanted to use averaging methods that require gridded data. To accommodate irregularly spaced data, they allocated each observation to the nearest grid point and proceeded by using the new location. Since their data field was based upon regional summaries rather than individual point observations, averaging at grid points was not necessary. Again, while facilitating application of a particular methodology, this alteration may affect the ability to detect small scale pattern.

Evaluation of reflexive nearest neighbors is another method that has been proposed for evaluating spatial point patterns (*e. g.*, Clark and Evans, 1955; Pielou, 1977; Cliff and

Ord, 1981; Diggle, 1983). This again relates to global data characterization rather than the evaluation of the remoteness of individual localities. In the case of clustered data, reflexive nearest neighbors will be evident. Pairs of points will be closer to each other than to any other points. The frequency of such point pairs is an index of localized clustering, or fragmentation, rather than changes in overall density of points. Outliers, in contrast, would be points that are far away from all other points, and would not have reflexive nearest neighbors.

There are a few simple ways of evaluating the isolation of an individual data point. First, one can calculate the distance from each observation to its nearest neighbor. A limitation of this approach for points located on the edge of a study area is that this distance is generally larger than for those that are internal. Second, one can calculate the average of the distances from one point to all other points. Third, for correlograms, Delaunay tessellations or other models with defined neighborhoods, one can calculate the number of points with which an observation is connected (6 per point, on average—Upton and Fingleton, 1985, p. 97). Cliff and Ord (1973) note that in conducting spatial autocorrelation analysis one should try to have equal numbers of connections for each point (sum of $w_i$s) lest one or a few points dominate the analysis. Fourth, one could calculate the area of the Thiessen (or otherwise specified) polygon surrounding each point. Each of these computations gives slightly different information about spatial isolation. Each can be studied by using conventional measures that detect aspatial outliers. To allow for comparisons among data sets, these indices should be scaled by the maximum possible values for a data set. In this paper I consider only the first three indices that I have proposed here.

### 3.3. Spatial Outliers and Influential Points

The third type of outlier I consider is the spatial outlier. This type consists of points that fall within the distribution of other observations but are unlike their neighbors. They influence statistics that assess spatial pattern of variate values because comparison with near neighbors show large differences. Unlike leverage points, influence points must be important in terms of both location and variate value.

There are (at least) three ways to investigate the consequences of spatial outliers. First, one can consider the extreme of deviations from an overall trend in the data. That is, if the data are non–stationary, one can model this effect and look for deviations from it. Rather than asking simply if the most extreme values are sufficiently far from the other values to be considered data from a separate population, one could ask if there are any values that are sufficiently far from a model of the geographic distribution of the data so as to be considered statistically different from the other observations; that is, are outliers or contaminants present with respect to overall geographic pattern? For example, along a linearly increasing trend of a specific variate, a new sample value equal to the overall maximum value of the spatial data series would be unusual if it were found at the lowest part of the trend. This point would not be an aspatial outlier, as it falls within the range of other observed values, but does represent a large deviation from the overall model of the data. One method of evaluation here is to fit a trend surface to the data set and then study the properties of the residuals. While this approach has some utility, large residuals may reflect unusual observations, nonlinear trends, data heterogeneity, or other complicated situations; residual analysis will not help distinguish between these causes.

In time series terminology, the analogue of the analysis of residuals for a fitted trend

surface is the application of a high–pass polynomial filter to the data. A generalization of this approach for spatial data would be to use a geographic high–pass filter on the data that is more general than a polynomial of the coordinate location (*e. g.*, Holloway, 1958; Tobler, 1969). Such a filter could remove low frequency patterns from the data (*i. e.*, trends) while leaving high frequency, local pattern. One example of such a filter is a first–difference filter in which one differences an observation from a weighted average of its near neighbors. This allows local variation to remain while removing large scale pattern, regardless of structure. Various weighting patterns and neighborhood sizes can be used in designing the filter to correspond to a particular type of long period, low frequency pattern. One then can evaluate the data that pass through the filter.

A nonparametric method of high–pass filtering used by Cressie (1986) is the decomposition of the surface by median polish (Emerson and Hoaglin, 1983; Mosteller and Tukey, 1977). In this method, one removes the median from each row and then from each column of a two–way table. One performs this removal repeatedly until no more changes result. The residuals in the two–way table represent the new data, and the values removed represent the row and column averages. This approach is less sensitive to individual, outlier values and edge effects than trend surface analysis, and yet also adjusts for global pattern. One limitation of this approach is that it presupposes gridded data. If the data are not gridded, modifications will be necessary.

A second approach for assessing local outliers is to subdivide a study region into a specified number of smaller subregions. Statistical features of each of these subregions can be assessed (*e. g.*, mean, median, variance, quartiles). One could compare the mean and median as an index of outlying observations, as suggested by Cressie (1984). Large differences suggest unusual distributions and probable outliers. The subregions can be defined as overlapping or non–overlapping. The results can be tabulated or can be plotted on the map of localities (as in Cressie, 1984). It is important that each box have similar numbers of points within it for comparable reliability.

Third, one could model local pattern and then look for outliers with respect to the local distribution. For example, for each point one could fit a trend surface to the $k$ nearest points (*e. g.*, $k = 6$ for the empirical analysis presented below). A value at the location of the locality under consideration could be predicted from the trend surface model and the locality value could be replaced by the difference between this predicted value and the corresponding observed value. These "local residuals" could be evaluated with unusually discrepant values (positive or negative) being indicative of outliers.

This treatment of residuals from locally fit polynomial trend surfaces can be thought of as the converse of the contouring methods using local polynomials (Czegledy, 1972). Rather than using local polynomials to smooth out irregular variations, we focus attention on the variations. Effective implementation is contingent upon the number and placement of control points in fitting the polynomials. In the exercises presented in this paper, I set an arbitrary number of control points and do not consider placement. This leads to decreased reliability in border or isolated locations. For routine use, I recommend limiting the maximum distance of any control point and making sure that the control points form a relatively even angular distribution around the point to be evaluated. Unwin and Wrigley (1987a; 1987b) address some of these issues, as noted above. However, their main concern is global rather than local trend surfaces. For comparability purposes residuals should be standardized.

Having removed bad or aberrant observations that are not representative of the data set as a whole, we now proceed to look at the influence of individual data points on a statistic of interest. In particular, we consider the impact of individual observations on the spatial autocorrelation index known as Moran's I (Cliff and Ord, 1981); similar considerations could be developed for other indices, such as Geary's $c$. Since this index, I, is calculated as the double sum of values over all possible point pairs, one can calculate how much each point, when considered with all other points, contributes to the statistic. I call this the index decomposition value. The sum of the individual values equals the statistic of interest. Similarly, one can calculate how much the index would change if one omitted a single observation. This is evaluated by calculating the index with all data points, and then calculating the index with one point omitted, doing so for each point in turn. The difference between the value with one point omitted and with all points included, times the sample size minus one, is called the sample influence function. I note that one of the assumptions on which the behavior of the sample influence function is based is that the observations are identically and independently distributed. This characteristic does not hold for these data, and hence compromises the statistical rigor of this approach. However, it is still useful as a descriptive index to assess apparent influence.

It may seem that the index decomposition and sample influence function are identical; they are not. For the index decomposition, one retains all of the data values in the data set but shows the contribution of each data point to the observed statistic. To calculate the sample influence function, one discards one data point from the data set and then recalculates the index. Since the discarded data point had been used in calculating both the mean and the variance of the data set as well as the spatial index, the new mean and variance differ from those calculated with the entire data set. This alone may affect results. Thus, this index reflects a slightly different property of the data than the decomposition index.

One can plot results of these analyses, looking for individual points that fail to make a uniform contribution to this index. Single points that contribute disproportionately to the statistic, or whose omission result in marked changes in the statistic, are not representative of the entire surface. They can be removed, the analyses redone and results presented both with and without the influential point(s). Since influential points are still a part of the data set, they should not be discarded entirely. Rather, the removal and impact must be included in the presentation of results.

In summary, I propose four measures to use for exploratory spatial data analysis. First, as is customary with any data set, one should look for aspatial outliers and unusual distributions. Their presence can overwhelm any spatial pattern. Then, one should conduct a local trend surface residual analysis. This will detect anomalous spatial values irrespective of the underlying pattern. Next, one can calculate spatial autocorrelation decomposition and influence indices. These reflect both the overall spatial pattern and the effects of local aberrant values. More specifically, unusual spatial decomposition values and nearly uniform influence values are found if there are outliers in a spatially autocorrelated surface. Unusual spatial decomposition values and unusual influence values are found when a surface has negligible spatial autocorrelation but outliers. Unusual influence values cannot occur if decomposition values are similar.

To summarize, one can use these indices to characterize different properties of spatial data. One can determine whether or not there are aspatial outliers, whether or not there are

spatial outliers, and whether or not there is overall spatial structure. If there is no overall spatial structure, but aspatial outliers are present (and, hence, also spatial outliers), the aspatial tests should reveal this. Some of the spatial measures, including autocorrelation indices, may show pattern as well, but this finding would be a reflection of the difference between the outlier and the rest of the data, rather than an index of pattern in the non-outlier points. Once outliers are removed, the autocorrelation indices should be near their expected values under the null hypothesis. If there is no overall spatial structure but spatial outliers are present, then the spatial tests should reveal this (*e. g.*, local trend surface residuals—see below), and again the spatial autocorrelation values also may appear to be large. Removal of these outliers also should return the spatial autocorrelation indices to their expected values. One can have a data set with spatial structure but no outliers. In this case the autocorrelation indices should show pattern while the outlier indices should not. Finally, one can have overall spatial pattern with outliers. Then both the outlier tests and autocorrelation values should exceed expectation, and removal of the outlier should not remove (although it may modify) the geographic structure detected by the autocorrelation analysis. It also is possible that in this last case the spatial autocorrelation index decomposition will show variability while the influence function does not. This outcome reflects the contribution of the outlier point while also showing that its removal does not eliminate all spatial pattern.

## 4. Data Sources

I use two different data sources to demonstrate the utility of the approach proposed above. I use a simulated one to demonstrate the utility of components of the approach. This allows me to construct situations that emphasize particular features of data distributions that are of interest. In addition, I use observed data to demonstrate the applicability of these methods to real situations. The limitation of using actual data is that a single data set rarely has all the features of interest. Indeed, I will characterize the data set to determine what class, from those listed above, it belongs to. Further application of the methodology to additional data sets (in other publications) will help provide more general guidance.

Two different types of data are used in this investigation: location and observed variate values. Accordingly, I have conducted two sets of simulations: one for the locational outliers and one for the spatial outliers. For the locational simulation, I positioned points on a unit square with the uniform pseudo-random number generator from Turbo Pascal. Then I calculated, for each data point, the nearest neighbor distance and the average (mean) distance to all other points. I report results of the shortest and farthest nearest neighbor distances as well as average distances for different numbers of localities at various quantiles. So that data can be compared regardless of the units of measurement, all distances are scaled as a proportion of the maximum distance observed. Table 1 lists all of these numerical tabulations, as well as the means of results for 100 simulations.

Other investigators have looked at the distribution of nearest neighbor distances for describing the pattern of geographic data (*e. g.*, Silverman and Brown, 1978; Ripley and Silverman, 1978; Saunders, Kryscio and Funk, 1982). Their goal was to evaluate whether or not clustering exists in a given data set, overall, and they used the mean first (or third) nearest neighbor distance as their index. I make similar inferences and consider the overall description of the inter-point spacing.

For the second set of simulations, I simulated the effect of a single outlier on stationary

TABLE 1
Locational Simulations

| Probability | n | Distance | | | |
|---|---|---|---|---|---|
| | | Shortest | | Longest | |
| | | Upper | Lower | Upper | Lower |
| Nearest Neighbor | | | | | |
| $< 0.01$ | 10 | 0.1843 | 0.0045 | 0.6298 | 0.1877 |
| | 20 | 0.0883 | 0.0026 | 0.4708 | 0.1369 |
| | 30 | 0.0519 | 0.0018 | 0.3425 | 0.1303 |
| | 40 | 0.0372 | 0.0009 | 0.3093 | 0.1152 |
| | 50 | 0.0295 | 0.0010 | 0.2701 | 0.1050 |
| | 75 | 0.0187 | 0.0006 | 0.2220 | 0.0917 |
| | 100 | 0.0142 | 0.0004 | 0.1992 | 0.0819 |
| $< 0.05$ | 10 | 0.1586 | 0.0143 | 0.5496 | 0.2169 |
| | 20 | 0.0728 | 0.0060 | 0.3953 | 0.1647 |
| | 30 | 0.0441 | 0.0035 | 0.3054 | 0.1435 |
| | 40 | 0.0330 | 0.0028 | 0.2677 | 0.1259 |
| | 50 | 0.0264 | 0.0020 | 0.2411 | 0.1160 |
| | 75 | 0.0161 | 0.0014 | 0.1878 | 0.0971 |
| | 100 | 0.0122 | 0.0009 | 0.1681 | 0.0884 |
| Average | | | | | |
| $< 0.01$ | 10 | 0.5335 | 0.2783 | 0.8304 | 0.5796 |
| | 20 | 0.4397 | 0.2658 | 0.7449 | 0.5657 |
| | 30 | 0.4080 | 0.2746 | 0.7096 | 0.5576 |
| | 40 | 0.3830 | 0.2704 | 0.6913 | 0.5557 |
| | 50 | 0.3737 | 0.2641 | 0.6674 | 0.5506 |
| | 75 | 0.3513 | 0.2630 | 0.6458 | 0.5478 |
| | 100 | 0.3414 | 0.2643 | 0.6295 | 0.5471 |
| $< 0.05$ | 10 | 0.4987 | 0.3036 | 0.7900 | 0.6075 |
| | 20 | 0.4208 | 0.2867 | 0.7306 | 0.5789 |
| | 30 | 0.3854 | 0.2885 | 0.6849 | 0.5670 |
| | 40 | 0.3708 | 0.2833 | 0.6693 | 0.5632 |
| | 50 | 0.3622 | 0.2781 | 0.6475 | 0.5584 |
| | 75 | 0.3399 | 0.2762 | 0.6299 | 0.5543 |
| | 100 | 0.3323 | 0.2741 | 0.6179 | 0.5533 |

and non–stationary surfaces. First, 81 random normal deviates were generated, having mean 0 and unit variance using the uniform pseudo–random number generator from Turbo Pascal and an inverse normal transformation (Abramovitz and Stegun, 1965), and then they were allocated to a 9–by–9 grid. Next, an outlier was added either to the corner or the central point of each surface, with the increment ranging between 0 and 9. For the simulation of non–stationary surfaces, for each row the value of the row index was added to all values in that row. This procedure yielded a simple cline of slope 1 and maximum displacement of 8. 100 replicates of each of these situations were run. Estimates of the mean, variance,

skewness and kurtosis, as well as results of the two additional aspatial outlier tests noted above and the residual from the trend surface at the location of the outlier, are reported in Table 2.

Following execution of the simulation experiments, one real data set has been analyzed: the 1961 populations of the counties of Ireland as a percent of their 1926 populations. These data are derived from a study on road accessibility by O'Sullivan, and already have been presented and analyzed by Cliff and Ord (1981, p.208). A map of these data is presented in Figure 1, and numerical tabulations are presented in Table 3.

## 5. Results and Discussion

### 5.1. Simulations

The first set of simulations evaluated the nearest neighbor distances and average distances for random point patterns. 1000 replicates were run for surfaces with 10, 20, 30, 40, 50, 75, and 100 localities. Quantiles at the 1% and 5% two–tailed levels are shown in Table 1. These data are not meaningful in and of themselves, but need to be evaluated with respect to real data locations. As expected, one sees a decrease in the interpoint distances as the number of points on the square increases. The nearest neighbor distances change markedly over the range of localities tested while the average distances are fairly stable.

To interpret the data, one compares both the near and far results. If even one small cluster exists, then the shortest nearest neighbor distance should be smaller than that in the table. If they are clustered more generally, then the shortest average distance should be smaller than some of those in the table. If there is an outlier, then the longest nearest neighbor distance should exceed that value appearing in the table, as should the longest average distance.

The second set of simulations was run to evaluate a variety of summary statistics for data with a single outlier in a non–stationary data field. The non–stationarity was an inclined plane of slope of 1 (total displacement of 8) and the outlier was placed either in the middle of the data field or at the lower corner. 100 replicates were run for each case for values of the outlier ranging from 0 to 9. The summary statistic results are reported in Table 2. For the stationary surface, spatial outliers are also aspatial outliers. All the indices, aspatial and spatial, detect the outliers.

For the non–stationary surface, except for one measure of skewness, none of the aspatial indices yield a statistically significant value. Even though we have added a large spatial outlier, in many cases larger than any other value in the data field, aspatial indices fail to detect it. The local trend surface residuals (LTSR), however, show a clear and consistent pattern increasing with the size of the outlier. Beginning with an outlier of 2 or 3 (depending on location), the LTSR method detects the outlier. Values greater than 2 seem to be indicative of unusual values. This threshold value is recommended as a rule of thumb. The LTSR is relatively sensitive to spatial outliers and has performed well with other data sets not reported here.

TABLE 2
Results of 100 Simulations of a Spatial Outlier

| Increment | Mean | Variance | Skewness | Kurtosis | N8 | N2 | LTSR |
|---|---|---|---|---|---|---|---|
| Stationary Background with Outlier at (1,1) | | | | | | | |
| 0 | −0.01 | 0.99 | 0.39 | 2.83 | 0.08 | 2.47 | 1.91 |
| 1 | 0.02 | 1.03 | 0.40 | 2.93 | 0.09 | 2.51 | 1.48 |
| 2 | 0.02 | 1.09 | 0.42 | 2.99 | 0.10 | 2.63 | 1.39 |
| 3 | 0.05 | 1.10 | 0.54* | 3.53 | 0.14 | 3.04** | 1.55 |
| 4 | 0.03 | 1.20 | 0.74** | 4.93** | 0.27* | 3.80** | 4.61 |
| 5 | 0.06 | 1.31 | 0.98** | 7.02** | 0.35** | 4.43** | 3.88 |
| 6 | 0.08 | 1.47 | 1.10** | 8.55** | 0.40** | 4.80** | 4.49 |
| 7 | 0.09 | 1.59 | 1.40** | 13.19** | 0.49** | 5.52** | 4.62 |
| 8 | 0.10 | 1.82 | 1.56** | 16.49** | 0.53** | 5.91** | 5.10 |
| 9 | 0.11 | 1.95 | 1.66** | 18.88** | 0.55** | 6.14** | 4.30 |
| Stationary Background with Outlier at (5,5) | | | | | | | |
| 0 | −0.01 | 0.99 | 0.39 | 2.83 | 0.08 | 2.47 | 1.91 |
| 1 | 0.02 | 1.03 | 0.42 | 2.94 | 0.09 | 2.50 | 0.69 |
| 2 | 0.02 | 1.08 | 0.44* | 2.99 | 0.10 | 2.60 | −0.29 |
| 3 | 0.05 | 1.10 | 0.54* | 3.61 | 0.15 | 3.06** | −1.12 |
| 4 | 0.03 | 1.17 | 0.73** | 4.66* | 0.25* | 3.65** | 1.85 |
| 5 | 0.06 | 1.29 | 0.94** | 6.75** | 0.34** | 4.34** | 0.54 |
| 6 | 0.08 | 1.50 | 1.14** | 9.19** | 0.41** | 4.90** | 0.36 |
| 7 | 0.09 | 1.59 | 1.40** | 13.10** | 0.49** | 5.52** | −0.27 |
| 8 | 0.10 | 1.78 | 1.53** | 15.70** | 0.52** | 5.82** | 0.35 |
| 9 | 0.11 | 2.02 | 1.71** | 20.08** | 0.57** | 6.26** | −1.52 |
| Non-Stationary Background with Outlier at (1,1) | | | | | | | |
| 0 | 4.98 | 7.66 | 0.26 | 1.98 | 0.05 | 1.98 | −1.75 |
| 1 | 5.03 | 7.63 | 0.27 | 2.02 | 0.05 | 2.02 | 1.19 |
| 2 | 5.01 | 7.63 | 0.28 | 2.01 | 0.04 | 1.99 | −1.14 |
| 3 | 5.04 | 7.55 | 0.27 | 2.04 | 0.05 | 2.02 | 2.08 |
| 4 | 5.06 | 7.52 | 0.29 | 2.03 | 0.05 | 2.02 | 2.53 |
| 5 | 5.07 | 7.51 | 0.28 | 2.00 | 0.05 | 1.98 | 4.94 |
| 6 | 5.08 | 7.68 | 0.27 | 1.99 | 0.04 | 1.99 | 2.98 |
| 7 | 5.09 | 7.77 | 0.23 | 1.99 | 0.05 | 1.99 | 3.85 |
| 8 | 5.09 | 7.67 | 0.25 | 1.96 | 0.05 | 2.00 | 5.66 |
| 9 | 5.12 | 7.80 | 0.28 | 2.03 | 0.05 | 2.02 | 5.69 |
| Non-Stationary Background with Outlier at (5,5) | | | | | | | |
| 0 | 4.98 | 7.66 | 0.26 | 1.98 | 0.05 | 1.98 | −0.49 |
| 1 | 5.03 | 7.72 | 0.27 | 2.01 | 0.05 | 2.02 | 0.30 |
| 2 | 5.01 | 7.83 | 0.27 | 1.98 | 0.04 | 1.97 | 2.21 |
| 3 | 5.04 | 7.83 | 0.28 | 1.98 | 0.05 | 2.00 | 3.43 |
| 4 | 5.06 | 7.91 | 0.27 | 1.96 | 0.05 | 1.99 | 1.97 |
| 5 | 5.07 | 8.00 | 0.27 | 1.96 | 0.05 | 1.99 | 4.07 |
| 6 | 5.08 | 8.26 | 0.27 | 2.02 | 0.07 | 2.11 | 4.29 |
| 7 | 5.09 | 8.47 | 0.33 | 2.15 | 0.12 | 2.40 | 4.77 |
| 8 | 5.09 | 8.53 | 0.39 | 2.37 | 0.20 | 2.75* | 6.17 |
| 9 | 5.12 | 8.72 | 0.49* | 2.69 | 0.24 | 3.03** | 5.80 |

* denotes a significant difference at the 0.05 level.

** denotes a significant difference at the 0.01 level.

**Figure 1.**

Chloropleth map of 1961 Irish population as a per cent of 1926 population.

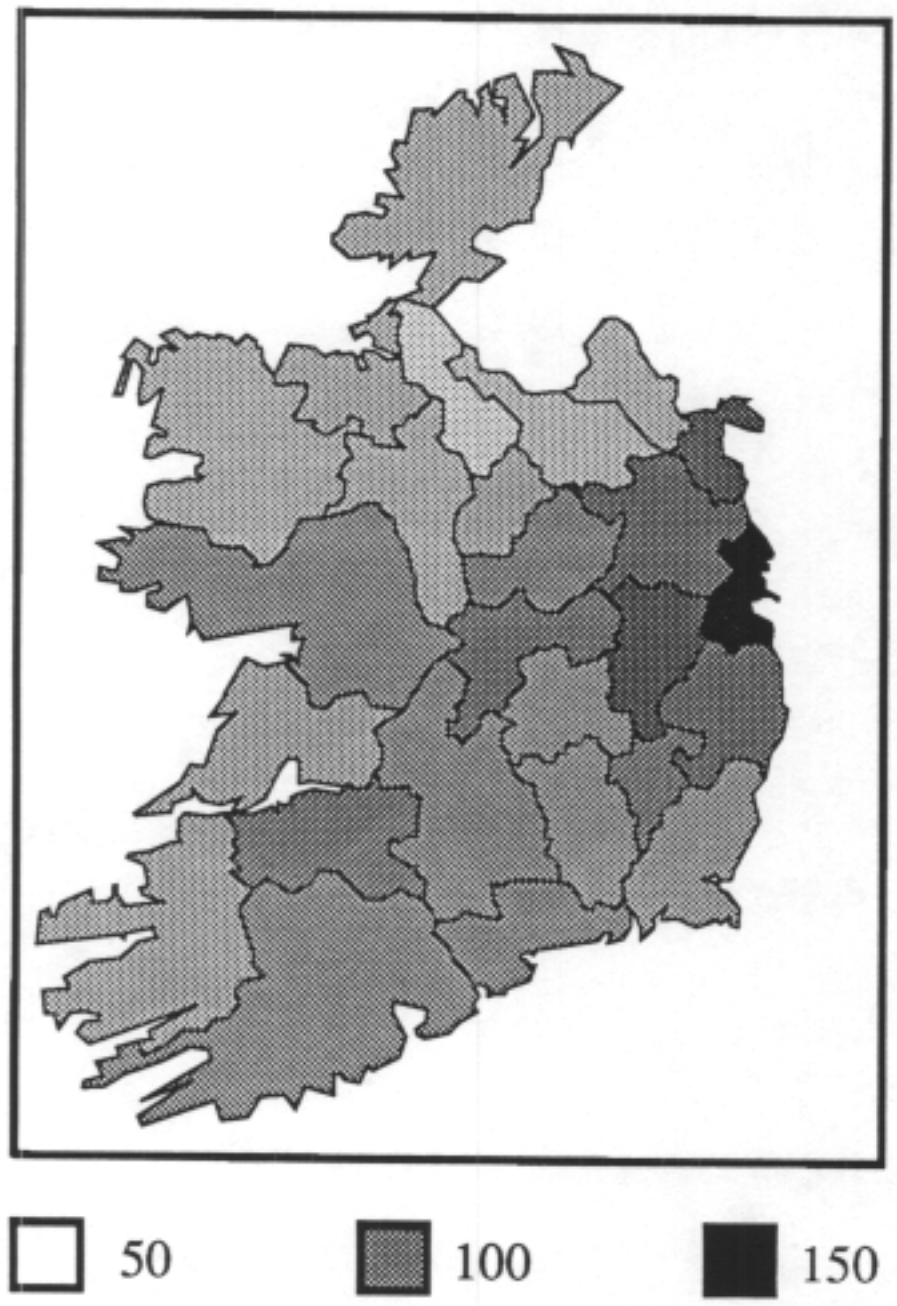


TABLE 3
1961 Irish Population as Percent of 1926 Population

**Table 3A: Aspatial Statistics**

| | N2 | N8 | Skewness | Kurtosis |
|---|---|---|---|---|
| Raw Data | 3.19** | 0.38* | 1.02* | 4.73* |
| Log Data | 2.71 | 0.29 | 0.64 | 3.27 |
| Data–Dublin | 1.91 | 0.08 | 0.18 | 2.04 |
| Log Data–Dublin | 1.72 | 0.06 | 0.49 | 2.18 |

**Table 3B: Local Trend Surface Residuals**

| Raw Data | Log Data | Raw Data–Dublin | Log Data–Dublin |
|---|---|---|---|
| 3.11 ( 6) | 2.29 ( 6) | 1.63 (15) | 1.85 ( 5) |
| −1.51 (26) | 1.73 ( 5) | 1.57 ( 5) | 1.65 (20) |
| 1.29 ( 5) | 1.54 (21) | 1.46 (20) | −1.55 (11) |

**Table 3C: Locational Statistics**

| Distances | | | |
|---|---|---|---|
| Nearest Neighbor | | Average | |
| Shortest | Longest | Shortest | Longest |
| 0.083 (14) | 0.257 (5) | 0.282 (19) | 0.590 (8) |

**Table 3D: Correlogram Results for Irish Population Data**

| Distance Class | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Raw Data | 0.45** | −0.07 | −0.21* | −0.16 | −0.20* |
| Log Data | 0.48** | −0.05 | −0.22* | −0.19 | −0.22* |
| Raw Data–Dublin | 0.45** | 0.00 | −0.15 | −0.27* | −0.24* |
| Log Data–Dublin | 0.45** | 0.00 | −0.15 | −0.27* | −0.24* |

* denotes a significant difference at the 0.05 level.

** denotes a significant difference at the 0.01 level.

### 5.2. The Real Data: The Irish Road Data

Finally, I analyze one data set to demonstrate how these methods work with actual observations. With this real data set, my goal is to determine whether or not there is any spatial structure in the data, and if there is, to describe it. The principal tool will be autocorrelation, although first one must consider the possibility and potential effects of outliers.

Results from aspatial tests (see Table 3A) show that this data set does not fit a normal distribution. There are two possible explanations. One is that the data fit a different distribution, such as an exponential. The other is that there is at least a single outlier that does not fit the distribution. To investigate these options, I can apply treatments for each effect. If I transform the data by taking logarithms of all the values to remove the effects of an exponential distribution, I see that the data now fit a normal distribution. Or, if I remove the single outlier (Dublin), these data also fit a normal distribution. Now, if I apply

the LTSR method to both raw and modified data sets, I find a spatial outlier for both the raw and log transformed data, namely Dublin (see Table 3B and Figure 2). However, if I remove Dublin from the raw data, there is neither an aspatial nor a spatial outlier. Thus, I conclude that Dublin is an outlier and ought to be removed, implying that the logarithmic transformation seems unnecessary.

Before performing spatial autocorrelation analysis, I investigated the spatial distribution of the observations (see Table 3C). Neither the shortest nor the longest nearest neighbor distance was remarkable. However, both the shortest and longest average interpoint distances fell at about the lower 5% level of the simulation sampling distributions. This outcome suggests that points are relative evenly spaced rather than random. This finding is not surprising as the data represent regional centroids rather than true point patterns.

Next, I determined appropriate distance class boundaries for correlogram analysis. I arbitrarily decided to use 5 distance classes, all with equal numbers of point pairs. Figure 3 shows the number of connections that each point has for each distance class. The expected number (if all connections were evenly distributed) is 5 joins per point per distance class. There is a moderate amount of variation with central localities taking on increased importance in the middle distance classes. One locality, Donegal, does not contribute to the first distance class at all.

Finally, I begin the spatial autocorrelation analysis by measuring the spatial autocorrelation for both the raw data and the logarithmically transformed data. The correlogram results are shown in Table 3D. The data exhibit a strong clinal pattern. The logarithmic transformation does not affect the correlogram markedly. It is reasonable to presume that there is a strong clinal pattern in the data.

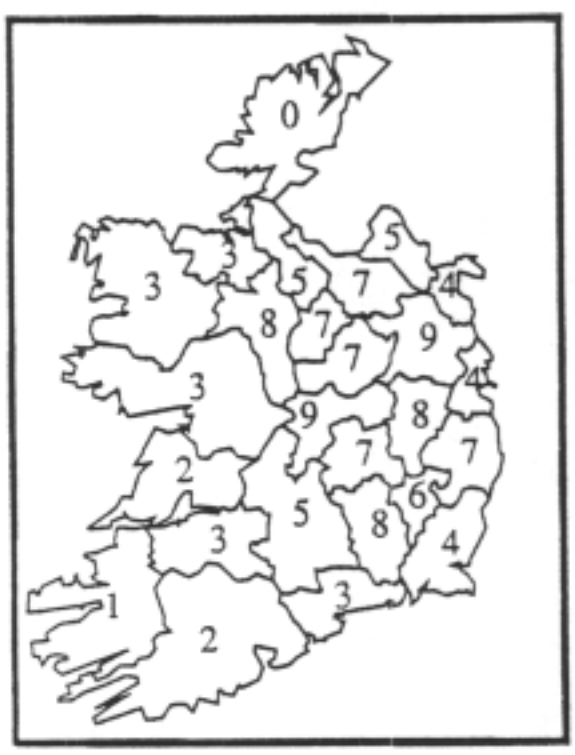
To investigate the impact of individual observations, I calculated the index decomposition and sample influence function. The resulting computations are summarized in Figure 4. The top set of circles display the index decomposition values. The left–most circle represents the first distance class, the next represents the second distance class, and so on. The solid circle is the expected value of each point's contribution to the statistic under the null hypothesis (approximately 0). The dashed circle shows the expected value of each point's contribution to the statistic, assuming each point contributed equally to the observed value of the statistic [1/(observed value)]. The larger the observed value of the statistic, the farther the dashed circle is from the solid circle. Rays projecting from the solid circle indicate the actual contribution of each point as measured from the center of the circle. They are plotted as the difference between the expected and the observed value. Values outside the solid circle are greater than zero and values inside the circle are less than zero. The first data point is represented by a ray at twelve o'clock, and subsequent data points by rays proceeding in a clockwise manner around the solid circle. For the left–most circle, which represents the first distance class, one sees that points 6 (Dublin), 9 (Kildare) and 12 (Leitrim) contribute the greatest amount to the index. Data point 6 contributes throughout all distance classes. It is an influential point.

The influence function is displayed in similar plots in the second row of Figure 4. In this set of plots, a solid circle represents the expected value of the statistic, a dashed circle represents the observed value of the statistic, and a ray represents the value of the statistic, with a given observation omitted. As the rays are of nearly the same length, the statistic is relatively insensitive to omission of data points.

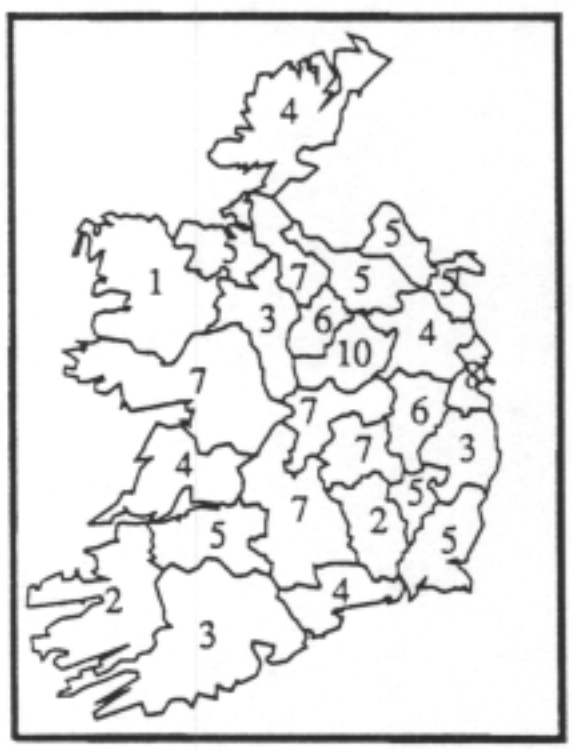
**Figure 2.**

Maps of the number of links emanating from each county of Ireland as used in the spatial autocorrelation analysis (see text). Each map represents a different distance class.

Distance Class 1

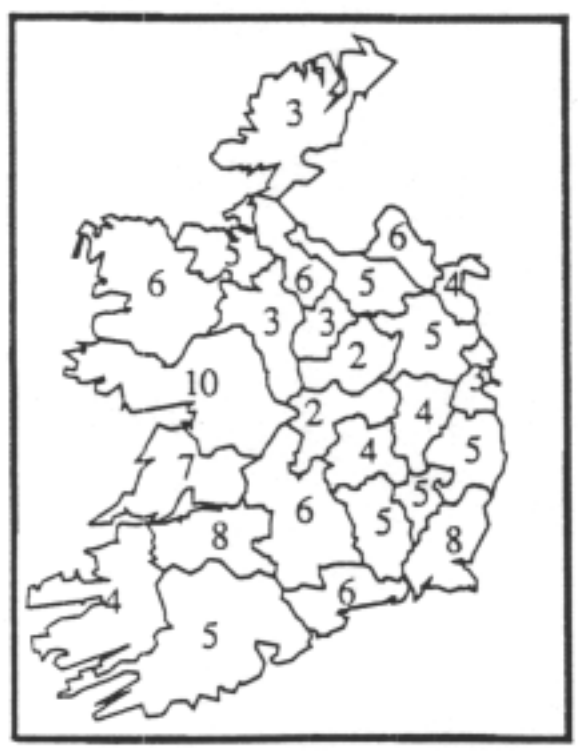


Distance Class 2



Distance Class 3

Distance Class 4



Distance Class 5

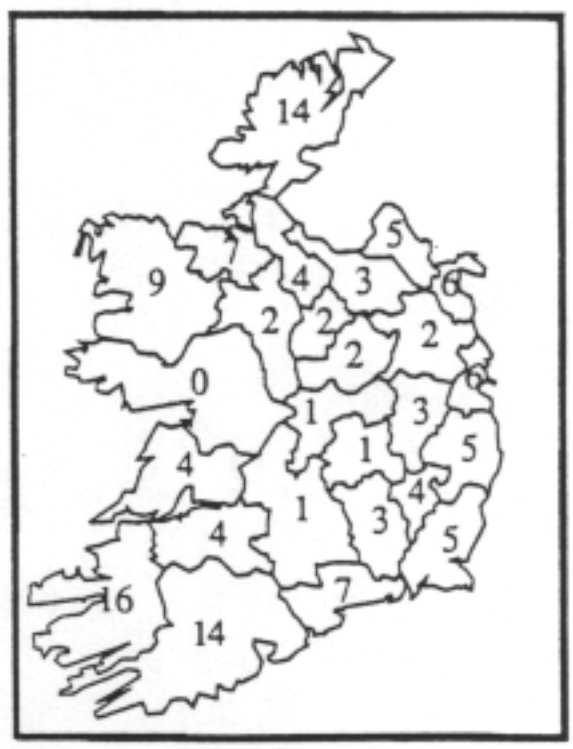


**Figure 3.**

Chloropleth map of local trend surface residuals (LTSR) of the 1961 population data.

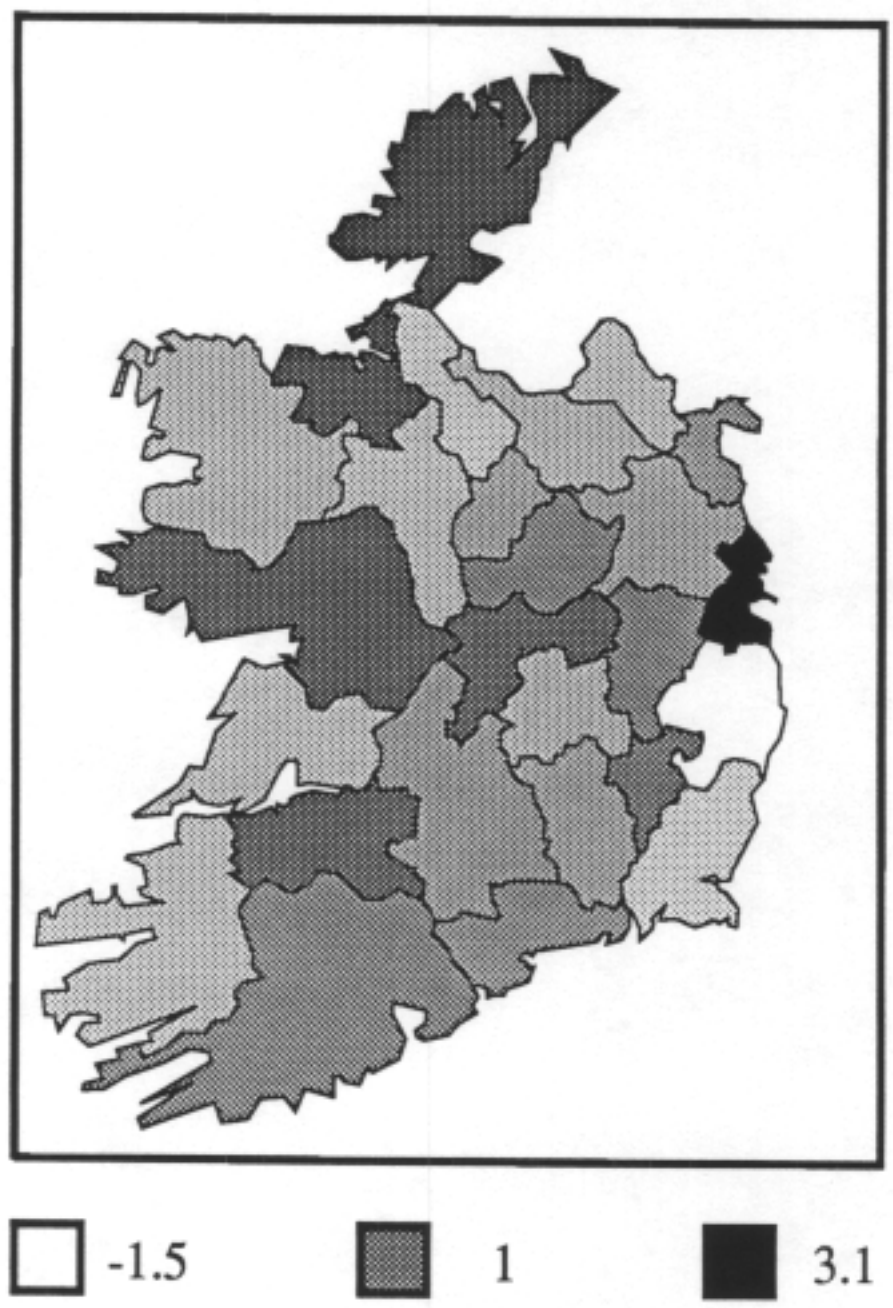
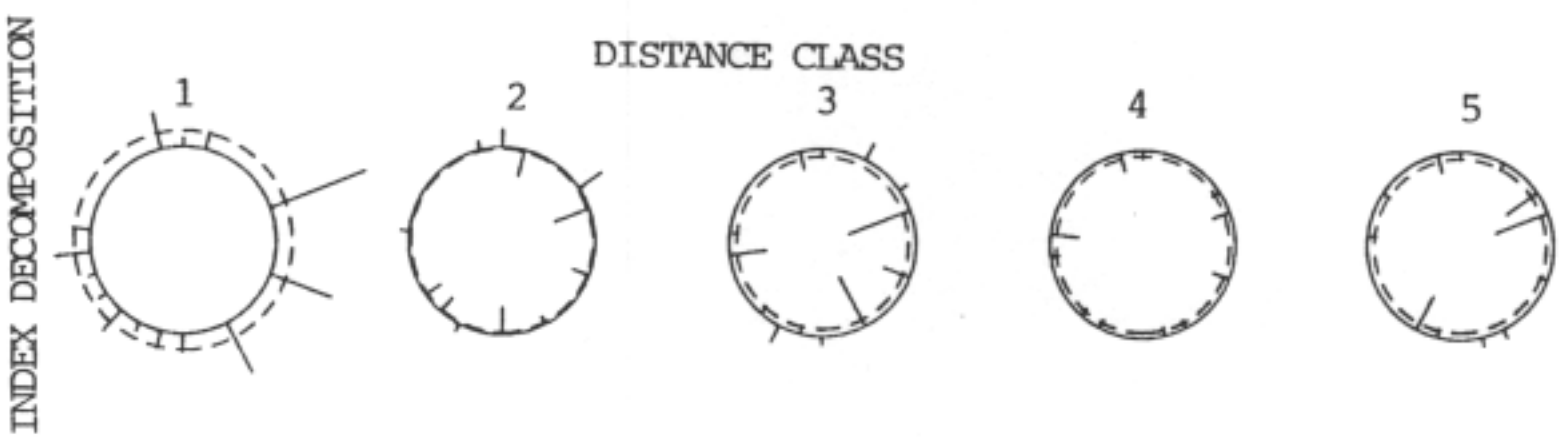
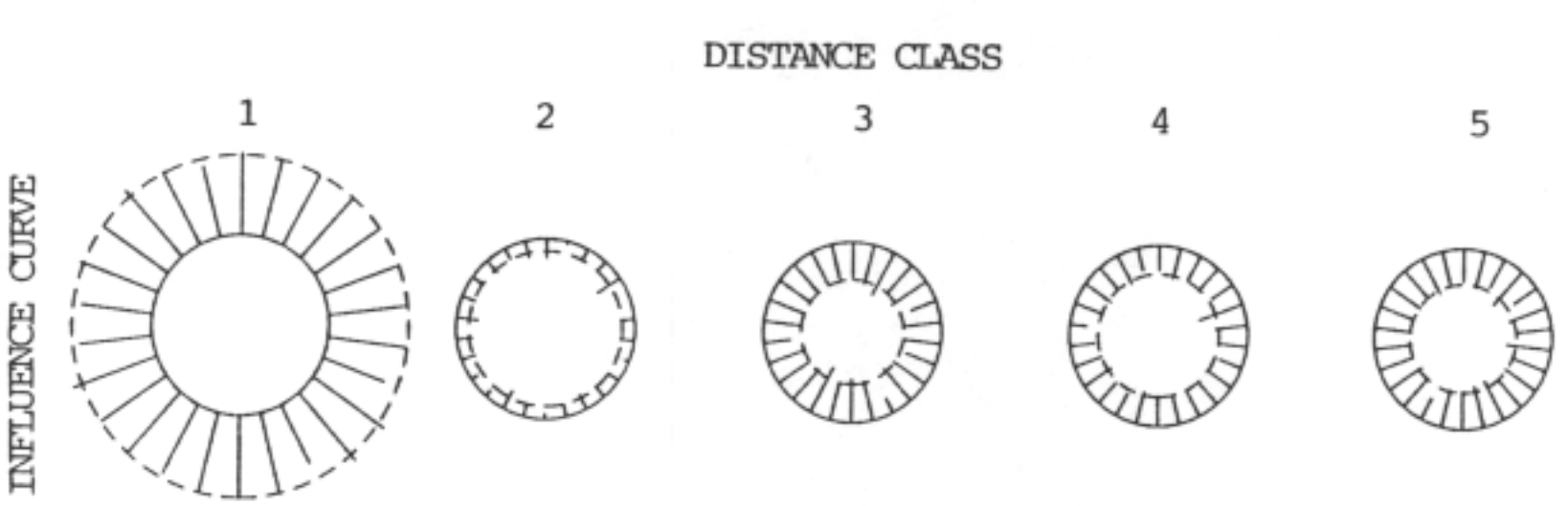


**Figure 4.**

Index decomposition and sample influence curve values for the Irish population data. Each circle represents a different distance class.



To summarize, the Irish population data have strong spatial pattern. Dublin is a large outlier, being larger than all other values and disproportionately greater than its neighbors. It contributes greatly to the observed spatial autocorrelation statistic although its omission does not affect the observed pattern greatly. Similarly, a few other points (Kildare, Leitrim) contribute disproportionately to the statistic, but their omission also does not greatly affect the overall statistic. Thus, the spatial pattern is spread over many points. In general, there is a clinal structure to the pattern of population. In the typology described above, the data correspond to the class of spatial structure with spatial outliers. If one were to model the pattern of these data, one would have to take into account both the outlier and the overall trend in the data.

## 6. Conclusions

Exploratory spatial analysis is a field that largely has been ignored. While much attention has been devoted to exploratory data analysis over the past number of years, investigators who study spatial phenomena have not adapted these methods for their own purposes. This paper proposes a few such methods and demonstrates that they can be effective through the employment of simulation experimentation.

Further, one can classify spatial data structure into four groups:

(1) aspatial outliers with no overall spatial pattern;

(2) aspatial outliers with overall spatial pattern;

(3) spatial outliers with no overall spatial pattern; and,

(4) spatial outliers with overall spatial pattern.

Using the indices proposed herein, one can classify real data sets into these different groupings. This classification exercise can be extremely useful in the study and evaluation of spatial process.

For example, I analyzed the spatial structure of 1961 Irish population as a percentage of the 1926 population. In this context, the goal is to determine whether or not there is any spatial structure in these data, and if there is, to describe it. Results indicate that both spatial outliers and spatial pattern exist. Because the data are regional summaries, consideration of locational outliers is not meaningful. Results obtained by spatial autocorrelation analysis without consideration of outliers or influential points is similar to that obtained after the identification and removal of such values. Indeed, Dublin, a major urban center, outstripped the growth of the rest of the country. Looking back to the original data (Figure 1), one can see such a pattern. However, if one proceeded with additional analyses of these data, such as regional pattern summarization or the regression work described by Cliff and Ord (1981), identification of these properties is extremely important. Residuals from regression, even though conducted in an aspatial context, were dominated by the influence of the Dublin. Removal of this value likely would have resulted in a more representative regression model.

The goal of this paper has been to propose some methods for the exploratory analysis of spatial data. These methods can be thought of as a series of pretreatments before rigorous statistical analysis. They are designed to give the investigator an intuitive understanding of the spatial structure of the data, and to assist in the design of subsequent statistical investigations. The methods, newly proposed herein, will require refinement and application if they are to become useful tools for the spatial analyst.

## 7. References


Abraham, B., and G. Box. (1979) Bayesian analysis of some outlier problems in time series. *Biometrika*, **66**, 229-236.

Abramovitz, M., and I. Stegun. (1965) *Handbook of mathematical functions*. New York: Dover.

Atkinson, A. (1985) *Plots, Transformations and Regression*. Oxford: Clarendon Press.

Bardossy, A. (1988) Notes on the robustness of the kriging system. *Journal of the International Association of Mathematical Geology*, **20**, 189-203.

Barnett, V., and T. Lewis. (1984) *Outliers in Statistical Data*, 2nd ed. New York: Wiley.

Belsey, D., E. Kuh, and R. Welsch. (1980) *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley.

Brooker, P. (1986) A parametric study of robustness of kriging variance as a function of range and relative nugget effect for a spherical semivariogram. *Journal of the International Association of Mathematical Geology*, **18**, 477-488.

Chorley, R., and P. Haggett. (1965) Trend surface mapping in geographical research. *Transactions*, Institute of British Geographers, **37**, 47-67.

Clark, P., and F. Evans. (1955) On some aspects of spatial pattern in biological populations. *Science*, **121**, 397-398.

Cliff, A., and J. Ord. (1973) *Spatial Autocorrelation*. London: Pion.

Cliff, A., and J. Ord. (1981) *Spatial Processes: Models and Applications*. London: Pion.

Cook, R., and S. Weisberg. (1982) *Residuals and Influence in Regression*. London: Chapman and Hall.

Cressie, N. (1984) Towards resistant geostatistics, in G. Verly, M. David, A. Journel, and A. Marechal (eds.), *Geostatistics for Natural Resource Characterisation*. Dordrecht: Reidel, pp. 21-44.

Cressie, N. (1986) Kriging nonstationary data. *Journal of the American Statistical Association*, 81: 625-634.

Cressie, N., and N. Chan. (1989) Spatial modeling of regional variables. *Journal of the American Statistical Association*, **84**, 393-401.

Cressie, N, and D. Hawkins. (1980) Robust estimation of the variogram. *Journal of the International Association of Mathematical Geology*, **12**, 115-126.

Cressie, N., and T. Read. (In press; cited in Cressie and Chan, 1989.) Spatial data analysis of regional counts. *Biometrical Journal*, **31**.

Czegledy, P. (1972) Efficiency of local polynomials in contour mapping. *Journal of the International Association of Mathematical Geology*, **4**, 291-305.

DeGruttola, V., J. Ware, and T. Louis. (1987) Influence analysis of generalized least squares estimators. *Journal of the American Statistical Association*, **82**, 911-917.

Denby, L, and R. Martin. (1979) Robust estimation of the first–order autoregressive parameter. *Journal of the American Statistical Association*, **74**, 140-146.

Diamond, P, and M. Armstrong. (1984) Robustness of variograms and conditioning of kriging matrices. *Journal of the International Association of Mathematical Geology*, **16**, 809-822.

Diggle, P. (1983) *The Analysis of Spatial Point Pattern*. New York: Wiley.

Dowd, P. (1984) The variogram and kriging: robust and resistant estimators, in G. Verly, M. David, A. Journel, and A. Marechal (eds.), *Geostatistics for Natural Resource Characterisation*. Dordrecht: Reidel, pp. 91-108.

Emerson, J., and D. Hoaglin. (1983) Analysis of two–way tables by medians, in D. Hoaglin, F. Mosteller, and J. Tukey (eds.), *Understanding Robust and Exploratory Data Analysis*. New York: Wiley, pp. 166-210.

Fox, A. (1972) Outliers in time series. *Journal of the Royal Statistical Society*, **34B**, 350-363.

Griffith, D. (1988) Interpretation of standard influential observations: regression diagnostics in the presence of spatial dependence, paper presented to the 35th annual North American Meeting of the Regional Science Association, Toronto.

Hampel, F., E. Ronchetti, P. Rousseeuw, and W. Stahel. (1986) *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley.

Hawkins, D. (1980) *Identification of Outliers*. New York: Chapman and Hall.

Hawkins, D., and N. Cressie. (1984) Robust kriging—a proposal. *Journal of the International Association of Mathematical Geology*, **16**, 3-18.

Hoaglin, D., and R. Welsch. (1978) The hat matrix in regression and anova. *American Statistician*, **32**, 17-22.

Holloway, J. (1958) Smoothing and filtering of time series and space fields. *Advances in Geophysics*, **4**, 351-389.

Huber, P. (1981) *Robust Statistics*. New York: Wiley.

Journel, A. (1983) Nonparametric estimation of spatial distributions. *Journal of the International Association of Mathematical Geology*, **15**, 445-468.

Kleiner, B., R. Martin, and D. Thomson. (1979) Robust estimation of power spectra. *Journal of the Royal Statistical Society*, **41B**, 313-351.

Künsch, H. (1984) Infinitesimal robustness for autoregressive processes. *Annals of Statistics*, **12**, 843-863.

Lee, A. (1988) Assessing partial influence in generalized linear models. *Biometrics*, **44**, 71-77.

Mather, P. (1977) Clustered data–point distributions in trend surface analysis. *Geographical Analysis*, **9**, 84-93.

Mosteller, F., and J. Tukey. (1977) *Data Analysis and Regression*. Reading, MA: Addison–Wesley.

Muirhead, C. (1986) Distinguishing outlier types in time series. *Journal of the Royal Statistical Society*, **48B**, 39-47.

Olmstead, P., and J. Tukey. (1947) A corner test for association. *Annals of Mathematical Statistics*, **18**, 496-513.

Omre, H. (1984) The variogram and its estimation, in G. Verly, M. David, A. Journel, and A. Marechal (eds.), *Geostatistics for Natural Resource Characterisation*. Dordrecht: Reidel, pp. 107-125.

Ord, J. (1990) Statistical methods for point pattern data, in D. Griffith (ed.), *Spatial Statistics: Past, Present and Future*. Ann Arbor, MI: Institute of Mathematical Geography,

(in press).

Pielou, E. (1977) *Mathematical Ecology*. New York: Wiley.

Pierce, D., and D. Schafer. (1986) Residuals in generalized linear model. *Journal of the American Statistical Association*, **81**, 977-986.

Putterman, M. (1988) Leverage and influence in autocorrelated regression models. *Applied Statistics*, **37**, 76-86.

Ripley, B. (1981) *Spatial Statistics*. New York: Wiley.

Ripley, B., and B. Silverman. (1978) Quick tests for spatial interaction. *Biometrika*, **65**, 641-642.

Robinson, G., and M. Mathias. (1972) On transforming to eliminate clusters. *Geographical Analysis*, **4**, 424-427.

Saunders, R., R. Kryscio, and G. Funk. (1982) Poisson limits for a hard–core clustering model. *Stochastic Processes and Their Applications*, **12**, 97-106.

Silverman, B., and T. Brown. (1978) Short distances, flat triangles and Poisson limits. *Journal of Applied Probability*, **15**, 815-825.

Tobler, W. (1969) Geographical filters and their inverses. *Geographical Analysis*, **1**, 234-253.

Tukey, J. (1949) One degree of freedom for non–additivity. *Biometrics*, **5**, 232-242.

Tukey, J. (1951) Quick and dirty methods in statistics, Part II, simple analyses for standard designs. *Proceedings of the 5th Annual Convention, American Society for Quality Control*, pp. 189-197.

Tukey, J. (1977) *EDA*. Reading, MA: Addison–Wesley.

Unwin, D., and N. Wrigley. (1987a) Control point distribution in trend surface modelling revisited: an application of the concept of leverage. *Transactions of the Institute of British Geographers*, New Series **12**, 147-160.

Unwin, D., and N. Wrigley. (1987b) Towards a general theory of control point distribution effects in trend surface models. *Computers and Geosciences*, **13**, 351-355.

Upton, G., and B. Fingleton. (1985) *Spatial Data Analysis by Example*, vol. 1, Point Pattern and Quantitative Data. New York: Wiley.

Welsch, R. (1985) An introduction to regression diagnostics. *Proceedings of the 13th Conference on the Design of Experiments in Army Research Development and Training*.

# DISCUSSION

## "Exploratory spatial analysis: outliers, leverage points, and influence functions"

## by Daniel Wartenberg

*Are outliers "bad or aberrant observations"?* The answer to this question has to be "Not necessarily"! It may well be the case that the data have been misreported or mistyped; in such cases, tests for outliers are useful diagnostic devices for locating such errors. However, when the data are free from errors, a test for an outlier should be regarded as a test for the validity of a model (often implicit) rather than a test of an observation *per se*.

The usual implicit model is that all the values being studied have resulted from a single distribution. The presence of an outlier implies that this assumption is faulty—the outlier is an observation from some other distribution. If the values under consideration are the residuals from some model, then an outlier residual implies that the model is inadequate with respect to the corresponding datum point.

In summary, therefore, the presence of an outlier usually should be regarded as pointing to a deficiency in the modelling process, rather than a deficiency in the data.

### Edge effects and boundaries.

In the analysis of small quantities of point pattern data, edge effects play a dominating role. All the points lie within some more or less well defined boundary. When the number of points is small, the proportion of "internal" points (those near the geometric center of the cluster of points) also will be small. Most points will have no points "outside" them (beyond them as one moves from the geometric center of the cluster outward). For example, of the 26 counties of Eire, only 9 (35%) are totally bordered by neighboring counties. The remainder have either the sea or Northern Ireland adjacent to their borders.

Influence, attributable to boundaries, upon the distribution of the popular Clark–Evans statistic is well known (*e. g.*, see Upton and Fingleton, 1985, p. 74). Required corrections to the mean and variance of the distribution of distances to nearest neighbors involves measures of both the perimeter and the area of the region under study. Doguwa and Upton (1988, 1989) have studied the corresponding "point–event" statistic, and find a need for similar corrections. The distribution function of nearest–neighbor distances is a powerful tool for detecting departures from randomness; but this, too, is complicated by the need to take boundaries into account. An improved estimator of this function is given by Doguwa and Upton (1990).

It follows from the above discussion that Wartenberg's simulated results, given in his Table 1, need to be treated with some care. He gives the upper and lower 1% and 5% significance points for the distance to the nearest neighbor, and for the average distance to the remaining $n - 1$ neighbors—but these results only apply to a square study region. It is easy to see that the results for a region such as the Florida Keys would be rather different!

As a check, I performed 99 simulations, representing Eire by an 11–sided polygon, retaining only those points that fell inside the polygon, and continuing until I had generated 26 randomly placed retained points for each simulation. For the shortest nearest neighbor distance, my simulations gave values between 0.0022 and 0.0486 (after scaling), compared

with an observed 0.0201 for Longford to Westmeath (Counties #14 and #24). My longest nearest neighbor distances varied between 0.2909 and 0.1165, compared with an observed 0.136 (Donegal and Leitrim Counties). Thus, neither observed value appears significant. Note that my observed values are very different from those of Wartenberg, partly (I conjecture) because of different scaling factors, and partly because the observed results are critically dependent upon the point positions taken as representative of the counties under study.

Wartenberg suggests scaling by the largest observed distance. I think it would be preferable to scale by the largest observable distance. The problem of representing areas by points is discussed in my own article in this volume. In my simulations, I evidently used rather different co–ordinates to those used by Wartenberg.

### Monte Carlo methods and simulation.

The rapid increase in easily available computing power during the last two decades has led to an increasing reliance on simulation as a means for determining the distributional properties of otherwise intractable statistics. There are many examples in the field of spatial statistics. However, there is no need to present simulated results when the theoretical results can be easily calculated, as is the case with the first four statistics reported in his Table 2. The results given there merely serve to confirm that the remaining results are plausible.

### Trend surfaces.

Wartenberg's analysis uses, I think, quadratic surfaces fitted to the nearest 6 neighbors of each point, with the value for a given point being omitted from its corresponding trend surface estimation. With either stationary or non–stationary linear backgrounds one would expect a "pimple" to appear as such, and I am therefore surprised at the entries in the second quarter of Table 2.

However, I confess that trend surfaces leave me uneasy! Although I have very little experience with them, I am very conscious of the potential differences that can arise as the degree of a surface is altered. It would be interesting to see the residuals that arise as surfaces of different orders are fitted to these artificial data sets.

### The diagrams.

I cannot let the diagram of distance classes pass by without querying its usefulness; to me it seems merely to confirm that distant points are indeed distant!

It is refreshing to see an entirely new method of presenting data being illustrated in Wartenberg's final figure. However, while I applaud the intention, I feel that its circular nature is totally misconceived, since there is nothing cyclic about the counties of Eire! A further problem arises because the diagrams are almost impossible to label effectively. A more successful display might be a dot diagram of the type advocated by Cleveland (1985), though with 26 counties this might not be feasible. It probably would be more useful simply to list the major departures from uniformity of contribution.

### Summary.

Much of the above has been critical in nature. However, the author is quite right to point to the need for spatial methods for exploring spatial data. Wartenberg raises an important and valid point when he argues that an obvious spatial outlier, as in the sequence

$\{8, 6, 4, 2, 0, -2, -4, -6, -8\}$, may appear to be entirely typical when the data are divorced from their spatial locations. Unfortunately, I do not believe that this paper has answered the question of how to identify such an outlier. I do think, however, that Professor Wartenberg has opened up a new and fruitful research area in the field of spatial statistics.

## References

Cleveland, W. (1985) *The Elements of Graphing Data*. Monterey, CA: Wadsworth.

Doguwa, S., and G. Upton. (1988) On edge corrections for the point–event analogue of the Clark–Evans statistic. *Biometrical Journal*, **30**, 957-963.

Doguwa, S., and G. Upton. (1989) Simulations to determine the mean and variance of the point–object analogue of the Clark–Evans statistic. *Biometrical Journal*, **31**, 163-170.

Doguwa, S., and G. Upton. (1990) On the estimation of the nearest neighbor distribution, G(t), for point processes. *Biometrical Journal*, **32**, (in press).

Upton, G., and B. Fingleton. (1985) *Spatial Data Analysis by Example*, vol. 1. Chichester: Wiley.

*Graham J. G. Upton*, University of Essex

## A REJOINDER TO UPTON'S DISCUSSION

**by Daniel Wartenberg**

New areas of research are always controversial. And, while EDA methodology has become widely accepted in statistical investigations, few attempts have been made to develop EDA methodology specifically for spatially dependent data. Thus, I am not surprised, although somewhat disheartened, by Upton's acerbic and contentious discussion of my paper. As he notes, there is yet a long way to go before this area of investigation develops into mature methodology that is routinely useful and diagnostic of spatial aberrations. But that does not diminish the value of initial innovations and first ideas. The proposals I put forth are meant to open a dialogue on these issues, rather than present definitive methodology. Toward that end, I address two specific issues Upton raises, with the goal of broadening the basis of discussion and stimulating further work. Page constraints preclude more comprehensive commentary.

The first issue is outliers, their definition, detection and interpretation. Upton notes that outlier tests are most useful as tests of model validity. The usual, implicit model employed is that observed data are from a single, statistical distribution. Indeed, while outlier tests may be useful for detecting data transcription or reporting errors, these are in the sphere of data processing rather than spatial statistics. Upton argues that outliers are diagnostic of "a deficiency in the modeling process, rather than a deficiency in the data". A still broader (and more appropriate) view is that outliers show an inconsistency between a model and the data, and that attribution of this inconsistency is not possible based on outlier detection alone. Substantive evaluation may help elucidate whether the problem resides in the data or the model.

In my paper, I use an implicit model of similarity among geographically proximate observations. Rather than being purely distributional, my implicit model accounts for spatial location. That is, I test the similarity or smooth variation of nearby values. It is common geographic knowledge that observations near one another are more similar than those widely separated, and the tests I propose exploit this property. And, when I make this model explicit by using local trend surface models, Upton dismisses the methodology out of hand. Numerous examples exist of useful applications of trend surface methodology, although it is an approach subject to misuse and misinterpretation. To improve on the specific application I propose, I encourage Upton to provide a more informative and easy to use model for local spatial structure! (I have experimented with trend surfaces of different orders, as Upton suggests, but these correspond to different spatial models with varying data requirements. A full discussion of this approach with surfaces of different orders and consideration of varying numbers and orientations of control points will be presented elsewhere.)

Upton also summarily dismisses the circle plots I propose because he believes circularity implies cyclicity and because of the difficulty in labeling specific localities on the circle. In preference he suggests dot diagrams or data listings. However, both of these have the limitation that they require more space on a printed page (which is always at a premium), and make it more difficult to compare the same object across sets of observations or distance classes. Circle plots exploit the human ability to juxtapose cyclic images on top of each other and compare objects at like positions (*e. g.*, comparing the length of objects at 3 o'clock on 5 different circles). This is one of a class of similar methods that display many variables simul-

taneously for many objects. Star plots (Chambers *et al.*, 1983), for example, are multivariate profiles plotted in polar coordinates for easier viewing. Experimentation with line plots used for regression diagnostics were noticeably more difficult to interpret. However, improvement of circle plots, or alternative representations that facilitate interpretation, would be useful.

In sum, the goal of my paper was to raise some questions of data quality, data consistency and diagnostic methodology. In view of the paucity of methods for addressing these issues, I have proposed a few. As developments continue in this area, new ideas, new methods and new interpretations likely will improve upon these preliminary explorations.

## References

Chambers, J., W. Cleveland, B. Kleiner, and P. Tukey. (1983) *Graphical Methods for Data Analysis*. Monterey, CA: Wadsworth.

## *PREAMBLE*

*The power of Thought — the magic of the Mind!*

Byron, **Corsair**

*Who was the innovator who first wrote of spatial econometrics? Probably hidden somewhere in the yellow-paged journals of yesteryear is a forgotten article, the first to break this barrier; if it exists, undoubtedly its discovery will occur at some future date. Certainly Paelinck was one of the first scholars to devote considerable thought to complications that lie dormant in traditional econometric analysis but become problematic when analyzing geo-referenced data. Over the past fifteen years he has repeatedly developed and/or modified econometric techniques in order to handle these complexities. As is characteristic of his earlier work, here he presents rationales, relevant properties, empirical examples, and possible extensions of estimators. In doing so, he highlights the difficulties of specification, interpretation, and computation. The purpose of this paper is to present the six new estimation techniques called simultaneous dynamic least squares, strictly positive conditional, linear logistic, least spheres, non-numerical regression, and distribution-free power. Acknowledging the innovativeness of Paelinck's work, Anselin emphasizes the focus of the six new estimators (i. e., problems of simultaneity in spatial modeling, data limitations, and complexities attributable to spatial interaction), as well as the technical issues of identifiability, distributional properties, and non-trivial implications associated with various approximations to non-linear estimators. All in all, this paper is as delightful an example of the spatial econometric viewpoint as can be found in the literature today.*

The Editor

# Some New Estimators in Spatial Econometrics

J. H. P. Paelinck

*Department of Theoretical Spatial Economics, Erasmus University, P. O. B. 1738, 3000 DR Rotterdam, The Netherlands*

**Overview:** The empirical study of spatial economic phenomena leads to a large number of specifically different problem settings. Reliable quantitative study of these problems often cannot proceed using standard econometric techniques, or approaches that initially were developed for other purposes. Moreover, better solutions can be obtained for these problems by modifying standard results in an appropriate way, or improving the properties of methods that already have been proposed. A number of new estimators are presented in this paper; it is believed that they will prove illuminating when applied to those spatial economic cases for which they have been developed. Without presenting an integrated body of econometric analysis—like k–class estimators, for instance—these new estimators represent a sample of spatial econometric estimation exercises that might usefully complement the body of knowledge already in existence.

## 1. Introduction

One important aspect of spatial econometrics is the development of estimators appropriate to given types of problems. In Paelinck and Klaassen (1979, Chapter 3) some special estimators, based on previous work, already have been presented. These include

- ⋆ estimators for a spatial income–generating model;
- ⋆ estimators for the interregional attraction model: MOLS (Multiregional Ordinary Least Squares), IOLS (Interregional Ordinary Least Squares), ISSML (Interregional Semi–Separable Maximum Likelihood); and,
- ⋆ estimation of threshold effects.

Estimators presented in Chapters 4 and 5 were original contributions at the time, and include

- ⋆ distribution–free testable spatial autocorrelation estimation;
- ⋆ component parameter estimation [for a recent application of this procedure, see Kuiper (1989)]; and,
- ⋆ fuzzy multiple regime estimation.

In this paper more recent materials that resulted from research undertaken since the publication of the aforementioned volume are presented. The organization of this presentation is as follows: rationale for the estimator, its presentation with relevant properties, an empirical example, and possible extensions. Additional aspects can be found in Ancot, Paelinck and Prins (1986), Ancot and Paelinck (1987) and Paelinck (1989).

## 2. Newcomers

### 2.1. Simultaneous Dynamic Least Squares (SDLS)

In Paelinck and Klaassen (1979, Chapter 7) it has been shown how SDLS estimation can simultaneously comply with synchronic, diachronic, sectoral and spatial interdependences. The model can indeed be written as

$$\mathbf{Ay} + \mathbf{Bx} = \boldsymbol{\xi}, \tag{2.1.1}$$

where matrix $\mathbf{A}$ represents the linkages between arbitrary endogenous variables $\mathbf{y}$ (spatialised or not, lagged or not), $\mathbf{Bx}$ the effects of exogenous shocks, and $\xi$ stochastic elements. [1] The SDLS estimator is derived from the optimization problem of

$$\min_{\mathbf{A},\mathbf{B}} (\mathbf{y} - \mathbf{A}^{-1}\mathbf{Bx})'(\mathbf{y} - \mathbf{A}^{-1}\mathbf{Bx}) \tag{2.1.2}$$

and computed from the vector–matrix transformation of equation (2.1.1)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\xi}, \tag{2.1.3}$$

with $\mathbf{y}$ and $\mathbf{X}$ respectively being a vector and a matrix of observed variables (one should note that vector $\mathbf{x}$ and matrix $\mathbf{X}$ are not the same), $\boldsymbol{\gamma}$ the vector of $\mathbf{A}$ and $\mathbf{B}$ coefficients, and (2.1.3) being in fact a so–called "normalised form" of equation (2.1.1). The estimator is given by

$$\boldsymbol{\gamma}^e = (\mathbf{X}'\mathbf{X}^e)^{-1}\mathbf{X}'\mathbf{y}, \tag{2.1.4}$$

where matrix $\mathbf{X}^e$ contains *estimated* endogenous variables. Numerical work conducted on this estimator has found that in practice convergence of a Gauss–Seidel nonlinear estimation procedure does occur (Prins, 1985).

Some properties of the equation (2.1.4) estimator are:

- ⋆ it is a generalized reduced form estimator;
- ⋆ if $\xi \sim \mathbf{N}(\mathbf{0}, \sigma^2\mathbf{I})$, then the estimator is a maximum likelihood (ML) one; and,
- ⋆ $\boldsymbol{\gamma}^e$ is consistent, with plim $\boldsymbol{\gamma}^e\boldsymbol{\gamma}^{e\prime} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

Some early applications of this estimator can be found in Ancot, Kuiper and Paelinck (1981). It has been applied more recently to the estimation of discrete versions of the Lotka–Volterra model (Bagchus a.o., 1985; Budding a.o., 1985; Dickmann and Spoorendonk, 1987; for the model itself, one is referred to Peschel and Mende, 1986), the latter being then specified as

$$\Delta'\ln(x_t) = a + bx_{t-1} + cy_{t-1} \tag{2.1.5a}$$

$$\Delta'\ln(y_t) = d + ex_{t-1} + fy_{t-1} \tag{2.1.5b}$$

where the variable x represents population and y income per capita. Applied to the city of Rotterdam over the period 1946-1978, this estimator has given those results appearing in Table 1 (the computer program is discussed in Schueren, 1986). These tabulated values have acceptable interpretations. Loo (1987) also has studied other dutch cities, the principal problems encountered being that of the availability, the quality and the comparability over time of income figures. One could consider introducing distributed lags into this model, too.

TABLE 1
PARAMETERS OF A LOTKA–VOLTERRA MODEL
FOR THE CITY OF ROTTERDAM

| Parameter | Estimated Value | Student's t–statistic |
|---|---|---|
| a | −0.8798 | −7.68 |
| b | 0.0711 | 7.33 |
| c | 0.3988 | 4.22 |
| d | 1.0870 | 9.49 |
| e | −0.8025 | −8.51 |
| f | −0.5355 | −5.67 |
| $a^{\star a}$ | 0.0362 | 1.64 |
| $d^{\star a}$ | 0.0538 | 2.43 |

[a] Optimisation parameters for the starting point of an endogenous simulation.

Simultaneous dynamic least squares also can be useful in studying spatial autocorrelation. Consider the model

$$\mathbf{y} = \rho\mathbf{C}\mathbf{y} + \mu\mathbf{i} + \xi, \tag{2.1.6}$$

where $\mathbf{C}$ is a geographic contiguity matrix, and $\mathbf{i}$ is a vector of ones. The SDLS estimator is generated by minimising w.r.t. $\rho$ and $\mu$ the expression

$$[\mathbf{y} - \mu(\mathbf{I} - \rho\mathbf{C})^{-1}\mathbf{i}]'[\mathbf{y} - \mu(\mathbf{I} - \rho\mathbf{C})^{-1}\mathbf{i}]. \tag{2.1.7}$$

Let us suppose that

$$|\rho\lambda(\mathbf{C})|_{\max} < 1, \tag{2.1.8}$$

so that one can consider an approximation supplied by only the linear terms of the spatial multiplier, $(\mathbf{I} + \rho\mathbf{C})$; equation (2.1.7) then can be rewritten as

$$[\mathbf{y} - \mu(\mathbf{I} + \rho\mathbf{C})\mathbf{i}]'[\mathbf{y} - \mu(\mathbf{I} + \rho\mathbf{C})\mathbf{i}], \text{ or} \tag{2.1.9}$$
$$[\mathbf{y} - \mu(\mathbf{i} + \rho\mathbf{n})]'[\mathbf{y} - \mu(\mathbf{i} + \rho\mathbf{n})], \tag{2.1.10}$$

where the vector $\mathbf{n}$ results from summing over the rows of matrix $\mathbf{C}$.

A first hypothesis can be that this summation gives a constant, $\nu$ such that

$$\mathbf{n} = \nu\mathbf{i}, \tag{2.1.11}$$

but differentiating equation (2.1.10) with respect to $\mu$ and $\rho$ gives one and the same equation. This outcome is due to the non–discriminating effect of an infinite spatial structure.

If one defines

$$n = \mathbf{i}'\mathbf{n}, \text{ and} \tag{2.1.12a}$$
$$n^{\star} = \mathbf{n}'\mathbf{n}, \tag{2.1.12b}$$

then the two parameter estimates become

$$\rho = (\mathbf{y}'\mathbf{n} - \mu n)/(\mu n^{\star}), \text{ and} \tag{2.1.13a}$$
$$\mu = (\mathbf{i}'\mathbf{y} + \rho\mathbf{y}'\mathbf{n})/(r + 2\rho n + \rho^2 n^{\star}), \tag{2.1.13b}$$

where $r$ is the number of spatial units. In equation (2.1.13), then, the non–spatial mean, $\mathbf{i}'\mathbf{y}/r$, is corrected for spatial autocorrelation.

From equations (2.1.13a) and (2.1.13b),

$$\rho = (n\mathbf{i}'\mathbf{y} - r\mathbf{y}'\mathbf{n})/(n\mathbf{y}'\mathbf{n} - n^{\star}\mathbf{i}'\mathbf{y}), \text{ and} \tag{2.1.14a}$$
$$\mu = (n\mathbf{y}'\mathbf{n} - n^{\star}\mathbf{i}'\mathbf{y})/(n^2 - n^{\star}r). \tag{2.1.14b}$$

These two expressions can be rewritten as

$$\rho = -(r/n)[(\mu^{\mathbf{s}}/\mu^{\star}) - 1]/[(\mu^{\mathbf{s}}/\mu^{\star}) - (n^{\star}/n)(n/r)], \text{ and} \tag{2.1.15a}$$
$$\mu = \mu^{\star}[(n/r)(\mu^{\mathbf{s}}/\mu^{\star}) - n^{\star}/n)]/(n/r - n^{\star}/n), \tag{2.1.15b}$$

with

$$\mu^{\star} \triangleq \mathbf{i}'\mathbf{y}/r, \text{ and} \tag{2.1.16a}$$
$$\mu^{\mathbf{s}} \triangleq \mathbf{n}'\mathbf{y}/n. \tag{2.1.16b}$$

From equation (2.1.15a) one finds that if $\mu^{\mathbf{s}} = \mu^{\star}$, then $\rho$ would be zero; the spatially corrected average does not add any information. Hence from equations (2.1.3b) or (2.1.15b) then, $\mu = \mu^{\star}$.

Study with respect to the critical values for $\rho$, namely 1 and $-1$, can proceed as follows. If $(n^{\star}/n)(n/r) - 1 = 1$, then $\rho = -r/n > -1$, so $r/n$ is a damping factor. Positive autocorrelation is to be expected with "skew" spatial structures defined as

$$(n^{\star}/n)(n/r)^{-1} > \mu^{\mathbf{s}}/\mu^{\star} > 1. \tag{2.1.17}$$

In the case of a constant number of first–order autoregressive links, $\nu$, the expression for $\rho$ becomes

$$\rho = (1/\nu)(\mu^{\star}/\mu - 1). \tag{2.1.18}$$

Supposing $\mu^{\star} > 0$ and $\mu > 0$, negative autocorrelation occurs for $\mu^{\star}/\mu < 1$, but will never be less than $-1$. Positive autocorrelation can only exceed $+1$ (*e. g.*, a non–stationary geographic process is operating) if $(\mu^{\star}/\mu - 1) > \nu$, which is a possibility for which the probability is unknown.

## 2.2. Strictly Positive Conditional Estimation (SPCE)

Ancot and Paelinck (1981) have drawn attention to the conceptual necessity of obtaining strictly positive values for certain parameters resulting from an a priori spatial theory. They have investigated the approach outlined here in the ensuing discussion. Let $\beta$ be a parameter of the equation

$$\xi_i \triangleq y_i - \beta x_i, \tag{2.2.1}$$

the probability of observing jointly $\xi_i$ and $\beta$ being written as

$$p(\xi_i, \beta) = p(\xi_i|\beta)p(\beta), \tag{2.2.2}$$

where $p(\xi_i|\beta)$ is given by equation (2.2.1) and $p(\beta)$ is a prior density for parameter $\beta$. One estimates $\beta$ under the hypothesis that over the observation period (or the observed regional

TABLE 2
FLEUR SECTOR NUMBER 28, PERIOD 1950–1960[a]

| Variables | With Elimination of Parameters With the Wrong Sign | | All Parameters | | 95% SPCE Bounds | | |
|---|---|---|---|---|---|---|---|
| | $\beta_{ols}$ | Student's t | $\beta_{ols}$ | Student's t | $\beta_{spce}$ | lower | upper |
| $X_1$ | 0.647 | 34.93 | 0.647 | 35.20 | 0.647 | 0.618 | 0.675 |
| $X_2$ | 0.832 | 5.31 | 0.799 | 5.10 | 0.802 | 0.530 | 1.034 |
| $X_3$ | 0.243 | 1.43 | 0.196 | 1.14 | 0.196 | 0.101 | 0.336 |
| $X_4$ | 0.759 | 4.75 | 0.705 | 4.34 | 0.706 | 0.444 | 0.941 |
| $X_5$ | 0.873 | 4.04 | 0.783 | 3.52 | 0.782 | 0.417 | 1.097 |
| $X_6$ | 0.215 | 6.06 | 0.216 | 6.15 | −0.100 | −0.120 | −0.084 |
| $X_7$ | 0.388 | 11.42 | 0.403 | 11.50 | 0.403 | 0.360 | 0.452 |
| $X_8$ | 0.434 | 20.52 | 0.424 | 19.41 | 0.424 | 0.396 | 0.452 |
| $X_9$ | ***** | ***** | −0.034 | −1.52 | 0.100 | 0.091 | 0.110 |
| $X_{10}$ | 0.119 | 1.76 | 0.110 | 1.64 | 0.110 | 0.083 | 0.144 |
| $X_{11}$ | 0.063 | 0.81 | 0.688 | 0.89 | 0.079 | 0.049 | 0.095 |
| $X_{12}$ | −0.121 | −1.16 | −0.158 | −1.49 | −0.158 | −0.293 | −0.095 |
| $X_{13}$ | 0.510 | 6.35 | 0.497 | 6.19 | 0.496 | 0.387 | 0.607 |
| $X_{14}$ | 0.330 | 8.03 | 0.331 | 8.11 | 0.331 | 0.287 | 0.376 |
| $X_{15}$ | 0.303 | 7.65 | 0.301 | 7.68 | 0.301 | 0.262 | 0.343 |
| $X_{16}$ | 0.134 | 3.53 | 0.129 | 3.39 | 0.129 | 0.110 | 0.150 |
| $X_{17}$ | 0.103 | 3.64 | 0.104 | 3.70 | 0.104 | 0.092 | 0.117 |
| | | $R^2 = 0.974$ | | $R^2 = 0.975$ | | | |
| | | $\text{MSE}^b = 1.542$ | | $\text{MSE} = 1.493$ | | $\text{MSE} = 10.340$ | |
| | | "price" of SPCE = 6.927 | | | | | |

[a] On the FLEUR model, see Ancot and Paelinck (1983).

[b] MSE is the residual variance.

system) $\beta$ has been constant. This estimation has been investigated for $\xi \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ and $\boldsymbol{\beta} \sim T(\boldsymbol{\beta}^*)$, where $T$ represents a Tanner distribution having estimates $\boldsymbol{\beta}^*$. The estimated value, $\boldsymbol{\beta}^e$, is

$$\boldsymbol{\beta}^e = \tilde{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) + 2n\sigma^2(\mathbf{X}'\mathbf{X})(\hat{\boldsymbol{\beta}}^e)^{-1}\mathbf{i}, \tag{2.2.3}$$

where $\hat{\boldsymbol{\beta}}$ is the ordinary least squares (OLS) estimator, and $n$ is the number of observations. One should note that

- ⋆ equation (2.2.3) has indeed a strictly positive (or strictly negative, if required) value in $\boldsymbol{\beta}^e$; and,
- ⋆ up to second–order $\sigma^2$, VAR($\boldsymbol{\beta}^e$) equals the OLS expression.

Table 2 summarizes an individual result borrowed from Enhus (1986) based upon this estimator.

TABLE 3
SYNTHESIS OF RESULTS OBTAINED BY ENHUS, 1986

| FLEUR Sector | Period | Number of Replaced Coefficients | "Price" |
|---|---|---|---|
| 19 | 1960–1970 | 2 | 6.337 |
| 28 | 1950–1960 | 2 | 6.927 |
| 37 | 1950–1960 | 2 | 7.559 |
| 19 | 1950–1960 | 2 | 8.912 |
| 53 | 1950–1960 | 2 | 14.643 |
| 28 | 1960–1970 | 1 | 17.168 |
| 37 | 1960–1970 | 2 | 54.603 |
| 7 | 1950–1960 | 3 | 229.399 |
| 7 | 1960–1970 | 3 | 566.485 |

Following Ancot and Paelinck (1981, p. 360, Property 3) the confidence intervals for the SPCEs have been assumed to be log–normal. The interpretation of the tabular results reported in Table 2 is obvious; the "price" to be paid for SPCE is the ratio of the residual variance SPCE/OLS (all parameters).

Table 3 presents an overview of those results obtained by Enhus (1986).

### 2.3. Linear Logistic Estimation (LLE)

In spatial analysis the presence of binary $0-1$ indicator variables that are to be predicted or statistically explained (presence or absence of certain elements) is frequent. Suppose the probability $p_{ijk}$ for a firm of type $i$ ($I$ characteristics of a "plant profile") of exporting product $j$ ($J$ characteristics of a "product profile") to country $k$ ($K$ characteristics of an "export profile") to be logistic. The three profiles are represented by a vector $\mathbf{x}$, with "more" of a characteristic increasing the probability of exporting according to the function

$$p_{ijk} = [1 + \exp(-\mathbf{a}'\mathbf{x})]^{-1}, \tag{2.3.1}$$

with

$$\mathbf{a} \geq \mathbf{0}, \tag{2.3.2}$$

and with the observations being 0 (no exports) or 1 (exports). Let two variables be defined, one for exporters as

$$d_{1i} \triangleq 1 - [1 + \exp(-\mathbf{a}'\mathbf{x}_i)]^{-1}, \tag{2.3.3a}$$

and one for non–exporters as

$$d_{2i} \triangleq [1 + \exp(-\mathbf{a}'\mathbf{x}_i)]^{-1}, \tag{2.3.3b}$$

Thus one can easily compute

$$\mathbf{a}'\mathbf{x}_{1i} = \ln(d_{1i}^{-1} - 1) \triangleq \delta_{1i} > 0, \text{ and} \tag{2.3.4a}$$

$$-\mathbf{a}'\mathbf{x}_{2i} = \ln(d_{2i}^{-1} - 1) \triangleq \delta_{2i} > 0. \tag{2.3.4b}$$

Maximising $\sum_{i=1}^{n}(\delta_{1i}+\delta_{2i})$, under a norm restriction, one obtains

$$\mathbf{a}^e = \lambda^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{i}, \text{ and} \tag{2.3.5a}$$

$$\text{VAR}(\mathbf{a}^e) = \lambda^{-2}(\mathbf{X}'\mathbf{X})^{-1}, \tag{2.3.5b}$$

which means that the ratio $a_k^e/\sigma(a_k^e)$ is independent of $\lambda$, so that the null hypothesis $\mathbf{a}=\mathbf{0}$ can be tested.

The numerical example presented in Table 4 has been explored here. Using those data, Table 5 compares the results from a classical "probit" analysis with that of LLE.

TABLE 4
DATA FOR L. L. E.

| | Variables | | |
|---|---|---|---|
| Values of $i$ | $Y$ | $X_1$ | $X_2$ |
| 1 | 1.0 | 6.0 | 4.0 |
| 2 | 1.0 | 8.0 | 2.0 |
| 3 | 1.0 | 4.0 | 3.0 |
| 4 | 1.0 | 7.0 | 0.0 |
| 5 | 1.0 | 9.0 | 1.0 |
| 6 | 0.0 | 2.0 | 6.0 |
| 7 | 0.0 | 3.0 | 9.0 |
| 8 | 0.0 | 1.0 | 4.0 |
| 9 | 0.0 | 0.0 | 8.0 |
| 10 | 0.0 | 1.0 | 7.0 |

TABLE 5
PROBIT AND L. L. E.

| Parameter | Probit | Student's t | L. L. E. | Student's t |
|---|---|---|---|---|
| of $X_1$ | 3.40 | $0.20\star10^{-02}$ | 0.20 | 1.18 |
| of $X_2$ | −1.72 | $-0.90\star10^{-03}$ | −0.13 | −0.70 |
| Constant | −2.54 | $-0.23\star10^{-03}$ | −0.26 | −0.18 |

One can see from these tabulated results that the signs as well as (for the parameters of $\mathbf{x_1}$ and $\mathbf{x_2}$) the ratios are consistent. The t statistics for the LLE estimators, however, are much less non–significant than are those for the probit estimators, the data obviously being ill–conditioned.

Recently this estimator has been extended to $0-z_i$ cases, where the $z_i$ are possibly all different real numbers. For the latter observations the vector $\mathbf{i}$ in equation (2.3.5a) is extended by $\lambda\xi$, with vector

$$\boldsymbol{\xi} \triangleq [\ln(2-z_i k_i^{-1}) - \ln\{(z_i k_i - 1)\exp(\mathbf{a}'\mathbf{x}_i)+1\}], \tag{2.3.6}$$

which reduces to the binary $0-1$ case for a choice of $\lambda \to 0$ (very small "distances" required) or $z_i k_i^{-1} \to 1$, $\forall i$ (a perfect fit, which is the equivalent result). The $k_i$s are variable asymptotes; for a specification of the form

$$k_i = \exp(\mathbf{b}'\mathbf{y}_i), \tag{2.3.7}$$

straightforward OLS (with an extra parameter for the $z_i = 0$ observations) allows the estimating of $\mathbf{b}$.

### 2.4. Least Spheres Estimation (LSE)

In some cases of spatial analysis, the presence of potentials (sums) can lead to multicollinearity; in such a case, another estimator, LSE, can bring relief, its objective function being

$$\Psi = [\sum_{i=1}^{n}\sum_{j=1}^{k}(a_j x_{ij} - y_i^\star)^2 + \sum_{i=1}^{n}\sum_{j=1}^{k}(x_{ij} - x_{ij}^\star)^2]/2, \tag{2.4.1}$$

where the $a_j$ and $x_{ij}$ are the endogenous variables to be estimated, and the starred variables $y_i^\star$ and $x_{ij}^\star$ are being observed (the number of degrees of freedom will remain $n-k$ in this case). More specifically, one minimises the sum of squares of the radii of the hyperspheres with centres $(y_i^\star, x_{ij}^\star)$, $\forall i$, that are tangent to the hyperplane $y_i = \sum_{j=1}^{k} a_j x_{ij}$, $\forall i$.

The estimators for $\mathbf{a} \stackrel{\Delta}{=} [a_j]$ are

$$\tilde{\mathbf{a}} = (\mathbf{X}^{\star\prime}\mathbf{X}^\star - \varepsilon^{-2} I)X^{\star\prime}y^\star,$$

the stars indicating exogenous variables, which is a curious rejoinder to ridge regression. As the $a_j$s are not inversely invariant with the measurement units of the $x_{ij}$s, $\varepsilon$ should be maximised, and to guarantee positive definiteness, its sign reversed (the mathematical justification for this approach appears in Paelinck and Klaassen, 1979, pp. 54-55). Table 6 reproduces the results for this estimator applied to a tourist model of Swiss data for the "canton du Valais" (Bailly and Paelinck, 1988). Significance tests for the estimated parameters are available, and their results are reported in Table 6.

### 2.5. Non–numerical Regression (QUALIREG)

Suppose one wants to explain a phenomenon on which only qualitative observations are available (for example a vector of ranked items $\mathbf{y}' = [+++, -, 0, ++, \ldots]$), with the same situation prevailing for the matrix of explanatory variables, $\mathbf{X}$:

$$\mathbf{X} = \begin{pmatrix} + & - & \cdot & \cdot & \cdot \\ ++ & 0 & \cdot & \cdot & \cdot \\ - & ++ & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix} \tag{2.5.1}$$

Suppose matrix $\mathbf{X}$ is of order $n$–by–$k$ ($n$ observations on $k$ explanatory variables). Such a situation is frequently encountered in spatial econometric analysis. The following programme gives a solution for the problem: find a vector of coefficients $\beta^e$ maximising

$$\tau(\mathbf{y}, \mathbf{y}^e), \tag{2.5.2}$$

where $\tau$ is Kendall's rank correlation coefficient [2] and $\mathbf{y}^e$ the vector of estimated ranks of $\mathbf{y}$. A normalisation of $\beta^e$ is necessary, which leads to the mathematical programme

$$\max_{\beta^e} = \max_{\beta^e} \beta^{e\prime}\tau^e \tag{2.5.3}$$

TABLE 6
PARAMETERS OF A TOURIST MODEL
APPLIED TO A "CANTON" IN SWITZERLAND

| Parameters | Types of Villages | | | |
|---|---|---|---|---|
| | Valley Locations | | Mountain Locations | |
| | French Speaking | German Speaking | French Speaking | German Speaking |
| Self–inducation/breaking | −0.074* | −0.760 | −0.722 | −0.236* |
| **Locations Potential:** | | | | |
| French Speaking Valley | 0.003* | 0.109* | 0.061* | 0.078* |
| German Speaking Valley | −0.069* | 0.111* | 0.006* | −0.002* |
| French Speaking Mountain | −0.056* | 0.111* | 0.016* | 0.013* |
| German Speaking Mountain | −0.010* | 0.112* | −0.019* | −0.038* |
| Optimised starting point of endogenous dynamics | 0.377* | −0.012* | 0.290 | 0.421* |
| Autonomous growth/ decline rate | 0.004* | 0.109* | 0.062* | 0.080* |
| Pseudo–$R^2$ | 0.241* | 0.563 | 0.460 | 0.416* |

NOTE: * denotes significance at the 95% confidence level using chi square and F test statistics.

subject to:

$$-\mathbf{i} \leq \boldsymbol{\beta}^e \leq \mathbf{i} \tag{2.5.4}$$

where $\boldsymbol{\tau}^e$ is the vector of Kendall $\tau$'s corresponding to the permutations of columns of $\mathbf{X}$ producing $\mathbf{y}^e$. "Multiple correlation" and $\boldsymbol{\beta}^e$ tests are available, as Table 7 shows.

This method has been applied to an explanatory relation of water discharge per province in the Netherlands (see Davelaar a. o., 1983). Table 7 gives the results of a comparative exercise with OLS–estimation; more elaborate commentary on these results appears in Ancot and Paelinck, forthcoming.

### 2.6. Distribution–Free Power Estimator (DFPE)

Relations in spatial econometrics are often of a highly non–linear nature (see Paelinck and Klaassen, 1979, pp. 6-9). Power parameter specifications can be useful to model such behaviour; an early application of this perspective to a so–called "multiple gap" investment model, using other solution methods, is reported on in Ancot e. a. (1978).

Generalised Box–Cox transformations (see Box and Cox, 1964) will be discussed next, together with a proposed procedure for nonparametric estimation. This latter procedure—for generalised Box–Cox transformations—can proceed in the following manner. Suppose a non–linear relation exists and may be specified as

$$y_i^{\sigma} = \sum_{j=1}^{k} a_j x_{ij}^{\rho_j} + \xi_i \tag{2.6.1}$$

TABLE 7
COMPARISON OF THE RESULTS OBTAINED BY QUALIREG AND BY OLS

| | Coefficients | | | | | | |
|---|---|---|---|---|---|---|---|
| | $b_1$ | Kendall's tau | $b_2$ | Kendall's tau | $b_3$ | Kendall's tau | Multiple Correlation |
| QUALIREG | | | | | | | |
| | 0.5 | 0.24 * | −0.4 | 0.31 | −0.3 | 0.16 | 0.636[a] |
| | 1.0 | 0.24 | −0.8 | 0.31 | −0.56 | 0.18 | |
| | 0.8 | 0.24 | −0.6 | 0.27 | −0.4 | 0.27 | 0.600 |
| | 1.0 | 0.20 | −0.6 | 0.35 | −0.6 | 0.09 | |

| | Coefficients | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $b_0$ | S–t | $b_1$ | S–t | $b_2$ | S–t | $b_3$ | S–t | Multiple Determination |
| OLS | | | | | | | | | |
| | 3535.02 | 1.65 | −1.50 | −1.26 | 0.17 | 3.02 | −103.35 | −1.33 | 0.875 |

"S–t" is short for "Student's t"

[a] Significant at the 5% level, in accordance with Kendall's tau, implies a value of 0.385 or more.

Note: * denotes significance at the 5% level, using Kendall's tau (critical region is −0.385 or less).

one of the $x_{ik}$s being equal to unity if necessary (the regression constant). The typical "normal" equations for this situation are as follows (from minimising $\sum_{i=1}^{n} \xi_i^2 \triangleq \Psi$) : [3]

$$\frac{\partial \Psi}{\partial \sigma} = \sum_{i=1}^{n} [y_i^{\sigma} - \sum_{j=1}^{k} a_j x_{ij}^{\rho_j}] y_i^{\sigma} \ln(y_i) = 0 \qquad (2.6.2a)$$

$$\frac{\partial \Psi}{\partial \rho_j} = \sum_{i=1}^{n} [y_i^{\sigma} - \sum_{j=1}^{k} a_j x_{ij}^{\rho_j}] x_{ij}^{\sigma_j} \ln(x_{ij}) = 0 \qquad (2.6.2b)$$

$$\frac{\partial \Psi}{\partial a_j} = \sum_{i=1}^{n} [y_i^{\sigma} - \sum_{j=1}^{k} a_j x_{ij}^{\rho_j}] x_{ij} = 0 \qquad (2.6.2c)$$

Equations (2.6.2a) and (2.6.2b) can be expressed in vector–matrix form as

$$\mathbf{y}_1(\boldsymbol{\rho}) = \mathbf{X}_1(\boldsymbol{\rho})\mathbf{a}, \qquad (2.6.3a)$$

and similarly for equations (2.6.2c),

$$\mathbf{y}_2(\boldsymbol{\rho}) = \mathbf{X}_2(\boldsymbol{\rho})\mathbf{a}, \qquad (2.6.3b)$$

where equation (2.6.3b) is a generalisation of the OLS normal equations.

This finding means that vector **a** can be eliminated by combining equations (2.6.3a) and (2.6.3b), yielding

$$\mathbf{y}_1(\boldsymbol{\rho}) = \mathbf{X}_1(\boldsymbol{\rho})\mathbf{X}_2^{-1}(\boldsymbol{\rho})\mathbf{y}_2(\boldsymbol{\rho}), \tag{2.6.4}$$

where the inverse term $\mathbf{X}_2^{-1}$ exists except in the presence of perfect multicollinearity or for $\boldsymbol{\rho} \equiv 0$. At least in principle, systems like equation (2.6.4) could be solved by Gauss–Seidel methods (see Hughes Hallett, 1984).

The specifications explored here are the following identical iterative equations, with $l$ denoting the iteration step, derived from equation (2.6.4):

$$\boldsymbol{\rho}_l = [\hat{\boldsymbol{\rho}}\hat{\mathbf{y}}_1^{-1}(\boldsymbol{\rho})X_1(\boldsymbol{\rho})X_2^{-1}(\boldsymbol{\rho})y_2(\boldsymbol{\rho})]_{l-1}, \tag{2.6.5a}$$

a multiplicative identity, and

$$\boldsymbol{\rho}_l = \boldsymbol{\rho}_{l-1} + [\mathbf{X}_1(\boldsymbol{\rho})X_2{-}1(\boldsymbol{\rho})y_2(\boldsymbol{\rho}) - \mathbf{y}_1(\boldsymbol{\rho})]_{l-1}, \tag{2.6.5b}$$

an additive identity.

The first estimation results obtained were disappointing.[4] Either different answers were converged upon, or no convergence occurred at all. A next step was to proceed directly with equations (2.6.2a) and (2.6.2b) in their identical additive form, the advantage being that they are mono–parametrical. This strategy rendered poor results, too. Hence the differential form of the latter equations, *i. e.*

$$\sum_{i=1}^{n} e_i y_i^{\sigma}[\sigma \ln(y_i) + 1]dy_i = 0, \tag{2.6.2a*}$$

and

$$\sum_{i=1}^{n} e_i x_{ij}^{e_j}[\rho_j \ln(x_{ij}) + 1]dx_{ij} = 0, \tag{2.6.2b*}$$

has not been investigated.

Inspection of the objective function $\boldsymbol{\Psi}$ shows that, given $\forall i, j$, $y_i$ and $x_{ij} > 1$, this function tends to zero for $\sigma$ and $\rho_j \to -\infty$. At best a local minimum can be found within given bounds. Perhaps the unit hypercube $[-\mathbf{i}, +\mathbf{i}]$ should be used as these bounds, given that it covers the interesting extreme cases, to wit linear regression $(\sigma, \rho_j = 1)$, double logarithmic regression $(\sigma, \rho_j \to 0)$, and inverse linear regression $(\sigma, \rho_j = -1)$. Then a global local optimum could be searched for by appropriate methods, *e. g.* simulated annealing (Laarhoven, 1988); other methods presently are being investigated.

## 3. Conclusions

Econometric estimation remains a difficult exercise. The first difficulty resides in *specifying* an approach that may lead to operational results, starting with the stated aim of an exercise. Common aims include better simulation, more precise estimations, the obtaining of parameter estimates with specific properties, exploitation of poor quality data, or handling non–linear or interdependent relationships. Econometric wisdom is useful, for attaining some of these objectives, but should be complemented by ingenious inventiveness to generate the mathematical entities that correspond to given requirements.

A second point of difficulty worth mentioning is that parameters to be estimated are not to be assumed as being independently given, from outside of the problem, but rather have to be viewed *relative to the use* that is to be made of their estimated values. Typical uses include prediction (again in a relative sense, and in terms of a specific purpose), simulation (again as specified with regard to its aims, such as analyzing system stability and sensitivity, or exploring consequences of either policy measures or exogenous shocks), hypothesis testing (here sometimes absolute parameter values are irrelevant, as can be seen in the foregoing discussion of QUALIREG and LLE).

A third, and final, difficulty acknowledged here concerns the *computation* of parameter values. A desirable specification renders simple, robust estimation procedures that produce achievable results. This extremely useful property should be welcomed, as well as sought in a chosen procedure.

Consequently, the spatial econometric findings reported in this paper, especially when couched in terms of the three difficulties outlined in this conclusion, illustrate the enormous possibilities that still remain for contributing to this subfield. The most beneficial contributions will combine spatial economic theory, general econometric wisdom, and numerical analysis, with each of these three fields continuing to hold immense potential for research and discovery.

## 4. References

Ancot, J–P., Iwema, R. et Paelinck, J. (1980) Test d'une hypothèse d'investissement à écarts multiples, *L'Actualité Economique*, **1**, 40-59.

Ancot, J–P., Kuiper, J. et Paelinck, J. (1981) Réflexions sur la simulation de modèles dynamiques, *Working Paper* No. 1981/2, Foundations of Empirical Economic Research Series, Netherlands Economic Institute, Rotterdam.

Ancot, J–P. and Paelinck, J. (1981) Recent research in spatial econometrics, in D. Griffith and R. MacKinnon (eds.), *Dynamic Spatial Models*, Sythoff & Noordhoff, Alphen a/d Rijn, pp. 344-364.

Ancot, J–P. and Paelinck, J. (1983) The spatial econometrics of the European FLEUR–model, in D. Griffith and A. Lea (eds.), *Evolving Geographical Structures*, Martinus Nijhoff Publishers, The Hague, pp. 229-246.

Ancot, J–P., Paelinck, J. and Prins, J. (1986) Some new estimators in spatial econometrics, *Economics Letters*, **21**, 245-250.

Ancot, J–P. et Paelinck, J. (1987) Nouveaux estimateurs en économétrie spatiale, *Working Paper* No. 1987/6, Foundations of Empirical Economic Research Series, Netherlands Economic Institute, Rotterdam.

Ancot, J–P. and Paelinck, J. Non–parametric estimation: the QUALIREG–model, its estimation and testing with spatial cross–sectional data, *Cahiers du GAMA*, forthcoming.

Bagchus, R., Nes, P. van en Zaan, A. van der. (1985) Stedelijke dynamiek, een schatting van het prooi–roofdier model voor Rotterdam (Urban dynamics, an estimation of the prey–predator model for the city of Rotterdam), *Working Paper*, Faculty of Economics, Department of Theoretical Spatial Economics, Erasmus University, Rotterdam.

Bailly, A. et Paelinck, J. (1988) Un modèle économétrique du développement socio–spatial

de régions touristiques, *Revue d'Economie Régionale et Urbaine*, **3**, 397-412.

Box, G. and Cox, D. (1964) An analysis of transformations, *Journal of the Royal Statistical Society*, **16B**, 211-243.

Budding, D., Cassa, H. en Hoek, P. van den. (1985) Het "prooi en roofdier"–model toegepast op de stedelijke bevolking (The prey–predator model applied to urban population), *Working Paper*, Faculty of Economics, Department of Theoretical Spatial Economics, Erasmus University, Rotterdam.

Davelaar, E., Jong, G. de en Koele, H. (1983) De methode `QUALIREG` (The `QUALIREG` method), *Working Paper*, Faculty of Economics, Department of Theoretical Spatial Economics, Erasmus University, Rotterdam.

Dickmann, A. en Spoorendonk, P. (1987) Een analyse van de stedelijke dynamiek van de gemeente Rotterdam (1946-1978) (An analysis of the urban dynamics of the city of Rotterdam), *Working Paper*, Faculty of Economics, Department of Theoretical Spatial Economics, Erasmus University, Rotterdam.

Enhus, J. (1986) Schatting van het `FLEUR`–model m.b.v. een Strictly Positive Conditional Estimator (Estimation of the `FLEUR` model by means of `SPCE`), unpublished Master's Thesis, Faculty of Economics, Department of Theoretical Spatial Economics, Erasmus University, Rotterdam.

Hughes Hallett, A. (1984) Simple and optimal extrapolations for first order iterations, *Journal of Computer Mathematics*, **15**, 309-318.

Kendall, M. (1962) *Rank Correlation Methods*, 3rd ed., Griffin, London.

Kuiper, J. (1989) Regional Analysis Using the Concept of Location Elasticities, *Revue d'Economie Régionale et Urbaine*, **3**, 363-391.

Laarhoven, P. van. (1988) *Theoretical and Computational Aspects of Simulated Annealing*, unpublished doctoral dissertation, Erasmus University, Rotterdam.

Loo, J. van. (1987) Schatten van een veralgemeend Lotka–Volterra model voor de stedelijke ontwikkeling in Nederland (Estimation of a generalised Lotka–Volterra model for the urban evolution in the Netherlands), unpublished Master's Thesis, Faculty of Economics, Department of Theoretical Spatial Economics, Erasmus University, Rotterdam.

Paelinck, J. and Klaassen, L. (1979) *Spatial Econometrics*, Saxon House, Farnborough; polish translation: *Ekonometria Przestrzenna*, Panstwowe Wydanictwo Naukowe, Warsawa, 1983.

Paelinck, J. (1989) Econométrie spatiale: contributions récentes après 20 ans d'histoire, *Cahiers Vilfredo Pareto*, forthcoming.

Peschel, M. and Mende, W. (1986) *The Predator–Prey Model*, Springer–Verlag, New York.

Prins, J. (1985) SDLS–schatting van een stedelijk model (SDLS–estimation of an urban model), unpublished Master's Thesis, Faculty of Economics, Department of Theoretical Spatial Economics, Erasmus University, Rotterdam.

Schueren, M. de van der (1986) Eindverslag SDLS schatting, stadsmodel en Lotka–Volterra (Final report SDLS estimation, urban model and Lotka–Volterra), *Working Paper*, Faculty of Economics, Department of Theoretical Spatial Economics, Erasmus University, Rotterdam.

## NOTES

1. Column–vectors are systematically represented by lower case bold letters, matrices by upper case bold letters except for diagonal matrices identified by a cap, and $'$ denotes matrix transposition.

2. $\tau = 1 - 4Q/[n(n-1)]$, where $Q$ is the number of so–called elementary permutations of $\mathbf{y}$ (see Kendall, 1962).

3. For small $\xi_i$, second order conditions for a minimum in $\sigma$ and $\rho_k$ will be satisfied for given $a_j$s.

4. Computer computations by Niek Mares are gratefully acknowledged.

5. Computations have shown the existence of various singular points satisfying equations (2.6.2); also see Footnote 3.

# DISCUSSION

## "Some new estimators in spatial econometrics"

## by J. H. P. Paelinck

In his chapter, Professor Paelinck presents six econometric estimators geared to problems encountered in spatial economics. They are the following: simultaneous dynamic least squares (SDLS), strictly positive conditional estimation (SPCE), linear logistic estimation (LLE), least spheres estimation (LSE), non–numerical regression (Qualireg), and distribution–free power estimation (DFPE). The latter is new, while the others are extensions and generalizations of earlier work, most notably in Ancot, Paelinck and Prins (1986), and Paelinck (1987).

Before I formulate some technical comments on the approaches suggested by Professor Paelinck, I would like to outline a more general organization of the six estimators based on three distinct perspectives or emphases, similar to the taxonomies suggested in my own chapter and in Anselin (1988a).

A first overall category is that of the particular perspective or paradigm represented in each method. Three distinct approaches are reflected in the six estimators. The classical perspective is taken for SDLS, LSE and LLE, which are examples of a maximum likelihood or pseudo–(quasi–)maximum likelihood (*e. g.*, Gourieroux *et al.*, 1984; White, 1984). The properties of these techniques are based upon asymptotic results, and hence they are alternative ways to solve estimation issues as problems of optimization (or best fit). A second approach is that reflected in the SPCE technique, where use is made of extraneous information to restrict the estimation, similar to Bayesian or Stein–like procedures (*e. g.*, Judge and Yancey, 1986). The crucial issue is how to derive the proper mix of data and prior information, which naturally leads to a concept of price associated with the imposed constraints. The third perspective is that of the nonparametric (robust) Qualireg and DFPE techniques, where limiting or otherwise unrealistic distributional assumptions are avoided.

A second important category of estimators involves the way in which specification issues particular to *spatial modeling* are taken into account. Foremost among these are the issues of spatial and space–time dependence as well as the various non–linearities associated with spatial interaction models (*e. g.*, distance decay functions, potentials). These topics are specifically addressed by SDLS, LSE and DFPE. A final category addresses the data limitations encountered in spatial analysis, namely problems of measurement (Qualireg) and positivity (SPCE).

In sum, the various new estimators suggested by Professor Paelinck focus on problems of simultaneity in spatial modeling, on data limitations encountered in spatial analysis, and on special complexities associated with spatial interaction. They compare to other recently advocated new directions, such as various shrinkage estimators (and the treatment of outliers), spatial adaptive filtering, and a spatial bootstrap estimator (for a review, see Anselin, 1988a).

From a technical standpoint, there are a number of issues raised by these new methods that merit closer scrutiny. First, the simultaneity expressed in most of the formulations (but especially for SDLS) raises questions of identifiability. In particular, the spatial structure inherent in the system under consideration somehow needs to be expressed in formal terms.

Whereas the SDLS approach is the most flexible one in this respect, as it avoids the familiar problem of assuming a weights matrix, it is conditional upon the availability of sufficient observations over space or space–time to allow for the identification of the structure of spatial interaction. In this respect, the re–introduction of a spatial weights matrix in the application of this technique to problems of spatial autocorrelation seems to be a step backward. Also, all six methods, to the extent that they deal with spatial dependence, presume spatial homogeneity, whereas spatial heterogeneity has been shown to be just as important a problem in empirical regional science (Anselin and Griffith, 1988).

The distributional properties of the various estimators are unclear. Based on asymptotic considerations, one may reasonably expect normality in most cases, but this is not necessarily reflected in the finite samples encountered in practice. The general issues involved in the trade–off between asymptotic normality and finite sample robustness are well–known, but the exact costs associated with each approach in practical empirical situations are less well understood. In addition, there is no unambiguous standard by which to compare the performance of the various new estimators in actual applications. Since most of the approaches are non–linear and do not necessarily result in residuals with a zero mean, the interpretation of the standard $R^2$ is not clear (Anselin, 1988b). The larger question is how to adequately summarize spatially differentiated or spatially dependent indicators of model accuracy. Unless this issue is addressed, there is really no standard by which to judge the superiority of these "new" approaches. Similarly, as yet there is no satisfactory way to assess the informational content of methods in which quantitative and qualitative measures are combined (as in Qualireg). Although the higher degree of realism expressed in the qualitative judgments in the data matrix is attractive, the degree of precision associated with the quantitative estimates remains unclear.

Finally, it may be interesting to compare some results for the spatial autoregressive model (2.1.6) between the SDLS approach and the more traditional regression approach. Using the notation of Paelinck, such a specification would be as follows:

$$\mathbf{y} = \rho\mathbf{C}\mathbf{y} + \mu\mathbf{i} + \xi$$

where $\mathbf{C}$ is the spatial weights matrix. With the simplifying condition (2.1.11), expression (2.1.18) finds the relation between $\rho$ and $\mu$ as:

$$\rho = (1/\nu)(\mu^*/\mu - 1).$$

As shown in Anselin (1988a), a conditional least squares estimate for $\mu$ is:

$$\mu = (\mathbf{i}'\mathbf{i})^{-1}\mathbf{i}'(\mathbf{I} - \rho\mathbf{C})\mathbf{y}$$

or, with $\mathbf{i}'\mathbf{i} = r$, and a symmetric matrix $\mathbf{C}$, $(\mathbf{C}\mathbf{i})' = \mathbf{i}'\mathbf{C} = (\nu\mathbf{i})'$,

$$\mu = (1/r)(\mathbf{i}'\mathbf{y} - \rho\mathbf{i}'\mathbf{C}\mathbf{y}),$$
$$\mu = (1 - \rho\nu)(\mathbf{i}'\mathbf{y}/r)$$

or, in Paelinck's notation, with $\mathbf{i}'\mathbf{y}/r = \mu^*$:

$$\mu = (1 - \rho\nu)\mu^* \qquad \text{and}$$
$$\rho = (1/\nu)(1 - \mu/\mu^*).$$

This result is not the same as expression (2.1.18), except for the uninteresting case where $\mu = \mu^\star$ and thus $\rho = 0$. This simple derivation illustrates that the various approximations implied by the non–linear estimators are not trivial, and may have significant consequences for the ultimate results. The exact nature of these consequences needs to be investigated in further detail. The innovative methods suggested by Professor Paelinck provide a challenge to other researchers in spatial statistics to pursue this more extensively, both from a formal as well as from an empirical viewpoint.

## References

Ancot, J-P., J. Paelinck, and J. Prins. (1986) Some new estimators in spatial econometrics. *Economics Letters*, **21**, 245-249.

Anselin, L. (1988a) *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer Academic Publishers.

Anselin, L. (1988b) Model validation in spatial econometrics: a review and evaluation of alternative approaches. *International Regional Science Review*, **11**, 279-316.

Anselin, L., and D. Griffith. (1988) Do spatial effects really matter in regression analysis? *Papers*, Regional Science Association, **65**, 11-34.

Gourieroux, C., A. Monfort, and A. Trognon. (1984) Pseudo maximum likelihood methods: theory. *Econometrica*, **52**, 681-700.

Judge, G., and T. Yancey. (1986) *Improved Methods of Inference in Econometrics*. Amsterdam: North Holland.

Paelinck, J. (1987) Empirical results with some new estimators in spatial econometrics, paper presented at the 27th Meeting of the Western Regional Science Association, Napa Valley, February.

White, H. (1984) *Asymptotic Theory for Econometricians*. New York: Academic Press.

*Luc Anselin*, University of California/Santa Barbara

## *PREAMBLE*

*'Tis a lesson you should heed:*
*Try, try, try again.*
*If at first you don't succeed,*
*Try, try, try again.*

W. E. Hickson, **Try and Try Again**

*In 1980 Griffith proposed the notion that the Jacobian term for spatial autoregressive models converges upon some constant as the sample size increases to infinity. Several subsequently published pieces severely criticized this idea, using very cogent arguments. Griffith's intuition led him to numerical investigations concerning this issue, and in 1988 he reported convincing but not totally conclusive results supporting it. The purpose of this paper is to report quite conclusive numerical results obtained from supercomputer experiments. Given this brief history, the author is truly optimistic about findings contained in this technical report, even though he constantly bumps into scholars who do not share his enthusiasm (for example, Martin's contribution to this volume). Some of both this enthusiastic and this disheartening viewpoint may be sensed in Ord's commentary on the paper, as he raises questions concerning the sensibleness of answers, on the one hand, and computational manageability, on the other hand.*

The Editor

# A Numerical Simplification for Estimating Parameters of Spatial Autoregressive Models

Daniel A. Griffith *

*Department of Geography, Syracuse University, Syracuse, NY, 13244-1160, U. S. A.*

**Overview:** The Jacobian term appears in likelihood functions to ensure that the use of variable transformations still leads to probability density functions whose complete integration yields unity. This term is particularly troublesome when dealing with spatial autoregressive models, since it does not disappear in the optimization process, and hence requires numerically intensive solutions to the parameter estimation problem. For these sorts of autoregressive models the Jacobian term is a function of the eigenvalues of the connectivity matrix that depicts the geographic configuration of those areal units under study. For a tessellation of $n$ areal units, then, the eigenvalues of an $n$–by–$n$ matrix need to be calculated. Ord has stated the equations for these eigenvalues when a regular lattice configuration is superimposed upon an infinite surface. Griffith has shown what the algebraic expression of the Jacobian term converges to for this same infinite surface situation. The problem addressed in this paper asks what implications these two simplifications have on parameter estimation for geographically referenced data.

## 1. Introduction

One reason spatial regression accommodating geographic dependence is so numerically intensive is that the Jacobian of the transformation from an autocorrelated space to an unautocorrelated space must be included in parameter estimation procedures. A Jacobian term is some function of the number of areal units as well as the degree of spatial dependence. Ord (1975) states that the eigenvalues of a binary configuration matrix for a regular lattice are given by the equation

$$\lambda_{kl} = 2\{\cos[k\pi/(n+1)] + \cos[l\pi/(n+1)]\} \tag{1.1}$$

More specifically, the spatial autoregressive parameter is a function of the geographic configuration characterized by this Jacobian, which in most popular models is written in terms of the eigenvalues of the $n^2$ matrix for this transformation determinant. Questions concerning the accuracy and feasibility of numerically extracted eigenvalues for a given Jacobian, derived from matrices of such large dimensions, have been posed by spatial analysts, and apparently represents a barrier to the dissemination of spatial statistics and spatial econometrics. A frequency distribution for geographic data set size would be sinusoidal or reverse–J shaped; there are numerous data sets where $n$ is quite small, virtually none where $n$ is of moderate

* This research was supported by NSF grant SES–87–22086, with much of its computing being cosponsored by an NSF grant to the National Center for Supercomputing Applications (NCSA), University of Illinois at Urbana–Champaign, for the 1989 Supercomputer Education Summer Workshop, July 10-21. This paper was presented to the Sixth European Colloquium on Theoretical and Quantitative Geography, Chantilly, France, September 5–9, 1989.

size, and some where $n$ is extremely large (most of which are generated from remotely sensed satellite images). The purpose of this paper is to explore ways of mathematically simplifying the calculation of this Jacobian term, especially to help in the analysis of intermediate and large size data sets, and is part of a comprehensive attempt to remove obstacles hindering the diffusion of spatial statistical technology.

### 1.1. Background

In some respects this present investigation is an extension of two previous undertakings. Griffith (1988a) began exploring possible mathematical simplifications of the Jacobian term for a simultaneous autoregressive model that is based upon a binary connectivity matrix and a regular square lattice configuration of areal units; his study was aimed at remotely sensed data situations. His findings included (1) that there are certain systematic regularities in the Jacobian term as $n$ increases, and (2) even using the analytical equations to compute eigenvalues resulted in considerable rounding error for $10,000$–by–$10,000$ lattices using double–precision FORTRAN on a DEC VAX mainframe. Based upon the numerical findings that were tabulated and reported by Griffith (1988a), searches for the analytical expression describing convergence of the Jacobian term could be restricted to three possible candidates; however, the serious rounding error that prevailed prevented identification of the correct expression from these three.

More recently Griffith (1990) has examined the computation of the Jacobian term, in some cases using double precision, on a Cray 2 supercomputer. His reported results show conclusively that (1) the eigenvalues of a matrix can be computed with a high degree of accuracy for at least $n = 100,000,000$ (this is equivalent to a $10,000$–by–$10,000$ regular square lattice), (2) for a regular square lattice, as $n$ goes to infinity, the Jacobian term for a conditional autoregressive model (which will be the subject of this paper) converges upon the expression

$$-\int_0^{\pi}\int_0^{\pi} \ln\{1 - 2\rho[\cos(\theta_1) + \cos(\theta_2)]\}/\pi^2 \; d\theta_1 \, d\theta_2, \tag{1.2}$$

which can be numerically integrated (see Table 1), and (3) that the parameter estimation impact of the Jacobian term diminishes in importance as $n$ goes to infinity. Equation (1.2) is the continuous version of and is converged upon by

$$-\sum_{k=1}^{m}\sum_{l=1}^{m} \ln\Big[1 - 2\rho\{\cos[k\pi/(m+1)] + \cos[l\pi/(m+1)]\}\Big]/n, \tag{1.3}$$

where $m^2 = n$ (or $m = \sqrt{n}$), which represents the Jacobian term for finite square lattices. As one can see from expressions (1.2) and (1.3), the Jacobian term is a mean; it should not be surprising from a statistical perspective, then, to find that this quantity converges as $n$ goes to infinity.

These two previous studies have set the stage for the analysis presented in this paper. Here attention will be restricted to the Jacobian term for a conditional autoregressive model.

TABLE 1
SELECTED RESULTS FOR THE NUMERICAL SOLUTION OF EQUATION (1.2)

| rho | integral value | error | rho | integral value | error |
|---|---|---|---|---|---|
| 0.025 | 0.0012535321 | 0.0000000000 | 0.150 | 0.0505218648 | 0.0000000000 |
| 0.050 | 0.0050573165 | 0.0000000000 | 0.200 | 0.1014553111 | 0.0000000000 |
| 0.100 | 0.0209735079 | 0.0000000000 | 0.250 | 0.2200507460 | 0.0000000003 |

NOTE: Numerical integration has been achieved with the IMSL10 routine E2LSF.

## 2. A Jacobian term equation with $\rho$ varying for selected $n$

Two interesting limiting cases of $n$ for which one might assess variation in the Jacobian term as the spatial autoregressive parameter $\rho$ changes are its lower limit, where for a regular square lattice $\sqrt{n} = 2$, and its upper limit, where for a regular square lattice $\sqrt{n} = \infty$. Five–hundred–and–one Jacobian terms were computed for each of these two cases, using values of $\rho$ that started with the limiting parameter space boundary $-0.25$, and were sequentially incremented by $0.001$, until $0.25$ was reached. An analysis of these two sets of results lead to the formulation of an equation describing how the Jacobian term changes over the possible natures and degrees of spatial dependence. Next, various intermediate values of $\sqrt{n}$ were studied, using twenty–one uniformly spaced values of $\rho$ across the feasible parameter space (namely, $\pm 0.25$, $\pm 0.225$, $\pm 0.2$, $\pm 0.175$, $\pm 0.15$, $\pm 0.125$, $\pm 0.10$, $\pm 0.075$, $\pm 0.5$, $\pm 0.025$, and $0.0$), and yielded the tabulated numerical results presented in Table 2. General tendencies present in this table include (1) a mean squared error value that increases with $n$, but never to a non–negligible level, (2) asymptotically converging estimates for the two parameters $\beta_n$ and $\gamma_n$, with very little difference in subsequent values for these parameters beyond $n = 900$ (the correlation between results for $n = 2^2$ and $n = \infty$ is $0.987$), and (3) a value of $\beta_n$ which is approximately twice the value of $\gamma_n$. To illustrate these findings within their equational context, the two limiting cases would yield the following equations:

$$\sqrt{n} = 2: \quad J = \ln(0.5)/2 - 0.25\,\ln(0.5+\rho) - 0.25\,\ln(0.5-\rho), \text{ and}$$
$$\sqrt{n} = \infty: \; J = -0.377580 - 0.150659\,\ln(0.285620+\rho) - 0.150659\,\ln(0.285620-\rho).$$

These findings imply that the Jacobian term is a concave–upwards function, whose general form is

$$J = 2\beta_n \ln(\gamma_n) - \beta_n \ln(\gamma_n + \rho) - \beta_n \ln(\gamma_n - \rho), \tag{2.1}$$

having an increasingly shallower trough as $n$ increases. This feature is consistent with the aforementioned contention that the importance of the Jacobian term diminishes as n increases. As is indicated by their subscripts, the parameters $\beta_n$ and $\gamma_n$ are functions of the number of areal units under study.

As $\sqrt{n}$ increases toward infinity, a slight bias seems to appear in the computations of $\beta_n$ and $\gamma_n$. This slight bias may well be attributable to the residual heteroscedasticity stemming from some systematic error component arising in the numerical eigenvalue extraction algorithm, or possibly from specification error.

Presumably for a regular rectangular lattice equation (2.1) would become

$$J = \alpha_n - \beta_n \ln(\gamma_n + \rho) - \beta_n \ln(\gamma_n - \rho), \tag{2.2}$$

TABLE 2
NONLINEAR REGRESSION PARAMETER ESTIMATES
OF SELECTED JACOBIAN TERMS

| $\sqrt{n}$ of square lattice | $\beta_n$ | $\gamma_n$ | MSE | $\sqrt{n}$ of square lattice | $\beta_n$ | $\gamma_n$ | MSE |
|---|---|---|---|---|---|---|---|
| 2 | 0.250000 | 0.500000 | 0.000000000 | 54 | 0.153596 | 0.289884 | 0.000001983 |
| 4 | 0.185791 | 0.354859 | 0.000000029 | 56 | 0.153498 | 0.289733 | 0.000002005 |
| 6 | 0.173667 | 0.327639 | 0.000000155 | 58 | 0.153405 | 0.289592 | 0.000002026 |
| 8 | 0.168056 | 0.315926 | 0.000000331 | 60 | 0.153318 | 0.289461 | 0.000002045 |
| 10 | 0.164745 | 0.309377 | 0.000000512 | 62 | 0.153237 | 0.289338 | 0.000002064 |
| 12 | 0.162536 | 0.305183 | 0.000000680 | 64 | 0.153162 | 0.289224 | 0.000002081 |
| 14 | 0.160951 | 0.302264 | 0.000000830 | 66 | 0.153091 | 0.289116 | 0.000002097 |
| 16 | 0.159752 | 0.300112 | 0.000000963 | 68 | 0.153023 | 0.289014 | 0.000002112 |
| 18 | 0.158816 | 0.298462 | 0.000001080 | 70 | 0.152958 | 0.288917 | 0.000002127 |
| 20 | 0.158058 | 0.297151 | 0.000001183 | 72 | 0.152900 | 0.288828 | 0.000002142 |
| 22 | 0.157433 | 0.296087 | 0.000001275 | 74 | 0.152842 | 0.288741 | 0.000002155 |
| 24 | 0.156909 | 0.295205 | 0.000001357 | 76 | 0.152787 | 0.288660 | 0.000002168 |
| 26 | 0.156462 | 0.294461 | 0.000001430 | 78 | 0.152736 | 0.288583 | 0.000002179 |
| 28 | 0.156077 | 0.293826 | 0.000001495 | 80 | 0.152687 | 0.288510 | 0.000002191 |
| 30 | 0.155742 | 0.293278 | 0.000001555 | 82 | 0.152640 | 0.288440 | 0.000002202 |
| 32 | 0.155446 | 0.292798 | 0.000001609 | 84 | 0.152594 | 0.288373 | 0.000002213 |
| 34 | 0.155184 | 0.292376 | 0.000001658 | 86 | 0.152551 | 0.288310 | 0.000002223 |
| 36 | 0.154950 | 0.292002 | 0.000001704 | 88 | 0.152512 | 0.288250 | 0.000002233 |
| 38 | 0.154737 | 0.291666 | 0.000001745 | 90 | 0.152470 | 0.288191 | 0.000002243 |
| 40 | 0.154547 | 0.291365 | 0.000001783 | 92 | 0.152434 | 0.288136 | 0.000002252 |
| 42 | 0.154375 | 0.291093 | 0.000001818 | 94 | 0.152397 | 0.288083 | 0.000002260 |
| 44 | 0.154217 | 0.290845 | 0.000001850 | 96 | 0.152363 | 0.288033 | 0.000002268 |
| 46 | 0.154072 | 0.290620 | 0.000001881 | 98 | 0.152330 | 0.287984 | 0.000002277 |
| 48 | 0.153940 | 0.290413 | 0.000001909 | 100 | 0.152298 | 0.287937 | 0.000002284 |
| 50 | 0.153816 | 0.290222 | 0.000001935 | 1000 | 0.150835 | 0.285857 | 0.000002664 |
| 52 | 0.153703 | 0.290047 | 0.000001960 | $\infty$ | 0.150659 | 0.285620 | 0.000002712 |

NOTE 1: The case of infinity had the Jacobian terms computed with IMSL10 subroutine E2LSF (see Table 1).

NOTE 2: Using a division of the feasible parameter space $[-0.25, 0.25]$ into 501 values resulted in almost exactly the same values for the case of $\sqrt{n} = 2$, but noticeably different values for $\sqrt{n} = \infty$ ($\beta_\infty = 0.163846$, $\gamma_\infty = 0.293988$, and MSE $= 0.000001457$). This latter inconsistency suggests that either the numerical integration, or the numerical eigenvalue extraction, is plagued with error.

NOTE 3: For the above cases $R^2 = 1.000$, the Wilk–Shapiro statistic $= 0.969$ for $\sqrt{n} = 2$ and asymptotically converges on 0.931 as $n$ increases, the Durbin–Watson statistic $= 2.82$ for $\sqrt{n} = 2$ and apparently converges in an oscillatory fashion on 2.10 as $n$ increases, and there is no apparent heteroscedasticity displayed by the residuals for $\sqrt{n} = 2$, with increasingly systematic, complex, nonlinear heteroscedasticity displayed by the residuals as $n$ increases, together with extreme values becoming influential estimation points as $n$ increases.

where the parameter $\alpha_n$ no longer is constrained to be a function of $\beta_n$ and $\gamma_n$ [such as the term $2\beta_n \ln(\gamma_n)$ appearing in equation (2.1)], and for an irregular lattice equation (2.1)

would become

$$J = \alpha_n - \beta_{1,n}\ln(\gamma_{1,n} + \rho) - \beta_{2,n}\ln(\gamma_{2,n} - \rho). \tag{2.3}$$

These last two conjectures require considerable subsequent investigation.

Given equation (2.1), the log–likelihood function, say $\ln(L)$, to be optimized when calculating a maximum likelihood estimate of the spatial autoregressive parameter $\rho$ becomes

$$\begin{aligned} K - (n/2)\ln(\sigma^2) - (n/2)[2\beta_n\gamma_n - \beta_n\ln(\gamma_n + \rho) - \beta_n\ln(\gamma_n - \rho)] \\ - (\mathbf{X} - \mu\mathbf{1})^t(\mathbf{I} - \rho\mathbf{C})(\mathbf{X} - \mu\mathbf{1})/(2\sigma^2), \end{aligned} \tag{2.4}$$

where $K = -(n/2)\ln(2\pi)$ is a constant, $\mathbf{X}$ is an $n$–by–1 data vector, $\mathbf{1}$ is an $n$–by–1 vector of ones, and $\mathbf{C}$ is an $n$–by–$n$ binary geographic configuration matrix (upon which the Jacobian term is based). Parameter estimation based upon equation (2.4) currently requires nonlinear optimization techniques.

## 3. Parameter estimation based upon the simplified Jacobian term

Four different estimation cases can be explored for the likelihood function portrayed by expression (2.4), each referring to a combinatorial possibility of unknown parameter values. In the first of these cases, suppose that only $\rho$ is unknown (in other words, let $\mu$ and $\sigma^2$ be known). Optimizing expression (2.4) with respect to $\rho$ yields

$$\frac{\partial\ln(L)}{\partial\rho} = -(n/2)[-\beta_n/(\gamma_n + \rho) + \beta_n/(\gamma_n - \rho)] + [(\mathbf{X} - \mu\mathbf{1})^t\mathbf{C}(\mathbf{X} - \mu\mathbf{1})]/(2\sigma^2) = 0,$$

which when solved produces a quadratic equation in $\rho$ having roots

$$\hat{\rho} = -n\sigma^2\beta_n/(\mathbf{X} - \mu\mathbf{1})^t\mathbf{C}(\mathbf{X} - \mu\mathbf{1}) \pm \{[-n\sigma^2\beta_n/(\mathbf{X} - \mu\mathbf{1})^t\mathbf{C}(\mathbf{X} - \mu\mathbf{1})]^2 + \gamma_n^2\}^{1/2}. \tag{3.1}$$

Consequently, the spatial autocorrelation parameter becomes an explicit function of the size of the geographic data series, as well as the configuration of the underlying areal unit surface partitioning. One should expect this definition of $\rho$ always to be real, and always to fall within the feasible parameter space region. This finding is particularly useful for remotely sensed data analysis, for once the parameters $\beta_n$ and $\gamma_n$ are established, then all one needs to know is the size of the regular square lattice partitioning in order to estimate $\rho$; numerical computation of eigenvalues no longer will be necessary.

In the second case one can assume that both $\mu$ and $\rho$ are unknown, and only $\sigma^2$ is known. Now optimizing expression (2.4) with respect to $\mu$ and $\rho$ yields the standard maximum likelihood estimation (MLE) result of

$$\hat{\mu} = \mathbf{1}^t(\mathbf{I} - \rho\mathbf{C})\mathbf{X}/\mathbf{1}^t(\mathbf{I} - \rho\mathbf{C})\mathbf{1}, \tag{3.2}$$

and hence the differential equation

$$\frac{\partial\ln(L)}{\partial\rho} = -(n/2)[-\beta_n/(\gamma_n + \rho) + \beta_n/(\gamma_n - \rho)] + [(\mathbf{X} - \hat{\mu}\mathbf{1})^t\mathbf{C}(\mathbf{X} - \hat{\mu}\mathbf{1})]/(2\sigma^2) = 0,$$

which when solved produces a quartic equation in $\rho$ of the form

$$
\begin{aligned}
&-[(1^t\mathbf{C}1\mathbf{X}^t - 1^t\mathbf{C}\mathbf{X}1^t)\mathbf{C}\mathbf{X}1^t\mathbf{C}1]\rho^4 \\
&+2[-n\beta_n\sigma^2(1^t\mathbf{C}1)^2 + n\mathbf{X}^t\mathbf{C}\mathbf{X}1^t\mathbf{C}1 - n(1^t\mathbf{C}\mathbf{X})^2]\rho^3 \\
&+\{(1^t\mathbf{C}1\mathbf{X}^t - 1^t\mathbf{C}\mathbf{X}1^t)\mathbf{C}\mathbf{X}1^t\mathbf{C}1\gamma_n^2 \\
&\quad +[4n^2\beta_n\sigma^2 - (1^t\mathbf{X})^2]1^t\mathbf{C}1 - n^2\mathbf{X}^t\mathbf{C}\mathbf{X} + 2n1^t\mathbf{C}\mathbf{X}1^t\mathbf{X}\}\rho^2 \\
&-2\{n^3\beta_n\sigma^2 + n[\mathbf{X}^t\mathbf{C}\mathbf{X}1^t\mathbf{C}1 - (1^t\mathbf{C}\mathbf{X})^2]\gamma_n^2\}\rho \\
&+[n^2\mathbf{X}^t\mathbf{C}\mathbf{X} + (1^t\mathbf{C}11^t\mathbf{X} - 2n1^t\mathbf{C}\mathbf{X})1^t\mathbf{X}]\gamma_n^2 = 0.
\end{aligned}
$$

The solution to a biquadratic equation of this kind is presented in theory of equation texts, such as the classic by Uspensky (1948, pp. 94-97). An algorithmic solution to solving the underlying pair of simultaneous differential equations also can be pursued, if one wishes to avoid extracting roots of a fourth–order equation. The iterative algorithm would be of the form

Step 1: let $\rho = 0 \Rightarrow \hat{\mu} = \mathbf{1}^t\mathbf{X}/n$ (for iteration $\tau = 0$);

Step 2: solve $\hat{\rho}$ for equation (3.1), in Case I;

Step 3: compute $\hat{\mu}_{\tau+1} = \mathbf{1}^t(\mathbf{I} - \hat{\rho}_\tau\mathbf{C})\mathbf{X}/\mathbf{1}^t(\mathbf{I} - \hat{\rho}_\tau\mathbf{C})\mathbf{1}$; and,

Step 4: iterate through Steps 2 and 3 until the parameter estimates converge (this and subsequent algorithms are believed to converge, although no proof of convergence is offered here; at worst they should be good heuristics).

For the third case consider both $\sigma^2$ and $\rho$ to be unknown, with only $\mu$ being known (for example, the case of regression residuals). Now optimizing expression (2.4) with respect to $\sigma^2$ and $\rho$ yields the standard MLE result of

$$\hat{\sigma}^2 = (\mathbf{X} - \mu\mathbf{1})^t(\mathbf{I} - \rho\mathbf{C})(\mathbf{X} - \mu\mathbf{1})/n, \tag{3.3}$$

and hence the differential equation

$$\frac{\partial \ln(L)}{\partial \rho} = -(n/2)[-\beta_n/(\gamma_n + \rho) + \beta_n/(\gamma_n - \rho)] + [(\mathbf{X} - \mu\mathbf{1})^t\mathbf{C}(\mathbf{X} - \mu\mathbf{1})]/(2\hat{\sigma}^2) = 0,$$

which when solved produces a quadratic equation in $\rho$ having roots

$$
\begin{aligned}
\hat{\rho} = &- [\beta_n/(1 - 2\beta_n)][(\mathbf{X} - \mu\mathbf{1})^t(\mathbf{X} - \mu\mathbf{1})/(\mathbf{X} - \mu\mathbf{1})^t\mathbf{C}(\mathbf{X} - \mu\mathbf{1})] \\
&\pm \{[\beta_n^2/(1 - 2\beta_n)^2][(\mathbf{X} - \mu\mathbf{1})^t(\mathbf{X} - \mu\mathbf{1})/(\mathbf{X} - \mu\mathbf{1})^t\mathbf{C}(\mathbf{X} - \mathbf{1})]^2 \\
&+ \gamma_n^2/(1 - 2\beta_n)\}^{1/2}.
\end{aligned} \tag{3.4}
$$

An algorithmic solution to solving the underlying pair of simultaneous differential equations can be pursued in this case, too, although it is doubtful if one ever would seriously wish to avoid calculating the pair of roots. The iterative algorithm would be of the form listed in the following steps:

1: let $\rho = 0 \Rightarrow \hat{\sigma}^2 = (\mathbf{X} - \mu\mathbf{1})^t(\mathbf{X} - \mu\mathbf{1})/n$ (for iteration $\tau = 0$);

2: solve $\hat{\rho}$ for equation (3.1), in Case I;

3: compute $\hat{\sigma}^2_{\tau+1} = (\mathbf{X} - \mu\mathbf{1})^t(\mathbf{I} - \hat{\rho}_\tau\mathbf{C})(\mathbf{X} - \mu\mathbf{1})/n$; and,

4: iterate through Steps 2 and 3 until the parameter estimates converge.

The fourth, and final, case to be treated here is the more likely situation that none of the parameters are known, or $\mu$, $\sigma^2$ and $\rho$ are unknown. Here optimizing expression (2.4) with respect to $\mu$, $\sigma^2$ and $\rho$ yields the standard MLE results appearing in equations (3.2) and (3.3), as well as the differential equation

$$\frac{\partial \ln(L)}{\partial \rho} = -(n/2)[-\beta_n/(\gamma_n+\rho) + \beta_n/(\gamma_n-\rho)] + [(\mathbf{X}-\hat{\mu}\mathbf{1})^{\mathbf{t}}\mathbf{C}(\mathbf{X}-\hat{\mu}\mathbf{1})]/(2\sigma^2) = 0,$$

which when solved produces a quartic equation in $\rho$ of the form

$$\begin{aligned}
&(2\beta_n - 1)[(\mathbf{1}^t\mathbf{C}\mathbf{1}\mathbf{X}^t - \mathbf{1}^t\mathbf{C}\mathbf{X}\mathbf{1}^t)\mathbf{C}\mathbf{X}\mathbf{1}^t\mathbf{C}\mathbf{1}]\rho^4 \\
&+ 2\{(2\beta_n - 1)n[(\mathbf{1}^t\mathbf{C}\mathbf{X})^2 - \mathbf{X}^t\mathbf{C}\mathbf{X}\mathbf{1}^t\mathbf{C}\mathbf{1}] \\
&\quad - \beta_n[(\mathbf{1}^t\mathbf{C}\mathbf{1})^2\mathbf{X}^t\mathbf{X} - 2\mathbf{1}^t\mathbf{C}\mathbf{1}\mathbf{1}^t\mathbf{X}\mathbf{1}^t\mathbf{C}\mathbf{X} + n(\mathbf{1}^t\mathbf{C}\mathbf{X})^2]\}\rho^3 \\
&+ \{4\beta_n\mathbf{1}^t\mathbf{C}\mathbf{1}[n\mathbf{X}^t\mathbf{X} - (\mathbf{1}^t\mathbf{X})^2] + \gamma_n^2\mathbf{1}^t\mathbf{C}\mathbf{1}[\mathbf{X}^t\mathbf{C}\mathbf{X}\mathbf{1}^t\mathbf{C}\mathbf{1} - (\mathbf{1}^t\mathbf{C}\mathbf{X})^2] \\
&\quad + (2\beta_n - 1)[n^2\mathbf{X}^t\mathbf{C}\mathbf{X} - 2n\mathbf{1}^t\mathbf{C}\mathbf{X}\mathbf{1}^t\mathbf{X} + \mathbf{1}^t\mathbf{C}\mathbf{1}(\mathbf{1}^t\mathbf{X})^2]\}\rho^2 \\
&+ 2n\{\beta_n[(\mathbf{1}^t\mathbf{X})^2 - n\mathbf{X}^t\mathbf{X}] + \gamma_n^2[(\mathbf{1}^t\mathbf{C}\mathbf{X})^2 - \mathbf{1}^t\mathbf{C}\mathbf{1}\mathbf{X}^t\mathbf{C}\mathbf{X}]\}\rho \\
&+ [n^2\mathbf{X}^t\mathbf{C}\mathbf{X} + (\mathbf{1}^t\mathbf{C}\mathbf{1}\mathbf{1}^t\mathbf{X} - 2n\mathbf{1}^t\mathbf{C}\mathbf{X})\mathbf{1}^t\mathbf{X}]\gamma_n^2 = 0. \qquad (3.5)
\end{aligned}$$

Two possible algorithmic solutions to solving the underlying triplet of simultaneous differential equations can be pursued in this case, if one wishes to avoid calculating the roots of a fourth–degree polynomial. One iterative algorithm could be of the form listed in the following steps:

A–1: let $\rho = 0 \Rightarrow \hat{\mu} = \mathbf{1}^t\mathbf{X}/n$ and $\hat{\sigma}^2 = (\mathbf{X}-\hat{\mu}\mathbf{1})^t(\mathbf{X}-\hat{\mu}\mathbf{1})/n$ (for iteration $\tau = 0$);

A–2: solve $\hat{\rho}$ for equation (3.1), in Case I;

A–3: first compute

$$\hat{\mu}_{\tau+1} = \mathbf{1}^t(\mathbf{I}-\hat{\rho}_\tau\mathbf{C})\mathbf{X}/\mathbf{1}^t(\mathbf{I}-\hat{\rho}_\tau\mathbf{C})\mathbf{1},$$

and then compute

$$\hat{\sigma}^2_{\tau+1} = (\mathbf{I}-\hat{\mu}_{\tau+1}\mathbf{1})^t(\mathbf{I}-\hat{\rho}_\tau\mathbf{C})(\mathbf{X}-\hat{\mu}_{\tau+1}\mathbf{1})/n;$$

and,

A–4: iterate through Steps A–2 and A–3 until the parameter estimates converge.

An alternative algorithm would be of the form listed in the following steps:

B–1: let $\rho = 0 \Rightarrow \hat{\mu} = \mathbf{1}^t\mathbf{X}/n$ (for iteration $\tau = 0$);

B–2: solve $\hat{\rho}$ for equation (3.4), in Case III;

B–3: compute $\hat{\mu}_{\tau+1} = \mathbf{1}^t(\mathbf{I}-\hat{\rho}_\tau\mathbf{C})\mathbf{X}/\mathbf{1}^t(\mathbf{I}-\hat{\rho}_\tau\mathbf{C})\mathbf{1}$;

B–4: iterate through Steps B–2 and B–3 until the parameter estimates converge; and,

B–5: $\hat{\sigma}^2 = (\mathbf{I}-\hat{\mu}\mathbf{1})^t(\mathbf{I}-\hat{\rho}\mathbf{C})(\mathbf{X}-\hat{\mu}\mathbf{1})/n$.

Intuitively speaking, for both Cases II and IV, at least two of the roots of their fourth–degree equations [such as (3.5)] must be real; accordingly, at most two roots can be complex.

An interesting mathematical exercise would be to prove this conjecture, perhaps using Ferrari's solution technique for biquadratic equations. At most, one of the roots should fall into the feasible parameter space; conditions governing the existence of this category of solution need to be established. The remaining four combinatorial possibilities for known and unknown parameters are of no interest here, since they do not involve the estimation of the spatial autoregressive parameter $\rho$. Finally, if the expression $(\mathbf{I} - \hat{\mu}\mathbf{1})^t\mathbf{C}(\mathbf{X} - \hat{\mu}\mathbf{1}) = 0$, then $\hat{\rho} = 0$.

## 4. Examples

A contrived example will be described in this section in order to illustrate the four estimation cases outlined in the preceding section. For this example imagine that $\mu = 0$, $\sigma^2 = 1$, $n = 4$ (so $\beta_2 = 0.25$ and $\gamma_2 = 0.5$; see Table 2), and the geographic distribution in question is

| | |
|---|---|
| 0 | 0 |
| 1 | -1 |

for which $\mathbf{1}^t\mathbf{X} = 0$, $\mathbf{X}^t\mathbf{X} = 2$, $\mathbf{1}^t\mathbf{C}\mathbf{X} = 0$, and $\mathbf{X}^t\mathbf{C}\mathbf{X} = -2$. The four eigenvalues of matrix $\mathbf{C}$ for this geographic configuration are $-2$, $0$, $0$, and $2$; hence, the spatial autoregressive parameter estimate has the restriction $-1/2 < \hat{\rho} < 1/2$.

Case I:

$$\hat{\rho} = 2\beta_n \pm (4\beta_n^2 + \gamma_n^2)^{1/2} = (1 \pm \sqrt{2})/2,$$

which means that $\hat{\rho} = (1 - \sqrt{2})/2$ satisfies the accompanying constraint;

Case II:

$$\begin{aligned} 4\rho^4 - 4(4\beta_n + 1)\rho^3 + (16\beta_n - 4\gamma_n^2 + 1)\rho^2 - 4(\beta_n - \gamma_n^2)\rho - \gamma_n^2 &= 0 \\ (\rho^2 - 4\beta_n\rho - \gamma_n^2)(4\rho^2 - 4\rho + 1) &= 0 \\ (\rho^2 - 4\beta_n\rho - \gamma_n^2) &= 0 \end{aligned}$$

is the equation for Case I, which is what would be expected since the sample mean, the MLE of the mean, and the population mean are identical;

Case III:

$$\begin{aligned} \hat{\rho} &= \beta_n/(1 - 2\beta_n) \pm [\beta_n^2/(1 - 2\beta_n)^2 + \gamma_n^2/(1 - 2\beta_n)]^{1/2} \\ &= (1 \pm \sqrt{3})/2, \end{aligned}$$

which means that $\hat{\rho} = (1 - \sqrt{3})/2$ satisfies the accompanying constraint; and,

Case IV:

$$\begin{aligned} 8(2\beta_n - 1)\rho^4 + 8\rho^3 + 2(4\gamma_n^2 - 6\beta_n - 1)\rho^2 + 4(\beta_n - 2\gamma_n^2)\rho + 2\gamma_n^2 &= 0 \\ [(2\beta_n - 1)\rho^2 + 2\beta_n\rho + \gamma_n^2](8\rho^2 - 8\rho + 2) &= 0 \\ (2\beta_n - 1)\rho^2 + 2\beta_n\rho + \gamma_n^2 &= 0 \end{aligned}$$

is the equation for Case III, which again is what would be expected since the sample mean, the MLE of the mean, and the population mean are identical.

Two interesting observations can be made about these findings. First, Case II will reduce to Case I, and Case IV will reduce to Case III, in selected situations. Second, different estimates of $\rho$ are obtained for different levels of ignorance (just like with the classical sample variance).

## 5. Simulation experiment results

As the reported error (MSE) in Table 2 indicates, while equation (2.1) furnishes an exact Jacobian term for the regular square lattice situation of $\sqrt{n} = 2$, all other square lattice sizes have some very small (and seemingly negligible) amount of error present. Since the preceding example is based upon this exact situation, a simple simulation experiment has been conducted for $\sqrt{n} = 4$ to explore whether or not the Jacobian approximation approach promoted in this paper accurately generalizes to larger square lattice cases.

The initial conditions of this simulation experiment are (1) 16 values were randomly generated with the MINITAB normal pseudo–random number generator, having $\mu = 0$ and $\sigma^2 = 1$, and (2) $\mathbf{1}^t\mathbf{C}\mathbf{1} = 48$, and $\beta_n = 0.185791$ and $\gamma_n = 0.354859$ (see Table 2). The resulting geographic distribution of generated sample values is

| | | | |
|---|---|---|---|
| -0.92327 | 1.31724 | -0.99017 | 1.07651 |
| -0.50171 | 1.82117 | -0.29935 | -1.13190 |
| 0.28974 | -0.98088 | -0.52315 | -1.10900 |
| 1.70435 | 0.81597 | 0.28311 | -1.47315 |

The sample statistic terms for this spatial arrangement are $\mathbf{1}^t\mathbf{X} = -0.62449$, $\mathbf{1}^t\mathbf{C}\mathbf{X} = -2.24012$, $\mathbf{X}^t\mathbf{X} = 18.03662$, and $\mathbf{X}^t\mathbf{C}\mathbf{X} = -0.18677$. The sixteen eigenvalues for this geographic configuration are 3.23607, 2.23607, 2.23607, 1.23607, 1.00000, 1.00000, 0.00000, 0.00000, 0.00000, 0.00000, −1.00000, −1.00000, −1.23607, −2.23607, −2.23607, and −3.23607. In addition, the classical statistics for this sample surface are $\bar{\mathbf{x}} = -0.03903$, $s = 1.09582$, and the modified Wilk–Shapiro = 0.96217. The traditional estimation procedure for obtaining $\hat{\rho}$, when all three of the parameters are unknown (Case IV above), solves the following optimization problem (see Upton and Fingleton, 1985):

$$\begin{aligned} MIN : & [\prod_{i=1}^{16}(1-\rho\lambda_i)]^{-1/16}(\mathbf{X}-\mu\mathbf{1})^t(\mathbf{I}-\rho\mathbf{C})(\mathbf{X}-\mu\mathbf{1}) \\ \text{st} : & -1/3.23607 < \hat{\rho} < 1/3.23607, \end{aligned} \tag{5.1}$$

where $\lambda_i$ $(i = 1, 2, \ldots, 16)$ are the sixteen aforementioned eigenvalues of the binary configuration matrix $\mathbf{C}$. The solution to this particular problem, using the IMSL10 subroutines

E2LSF (to extract eigenvalues) and UVMID (to achieve univariate nonlinear optimization) [in single precision on a VAX mainframe], yielded $\hat{\rho} = -0.00541$, and $\hat{\mu} = -0.03915$.

Appropriate substitutions into equation (3.5) produce the quartic equation

$$421.7902\rho^4 - 15702.7248\rho^3 + 10242.3840\rho^2 - 1657.0680\rho - 9.3008 = 0. \qquad (5.2)$$

The roots of equation (5.2) have been extracted using the IMSL10 subroutine ZPORC (in single precision on a VAX mainframe), and are $-0.00543$, $0.33189$, $0.33467$, and $36.56763$. Of these four roots, the only one that falls within the feasible parameter space interval $(-0.30930, 0.30930)$ is $-0.00543$, which is equivalent to that obtained with the nonlinear optimization of expression (5.1), except for rounding error. This illustration demonstrates that, indeed, the set of equations (3.2), (3.4), and (3.5), involving Jacobian term approximations, do render very accurate estimates, and dramatically reduce the numerical intensity of spatial autoregression analysis.

## 6. Concluding comments and implications

The Jacobian term appears in likelihood functions to ensure that principal components types of transformations still lead to probability density functions whose complete integration yields unity. This term is particularly troublesome when dealing with spatial autoregressive models, since it is a function of the prevailing nature and degree of spatial dependence, becomes complex because of the multi–directional and two–dimensional interdependence involved, and thus does not disappear in the optimization process. Historically this term has required numerically intensive, and perhaps often computationally inaccurate, solutions to the parameter estimation problem. Findings reported in this paper suggest that at least a closed form approximation to this Jacobian may exist. The form of this approximation for rectangular regular lattices, and irregular lattices, still needs to be identified. The accuracy of this approximation remains to be comprehensively studied.

Having an approximation that is relatively simple in form, like equations (3.1) and (3.4), should allow a more careful and clearer investigation of the statistical properties of bias, sufficiency, consistency, and efficiency, for the parameter estimate $\hat{\rho}$. Ord (1975) already has reported some findings pertaining to these characteristics. The formulation presented here also will facilitate a better comparison between ordinary and generalized least squares parameter estimates for spatial regression models. Hopefully the formulations uncovered here will afford deeper insights into these statistical properties. In addition, equations need to be established depicting the convergence, as $n$ increases, of the Jacobian term parameters. Attempts thus far to achieve this goal have failed, but were for the Jacobian terms themselves. Finally, direct extensions to the moving average and simultaneous autoregressive model, as well as to stochastic versions of the geographic configuration matrix $\mathbf{C}$, merit careful attention.

Meanwhile, interfacing these findings with previous projects has some interesting implications. Griffith (1988b, 1989) has made a concerted effort to translate spatial regression techniques into algebraic language that is compatible with commercial statistical software packages. The general approach employed is to attach weights when writing regression equations, in much the same way that weighted least squares regression does. The weight that is attached is a function of the Jacobian term studied in this paper. By being able to write

this term in a simplified and consolidated form, rather than as a sum of eigenvalue expressions, these sorts of efforts with commercial packages will be further enhanced. The principal drawback here, though, is that the theory upon which optimization is based in these reformulated situations may be inapplicable; Warnes and Ripley (1987) have cast some doubt on the soundness of this approach, although they comment primarily on the parametric covariance function rather than an autoregressive formulation. Nevertheless, at least a good and useful first–approximation may be obtainable from these techniques.

## 7. References

Griffith, D. (1988a) Jacobian term specification and parameter estimation for spatial autoregressive models. Paper presented to the Association of American Geographers, 84th Annual Meeting, Phoenix.

Griffith, D. (1988b) Estimating spatial autoregressive model parameters with commercial statistical packages. *Geographical Analysis*. **20**, 176-186.

Griffith, D. (1989) Spatial regression analysis on the PC: spatial statistics using `MINITAB`. *Discussion Paper #1*. Ann Arbor: Institute of Mathematical Geography.

Griffith, D. (1990) Supercomputing and spatial statistics: a reconnaissance. *The Professional Geographer*. **42**, forthcoming.

Ord, J. (1975) Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*. **70**, 120-126.

Upton, G., and B. Fingleton (1985) *Spatial Data Analysis by Example*. Vol. 1. New York: Wiley.

Uspensky, J. (1948) *Theory of Equations*. New York: McGraw–Hill.

Warnes, J. and B. Ripley (1987) Problems with likelihood estimation of covariance functions of spatial Gaussian processes. *Biometrika*. **74**, 640-642.

# DISCUSSION

## "A numerical simplification for estimating parameters of spatial autoregressive models"

**by Daniel A. Griffith**

Estimation of spatial autoregressive (AR) processes has proved to be a very awkward statistical problem. For small or moderate numbers of areal units, the computational aspects are manageable, although the performance of the maximum likelihood procedure may be disappointing, manifested in a wide confidence band or a very flat likelihood function (LF). Worse yet, when a parametric covariance function is used, Warnes and Ripley (1987) show that the LF may have multiple maxima and that these need not relate well to the true parameter values, as is illustrated by simulation results in Ripley (1988, pp. 15-19). The reasons for this are not fully understood, but seem to be bound up with the scale of the process. That is, the AR structure is modeled primarily as a local phenomenon, yet longer range effects may have a major impact on the estimation process.

A second aspect of this is the pattern of sample covariances, first noted in the time series context by Anderson (1981), but likely to apply *a fortiori* in the spatial case. Suppose that there are n observations, yielding $\mathbf{N} = \binom{\mathbf{n}}{\mathbf{2}}$ pairings classified into $\mathbf{K}$ groups. For example, we may group by distances; if $\mathbf{d(i,j)}$ represents the distance between the pair of locations $\mathbf{(i,j)}$, the $\mathbf{k^{th}}$ group may be defined as $\mathbf{d_{k-1} < d(i,j) \leq d_k}, \quad \mathbf{k = 1, \ldots, K}$. Typically, $\mathbf{d_0 = 0}$ and $\mathbf{d_K}$ is the maximum distance between locations in the study area. Define

$$\mathbf{c_k = \sum (x_i - \bar{x})(x_j - \bar{x})/N_k, \quad k = 1, \ldots, K,} \tag{1}$$

where the sum is taken over pairs $\mathbf{(i,j)}$ in class $\mathbf{k}$, and $\mathbf{N_k}$ denotes the number of such pairs. Then it is easily shown that

$$\mathbf{nc_0 + 2 \sum N_k c_k = \sum_i \sum_j (x_i - \bar{x})(x_j - \bar{x}),} \tag{2}$$

where $\mathbf{nc_0 = \sum (x_i - \bar{x})^2}$. But the right hand side of equation (2) is identically zero so that the autocorrelations, $\mathbf{r_k = c_k / c_0}$ satisfy

$$\mathbf{\sum m_k r_k = -1/2,} \tag{3}$$

where $\mathbf{m_k = N_k / n}$. Thus, even when the theoretical ACF is nonnegative, as is often assumed, equation (3) *requires* that some of the sample values are negative. Since $\mathbf{K}$ is of the order of $\mathbf{n^{1/2}}$ for two–dimensional processes (against $\mathbf{K = n - 1}$ for time series), this poses real problems. Combined with the knowledge that $\rho$ rapidly approaches its upper bound as the autocorrelation increases (cf. Bartlett, 1975, pp. 82-83), the conclusions must be that the sample ACF is rather uninformative and the ML estimators may not be reliable either.

For the very large samples often encountered in image processing, the dominant problem is computational, although one must still worry whether the global assumptions of stationarity is justified; weaker assumptions such as the existence of the variogram seem easier to sustain.

Where then do these comments lead us? In many cases, $\rho$ will be close to its upper bound and so an approximation to the Jacobian element of the likelihood that identifies the bound will often lead to an estimate close to the actual MLE. The approximate large sample variance may be inaccurate, but this could be improved by examining approximations to the second derivative of expression (1.2) in Griffith's paper. Thus, for larger samples, the computational burden is greatly eased; for smaller samples, I suspect that when we are close to the boundary of the parameter space, nothing will be of much help.

If nonstationarity is suspected for large samples, the study area may be partitioned and the estimates obtained for each subarea. Griffith's proposals make such exploratory analyses much more accessible. Overall, past results lead us to be cautious in introducing approximate methods, but such an approach may lead to better data analysis.

We now turn to the particular approximations suggested by Griffith. Rather than the computer intensive search process suggested above his equation (2.1) we may use the inequality

$$\text{largest eigenvalue} = \lambda_1 \geq \mathbf{u}^T \mathbf{C}\, \mathbf{u}, \tag{4}$$

where $\mathbf{u}$ is any vector such that $\mathbf{u}^T\mathbf{u} = \mathbf{1}$. The equality holds if and only if $\mathbf{u}$ is the eigenvector corresponding to $\lambda_1$. Therefore, any choice of $\mathbf{u}$ will give a lower bound for $\lambda_1$ and thus a lower bound for $\rho^{-1}$. For regular lattices $\mathbf{u} = \mathbf{n^{-1/2}\,1}$ will typically be a good choice. For the square lattice:

$$\lambda_1^{-1} \leq 0.500 \quad \text{when} \quad \sqrt{n} = 2$$
$$\lambda_1^{-1} \leq 0.259 \quad \text{when} \quad \sqrt{n} = 30$$

suggesting a rather faster rate of convergence to 0.25 than Griffith's results. For smaller lattices, the exact value of $\lambda_1$ is readily computed; one should note that the smallest eigvenvalue, $\lambda_n$, satisfies

$$\lambda_n \leq \mathbf{u}^T\mathbf{C}\mathbf{u}, \quad \text{for any choice of } \mathbf{u} \text{ with } \quad \mathbf{u}^T\mathbf{u} = \mathbf{1}.$$

However, an effective intuitive choice for $\mathbf{u}$ is more difficult here. For regular lattices, $\lambda_n = -\lambda_1$ works well, as can be seen from equation (1.1); this choice is also made by Griffith. Fortunately whenever $\rho > 0$, as is usually the case, the exact choice for $\lambda_n$ has little impact on the estimation process.

Approximations to the rest of the Jacobian function are still required, particularly to ensure the accurate assessment of the large sample variance.

Finally, it should be noted that the single parameter case is tractable because of the eigenvalue approach, but that this approach fails for two or more parameters (unless the weighting matrices are orthogonal). However, some progress may be possible using (4). Let the inverse of the covariance matrix be

$$\mathbf{B} = \mathbf{I} - \rho_1\mathbf{C}_1 - \rho_2\mathbf{C}_2.$$

We know that $\mathbf{u^T B u} > \mathbf{0}$ so that approximations to the determinant might be feasible by selecting suitable $\mathbf{u}$ to generate factors like

$$(\mathbf{1} - \rho_1\mathbf{c}_1 - \rho_2\mathbf{c}_2).$$

We know that the determinant is the product of $\mathbf{n}$ such factors. For example, using

$$\mathbf{u} = m^{-1}\mathbf{1}$$

on a regular square lattice with $\mathbf{n} = \mathbf{m}^2$, where $\mathbf{C}_1$ denotes East–West links and $\mathbf{C}_2$ denotes North–South links, produces the factor

$$\mathbf{1} - [\mathbf{2(m-1)/m}]\rho_1 - [\mathbf{2(m-1)/m}]\rho_2,$$

implying, for large $\mathbf{m}$,

$$\rho_1 + \rho_2 \leq 1/2.$$

Whether such an approach produces sensible answers remains to be seen, but Griffith's paper has opened the door to new lines of attack.

## References

Anderson, O. (1981) Serial dependence properties of linear processes. *Journal of the Operations Research Society*, **31**, 905-917.

Bartlett, M. (1975) *The Statistical Analysis of Spatial Pattern*. London: Chapman and Hall.

Ripley, B. (1988) *Inference for Spatial Processes*. Cambridge: University Press.

*J. Keith Ord*, The Pennsylvania State University

*PREAMBLE*

*Knowledge comes, but wisdom lingers.*

Tennyson, **Locksley Hall**

*In the ideal anthology dealing with contemporary spatial statistics, one would like to include a paper by Besag, by Mardia, by Ord, and by Ripley; I am glad that the present volume is able to contain a contribution by at least three of these four statisticians. In terms of Mardia's submission, more than just a knowledge of spatial statistical modelling is imparted. The benefits of his accumulated wisdom are presented, with an emphasis on his spatial work during the past decade. The purpose of this paper is to outline maximum likelihood estimation methods for spatial linear models, with the presentation being enhanced by the inclusion of numerous abstract examples and data set analyses. Much guidance is offered here, for the researcher who is looking for a comprehensive treatment of spatial statistical modelling. Once more computational issues are raised, both by Mardia and by his discussant. Paelinck emphasizes the points raised by Mardia, and agrees with Mardia's contention that major problems meriting subsequent research attention include model identifiability, anisotropy, and second-order conditions.*

The Editor

# Maximum Likelihood Estimation for Spatial Models

Kanti V. Mardia*

*Department of Statistics, University of Leeds, Leeds, LS2 9JT, UK*

**Overview:** The paper gives maximum likelihood (ML) estimation methods for spatial linear models in three forms: Direct Representation (DR), Conditional Autoregressive (CAR) Models and Simultaneous Autoregressive (SAR) Models for the Gaussian Case. We also discuss the computational aspects of the methods. The problem of asymptotic bias is considered for the DR. For the intrinsic random field we obtain the maximum likelihood estimators (MLEs) and indicate a relationship with marginal likelihood. We give the exact MLEs for CAR models on a circle and a torus together with some properties. It is indicated how the same approach applies to the SAR models. For the stationary random field, we discuss the Whittle approximation. We also consider the MLE for intrinsic CAR. We then describe the estimation of a nugget parameter and its asymptotic distribution. We indicate the extension of the method to the multivariate case, block data, missing values in lattice, designs under spatial correlation, and the such. Finally, a general discussion is given.

## 1. Introduction

Spatial models have become increasingly used for image analysis (see, Mardia, 1989 for references). Previous applications were more in agriculture, forestry, ecology, geography, geology, to name a few disciplines. We mainly study ML estimation for three forms of spatial linear models: Direct Representation (DR), Conditional Auto–regressive (CAR) and Simultaneous Auto–regressive (SAR) when the random field is Gaussian. Finite range covariance schemes lead to a sparse covariance matrix $\Sigma$ of the random field in DR, whereas CAR and SAR lead to sparse $\Sigma^{-1}$. However, the models are interrelated; see Section 2.

Section 3 gives the MLEs and their problems for DR. Section 4 gives the MLEs for the intrinsic model. Section 5 gives CAR models with some comments on the SAR case. Section 6 discusses the problem for a nugget parameter or errors in variable model. Section 7 considers some other uses and extensions such as block data, missing values in a lattice and the use of these methods in experimental design. The last section gives a discussion.

* The author is grateful to Dan Griffith for inviting him to participate in this Symposium. He is also grateful to John Kent, Tim Hainsworth and Alan Watkins for their helpful comments. This work is supported by NSF grant DMS-8803207 to Professor Watson, Princeton University.

## 2. The model and its three forms

### 2.1. The spatial linear model

Let $\{X(\mathbf{t})\}$ be a stochastic process where $\mathbf{t}$ represents a point in $d$–dimensional space where we write $T = R^d$ for the Euclidean space and $T = Z^d$ for points on a regular lattice. Suppose the process is sampled at points $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_n$ to give the sample vector

$$\mathbf{X} = \{X(\mathbf{t}_1), X(\mathbf{t}_2), \ldots, X(\mathbf{t}_n)\}'.$$

Suppose that $\{X(\mathbf{t})\}$ is Gaussian and that

$$\mu(\mathbf{t}) = E\{X(\mathbf{t})\} = \mathbf{f}(\mathbf{t})'\boldsymbol{\beta}, \tag{2.1.1}$$

where $\boldsymbol{\beta}$ is a $q$–by–1 parameter vector and $\mathbf{f}(\mathbf{t})$ is a vector of known functions, possibly monomials, which may describe a trend. If there is no confusion we will write

$$\mathbf{X} = (X_1, X_2, \ldots, X_n)' \text{ and } \boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)'.$$

Suppose that

$$\{X(\mathbf{t}) - \mu(\mathbf{t})\}$$

is second–order stationary with

$$\mathrm{Cov}\{X(\mathbf{t}),\, X(\mathbf{t}+\mathbf{h})\} = \sigma(\mathbf{h}; \boldsymbol{\theta}), \tag{2.1.2}$$

where $\sigma(\cdot;\, \boldsymbol{\theta})$ is a positive definite function of $\mathbf{h}$, assumed known apart for a $p$–by–1 vector of parameters $\boldsymbol{\theta}$. [We will discuss some restrictions on $\sigma(\cdot;\, \boldsymbol{\theta})$ for the asymptotic theory in Section 3.] Thus from (2.1.1),

$$E(\mathbf{X}) = \mathbf{F}\boldsymbol{\beta}, \tag{2.1.3}$$

where $\mathbf{F} = \{\mathbf{f}(\mathbf{t}_1), \mathbf{f}(\mathbf{t}_2), \ldots, \mathbf{f}(\mathbf{t}_n)\}'$. Let the covariance matrix of $\mathbf{X}$ be $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$ with

$$\sigma_{ij} = (\boldsymbol{\Sigma})_{ij} = \{\sigma(\mathbf{t}_i - \mathbf{t}_j;\, \boldsymbol{\theta})\}.$$

Thus our model is of the form

$$\text{observation} = \text{deterministic trend} + \text{stochastic fluctuation}$$

where trend measures the long–term variation whereas the stochastic fluctuation measures the short–term or the local variation.

### 2.2. The direct representation

We can specify $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ by modelling $\sigma(\mathbf{h}; \boldsymbol{\theta})$ directly. With this formulation, we will call the spatial linear model the Direct Representation (DR).

Consider the following example. Table 1 gives a topographic data set consisting of 52 points from Davis (1973, pp. 313-314) with $T = R^2$. Figure 1 plots the data, and looking at the values closely, there is some indication of trend. At a smaller scale, we will expect local variation. One way to model the local variation, is to take $\sigma(\mathbf{h}; \boldsymbol{\theta})$ with finite *range*

TABLE 1
GEOGRAPHIC COORDINATES
AND ELEVATIONS OF CONTROL POINTS
FOR EXAMPLE SURVEYING PROBLEM

| E–W Coordinate $\mathbf{t}_1$ | N–S Coordinate $\mathbf{t}_2$ | Elevation $X(\mathbf{t}_1, \mathbf{t}_2)$ | E–W Coordinate $\mathbf{t}_1$ | N–S Coordinate $\mathbf{t}_2$ | Elevation $X(\mathbf{t}_1, \mathbf{t}_2)$ |
|---|---|---|---|---|---|
| 0.3 | 6.1 | 870 | 5.2 | 3.2 | 805 |
| 1.4 | 6.2 | 793 | 6.3 | 3.4 | 840 |
| 2.4 | 6.1 | 755 | 0.3 | 2.4 | 890 |
| 3.6 | 6.2 | 690 | 2.0 | 2.7 | 820 |
| 5.7 | 6.2 | 800 | 3.8 | 2.3 | 873 |
| 1.6 | 5.2 | 800 | 6.3 | 2.2 | 875 |
| 2.9 | 5.1 | 730 | 0.6 | 1.7 | 873 |
| 3.4 | 5.3 | 728 | 1.5 | 1.8 | 865 |
| 3.4 | 5.7 | 710 | 2.1 | 1.8 | 841 |
| 4.8 | 5.6 | 780 | 2.1 | 1.1 | 862 |
| 5.3 | 5.0 | 804 | 3.1 | 1.1 | 908 |
| 6.2 | 5.2 | 855 | 4.5 | 1.8 | 855 |
| 0.2 | 4.3 | 830 | 5.5 | 1.7 | 850 |
| 0.9 | 4.2 | 813 | 5.7 | 1.0 | 882 |
| 2.3 | 4.8 | 762 | 6.2 | 1.0 | 910 |
| 2.5 | 4.5 | 765 | 0.4 | 0.5 | 940 |
| 3.0 | 4.5 | 740 | 1.4 | 0.6 | 915 |
| 3.5 | 4.5 | 765 | 1.4 | 0.1 | 890 |
| 4.1 | 4.6 | 760 | 2.1 | 0.7 | 880 |
| 4.9 | 4.2 | 790 | 2.3 | 0.3 | 870 |
| 6.3 | 4.3 | 820 | 3.1 | 0.0 | 880 |
| 0.9 | 3.2 | 855 | 4.1 | 0.8 | 960 |
| 1.7 | 3.8 | 812 | 5.4 | 0.4 | 890 |
| 2.4 | 3.8 | 773 | 6.0 | 0.1 | 860 |
| 3.7 | 3.5 | 812 | 5.7 | 3.0 | 830 |
| 4.5 | 3.2 | 827 | 3.6 | 6.0 | 705 |

Elevation is measured in feet above sea level. Coordinates are expressed in 50–feet units measured from an arbitrary origin located in the southwest corner, with $\mathbf{t}_1$ being the East–West coordinate and $\mathbf{t}_2$ being the North–South coordinate (from Davis, 1973).

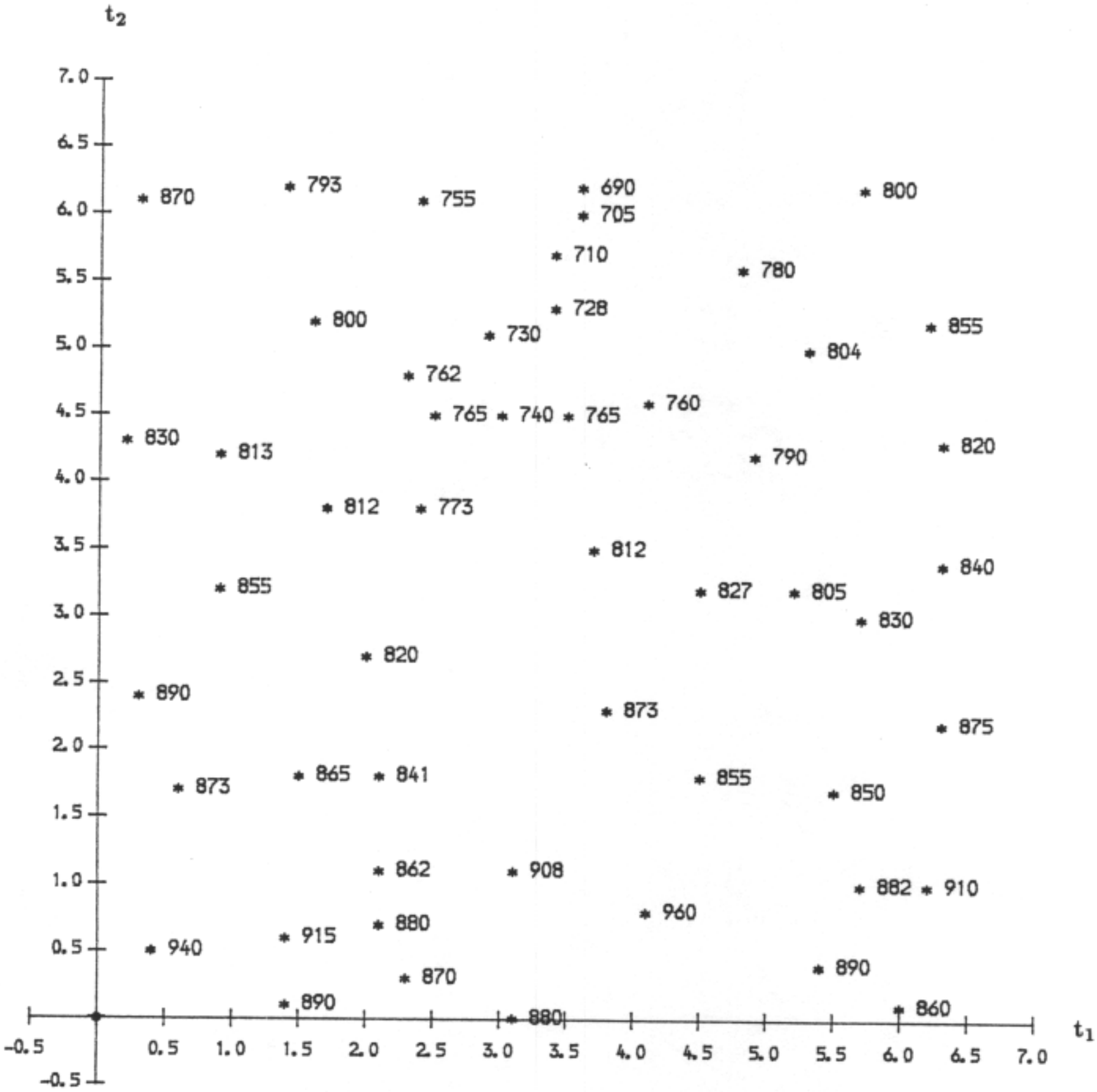
$\alpha$ so that $\sigma(\mathbf{h}; \boldsymbol{\theta}) = 0$ for $|\mathbf{h}| > \alpha$. For example, we could use the power scheme where the covariance function is given by (Mardia and Watkins, 1989)

$$\sigma(\mathbf{h}; \boldsymbol{\theta}) = \sigma^2(1 - \alpha^{-1}|\mathbf{h}|)^4, \quad |\mathbf{h}| < \alpha; \quad = 0 \text{ otherwise} \tag{2.2.1}$$

where $\boldsymbol{\theta} = (\sigma^2, \alpha)'$. We note that $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is sparse if $\alpha$ is less than the maximum distance between the points. Further, the sites can be irregularly distributed.

**Figure 1.**

Spatial distribution of elevations (Davis, 1973).



### 2.3. The conditional autoregressive model

Another way to model the spatial linear model is to specify $\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}$ through a conditional autoregression (CAR) model. We will mainly concentrate here on the case when all the $n$ sites are on a regular lattice. On an infinite regular lattice $Z^d$, the CAR model is defined by (Besag, 1974)

$$E(X_i \,|\, \text{rest}) = \mu_i + \sum_{j \neq i} \phi_{ij}(\boldsymbol{\theta})(x_j - \mu_j), \tag{2.3.1}$$

$$\text{Var}(X_i|\text{rest}) = \tau^2, \tag{2.3.2}$$

where 'rest' denotes all points $j \in Z^d, j \neq i$, and $\boldsymbol{\Phi}(\boldsymbol{\theta}) = [\phi_{ij}(\boldsymbol{\theta})]$ is a symmetric matrix with $\phi_{ii}(\boldsymbol{\theta}) = 0$, and $\phi_{ij}(\boldsymbol{\theta})$ a function such that the process $\{X(\mathbf{t})\}$ is covariance stationary. In particular, we take

$$\begin{aligned} \phi_{ij}(\boldsymbol{\theta}) &= \theta_{i-j} \text{ for } i - j \in N, \\ &= 0 \text{ otherwise} \end{aligned} \tag{2.3.3}$$

where $N$ denotes a finite symmetric neighbourhood of the origin and $\theta_{i-j} = \theta_{j-i}$. For example, for the first–order neighbourhood in 2–dimensions,

$$N = \{(-1, 0), (1, 0), (0, 1), (0, -1)\}, \tag{2.3.4}$$

and there are two parameters in $\boldsymbol{\Phi}(\boldsymbol{\theta})$, $\theta_{10}$ and $\theta_{01}$ which we will write as $\theta_1$ and $\theta_2$ respectively. Thus for $\mu_i = 0$; the CAR is defined by

$$E(X_i \,|\, \text{rest}) = \sum_{j \in N} \theta_{i-j} x_j, \; \text{Var}(X_i \,|\, \text{rest}) = \tau^2. \tag{2.3.5}$$

Sometimes it will be convenient to write the parameters as

$$\phi_h = \tau^{-2}, \; \mathbf{h} = \mathbf{0}; \; \phi_h = \tau^{-2}\theta_h, \; \mathbf{h} \neq \mathbf{0}. \tag{2.3.6}$$

The simplest example of (2.3.5) is when $\theta_{i-j} = \theta$ so that

$$\boldsymbol{\Phi}(\boldsymbol{\theta}) = \theta \mathbf{W}, \; E(X_i \,|\, \text{rest}) = \theta \sum_{j \in N} x_{i+j}, \tag{2.3.7}$$

where $(\mathbf{W})_{ij} = 1$ if $i - j \in N$; $= 0$ otherwise. $\mathbf{W}$ is called the adjacency matrix. We will call this the basic CAR model.

If $\mathbf{X}$ is $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we have

$$E(X_i \,|\, \text{rest}) = \mu_i + \sum_{j \neq i} (\sigma^{ij}/\sigma^{ii})(x_j - \mu_j), \text{ and } \text{Var}(X_i \,|\, \text{rest}) = 1/\sigma^{ii}.$$

On identifying these two expressions with (2.3.1) and (2.3.2) respectively, we obtain

$$\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} = \tau^{-2}[\mathbf{I} - \boldsymbol{\Phi}(\boldsymbol{\theta})]. \tag{2.3.8}$$

It should be noted that the infinite matrices $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}$ have to be absolutely convergent and positive definite. $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is certainly p.d. if $|1 - \sum_{h \in N} \theta_h \cos(\boldsymbol{\omega}'\mathbf{h})| < 1$. This holds if

$$\sum_{h \in N} |\theta_h| < 1 \tag{2.3.9}$$

[since $|\cos(\boldsymbol{\omega}'\mathbf{h})| \leq 1$] and therefore in the basic model $|\theta| < 1/\nu$ where $\nu$ is the number of neighbours.

In general, it should be noted that for the stationary CAR,

$$\sigma(\mathbf{h};\boldsymbol{\theta}) = (2\pi)^{-d} \int_{(-\pi,\pi)^d} [\exp(i\boldsymbol{\omega}'\mathbf{h})]/[\sum_{s \in N_0} \phi_s \cos(\boldsymbol{\omega}'\mathbf{s})] d\boldsymbol{\omega}, \tag{2.3.10}$$

where the set $N_0$ is $N$ with the origin included. Hence the class of stationary process includes the class of CARs. We can use $\sigma(\mathbf{h};\boldsymbol{\theta})$ in the DR but here $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is complicated. However, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is simple and sparse.

So far we have discussed the CAR on infinite lattice $Z^2$ but in practice, our sites are on a finite lattice $D$. Let $C = Z^2 - D$. We first obtain the inverse covariance matrix of $\{X(\mathbf{t})\}, t \in D$.

For the infinite lattice, we can write the inverse covariance matrix (2.3.8) of $\mathbf{X}$ as

$$\tau^2 \boldsymbol{\Sigma}_\infty^{-1} = \mathbf{I}_\infty - \boldsymbol{\Phi}_\infty \equiv \begin{pmatrix} \mathbf{I}_D - \boldsymbol{\Phi}_D & \mathbf{B} \\ \mathbf{B}' & \mathbf{I}_c - \boldsymbol{\Phi}_c \end{pmatrix},$$

where $\mathbf{I}_D$ and $\mathbf{I}_c$ are the identity matrix, $\boldsymbol{\Phi}_D$ matrix for the finite lattice $D$ and so forth. Then Künsch (1983) has shown that the inverse covariance matrix for the finite lattice $D$ is

$$\boldsymbol{\Sigma}_D^{-1} = \mathbf{I}_D - \boldsymbol{\Phi}_D - \boldsymbol{\Gamma}_D,$$

where

$$\boldsymbol{\Gamma}_D = \mathbf{B}_D (\mathbf{I}_c - \boldsymbol{\Phi}_c)^{-1} \mathbf{B}'.$$

provided all matrices converge absolutely. Thus (2.3.8) is not valid for $D$ unless $\boldsymbol{\Gamma}_D = \mathbf{0}$. However, the exact $\boldsymbol{\Sigma}_D^{-1}$ can be obtained through $\boldsymbol{\Sigma}_D$ whose elements are obtained from (2.3.10).

In fact, $\boldsymbol{\Sigma}_D(\boldsymbol{\theta})$ is the covariance function of the marginal distribution of $\{X(\mathbf{t})\}$ on $\mathbf{t} \in D$, and therefore the process is stationary on $\mathbf{t} \in D$ with $\sigma(\mathbf{h};\boldsymbol{\theta})$ given by (2.3.10). We will call this process an M–CAR. However, $\boldsymbol{\Sigma}_D(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}_D(\boldsymbol{\theta})^{-1}$ are both complicated, unlike $\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}$ given by (2.3.8) with $\boldsymbol{\Phi}(\boldsymbol{\theta})$ defined at (2.3.3). We can achieve some simplicity by making boundary adjustments in the following two ways:

(i) T–CAR. Wrap the CAR on a torus.

(ii) C–CAR. For $\mu(\mathbf{t}) = 0$, use the conditional distribution of $\{X(\mathbf{t})\}$ on $D$ given $X(\mathbf{t}) = 0$, $\mathbf{t} \notin D$, *i. e.* use the free boundaries.

Under the C–CAR, the CAR representation (2.3.5) is preserved but we do not have stationarity. Under the T–CAR, we have stationarity as well as the CAR representation but the periodic boundaries are not realistic.

Using the C–CAR can lead to serious bias in estimation for large $n$. The main reason is that the boundary for $d = 2$ is of order $n^{1/2}$ and the effect of neglecting it can be of order higher than $n^{-1}$. For some practical examples see Guyon (1982). Martin (1987) has highlighted some confusion in this area. The above result (2.3.8) requires basic knowledge of conditional distributions for the multivariate normal case, especially Theorems 3.2.3 and 3.2.4 in Mardia, Kent and Bibby (1989).

Suppose the sites are irregularly distributed; then the extension of the CAR model is not straightforward, in general (see, Besag, 1975). An interesting particular case is for $\boldsymbol{\Phi}(\boldsymbol{\theta})$ given by (2.3.7), where $W_{ij} = 1$ if $i$ and $j$ are nearest neighbours, and zero otherwise. Also for the regularized process, equation (2.3.7) can be used, where now

$W_{ij} = 0$ if areas are not contiguous, and

$\propto$ a monotonic function of the length of the common boundary otherwise.

If $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $\mathbf{W}$, with $\lambda_1 < \cdots < \lambda_n$, then $\boldsymbol{\Sigma}$ is positive definite if $0 \leq \theta < 1/\lambda_n$.

### 2.4. The simultaneous autoregressive model

For simplicity let us assume the finite lattice case. We have

$$\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\psi}(\boldsymbol{\theta})(\mathbf{X} - \boldsymbol{\mu}) + \boldsymbol{\varepsilon}, \;\; \boldsymbol{\varepsilon} \sim \mathbf{N}(\mathbf{0}, \sigma^2\mathbf{I}).$$

Thus with $\boldsymbol{\psi} \equiv \boldsymbol{\psi}(\boldsymbol{\theta})$,

$$\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} = \sigma^{-2}(\mathbf{I} - \boldsymbol{\psi}')(\mathbf{I} - \boldsymbol{\psi}), \text{ or } \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \sigma^2(\mathbf{I} - \boldsymbol{\psi})^{-1}(\mathbf{I} - \boldsymbol{\psi}')^{-1}, \tag{2.4.1}$$

where the defining property requires $|\mathbf{I} - \boldsymbol{\psi}|$ to be non–singular. As in the CAR case, we can take a particular case as $\boldsymbol{\psi} = \theta\mathbf{W}$ so that $\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}$ is sparse. Here $\mathbf{W}$ need not be a symmetric matrix. For almost all values of $\theta$, the class of the SARs is included in the class of the CARs on the infinite lattice. However, note that $\boldsymbol{\psi}$ is not uniquely determined by a given $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, unlike the CAR. We will not discuss the estimation problems for the SAR in detail.

### 2.5. Particular cases

We now give these three representations for the Geometric Scheme, where the correlation function is

$$\rho(\mathbf{h}; \lambda, \nu) = \lambda^{|h_1|}\nu^{|h_2|}. \tag{2.5.1}$$

We have its SAR and CAR representations on the infinite lattice as

SAR: $X_{ij} = \lambda X_{i-1,j} + \nu X_{i,j-1} - \lambda\nu X_{i-1,j-1} + \varepsilon_{ij}$,

CAR: $E(X_{ij}|\cdot) = \alpha(x_{i-1,j} + x_{i+1,j}) + \beta(x_{i,j-1} + x_{i,j+1}) - \alpha\beta(x_{i-1,j-1} + x_{i-1,j+1} + x_{i+1,j-1} + x_{i+1,j+1})$,

Var $(X_{ij}|\cdot) = \sigma^2/(1+\lambda^2)(1+\nu^2)$,

where $\alpha = \lambda/(1+\lambda^2)$ and $\beta = \nu/(1+\nu^2)$. From $E(X_{ij}|\cdot)$, it follows that the neighbourhood is of the second order.

For the first–order neighbourhood for the CAR in $Z^2$, we have from (2.3.10)

$$\sigma(h_1, h_2) = \tau^2(2\pi)^{-2} \int_{(-\pi,\pi)^2} \frac{\cos(\omega_1 h_1)\cos(\omega_2 h_2)}{1 - 2\theta_1\cos(\omega_1) - 2\theta_2\cos(\omega_2)}\, d\omega_1 d\omega_2. \tag{2.5.2}$$

Besag (1981) shows that for $\theta_1 + \theta_2 > 0.48$ and $(h_1, h_2) \neq (0,0)$,

$$\sigma(h_1, h_2) \simeq \tau^2 \{2\pi(\theta_1\theta_2)^{1/2}\}^{-1} K_0\left((1 - 2\theta_1 - 2\theta_2)^{1/2}[\frac{h_1^2}{\theta_1^2} + \frac{h_2^2}{\theta_2^2}]^{1/2}\right),$$

where $K_0(\cdot)$ is the modified Bessel function of the second kind and order zero. In particular, for $\theta_1 = \theta_2 = \theta$, $\sigma(h_1, h_2)$ is closely approximated by a monotonic decreasing function of $|\mathbf{h}| = (h_1^2 + h_2^2)^{1/2}$, so that it is almost an isotropic scheme. Further, there is very slow decay of $\rho(\mathbf{h})$ with increasing $|\mathbf{h}|$ whenever $\rho(1,0)$ is moderately high; *e. g.*, if $\rho(1,0) = 0.85$, then $|\mathbf{h}|$ must exceed 2000 before $\rho(\mathbf{h}) < 0.1$. Note that the Geometric Scheme is highly anisotropic and cannot display the type of slow decay for the first-order CAR scheme considered here.

## 3. ML estimation for DR

### 3.1. ML equations

In this Section, we follow Mardia and Marshall (1984). Since from Section 2.1 $\mathbf{X}$ is multivariate normal, the log-likelihood function of $\mathbf{X}$ with the parameters $(\boldsymbol{\beta}, \boldsymbol{\theta})$ is

$$\ell = \ell(\mathbf{X}; \boldsymbol{\beta}, \boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| - \frac{1}{2}(\mathbf{X} - \mathbf{F}\boldsymbol{\beta})'[\boldsymbol{\Sigma}(\boldsymbol{\theta})]^{-1}(\mathbf{X} - \mathbf{F}\boldsymbol{\beta}). \tag{3.1.1}$$

On differentiating (3.1.1) with respect to $\boldsymbol{\beta}$, with the help of $\frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}$ we get

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{X} - (\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{F})\boldsymbol{\beta}. \tag{3.1.2}$$

Hence, the MLE of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{F}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{F})^{-1}\mathbf{F}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}, \tag{3.1.3}$$

where $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$, $\hat{\boldsymbol{\theta}}$ being the MLE of $\boldsymbol{\theta}$. To differentiate (3.1.1) with respect to $\boldsymbol{\theta}$, we first note the following two results:

$$\frac{\partial \log|\boldsymbol{\Sigma}|}{\partial \theta_i} = \mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_i), \quad \frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial \theta_i} = -\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}^{-1},$$

where $\boldsymbol{\Sigma}_i = \frac{\partial \boldsymbol{\Sigma}}{\partial \theta_i}$. Hence differentiating with respect to $\theta_i$, with the help of these two results, we have

$$\frac{\partial \ell}{\partial \theta_i} = -\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_i) + \frac{1}{2}\mathbf{w}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}^{-1}\mathbf{w}, \quad i = 1, \ldots, p, \tag{3.1.4}$$

where $\mathbf{w} = \mathbf{X} - \mathbf{F}\boldsymbol{\beta}$. Thus the $p$ equations for $\hat{\boldsymbol{\theta}}$ are

$$\hat{\mathbf{w}}'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Sigma}}_i\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mathbf{w}} = \mathrm{tr}(\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Sigma}}_i), i = 1, \ldots, p, \tag{3.1.5}$$

with $\hat{\mathbf{w}} = \mathbf{X} - \mathbf{F}\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$. We note that the equations hold even when $\{X(\mathbf{t})\}$ is not covariance stationary.

One does not see an analytical solution to the ML equations (3.1.3) and (3.1.5). However, for a nested scheme such that

$$\sigma(\mathbf{h};\boldsymbol{\theta}) = \sum_{i=1}^{p} \theta_i \sigma_i(|\mathbf{h}|) \tag{3.1.6}$$

some progress can be made, since $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is of the form $\Sigma \mathbf{K}_i \theta_i$, where the matrices $\mathbf{K}_i$ are fixed. These are realistic models in the Analysis of Variance (see, Hocking, 1984), but not so realistic in Spatial Statistics. However, we will just give one example for its mathematical contents, namely the variance component scheme. Suppose the process is stationary, with

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \theta_1 \mathbf{I}_n + \theta_2 (\mathbf{I}_r \otimes \mathbf{E}_s), \tag{3.1.7}$$

where $n = rs$, and $\mathbf{E}_s$ is an $s$-by-$s$ matrix of 1s. Then $\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}}$ and the MLEs of $\theta_1$ and $\theta_2$ are obtained from the solution to

$$r(s-1)\lambda_0^{-1} + (r-1)\lambda_1^{-1} + \lambda_2^{-1} = q_0\lambda_0^{-2} + q_1\lambda_1^{-2},\ s(r-1)\lambda_1^{-1} + n\lambda_2^{-1} = sq_1\lambda_1^{-2},$$

where

$$q_i = \mathbf{X}'\mathbf{H}\mathbf{A}_i\mathbf{H}\mathbf{X}\ (i = 0, 1),\ \lambda_0 = \theta_1,\ \lambda_1 = \theta_1 + s\theta_2,\ \lambda_2 = \theta_1 + n\theta_2, \tag{3.1.8}$$

and

$$\mathbf{H}_n = \mathbf{I}_n - n^{-1}\mathbf{E}_n,\ \mathbf{A}_0 = \mathbf{I}_r \otimes \mathbf{H}_s,\ \text{and}\ \mathbf{A}_1 = s^{-1}\mathbf{H}_r \otimes \mathbf{E}_s.$$

The ML equations can be given in closed form since here $\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}$ can be obtained explicitly.

Also, some progress can be made for a finite lattice with the doubly geometric scheme given by (2.5.1). To obtain numerical solutions in general, we give the standard solving method involving the information matrix, which we now derive.

### 3.2. The information matrix and asymptotic normality

On differentiating (3.1.4) with respect to $\theta_j$, we find that

$$2\frac{\partial^2 \ell}{\partial\theta_i \partial\theta_j} = -\mathrm{tr}(\mathbf{R}_{ij} - \mathbf{S}_{ij}) - \mathbf{w}'(\mathbf{S}_{ij} + \mathbf{S}_{ji} - \mathbf{R}_{ij})\boldsymbol{\Sigma}^{-1}\mathbf{w}, \tag{3.2.1}$$

where

$$\mathbf{R}_{ij} = \boldsymbol{\Sigma}_{ij}\boldsymbol{\Sigma}^{-1},\ \mathbf{S}_{ij} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_j, \boldsymbol{\Sigma}_{ij} = \frac{\partial^2\boldsymbol{\Sigma}}{\partial\theta_i\partial\theta_j} = \frac{\partial\boldsymbol{\Sigma}_i}{\partial\theta_j}. \tag{3.2.2}$$

Since $E(\mathbf{w}'\mathbf{X}\mathbf{w}) = \mathrm{tr}(\boldsymbol{\Sigma}\mathbf{X})$, after taking the expectation of (3.2.1) we obtain

$$E[-\frac{\partial^2\ell}{\partial\theta_i\partial\theta_j}] = (1/2)\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_j) = a_{ij},\ \text{say.} \tag{3.2.3}$$

We also have, on differentiating (3.1.2) with respect to $\theta_j$, $\frac{\partial^2\ell}{\partial\beta\,\partial\theta_j} = -\mathbf{F}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_i\mathbf{w}$. Since $E(\mathbf{w}) = \mathbf{0}$, we find that

$$E[\frac{\partial^2\ell}{\partial\beta\,\partial\theta_j}] = 0.$$

Furthermore, from (3.1.2) we also get $\frac{\partial^2 \ell}{\partial \beta^2} = -\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{F}$. Hence the information matrix for $(\boldsymbol{\beta}, \boldsymbol{\theta})$ is

$$\mathbf{B}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \begin{bmatrix} \mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{F} & \mathbf{0} \\ \mathbf{0} & \mathbf{A} \end{bmatrix} = \text{diag}(\mathbf{B}_\beta, \mathbf{B}_\theta), \text{ say,} \tag{3.2.4}$$

where $\mathbf{A} = (a_{ij})$ is defined by (3.2.3).

Under certain regularity conditions, including differentiability of $\sigma(\mathbf{h}; \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, it can be shown that (Mardia and Marshall, 1984)

$$\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\theta}} \end{pmatrix} \sim N\left[\begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\theta} \end{pmatrix}, \begin{pmatrix} (\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{F})^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}^{-1} \end{pmatrix}\right]. \tag{3.2.5}$$

Hence $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ are asymptotically independent. Further, the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is $(\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{F})^{-1}$, and of $\hat{\boldsymbol{\theta}}$ is $\mathbf{A}^{-1}$.

Note that for asymptotic normality we require continuity, growth and convergence of the observed information matrix. Mardia and Marshall (1984) assume, among other conditions, that the sample set grows (*e. g.*, $|\mathbf{t}_i - \mathbf{t}_j| \geq c > 0$) in such a way that the sampling domain increases in extent as $n$ increases. The sufficient conditions for the asymptotic normality and weak consistency of $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ are given in Mardia and Marshall (1984). One of the conditions is that $\sigma(\cdot; \boldsymbol{\theta})$ is twice differentiable, with continuous second derivatives.

Note that the above condition is not satisfied for the spherical scheme with range parameter, *e. g.* at $|\mathbf{h}| = \alpha$. Moreover,

$$\sigma(\mathbf{h}; \alpha) = 1 - \frac{3}{2}\frac{|\mathbf{h}|}{\alpha} + \frac{1}{2}\frac{|\mathbf{h}|^3}{\alpha^3}, \quad |\mathbf{h}| < \alpha; \ = 0 \text{ otherwise,}$$

is not twice differentiable (see Mardia and Watkins, 1989). Note that this scheme is commonly used in geostatistics (Matheron, 1971); but, in particular, the asymptotic standard error (SE) written from the information matrix may not be valid for this scheme.

The problem of estimation in a bounded region $D \subset R^d$ with sampling increasingly dense in $D$ was recognized but was excluded from their discussion. Subsequently Stein (1987, 1988) has investigated such problems effectively.

Next consider the scale parameter. Note that we could write $\boldsymbol{\theta}' = (\theta_1, \boldsymbol{\theta}_2')$, where $\theta_1 = \sigma^2$ is such that $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \sigma^2\mathbf{P}(\boldsymbol{\theta}_2)$ such that $P(\boldsymbol{\theta}_2)$ represents a correlation matrix. Then we have

$$\boldsymbol{\Sigma}_i(\boldsymbol{\theta}) = \mathbf{P}(\boldsymbol{\theta}_2) \ (i = 1); \ = \sigma^2\mathbf{P}_i(\boldsymbol{\theta}_2) \ (i \neq 1),$$

where $\mathbf{P}_i(\boldsymbol{\theta}_2) = \frac{\partial \mathbf{P}(\boldsymbol{\theta}_2)}{\partial \theta_i}$ $(i \neq 1)$. Hence, the ML equations from (3.1.3) and (3.1.4) are

$$\hat{\boldsymbol{\beta}} = (\mathbf{F}'\hat{\mathbf{P}}^{-1}\mathbf{F})^{-1}\mathbf{F}'\hat{\mathbf{P}}^{-1}\mathbf{X}, \tag{3.2.6}$$

$$\hat{\sigma}^2 = [(\mathbf{X} - \mathbf{F}\hat{\boldsymbol{\beta}})'\hat{\mathbf{P}}^{-1}(\mathbf{X} - \mathbf{F}\hat{\boldsymbol{\beta}})]/n, \tag{3.2.7}$$

and

$$\hat{\sigma}^2 \text{tr}(\hat{\mathbf{P}}^{-1}\hat{\mathbf{P}}_i) = \mathbf{w}'\hat{\mathbf{P}}^{-1}\hat{\mathbf{P}}_i\hat{\mathbf{P}}^{-1}\mathbf{w}.$$

Further, the information matrix has elements

$$a_{11} = (n/2\sigma^4),$$
$$a_{1i} = \frac{1}{2}\,\mathrm{tr}(\mathbf{P}^{-1}\mathbf{P}_i)/\sigma^2 \ (i \neq 1)$$
$$a_{ij} = \frac{1}{2}\,\mathrm{tr}(\mathbf{P}^{-1}\mathbf{P}_i\mathbf{P}^{-1}\mathbf{P}_j) \ (i, j \neq 1). \tag{3.2.8}$$

However, mathematically it is simpler to consider $\boldsymbol{\theta}$ itself.

## 3.3. Computational aspects

*3.3.1 The scoring method.*

From (3.2.4), the method implies updating $(\boldsymbol{\beta}_{k+1}, \boldsymbol{\theta}_{k+1})$ at stage $(k+1)$ using

$$\begin{pmatrix}\boldsymbol{\beta}_{k+1}\\ \boldsymbol{\theta}_{k+1}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\beta}_{k}\\ \boldsymbol{\theta}_{k}\end{pmatrix} + \mathrm{diag}(\mathbf{B}_{\beta_k}^{-1}, \mathbf{B}_{\theta_k}^{-1}) \begin{pmatrix}\ell_{\beta k}\\ \ell_{\theta k}\end{pmatrix},$$

where $\ell_{\beta_k}$, and $\ell_{\theta k}$ are the derivative vectors of the log–likelihood $\ell$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, respectively at $\boldsymbol{\beta} = \boldsymbol{\beta}_k, \boldsymbol{\theta} = \boldsymbol{\theta}_k$. This implies equivalence with the use of

$$\boldsymbol{\beta}_k = [\mathbf{F}'\boldsymbol{\Sigma}(\boldsymbol{\theta}_k)^{-1}\mathbf{F}]^{-1}\mathbf{F}'\boldsymbol{\Sigma}(\boldsymbol{\theta}_k)^{-1}\mathbf{X},$$

and

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \mathbf{B}_{\theta_k}^{-1}\ell_{\theta_k}, \tag{3.3.1}$$

where

$$(\ell_{\theta_k})_i = -\frac{1}{2}\,\mathrm{tr}\{[\boldsymbol{\Sigma}(\boldsymbol{\theta}_k)^{-1}\boldsymbol{\Sigma}_i(\boldsymbol{\theta}_k)] - (\mathbf{X} - \mathbf{F}\boldsymbol{\beta}_k)'\boldsymbol{\Sigma}(\boldsymbol{\theta}_k)^{-1}\boldsymbol{\Sigma}_i(\boldsymbol{\theta}_k)\boldsymbol{\Sigma}(\boldsymbol{\theta}_k)^{-1}(\mathbf{X} - \mathbf{F}\boldsymbol{\beta}_k)\}.$$

For example, one can start with $\boldsymbol{\theta}$ based on a graphical method and then update $\boldsymbol{\beta}$, or start with $\boldsymbol{\beta}$ as the least squares solution. However, for large $n$ the numerical problem is formidable, and various approximation methods or modifications are put forward (see, for example, Mardia and Marshall, 1984). Also, the likelihood may be multimodal for small samples, causing the scoring method to lead to any stationary point (see Warnes and Ripley, 1987; Mardia and Watkins, 1989). In the above procedure, we also can use the observed Fisher information, since the second derivatives of the likelihood are known (from Section 3.2). Kitanidis and Lane (1985) fully discuss the Gauss–Newton methods (for a general discussion of this topic also see Kitanidis, 1987). However, using the observed information implies calculation of several second derivatives.

A computationally simpler method is given by Vecchia (1988). We can approximate the likelihood function by a partial likelihood function of the form

$$L_m(\mathbf{X}) = \prod_{i=1}^{n} P(X_i|\{X_{im}\}),$$

where $\{X_{im}\}$ is an array consisting of MIN$(i-1, m)$ observations from among $X_1, \ldots, X_{i-1}$ that are closest to $X_i$, as measured by the distances $|\mathbf{t}_i - \mathbf{t}_j|$. As $m$ approaches $n$, $L_m(\mathbf{X})$ approaches the likelihood function; but, $L_m(\mathbf{X})$ is very easy to compute for small $m$. Vecchia (1988) gives an iterative procedure whereby estimates based on $L_1(\mathbf{X})$ are used as initial values of estimates based upon $L_2(\mathbf{X})$, and so on. A statistic is computed at each step of the iterative procedure in order to assess the convergence of the iterative estimates.

*3.3.2 Profile likelihood.*

One method to check that a solution to the ML equations is really the global maximum is to plot the profile likelihood when the correlation parameters are $\leq 2$. Substituting $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ from (3.2.6) and (3.2.7) into the log–likelihood (3.1.1), we get the *profile likelihood*

$$\ell_p(\mathbf{X};\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{P}(\theta)| - \frac{1}{2}\log\{[\mathbf{X} - \mathbf{F}\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})]'\mathbf{P}(\boldsymbol{\theta})^{-1}[\mathbf{X} - \mathbf{F}\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})]\}, \quad (3.3.2)$$

where

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = [\mathbf{F}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{F}]^{-1}\mathbf{F}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{X}.$$

If $\hat{\boldsymbol{\theta}}$ maximizes $\ell_p(\mathbf{X};\theta)$, then $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}})$ and $\hat{\sigma}^2(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ also maximize this likelihood. Usually $\boldsymbol{\theta}$ is a scalar, especially for the stationary case, so that it is relatively easy to plot $\ell_p(\mathbf{X};\boldsymbol{\theta})$ rather than $\ell(\mathbf{X};\boldsymbol{\theta},\sigma^2)$ (see Mardia and Watkins, 1989; cf. Warnes and Ripley, 1987). Also, we then can obtain $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}})$ and $\hat{\sigma}^2 = \hat{\sigma}^2(\hat{\boldsymbol{\theta}})$ from (3.2.6) and (3.2.7), respectively.

## 3.4. Analysing topographic data

We now consider the analysis of the topographic data of Table 1. Instead of plotting the sample covariance function, we plot the semi–variogram given by

$$2g(\mathbf{h}) = \Sigma(X_i - X_{i+h})^2/n,$$

where

$$(t_i, t_{i+h}) \in D, \quad \mathrm{t}_1 = |\mathbf{h}|\cos(\theta), \ \mathrm{t}_2 = |\mathbf{h}|\sin(\theta).$$

Figure 2 shows the semi–variogram in four directions, namely $\theta = 0°$, $45°$, $90°$, and $135°$, measured along the $t_1$ axis in the clockwise direction. Note that if there is a trend then it can be shown that the population semi–variogram, $2\gamma(\mathbf{h}) = E(X_t - X_{t+h})^2$, is proportional to $|\mathbf{h}|^2$ for small $\mathbf{h}$. Thus there is an indication of trend. Also, since the semi–variograms are different for different directions, the data tabulated in Table 1 are not isotropic.

For simplicity, we fit the stationary model with mean $\boldsymbol{\beta}$ and the power covariance function given by (2.2.1). Having the parameters as range $\alpha$ and variance $\sigma^2$, we find that [with the asymptotic SE in brackets obtained from (3.2.4)]

$$\hat{\alpha} = 18.6(6.4), \ \hat{\sigma}^2 = 3103.4(1147.7), \ \hat{\beta} = 860.9(33.8),$$

$$\mathrm{CORR}(\hat{\alpha}, \hat{\sigma}^2) \simeq 0.85, \ \log(L) = -244.3.$$

Thus, $\hat{\alpha}$ and $\hat{\sigma}^2$ have large variance and are highly correlated.

The profile log–likelihood for $\alpha$ from (3.3.2) is shown in Figure 3, which has a unique mode at $\hat{\alpha} = 18.6$. Figure 4 shows the contour obtained from the ML predictor (Mardia and Marshall, 1984), which is given by

$$\hat{X}(\mathrm{t}) = \{\mathbf{f}(\mathrm{t})\}'\hat{\boldsymbol{\beta}} + \boldsymbol{\sigma}_X'\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{X} - \mathbf{F}\hat{\boldsymbol{\beta}}). \quad (3.4.1)$$

Here, we have $\mathbf{f}(\mathrm{t}) = 1$ and $\mathbf{F} = 1$. The contour is similar to Davis (1973, p. 322, Figure 6.9), except that there are slight differences near the edges. The contour indicates that the

**Figure 2.**

Semi–variogram of topographic data in four directions with fitted power covariance scheme for the stationary case.

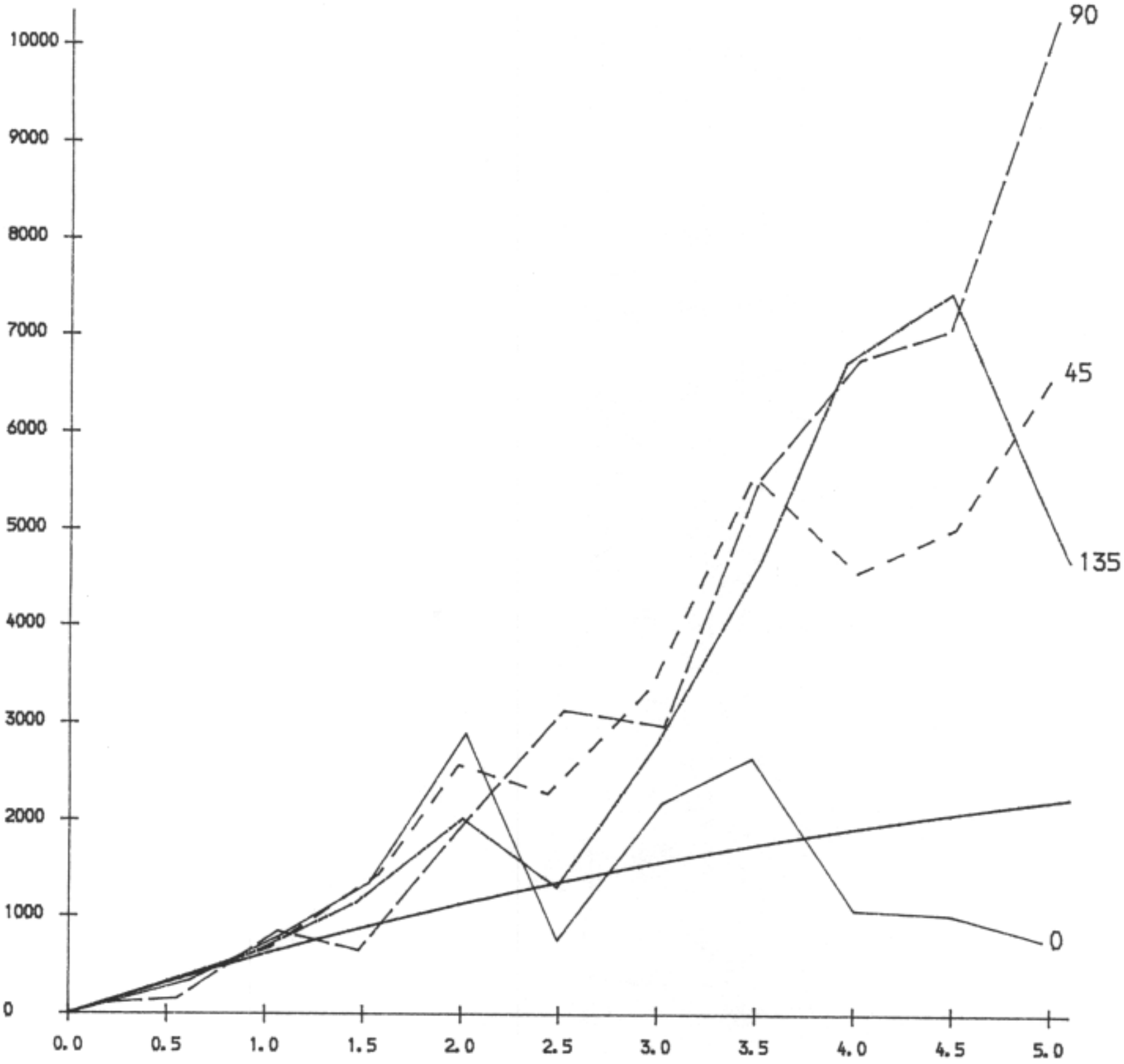


**Figure 3.**

Profile likelihood under stationary model for the topographic data, with power scheme with range $\alpha$

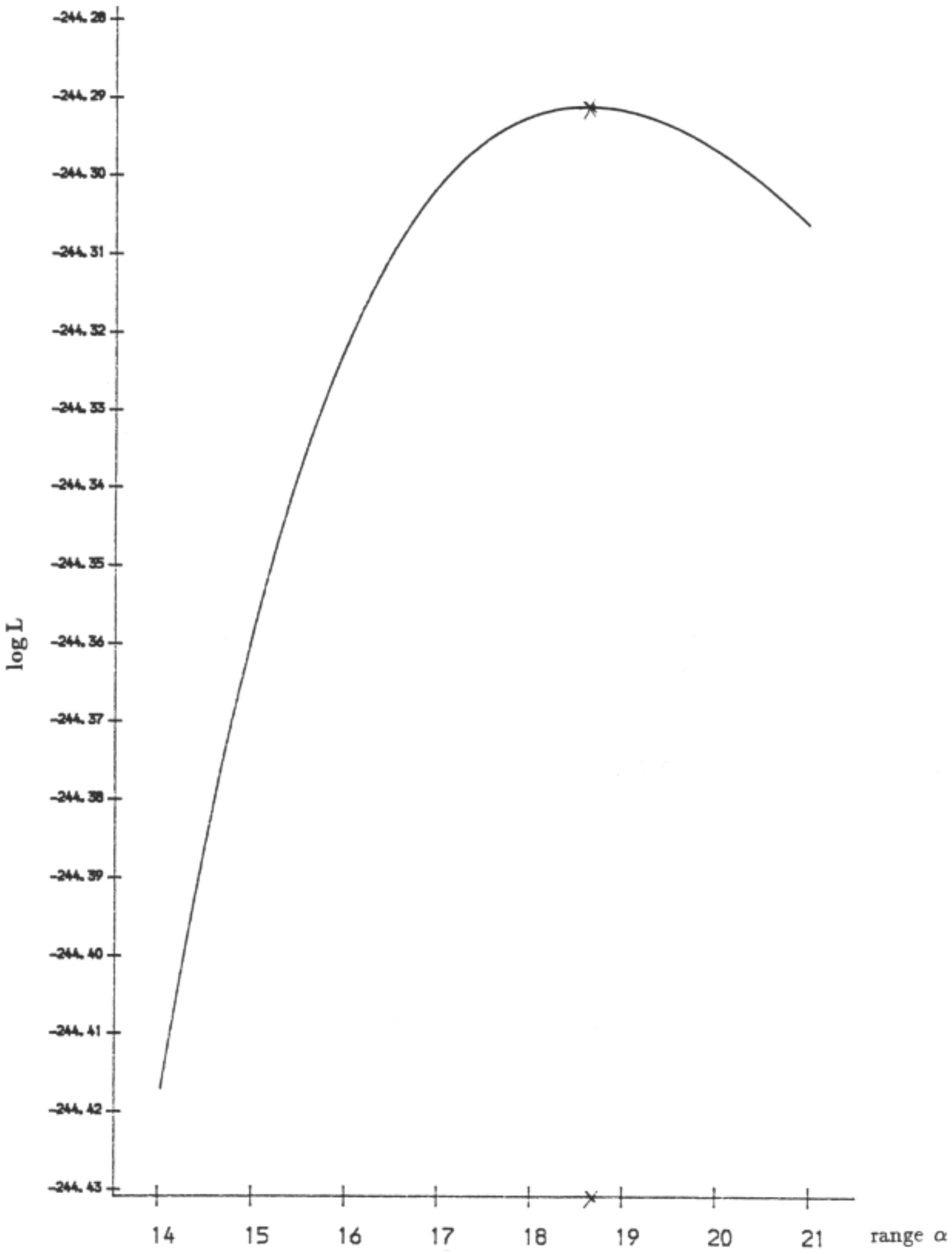
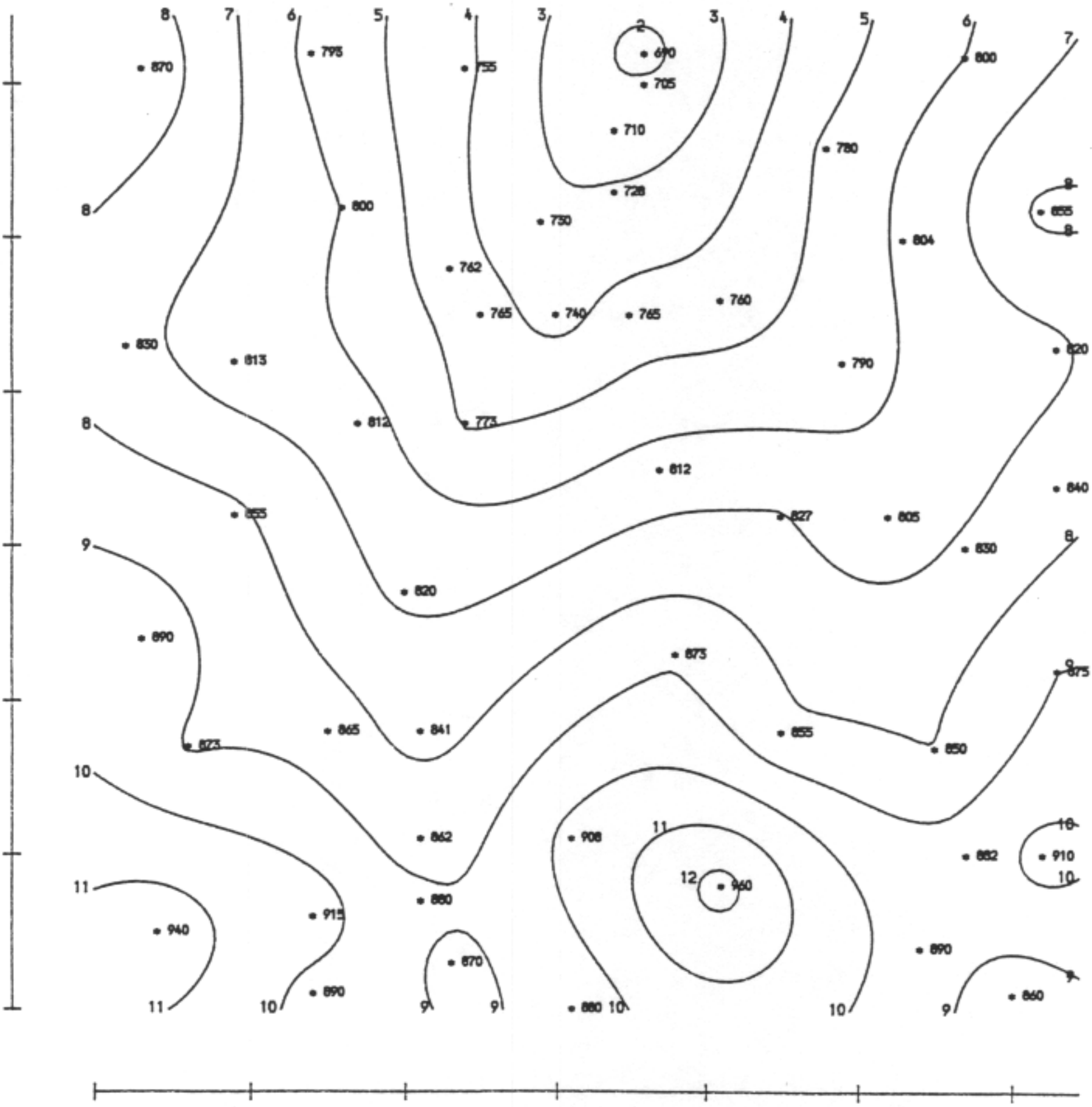


**Figure 4.**

Contours from the maximum likelihood predictor with stationary model and the long–range correlation. [Contours: $1 = 675(25), \ldots, 13 = 975$.]



data have a basin-shape. Note that the non-differentiability at $\mathbf{h} = \mathbf{0}$ of $\sigma(\mathbf{h}; \boldsymbol{\theta})$ produces spikes when the predictor passes through the data points.

If we fit a quadratic trend, then we find that the MLEs, with asymptotic SE in the brackets, are

$$\hat{\alpha} = 5.2(1.6),\ \hat{\sigma}^2 = 812(225.9),\ \hat{\boldsymbol{\beta}}' = (960.12, -50.38, -19.85, 6.88, 0.28, -0.2), \qquad (3.4.2)$$

which are the coefficients of 1, $t_1$, $t_2$, ${t_1}^2$, $t_1t_2$, and ${t_2}^2$, respectively. Their SEs are (30.2, 13.8, 13.1, 1.2, 1.6, 1.8), and $\mathrm{CORR}(\hat{\alpha}, \hat{\sigma}^2) = 0.71$. The asymptotic correlation matrix of the parameter estimates $\hat{\boldsymbol{\beta}}$ is

$$\begin{pmatrix} +1.000 & -0.708 & -0.655 & +0.429 & +0.614 & +0.350 \\ & +1.000 & +0.253 & -0.888 & -0.446 & +0.081 \\ & & +1.000 & -0.099 & -0.454 & +0.872 \\ & & & +1.000 & +0.078 & +0.077 \\ & & & & +1.000 & +0.060 \\ & & & & & +1.000 \end{pmatrix}$$

The correlations between the regression coefficients of ${t_1}^2$, $t_1t_2$, and ${t_2}^2$ are very small. The maximum log-likelihood is $\ell = -236.45$. The least squares estimates are

$$\hat{\boldsymbol{\beta}}' = (997.1, -51.9, -30.1, 7.3, 0.4, 0.8),\ \hat{\sigma}^2 = 784.0,$$

which are very similar to those reported in (3.4.2). Figure 5 shows the profile likelihood with quadratic trend, which is also unimodal. Figure 6 shows the ML predictor with the fitted quadratic trend. This is now more similar to the Davis contour plots, even at the edges. However, there is hardly any significant difference between the contour plots for constant trend (Figure 4) and the contour plots for quadratic trend (Figure 6). The stationary case depicts the trend as a long-range correlation, whereas in Case 2 the range is small, depicting small scale variations. Thus there is this a non-identifiability problem in modelling.

However note that the value of the Akaike Criterion $2(-\ell +$ the number of parameters) for the first case is 495, whereas for the second case it is 489, hence indicating that the trend model is better. Figure 7 shows the semi-variogram plots together with the fitted power covariance scheme after removing the trend. On comparison with Figure 2, we again note that the trend model is better. Therefore, both of these factors lead us to recommend the trend model.

We note that for the power scheme, the parameter $\alpha$ does not reflect the true range, unlike for the spherical scheme (see, Section 3.2). A qualitative feel for the observed range can be gained by identifying the values of $\frac{\partial \gamma(h)}{\partial h}$ at $h = 0+$ for the two schemes, which leads to using the true range for the power scheme of $3\alpha/8$ in place of $\alpha$.

Numerically, one can optimize the profile likelihood first with respect to $\alpha$ through the profile likelihood and then obtain the other estimates from (3.2.6) and (3.2.7), as we have done here.

**Figure 5.**

Profile for the topographic data with quadratic trend.

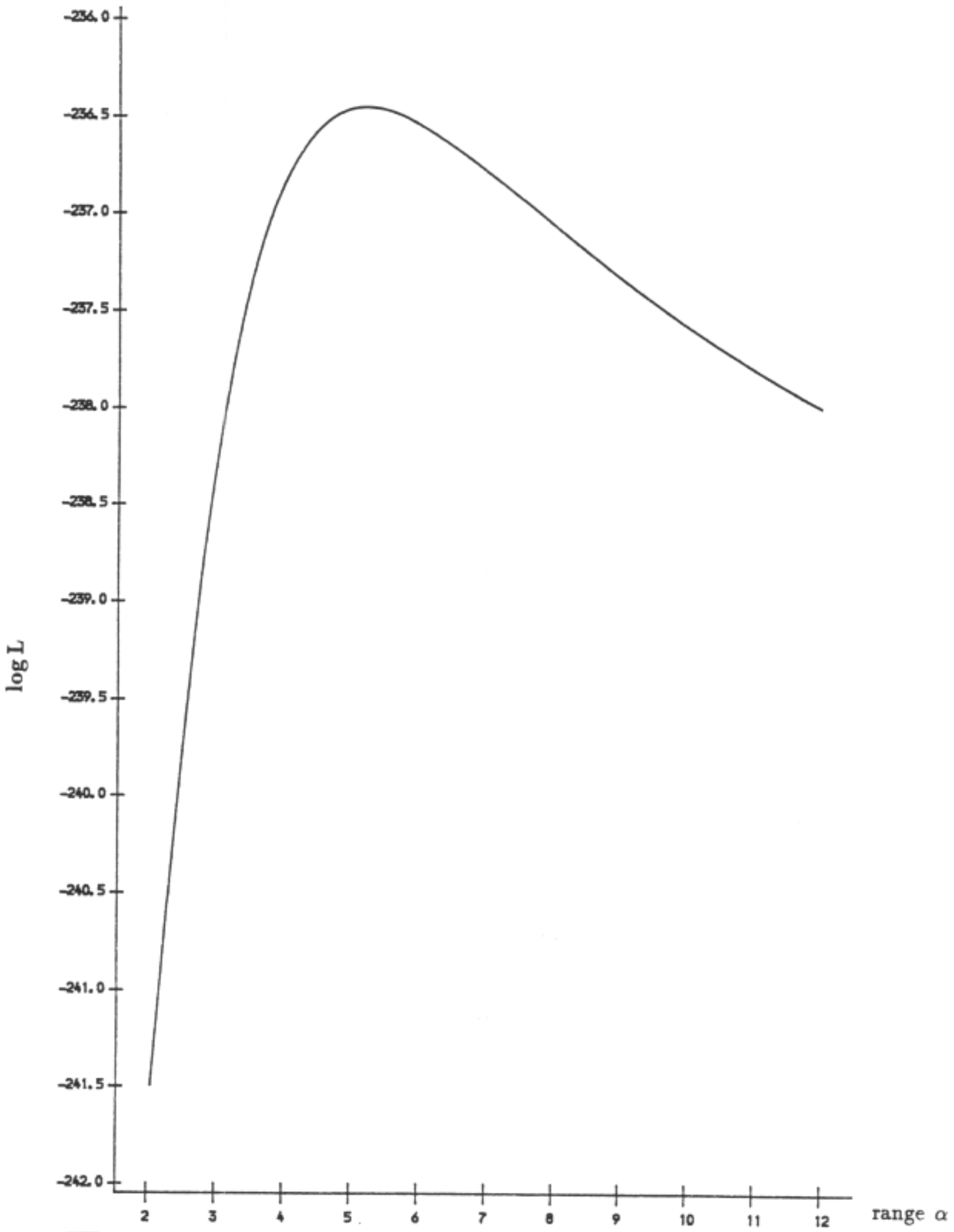


**Figure 6.**

Contours from the maximum likelihood predictor with quadratic trend for topographic data. [Contours: $1 = 675(25), \ldots, 13 = 975$.]

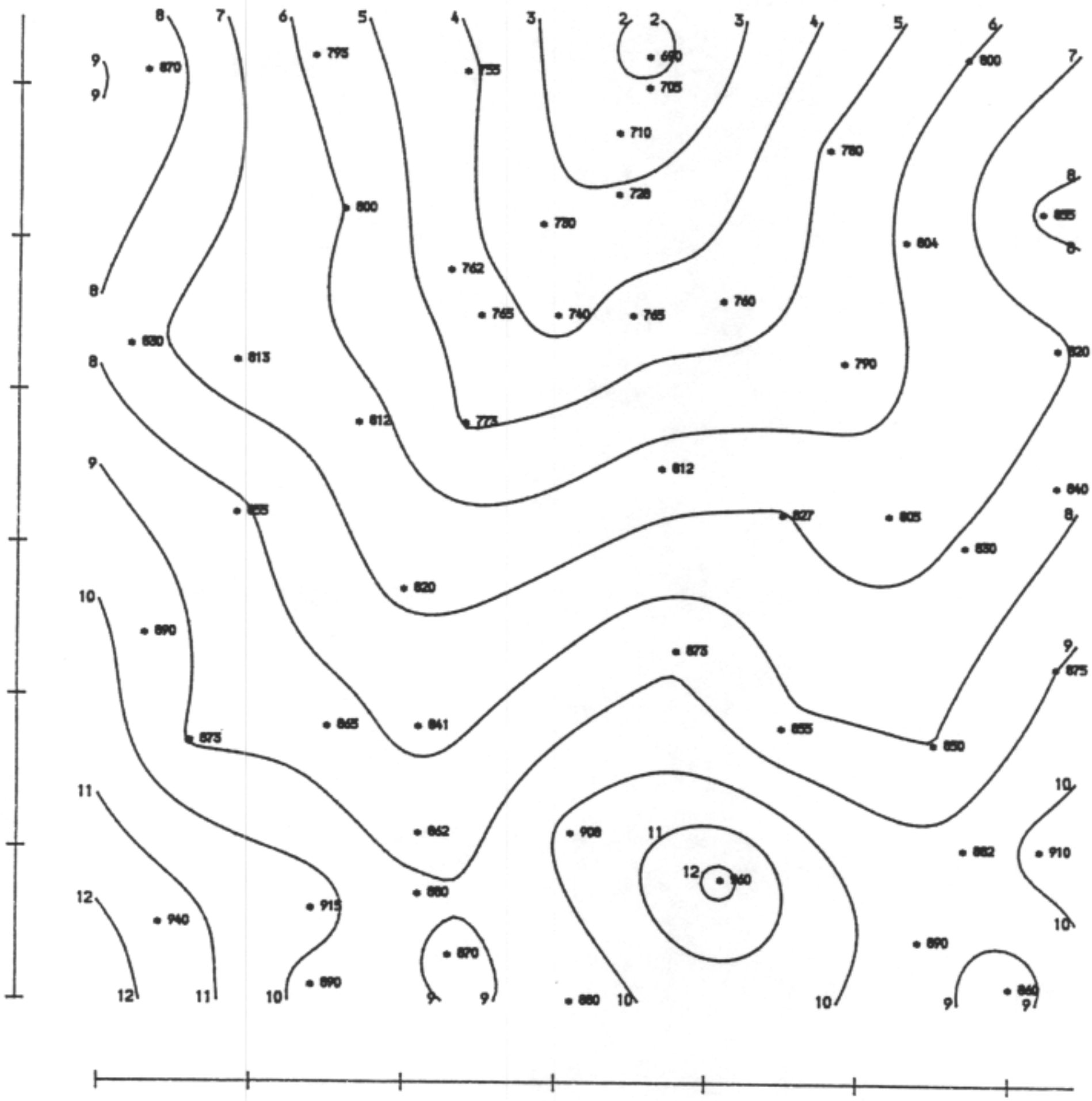
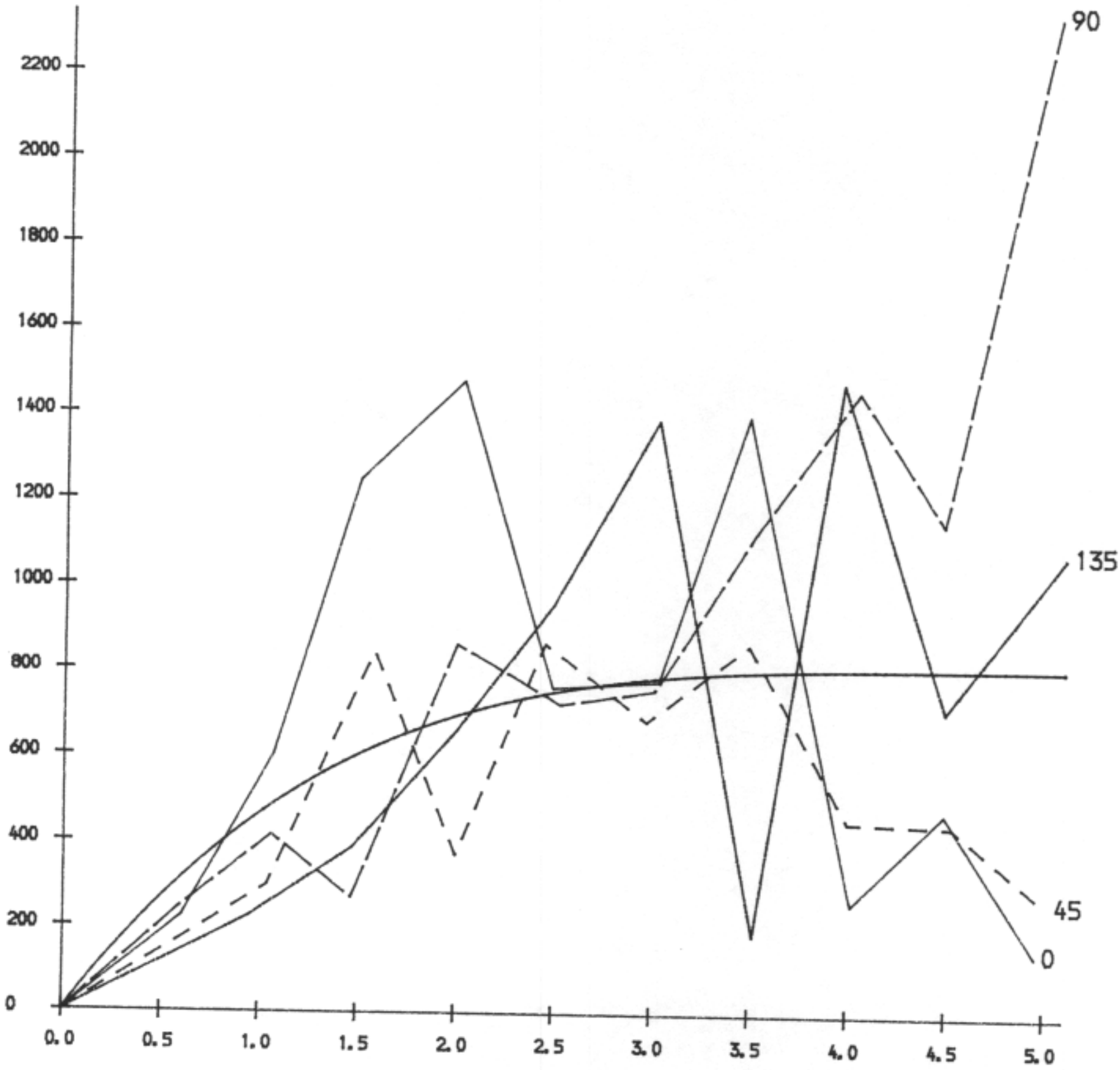


**Figure 7.**

Semi–variogram after removing the quadratic drift from least squares and fitted power variogram.



### 3.5. Bias in the estimators

Let $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ be the MLE of $(\boldsymbol{\beta}, \boldsymbol{\theta})$, then under some mild conditions it can be shown that (Watkins and Mardia, 1989)

$$\begin{aligned} E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= o(n^{-1}), \qquad \text{whereas} \\ E(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) &= (\mathbf{B}_{\theta}^{-1})\mathbf{C}_{\theta} + o(n^{-1}), \end{aligned} \tag{3.5.1}$$

where $(\mathbf{C}_{\boldsymbol{\theta}})_i = (1/2)\mathrm{tr}(\mathbf{B}_{\beta}^{-1}\mathbf{B}_{\beta i}) + (1/2)\mathrm{tr}(\mathbf{B}_{\theta}^{-1}\mathbf{M}_i)$, with $\mathbf{B}_{\beta i} = \frac{\partial \mathbf{B}_{\beta}}{\partial \theta_i}$ and

$$(\mathbf{M}_i)_{jk} = (1/2)\mathrm{tr}(\boldsymbol{\Sigma}_{ij}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_k\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}_{ik}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_j\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}_{jk}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}^{-1}), \tag{3.5.2}$$

for $i, j, k = 1, 2, \ldots, p$, and $(\mathbf{B}_{\beta}, \mathbf{B}_{\theta})$ are defined by (3.2.4). The bias is typically of order $1/n$. We now consider some particular cases.

Case 1: If $\boldsymbol{\theta}$ is known *a priori*, then $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, so that there is no bias.

Case 2: If $q = 0$, $p = 1$, then $\mathbf{M}_1 = (1/2)\mathrm{tr}(\boldsymbol{\Sigma}_1\ \boldsymbol{\Sigma}^1)$ and $\mathbf{B}_{\theta} = (1/2)\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_1)$, where $\boldsymbol{\Sigma}^1 = \frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial \boldsymbol{\theta}_1}$, so that the bias in (3.5.1) becomes

$$[\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_1)]^{-2}\mathrm{tr}(\boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}^1). \tag{3.5.3}$$

Case 3: For $p = 1$ with $\boldsymbol{\Sigma} = \theta_1 \mathbf{P}$, we have $\boldsymbol{\Sigma}_1 = \mathbf{P}$ and $\boldsymbol{\Sigma}_{11} = 0$ so that $\mathbf{M}_1 = 0$. But

$$\mathbf{B}_{\beta} = \theta_1^{-1}(\mathbf{F}'\mathbf{P}^{-1}\mathbf{F}) \text{ and } \mathbf{B}_{\beta 1} = -\theta_1^{-2}(\mathbf{F}'\mathbf{P}^{-1}\mathbf{F}),$$

so that $\mathrm{tr}(\mathbf{B}_{\beta}^{-1}\mathbf{B}_{\beta 1}) = -q\theta_1^{-1}$. Also $\mathbf{B}_{\theta} = \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_1) = \frac{1}{2}n\theta_1^{-2}$. Hence the bias is

$$-q\theta_1/n,$$

as can be expected, considering the independent and identically distributed (i.i.d) case. There is no bias in $\hat{\boldsymbol{\beta}}$, since it does not depend on $\theta_1$.

Case 4: Consider $q = 0$, $t \in Z^1$, $\sigma(\mathbf{h}) = \lambda^{|\mathbf{h}|}$, and $|\mathbf{h}| \in Z^1$. Then the bias is zero for $\hat{\lambda}$; but, if we take

$$\sigma(\mathbf{h}) = (1 - \lambda^2)^{-1}\lambda^{|\mathbf{h}|},$$

then the bias is $-2\lambda/n$.

This difference in behaviour of the bias can be explained. In the first case, $\sigma(\mathbf{h})$ parameterizes only the correlation structure, whereas in the second case we are taking it as the variance component of the process also. This work can be extended to the doubly geometric scheme given by (2.5.1).

## 4. Intrinsic models

### 4.1. Introduction

Consider the simple random walk along a line. Suppose that the steps $\varepsilon_i$ are independently distributed as $N(0, \sigma^2), i = 1, 2, \ldots$. Let $X_i$ be the distance covered after the $i$th step. Then

$$X_i = \varepsilon_1 + \cdots + \varepsilon_i,$$

and $\mathrm{Var}(X_i) = i\sigma^2 \to \infty$ as $i \to \infty$, so the process is non-stationary. However, the process $\{X_{i+h} - X_i\}$ is stationary with $\mathrm{Var}(X_{i+h} - X_i) = |\mathrm{h}|\sigma^2$. Thus, the semi-variogram is defined for all $i$. The process, which is increment stationary, will be called an intrinsic process of order 0, or the intrinsic random function (IRF) of order 0. In general, we define the intrinsic process of order $k$ below, following Matheron (1971). It should be noted that every stationary process is intrinsic, while the converse is not true. Also, note that the class of variogram schemes for the intrinsic case is larger; compare on the line the semi-variogram

$$\gamma(\mathrm{h}; \theta) = |\mathrm{h}|^\theta, \quad 0 < \theta < 2, \tag{4.1.1}$$

with the covariance function

$$\sigma(\mathrm{h}; \theta) = 1 - |\mathrm{h}|^\theta, \ 0 < \theta < 1. \tag{4.1.2}$$

### 4.2. Estimation

Let us take $\mathbf{X}$ ($n$-by-1) as data at the points $\mathbf{t}_i, i = 1, \ldots, n$, from $N_n(\mathbf{0}, \boldsymbol{\Sigma})$, $d = 2$, say. Let

$$\mathbf{H} = \mathbf{I} - \mathbf{T}(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}', \tag{4.2.1}$$

where the matrix $\mathbf{T}$ ($n$-by-$p$) has its $p$-columns, $p = (k+1)(k+2)/2$, as

$$\mathbf{t}_{ij} = (\mathrm{t}_1[1]^i \mathrm{t}_1[2]^j, \mathrm{t}_2[1]^i \mathrm{t}_2[2]^j, \ldots, \mathrm{t}_n[1]^i \mathrm{t}_n[2]^j)', \ 0 \le i + j \le k,$$

and $(\mathrm{t}_i[1], \mathrm{t}_i[2])$ denotes the coordinates of the $i$th site, $i = 1, 2, \ldots, n$. For example, $t_{00} = (1, 1, \ldots, 1)$, and $t_{10} = (\mathrm{t}_1[1], \mathrm{t}_2[2], \ldots, \mathrm{t}_n[1])$. Thus here $\mathbf{f}(\cdot)$ is specified as the full polynomial of degree $k$. It should be noted that $\mathbf{H}$ is a singular and idempotent matrix of rank $n - p$. Hence

$$\mathbf{Y} = \mathbf{HX} \tag{4.2.2}$$

defines the increment of the order $k$. Note that

$$Y_i = b_0 - b_{11}t_{1i} - b_{12}t_{2i} - \cdots - b_{kk}t_{1i}^k - \cdots,$$

where $\mathbf{b} = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{X}$.

If $E(\mathbf{X}) = \mathbf{T}\boldsymbol{\beta}$, then $\mathbf{b}$ is the least squares estimator of $\boldsymbol{\beta}$. For IRF-0, a constant is filtered out and we work on $X_i - \bar{X}, i = 1, 2, \ldots, n$, where $\bar{X}$ is the mean of $X_1, X_2, \ldots, X_n$. Note that for IRF-k, polynomials of degree $\le 2k$ will be filtered out. Let $\sigma_k(\mathbf{h})$ be the generalized covariance function for the IRF-k process, defined such that $\mathbf{H\Sigma H}$ is positive definite for all choices of sites and all $n$. Thus $\{\sigma_k(\cdot) + \mathbf{P}(\cdot) : \mathbf{P}(\cdot)\text{is a polynomial of degree } \le 2k\}$ forms an equivalence class. From (4.2.2) we have $\mathrm{Cov}(\mathbf{HX}) = \mathbf{H\Sigma H} = \mathbf{A}$, say, which is a singular matrix of rank $n - p$. We have

$$\mathbf{HAH} = \mathbf{A}. \tag{4.2.3}$$

Let $\mathbf{A}^-$ be the symmetric generalized inverse of $\mathbf{A}$, such that

$$\mathbf{H}\mathbf{A}^-\mathbf{H} = \mathbf{A}^-. \tag{4.2.4}$$

We will give a construction of $\mathbf{A}^-$ below. We first give an important result for a single parameter $\boldsymbol{\theta}$ in $\sigma(\mathbf{h};\boldsymbol{\theta})$; this result can be extended to the multi-parameter case.

Let $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ be the covariance matrix for the above case. Let $\lambda_i(\mathbf{A})$ be non-zero eigenvalues of $\mathbf{A}$. If $L(\boldsymbol{\theta})$ is the likelihood, then

$$-2\log_e[L(\boldsymbol{\theta})] = \text{constant} + \Pi\lambda_i(\mathbf{A}) + \mathbf{X}'\mathbf{A}^-\mathbf{X}, \tag{4.2.5}$$

where $\lambda_i(\mathbf{A})$ are non-zero eigenvalues of $\mathbf{A}$. Further, its derivative with respect to $\boldsymbol{\theta}$ is

$$\text{tr}(\mathbf{A}^-\mathbf{A}_\theta) + \mathbf{X}'\mathbf{A}^-\mathbf{A}_\theta\mathbf{A}^-\mathbf{X}, \tag{4.2.6}$$

where $\mathbf{A}_\theta = \frac{\partial \mathbf{A}}{\partial \boldsymbol{\theta}}$. A proof is as follows. We can find a matrix $\mathbf{D}$ $[(n-p)\text{-by-}n]$ such that

$$\mathbf{D}\mathbf{D}' = \mathbf{I}_{n-p} \quad \text{and} \quad \mathbf{D}'\mathbf{D} = \mathbf{H}, \tag{4.2.7}$$

by constructing an orthogonal matrix $\mathbf{C}$ such that

$$\mathbf{C} = \begin{pmatrix} (\mathbf{T}'\mathbf{T})^{-\frac{1}{2}}\mathbf{T} \\ \mathbf{D} \end{pmatrix}.$$

Set

$$\mathbf{Y} = \mathbf{D}\mathbf{X}; \text{ then} \tag{4.2.8}$$

$$\text{Cov}(\mathbf{Y}) = \mathbf{D}\mathbf{A}\mathbf{D}' = \mathbf{B}, \text{ say,} \tag{4.2.9}$$

which is a non-singular matrix of order $(n-p)$-by-$(n-p)$, and $E(\mathbf{Y}) = \mathbf{0}$. Note that (4.2.8) implies

$$\mathbf{X} = \mathbf{D}'\mathbf{Y} + \text{polynomial of degree k.}$$

Furthermore, from (4.2.3) and (4.2.7) we get

$$\mathbf{A} = \mathbf{D}'\mathbf{B}\mathbf{D} \text{ and } \mathbf{A}^- = \mathbf{D}'\mathbf{B}^{-1}\mathbf{D}, \tag{4.2.10}$$

as $\mathbf{A}^-$ obviously satisfies (4.2.4). Since $\mathbf{B}$ is non-singular, $-2\log_e[L(\boldsymbol{\theta})]$ is

$$\text{constant} + \log(|\mathbf{B}|) + \mathbf{Y}'\mathbf{B}^{-1}\mathbf{Y}, \tag{4.2.11}$$

and its derivative, as given in Section 3, is

$$\text{tr}(\mathbf{B}^{-1}\mathbf{B}_{\boldsymbol{\theta}}) + \mathbf{Y}'\mathbf{B}^{-1}\mathbf{B}_{\boldsymbol{\theta}}\mathbf{B}^{-1}\mathbf{Y}. \tag{4.2.12}$$

Using (4.2.8) and (4.2.10) with (4.2.11) we immediately obtain (4.2.5). From (4.2.7) we can write (4.2.12)

$$\text{tr}(\mathbf{B}^{-1}\mathbf{D}\mathbf{D}'\mathbf{B}_{\boldsymbol{\theta}}\mathbf{D}\mathbf{D}') + \mathbf{Y}'\mathbf{D}\mathbf{D}'\mathbf{B}^{-1}\mathbf{D}\mathbf{D}'\mathbf{B}^{-1}\mathbf{D}\mathbf{D}'\mathbf{Y},$$

which, using (4.2.8) and (4.2.10), leads to (4.2.6).

In a similar fashion, higher derivatives including the information matrix can be written simply by replacing $\mathbf{\Sigma}^{-1}$ by $\mathbf{A}^{-}$ in previous results. It should be noted that we could have used the algebraically "independent" variables out of $\mathbf{Y}$ from (4.2.2), and then use our previous results (see Kitanidis, 1983); but this approach destroys the symmetry achieved in (4.2.5).

Another approach is to assume that $\mathbf{\Sigma}^{-1}$ exists. On exploiting the equivalence of $\sigma_k(\mathbf{h})$, we always can obtain $\mathbf{\Sigma}$ so that $\mathbf{\Sigma}^{-1}$ exists. Then we have explicitly for the likelihood equation

$$\begin{aligned} \mathbf{A}^{-} &= \mathbf{G}'\mathbf{\Sigma}^{-1}\mathbf{G}, \text{ and} \\ \Pi\lambda_i(\mathbf{A}) &= |\mathbf{T}'\mathbf{\Sigma}^{-1}\mathbf{T}||\mathbf{\Sigma}|/|\mathbf{T}'\mathbf{T}|, \text{ where} \\ \mathbf{G} &= \mathbf{I} - \mathbf{T}(\mathbf{T}'\mathbf{\Sigma}^{-1}\mathbf{T})^{-1}\mathbf{T}'\mathbf{\Sigma}^{-1}. \end{aligned} \tag{4.2.13}$$

Thus the p.d.f. is proportional to

$$|\mathbf{T}'\mathbf{T}|^{\frac{1}{2}}[|\mathbf{\Sigma}||\mathbf{T}'\mathbf{\Sigma}^{-1}\mathbf{T}|]^{-\frac{1}{2}} \exp\{-\frac{1}{2}\mathbf{X}'\mathbf{G}'\mathbf{\Sigma}^{-1}\mathbf{G}\mathbf{X}\}, \tag{4.2.14}$$

where $\mathbf{G}$ is defined by (4.2.13). Note that $\mathbf{G}$ depends on $\mathbf{\Sigma}$, unlike in our preceding work. Nevertheless, $\mathbf{GH} = \mathbf{H}$ and $\mathbf{HG}' = \mathbf{H}$, so that $\mathbf{HA}^{-}\mathbf{H} = \mathbf{A}^{-}$ is satisfied.

The function

$$\sigma_k(\mathbf{h}) = \sum_{p=0}^{k}(-1)^{p+1}|\mathbf{h}|^{2p+1}\frac{a_p}{[(2p+1)!]}\frac{\Gamma(d/2)p!}{\sqrt{\pi}\Gamma\{p+[(d+1)/2]\}}, \tag{4.2.15}$$

defined a covariance function for the IRF-k (Delfiner, 1976) provided the $a_p$s satisfy

$$\sum_{p=0}^{k} a_p t^{k-p} \geq 0 \text{ for any } t > 0.$$

Thus we have, for example,

$$\sigma_0(\mathbf{h}) = -\theta|\mathbf{h}|,\ \theta > 0, \quad \text{and} \quad \sigma_1(\mathbf{h}) = \theta|\mathbf{h}|^3,\ \theta > 0,$$

defining valid intrinsic covariance functions.

### 4.3. Regression and the IRF-k

Let

$$X(\mathbf{t}) = \boldsymbol{\beta}_1'\mathbf{f}_1(\mathbf{t}) + \boldsymbol{\beta}_2'\mathbf{f}_2(\mathbf{t}) + \varepsilon(\mathbf{t}). \tag{4.3.1}$$

Suppose that $\boldsymbol{\beta}_1'\mathbf{f}_1(\mathbf{t})$ is a full polynomial of degree $k$, and $\boldsymbol{\beta}_2'\mathbf{f}_2(\mathbf{t})$ is a polynomial with no terms of degree less than $k$. Then under an IRF-k model for $\boldsymbol{\varepsilon}(\mathbf{t})$ we have

$$\hat{\boldsymbol{\beta}}_2' = (\mathbf{F}_2'\mathbf{\Sigma}^{-}\mathbf{F}_2)^{-1}\mathbf{F}_2'\mathbf{\Sigma}^{-}\mathbf{X}, \tag{4.3.2}$$

where $\mathbf{F}_2' = [\mathbf{f}_2(\mathbf{t}_1), \mathbf{f}_2(\mathbf{t}_2), \ldots, \mathbf{f}_2(\mathbf{t}_n)]$ since the quadratic form of importance is

$$(\mathbf{X} - \mathbf{F}_2\boldsymbol{\beta}_2)'\mathbf{\Sigma}^{-}(\mathbf{X} - \mathbf{F}_2\boldsymbol{\beta}_2)$$

### 4.4. Marginal likelihood and IRF-k

Let $\sigma(\mathbf{h};\boldsymbol{\theta})$ be the covariance function. Consider two models.

Model 1:

$$\mathbf{X} \sim N[\boldsymbol{\beta}\mathbf{T},\ \boldsymbol{\Sigma}(\boldsymbol{\theta})];$$

Model 2:

treat $\sigma(\mathbf{h};\boldsymbol{\theta})$ as a covariance function for an IRF-k.

We show that the marginal likelihood under Model 1 with nuisance parameter $\boldsymbol{\beta}$ is the same as the likelihood under Model 2. We have

$$\hat{\boldsymbol{\beta}} = (\mathbf{T}'\boldsymbol{\Sigma}^{-1}\mathbf{T})^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{T}'\mathbf{X}.$$

Let $\mathbf{Y} = \mathbf{R}\mathbf{X}$, where $\mathbf{R}$ is a singular matrix,

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_1 \\ \mathbf{R}_2 \end{pmatrix},$$

$$\mathbf{R}_1 = \mathbf{H}, \text{ and } \mathbf{R}_2 = (\mathbf{T}'\boldsymbol{\Sigma}^{-1}\mathbf{T})^{-1}\mathbf{T}'\boldsymbol{\Sigma}^{-1},$$

where $\mathbf{H}$ is given at (4.2.1). Now $\mathbf{Y} \sim N(\mathbf{R}\mathbf{T}\boldsymbol{\beta},\ \mathbf{R}\boldsymbol{\Sigma}\mathbf{R}')$, with

$$\mathbf{R}\mathbf{T}\boldsymbol{\beta} = (0, \boldsymbol{\beta}')', \text{ and } \mathbf{R}\boldsymbol{\Sigma}\mathbf{R}' = \text{block diag}\,[\mathbf{R}_1\boldsymbol{\Sigma}\mathbf{R}_1', (\mathbf{T}'\boldsymbol{\Sigma}^{-1}\mathbf{T})^{-1}].$$

Let $\mathbf{Y} = (\mathbf{Y}_1', \mathbf{Y}_2')'$. Note that $\mathbf{Y}_2 = \hat{\boldsymbol{\beta}}$. Further, $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are independent, and

$$\mathbf{Y}_1 \sim N(\mathbf{0},\ \mathbf{R}_1\boldsymbol{\Sigma}\mathbf{R}_1') \text{ and } \mathbf{Y}_2 \sim N[\boldsymbol{\beta},\ (\mathbf{T}'\boldsymbol{\Sigma}^{-1}\mathbf{T})^{-1}]. \tag{4.4.1}$$

Hence the marginal likelihood is the p.d.f. of $\mathbf{Y}_1$, which is precisely the likelihood of the IRF given before. We should note that the marginal likelihood principle is expounded in Kalbfleisch and Sprott (1970); Patterson and Thompson (1975) and Harville (1977) proposed such a procedure in the context of variance components modelling. Tunnicliffe–Wilson (1989) has given the use of this principle for time series analysis.

Our recommendation is that the polynomial of degree greater than $k$ and $\boldsymbol{\theta}$ be estimated first by the marginal likelihood. (The polynomial of degree $k$ is not modelled.) The estimate of $\boldsymbol{\theta}$ then is used to estimate the full polynomial of degree $k$.

## 5. Estimation for the CAR model

### 5.1. The general case

We now consider the CAR model of Section 2.3. The key point is that $\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}$ is given by (2.3.5). Consequently we could write the MLE for $\boldsymbol{\theta}$ given by (3.1.5) in relation to $\boldsymbol{\Sigma}^i = \frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\theta_i}$ using $\boldsymbol{\Sigma}_i = -\boldsymbol{\Sigma}\boldsymbol{\Sigma}^i\boldsymbol{\Sigma}$. Thus the ML equation for $\boldsymbol{\theta}_i$ given by (3.1.5) becomes simply

$$\hat{\mathbf{w}}'\hat{\boldsymbol{\Sigma}}^i\hat{\mathbf{w}} = \text{tr}(\hat{\boldsymbol{\Sigma}}^i\hat{\boldsymbol{\Sigma}}). \tag{5.1.1}$$

Also note for the matrix $\mathbf{A}$ in the information matrix (3.2.4) we can use that

$$2a_{ij} = \text{tr}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^i\boldsymbol{\Sigma}\boldsymbol{\Sigma}^j).$$

We now consider a particular case of Section 2.3 with

$$\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} = (\mathbf{I} - \theta\mathbf{W})/\tau^2, \tag{5.1.2}$$

where $\mathbf{W}$ is a given matrix and $\boldsymbol{\theta}' = (\theta, \tau^2)$ are the parameters. Implicitly, we are neglecting the boundary effects (see Section 2.3). It should be noted that the log–likelihood becomes

$$\text{constant} - \frac{1}{2}\log|\mathbf{I} - \theta\mathbf{W}| - \frac{n}{2}\log(\tau^2) - (\mathbf{X} - \mathbf{F}\boldsymbol{\beta})'(\mathbf{I} - \theta\mathbf{W})(\mathbf{X} - \mathbf{F}\boldsymbol{\beta})/(2\tau^2) \tag{5.1.3}$$

Further, from (3.2.6)–(3.2.8), we obtain

$$\hat{\boldsymbol{\beta}} = (\mathbf{F}'\mathbf{F})^{-1}(\mathbf{F}'\mathbf{X} - \hat{\theta}\mathbf{F}'\mathbf{W}'\hat{\mathbf{w}}), \tag{5.1.4}$$

$$\hat{\tau}^2 = \hat{\mathbf{w}}'(\mathbf{I} - \hat{\theta}\mathbf{W})\hat{\mathbf{w}}/n, \tag{5.1.5}$$

and

$$\hat{\mathbf{w}}'\mathbf{W}\hat{\mathbf{w}} = \hat{\tau}^2\text{tr}[\mathbf{W}(\mathbf{I} - \hat{\theta}\mathbf{W})^{-1}], \tag{5.1.6}$$

where $\hat{\mathbf{w}} = \mathbf{X} - \mathbf{F}\hat{\boldsymbol{\beta}}$. In Design of Experiments, $\mathbf{F}'\mathbf{F}$ is of a simple form so that (5.1.4) can be simplified further (see Section 7). Some alternative iterative procedure can be suggested, *e. g.* for given $\hat{\boldsymbol{\beta}}$ (or from the least square estimation), we can obtain $\hat{\theta}$ from (5.1.4) which when substituted into (5.1.5) leads to $\hat{\tau}^2$. Also note that the profile likelihood can be simplified through (3.3.2).

## 5.2. Properties

### *5.2.1. Unimodality.*

For $\mu = 0$, we can write the likelihood for the basic CAR, *i. e.* with $\boldsymbol{\Sigma} = \mathbf{I} - \theta\mathbf{W}$, as

$$\psi(\theta_1, \theta_2)\exp(\theta_1 T_1 + \theta_2 T_2), \tag{5.2.1}$$

where $\theta_1 = -\frac{1}{2}\tau^2$, $\theta_2 = -\frac{1}{2}\theta\tau^{-2}$, $T_1 = \mathbf{x}'\mathbf{x}$, and $T_2 = \mathbf{x}'\mathbf{W}\mathbf{x}$.

Hence the density (5.2.1) belongs to the canonical exponential family, and therefore various well–known results for this family apply. In particular, except when the sufficient statistics $T_1$ and $T_2$ are on the boundary of $(\theta_1, \theta_2)$, the MLEs of $\theta_1$ and $\theta_2$ exist and are unique. Hence the likelihood will be well–behaved (*e. g.* the log–likelihood will be concave) and unimodal. The exceptional cases are when $x_i$ equals a constant or $\mathbf{x}$ is an eigenvector of $\mathbf{W}$. In the latter case, $T_2$ is a constant. For an alternative treatment, see Ripley (1988).

For the uniqueness of the MLE for the general CAR on a lattice, see Künsch (1981); for $d = 2$, the proof of concavity of the log–likelihood is simpler.

### *5.2.2. Asymptotic normality.*

The asymptotic normality of the MLEs for the Gaussian CAR follows from Section 3.2. Further, Künsch (1983) proves the following result. Let $f(x)$ be the spectral density, and let, for a fixed subset, $\mathbf{C}^\star$ be the vector of the sample covariance functions with the divisor as the number of terms in the product. Then under certain regularity conditions, we have

$$(2n+1)^{d/2}[\mathbf{C}^\star - E(\mathbf{C}^\star)] \sim N[\mathbf{0}, 2(2\pi)^{-d}\boldsymbol{\Sigma}^\star],$$

where

$$E(\mathbf{C}^*)_h = (2\pi)^{-d} \int \cos(h\omega) f(\omega) d\omega, \text{ and}$$

$$(\mathbf{\Sigma}^*)_{g,h} = \int \cos(g\omega)\cos(h\omega)[f(\omega)]^2 d\omega.$$

We have only considered the stationary case but for the trend case, similar behaviour is expected. For small $\theta$, the CAR and the SAR are similar and therefore the MLEs are expected to behave the same way (see, Cliff and Ord, 1981; Griffith, 1988).

### 5.3. Other estimators

Another way of estimating the parameters is to use pseudo–likelihood which maximises the conditional probabilities

$$\prod f(\mathbf{x}_i \,|\, \mathbf{x}_j, j \neq i).$$

This leads, after some algebra, to the ordinary least squares estimation. Another approach is to use coding methods, where some sites are coded in a region so that no two encoded sites are to be neighbours of each other. Then the coded variables, given the rest of the sites, are mutually independent, and again we can use the least squares estimation. A simple example is for the first–order CAR on a line, where even sites are encoded given the odd sites, and vice–versa. We then can use the mean from the two estimates as the final estimator (also see Plackett, 1960).

Besag and Moran (1975) have emphasized that the coding technique always provides unbiased estimators and exact tests of significance. Also Besag (1975) has shown that under certain regularity conditions, the pseudo–likelihood estimators are consistent. In general, though, these estimators will not be as efficient as the MLE.

Besag (1977a) has given examples where the pseudo–likelihood estimators are more efficient than the coding estimators. For our discussion, consider the basic CAR given by (2.3.7). The pseudo–likelihood estimators $\{\theta^*, \tau^{*2}\}$ of $\{\theta, \tau^2\}$ can be obtained. It is found that

$$\text{Var}(\theta^*) \sim 2\theta^2[(1-\nu\rho_1)^2/n]\nu\rho_1^2, \quad \text{Var}(\tau^*) \sim \tau^2(1+\nu\theta^2)/n,$$

where $\nu$ is the number of neighbours and $\rho_1$ is the correlation between variates at the neighbouring sites. Further, if $f$ denotes the fraction of sites used in estimation (*i. e.*, coded) then the coding estimators $\tilde{\theta}$ and $\tilde{\tau}^2$ of $\theta$ and $\tau^2$ have

$$\text{Var}(\tilde{\theta}) = \theta(1-\nu\theta\rho_1)/(fn\nu\rho_1), \quad \text{Var}(\tilde{\tau}) = 2\tau^2/(fn).$$

Hence the pseudo–likelihood estimators are more efficient than the coding estimators.

### 5.4. Estimation for the T–CAR

*5.4.1. The circle case.*

Let $x_t$ be on a circle, $t = 0, 1, \ldots, n-1$, with $x_{n-1}$ being "next to" $x_0$. Initially let us assume that $\mu = 0$ so that the population covariance function is

$$\sigma_h = E\{x_t \; x_{(t+h)\text{mod}(n)}\},$$

whereas the sample covariance function at lag $h$ is

$$C(h) = n^{-1} \sum_{t=0}^{n-1} x_t \; x_{(t+h) \bmod (n)}.$$

The periodogram of $x_t$ is defined by

$$I(\omega) = \sum_{h=0}^{n-1} \exp(i\omega h) C(h). \tag{5.4.1}$$

Define from $\sigma_h, h = 0, 1, \ldots, n-1$, a circulant matrix $\mathbf{\Sigma}$ with

$$\sigma_{st} = \sigma_{(s-t) \bmod (n)} = \sigma_{(t-s) \bmod (n)}.$$

Then $\mathbf{\Sigma}$ has eigenvalues

$$\lambda_j = \sum_{h=0}^{n-1} \exp\{-2\pi i j h/n\} \sigma_h, \;\; j = 0, 1, \ldots, n-1, \tag{5.4.2}$$

and the eigenvectors $\mathbf{w}_j (n \times 1)$, say, $j = 0, 1, \ldots, n-1$, where

$$\mathbf{w}_j = [(1/\sqrt{n}) \exp\{2\pi j h/n\}, \;\; h = 0, 1, \ldots, n-1]'.$$

Thus we have a spectral decomposition of $\mathbf{\Sigma}, \mathbf{\Sigma} = \mathbf{W \Lambda W^*}$, where $\mathbf{W^*}$ is the complex conjugate transpose of $\mathbf{W}$, has columns $\mathbf{w}_j$, and where $\mathbf{\Lambda} = \text{diag}(\lambda_j)$. Hence it can be shown that the log–likelihood function simplifies to

$$\text{constant} + \frac{1}{2} \sum_{j=0}^{n-1} \log \lambda_j^{-1} - \frac{n}{2} \sum_{j=0}^{n-1} I(2\pi j/n)/\lambda_j. \tag{5.4.3}$$

From (5.4.2) we have the Fourier expansion of $\sigma_h$ as

$$\Sigma \lambda_j \cos(2\pi j h/n) = \sigma_h. \tag{5.4.4}$$

Using $\mathbf{\Lambda^{-1}} = \mathbf{W \Sigma^{-1} W^*}$, we can express $\lambda_j^{-1}$ in the form

$$\lambda_j^{-1} = \sum_{h=0}^{n-1} \phi_h \cos(2\pi j h/n), \tag{5.4.5}$$

where $\phi_h$ from (5.4.3) must satisfy

$$\sigma_h = (1/n) \sum_{j=0}^{n-1} \cos(2\pi j h/n) / \sum_{h=0}^{n-1} \phi_h \cos(2\pi j h/n). \tag{5.4.6}$$

Note that $\phi_h = \phi_{n-h}(= \phi_{-h})$; otherwise $\phi_h$ are arbitrary parameters with $\lambda_j^{-1} > 0$. If we substitute $I(\cdot)$ and $\lambda_j^{-1}$ from (5.4.1) and (5.4.5), respectively, into (5.4.3), we obtain

$$n\sum_{j=0}^{n-1}\lambda_j^{-1}I(2\pi j/n) = \sum_{h=0}^{n-1}\phi_h C(h).$$

Hence the log–likelihood from (5.4.3) becomes, except for a constant,

$$\frac{1}{2}\sum_{j=0}^{n-1}\log(\lambda_j^{-1}) - \frac{1}{2}\sum_{h=0}^{n-1}\phi_h C(h). \tag{5.4.7}$$

Now from (5.4.5), $\frac{\partial\lambda_j}{\partial\phi_h} = \cos(2\pi jh/n)$, so that on differentiating (5.4.7) with respect to $\phi_h$ we obtain the ML equation

$$\Sigma\lambda_j\cos(2\pi jh/n) = C(h).$$

The left–hand side is simply $n\sigma_h$ from (5.4.4), and thus we obtain

$$\hat{\sigma}_h = C(h). \tag{5.4.8}$$

That the MLEs of $\phi_h$ coincide with the moment estimators of covariance function $\phi_h$ is an important result. Usually, the number of $\phi_h$s is limited by $N$.

In fact, we could have started with a Gaussian CAR on a circle, with

$$E(X_t|\text{rest}) = \sum_{s\neq t}\theta_t \quad x_{(t+s)\bmod(n)}, \quad \text{Var}(X_t|\text{rest}) = \tau^2. \tag{5.4.9}$$

Then the MLEs of $\theta_h$ from (5.4.6) and (5.4.8) are the solutions to

$$C(h) = (1/n)\sum_{j=0}^{n-1}[\cos(2\pi jh/n)]/[\sum_{h=0}^{n-1}\phi_h\cos(2\pi jh/n)], \tag{5.4.10}$$

where $\phi_0 = \tau^{-2}$ and $\phi_h = -\tau^{-2}\theta$, $h \neq 0$. Analogous to (5.4.3), we can approximate the log–likelihood for large $n$ by

$$\text{constant} + \frac{1}{2}\{\int \log[\Sigma\phi_s\cos(s\omega)]^{-1}d\omega - \Sigma\phi_h C(h)\}, \tag{5.4.11}$$

which leads to the approximate ML equation (5.4.8), since

$$C(h) = (2\pi)^{-d}\int[\cos(h\omega)][\sum_s \hat{\phi}_s\cos(s\omega)]^{-1}d\omega, \tag{5.4.12}$$

where the right–hand side is precisely $\hat{\sigma}(h)$. Further, we also can express (5.4.11) as

$$\text{constant} + (n/2)\int \log[f(\omega)]^{-1}d\omega - (n/2)\int[I(\omega)/f(w)]d\omega, \tag{5.4.13}$$

where $f(w) = [\Sigma\phi_s\cos(s\omega)]^{-1}$. The importance of this result will become apparent later on, when we discuss the Whittle approximation.

*5.4.2. The torus case.*

Consider the discrete torus

$$T = \prod_{\ell=1}^{d} \{0, 1, \ldots, n[\ell] - 1\} \subset Z^d,$$

with opposite faces identified. Write $s = t \bmod(T)$ if $s[\ell] - t[\ell]$ is an integer multiple of $n[\ell]$, for each $\ell = 1, 2, \ldots, d$. Then the preceding section carries over to the torus case directly. For example, the sums over $s = 0, 1, \ldots, n-1$ are now over $s \in T$. The eigenvectors $\mathbf{w}_j, j \in T$, have entries

$$\frac{1}{\sqrt{|T|}} \exp\{2\pi i \sum_{\ell} j[\ell]h[\ell]/n[\ell]\}, \; h \in D.$$

Let $\sigma_h$, $h \in T$, be a covariance function; then the covariance matrix $\boldsymbol{\Sigma} = (\boldsymbol{\sigma}_{st})$ has entries $\boldsymbol{\sigma}_{st} = \boldsymbol{\sigma}_{s-t}$. The matrix $\boldsymbol{\Sigma}$ is called a block circulant matrix. We have the eigenvalues

$$\lambda_j = \Sigma_h \exp(-2\pi i \mathbf{j}'\mathbf{h}/n)\sigma_h. \tag{5.4.14}$$

To represent $\boldsymbol{\Sigma}$ it is necessary to arrange the elements of $\sigma_h, \mathbf{h} \in T$, in some order, *e. g.* lexicographic, though the results do not depend upon the order chosen. Again we have a spectral decomposition of $\boldsymbol{\Sigma}$, and with

$$I(\boldsymbol{\omega}) = \sum_{h \in T} \exp(i\boldsymbol{\omega}'\mathbf{h})C(\mathbf{h}), C(\mathbf{h}) = (1/n) \sum_{t, t+h \in T} x_t \; x_{(t+h) \bmod n}. \tag{5.4.15}$$

We now can write the log–likelihood, which is the same as (5.4.3) but with the above modification. In addition, for the Gaussian CAR with the representation similar to (5.4.9),

$$\lambda_j^{-1} = \sum_h \phi_h \cos(2\pi \mathbf{j}'\mathbf{h}/n),$$

and it is found that the MLEs of $\sigma_h$ are given by

$$\hat{\sigma}_h = C(\mathbf{h}), \tag{5.4.16}$$

where $\sigma_h$ satisfy an equation similar to (2.3.10). If there are restrictions on $\phi_h$ (other than $\phi_h = \phi_{-h}$, $\mathbf{h} \neq \mathbf{0}$), then we can modify the ML equations (see below). We also should note that if $\mu$ is the population mean, then from $\hat{\mu} = \mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{X}/\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}$ we get

$$\hat{\mu} = \bar{x},$$

as $\mathbf{1}$ is an eigenvector of $\boldsymbol{\Sigma}$. Also note that the ML equation and estimate also can be written from the general case, since

$$\boldsymbol{\Sigma} = \mathbf{W}\boldsymbol{\Lambda}\mathbf{W}^{\star} \text{ and } \boldsymbol{\Sigma}^{-1} = \mathbf{W}\boldsymbol{\Lambda}^{-1}\mathbf{W}^{\star}, \text{ with}$$
$$|\boldsymbol{\Sigma}| = \prod \lambda_i \text{ and } \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X} = (\mathbf{W}^{\star}\mathbf{X})'\boldsymbol{\Lambda}^{-1}(\mathbf{W}\mathbf{X})^{\cdot}$$

The key point is that *only* $\lambda_i$ depends on the parameters. Thus, for example, we can write for the Fisher information matrix for $\boldsymbol{\theta}$ (or $\phi_h$), from (3.2.4),

$$(\mathbf{B}_\theta)_{ij} = \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_j) = \frac{1}{2}\sum_{r=1}^{n}\frac{1}{\lambda_r^2}\frac{\partial\lambda_r}{\partial\theta_i}\frac{\partial\lambda_r}{\partial\theta_j}. \tag{5.4.17}$$

We also note that we can wrap (unnaturally) on the torus if there is a trend (see Mardia and Marshall, 1984).

We now consider the Gaussian first–order T–CAR with $d = 2$ given by

$$E(X_{rs}\,|\,\text{rest}) = \mu + \theta(x_{r-1,s} + x_{r+1,s} + x_{r,s-1} + x_{r,s+1} - 4\mu),$$

where $X_{rs}$ is the $(r,s)$-th observation on a torus, and $|\theta| < 1/4$. We have $\hat{\mu} = \bar{x}$.

Let $\mathrm{Var}(X_{rs}|\text{rest}) = \tau$, and let us write

$$f(\theta) = (2\pi)^{-2}\int_{(-\pi,\pi)^2}\{1 - 2\theta[\cos(x) + \cos(y)]\}^{-1}dx\,dy.$$

The ML equations for $\hat{\tau}$ and $\hat{\theta}$ are $\hat{\sigma}(0,0) = C(0,0)$, and $\hat{\sigma}(1,0)+\hat{\sigma}(0,1) = C(0,1)+C(1,0)$. For the circle, the ML equations for $\hat{\tau}$ and $\hat{\theta}$ are

$$\hat{\tau} = C(0,0) - 2\hat{\theta}[C(1,0) + C(0,1)], \text{ and } \{C(0,0) - 2\hat{\theta}[C(1,0) + C(0,1)]\}f(\hat{\theta}) = C(0,0).$$

The reason is that for the T–CAR we have $\sigma(0,0) - 2\theta[\sigma(0,1) + \sigma(1,0)] = \tau$.

We also can simplify the information matrix of $(\mu, \tau, \theta)$ from (5.4.14) and (5.4.17) (see Besag and Moran, 1975) for the simplified expressions and their numerical calculations.

## 5.5. The Whittle approximation

Let us consider a lattice for the stationary case with zero mean. Consider data $(X_t)$ available in a region $D \subset Z^d$, where $D$ is typically rectangular. Let $\{\sigma_h, \mathbf{h} \in Z^d\}$ be a covariance function with spectral density $f(\boldsymbol{\omega})$, $\boldsymbol{\omega} \in (-\pi,\pi)^d$. Then, motivated by the torus case [see equation (5.4.13) for the circle case], we consider a spectral approximation to the log–likelihood $\ell$ given by

$$2\ell = \text{constant} + n\int_{(-\pi,\pi)^d}\log[f(\boldsymbol{\omega})^{-1}]d\boldsymbol{\omega} - n\int_{(-\pi,\pi)^d}[I(\boldsymbol{\omega})/f(\boldsymbol{\omega})]d\boldsymbol{\omega}, \tag{5.5.1}$$

where

$$I(\boldsymbol{\omega}) = \sum_h \exp(i\mathbf{h}'\boldsymbol{\omega})C(\mathbf{h}),\ f(\boldsymbol{\omega}) = (2\pi)^{-d}\sum\exp(i\mathbf{h}'\boldsymbol{\omega})\sigma_h, \tag{5.5.2}$$

and $C(\mathbf{h})$ is some estimate of the covariance at lag $h$. We now outline some estimates.

Let $D_h = \{t : (t, t+h) \in D\}$. Whittle (1954) recommended using

$$C_W(\mathbf{h}) = \frac{1}{|D|}\sum_{\mathbf{t}\in D_h} x_t x_{t+h},\ \mathbf{h} \in D;\ = 0, \mathbf{h} \notin D,$$

which essentially amounts to taking $X_t = 0$ outside $D$. It is noted by Guyon (1982) that this leads to bias asymptotically in estimating the MLE. He recommends using

$$C_G(\mathbf{h}) = \frac{1}{|D_h|} \sum_{\mathbf{t} \in D_h} x_t x_{t+h}, \mathbf{h} \in D; \ = 0, \mathbf{h} \notin D,$$

where $|D_h| = \prod (n_i - |h_i|)^{-1}$. While this modification removes the bias in the MLE, $I(\omega)$ now can be less than zero. Its effect on the variance remains unclear.

Dahlhaus and Künsch (1987) recommend using $C_K(\mathbf{h})$ in place of $C(\mathbf{h})$ in (5.5.2), where

$$C_K(\mathbf{h}) = \frac{\sum_{(t,\, t+h) \in D} x_t x_{t+h} w_t w_{t+h}}{\sum_{(t,\, t+h) \in D} w_t^2},$$

with

$$w_t = \prod_{i=1}^{d} u[(t_i - \frac{1}{2}) n_i],$$

and

$$u(y) = w(2y/\rho), \ 0 \le y \le \rho/2; \ = 1, \ \rho/2 \le y \le 1/2; \text{ and, } = u(1-y), \ 1/2 \le y \le 1,$$

where $\rho$ is a smoothing parameter. The tapering pulls the data toward zero near the boundary, while keeping exactly the same value at the centre. A common taper in time series analysis is the Tukey–Hanning taper, with $w(u) = [1 - \cos(u\pi)]/2$. Dahlhaus and Künsch (1987) have shown that the bias is asymptotically negligible in estimating $\theta$ by the MLE using this approximation, and the estimator is asymptotically efficient. Also $I(\omega) > 0$.

For the M–CAR we have from (2.3.10), and $f(\omega) = 1/\sum_{s \in N_0} \phi_s \cos(\omega' \mathbf{h})$. Thus the Whittle approximation (5.5.1) leads to the moment estimates $\hat{\sigma}_h = C(\mathbf{h})$, where $\sigma_h$ is given by (2.3.10). It should be noted that $C_W(\mathbf{h})$ and $C_G(\mathbf{h})$ use the free boundaries of the C–CAR, and therefore asymptotically they may experience some loss of efficiency. Again $C_K(\mathbf{h})$ is recommended.

Assuming that the data are from the infinite CAR, then the MLE based on the C–CAR and the T–CAR will lead to a biased estimator, in general, unless these are adjusted as above. This problem does not arise for the estimates by the M–CAR.

### 5.6. The intrinsic CAR (IRF-0)

Let $f(\omega)$ be the spectral density of a process. For intrinsic processes we have

$$\int_{(-\pi,\pi)^d} f(\boldsymbol{\omega}) d\boldsymbol{\omega} = \infty \text{ and } \int_{(-\pi,\pi)^d} |\boldsymbol{\omega}|^2 f(\boldsymbol{\omega}) d\boldsymbol{\omega} < \infty.$$

Thus for the CAR given by (2.3.5) and (2.3.9), we have

$$\sum_{h \in N} \theta_h = 1 \text{ or } \sum_{h \in N} \phi_h = 0. \tag{5.6.1}$$

Given this restriction, we can proceed as before (see, Künsch, 1987). For example, consider the approximate log–likelihood given by (5.4.11), which under restriction (5.6.1) leads to

$$\hat{\gamma}_h = g(\mathbf{h}), \tag{5.6.2}$$

where $g(\mathbf{h})$ is the sample semi–variogram and $\gamma_h$ is the population semi–variogram on an infinite lattice, given by

$$\gamma_h = (2\pi)^{-d} \int [1 - \cos(\mathbf{h}'\boldsymbol{\omega})]/\{\sum_s \phi_s[1 - \cos(\mathbf{s}'\boldsymbol{\omega})]\} d\boldsymbol{\omega}. \tag{5.6.3}$$

Since in the intrinsic case the generalized covariance function is defined only up to an additive constant, equation (5.6.2) makes sense. It simply is an expression of the moment estimators

$$\sigma_{(\mathbf{0})} - \sigma_h = C(\mathbf{0}) - C(\mathbf{h})$$

from (5.4.8), when $\sigma(\mathbf{0})$ exists. Of course, this is completely analogous to (5.4.11). Further, equation (5.6.2) is exact for the T–CAR.

A particular "intrinsic" CAR of interest for the finite lattice is

$$E(X_i|\text{rest}) = \bar{x}_i \text{ and } \text{Var}(X_i|\text{rest}) = \tau^2/\nu_i, \tag{5.6.4}$$

where $\nu_i$ is the number of neighbours of $i$, and $\bar{x}_i$ is the mean of the neighbouring values of $i$. This specification suggests a neat way of defining boundary corrections for finite lattices (Besag, 1989). It has the joint density (Künsch, 1987)

$$\text{constant } \tau^{-n} \exp[-\frac{1}{2\tau^2} \sum_{i \sim j} (x_i - x_j)^2], \tag{5.6.5}$$

where the sum is over all $i \sim j$ that are neighbours. A practical example appears in Kent and Mardia (1988). Note that the density is singular, but the MLE of $\tau^2$ is straightforward and it is clearly well–behaved. For the infinite lattice, this becomes a particular case of the basic CAR defined by (2.3.7).

### 5.7. Landsat data

We consider Switzer's Landsat data of three rock–types on one of four spectral bands (namely, red). The assumption of stationarity is inadequate since the mean depends upon the rock type. The simple model that $x_i$ is a function only of rock type plus white noise also is not adequate, since the values in two regions of the same rock type differ considerably; this also may be due to some blurring and other features of the ground, such as texture and orientation, contributing to the signal. It seems that fitting an overall trend or fitting an intrinsic model may prove adequate. The data are on a 16–by–25 lattice, and have been analyzed by Künsch (1987) through IRF-0. Table 2 shows the MLE with various neighbourhood schemes, utilizing a Whittle–type approximation to the IRF.

TABLE 2
ESTIMATED PARAMETERS OF AND INTRINSIC CAR OF ORDER 0 FOR LANDSAT DATA

| Number of Neighbors | $\hat{\theta}_h$ with $h$ specified (0,1) | (0,2) | (1,−1) | (1,0) | (1,1) | (2,0) | $\hat{\tau}^2$ | 400L |
|---|---|---|---|---|---|---|---|---|
| 4 | 0.346 | | | 0.154 | | | 7.35 | 931.7 |
| 8 | 0.385 | | −0.014 | 0.267 | −0.138 | | 6.05 | 905.0 |
| 12 | 0.442 | −0.057 | −0.002 | 0.206 | −0.104 | 0.015 | 5.74 | 897.8 |

Table 2 gives estimated parameters for the first, second, and third–order neighbourhoods, based upon solving (5.6.2) numerically. It also gives $L = -(2) \times (400) \times$ log–likelihood, which has been approximated, where 400 is the number of observations. There is a clear anisotropy in the data, since $\hat{\theta}_{1,1}$ and $\hat{\theta}_{1,-1}$ are different.

One should note that the Akaike criterion,

$$L + 2 \times \text{(the number of parameters)},$$

gives values for $400L$ of

$$937.7 \quad 915.0 \quad 911.8.$$

These values indicate that it is better to use at least the second–order neighbourhoods. There is only a slight gain in using the third–order neighbourhood.

We now check how the model fits the data and how it compares with the corresponding stationary models. If we use the third–order neighbourhood, the sum of regression coefficients $\Sigma\theta_h = 0.9954$, so that we are on the boundary of the parameter space (*i. e.*, it indicates an intrinsic model). Figure 8 shows the theoretically fitted combined semi–variograms for (i) an intrinsic CAR with third–order neighbours, (ii) an ordinary CAR with third–order neighbours, and (iii) an empirical variogram. Figure 8 also shows that the agreement is good for both models, for small lags, while the intrinsic model fits better for larger lags.

These findings imply that the simple discrimination technique using a constant mean for each rock type cannot work; but, an intrinsic model with constant mean in each region is plausible. Kent and Mardia (1988) have given another analysis of the same data.

## 6. The errors in variable model

### 6.1. The direct model case

Consider

$$\text{Cov}(\mathbf{X}) = \sigma^2\mathbf{P} + \psi^2\mathbf{I}, \tag{6.1.1}$$

where $\mathbf{P}$ is a correlation matrix, or

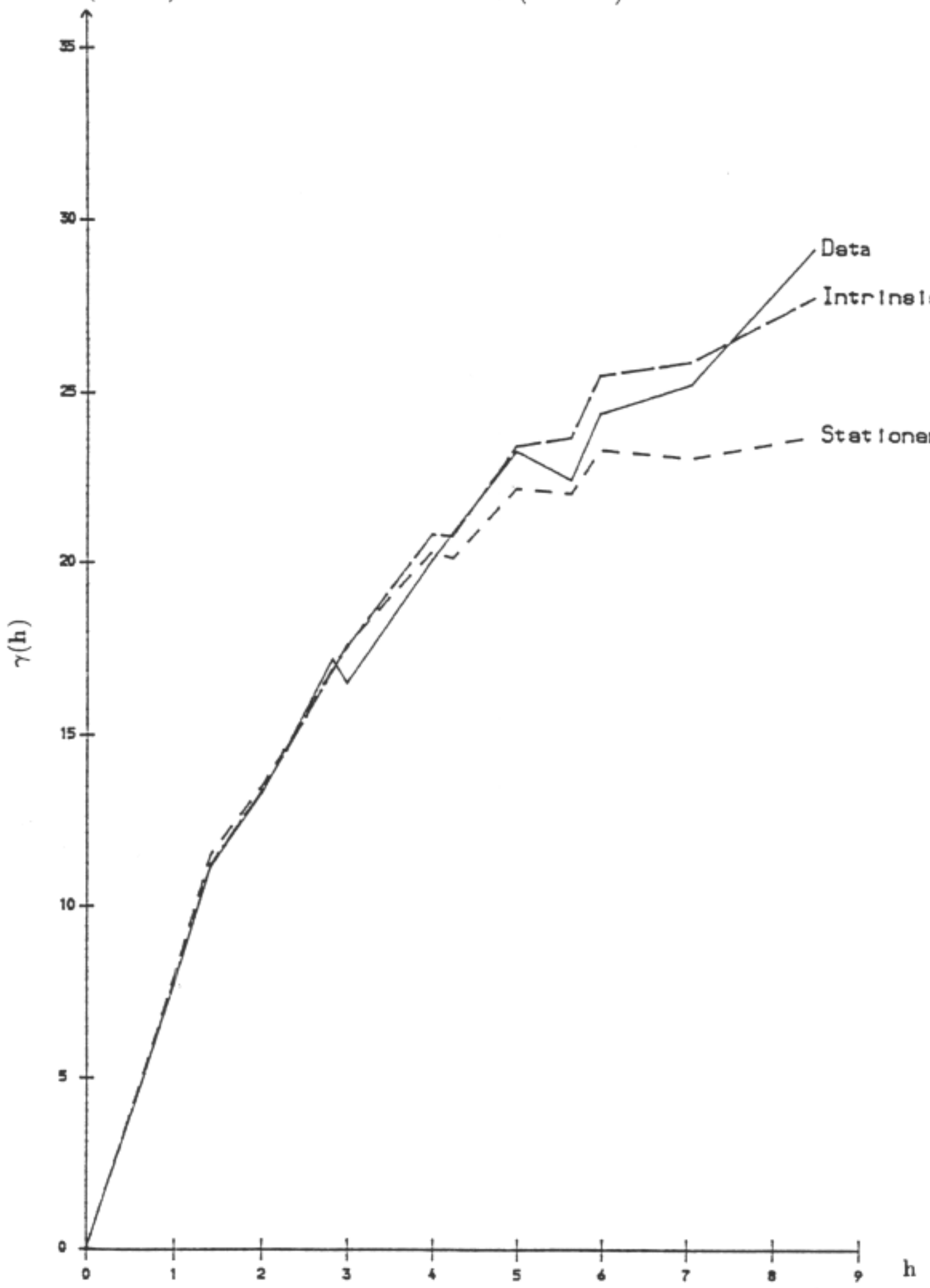
$$\text{Cov}(X_i, X_j) = \sigma^2\rho_{ij} + \psi^2\delta_{ij}, \ \rho_{ij} = \rho(\mathbf{t}_i - \mathbf{t}_j),$$

with

$$\rho(0) = 1, \text{ and } \delta_{ij} = 1 \text{ if } i = j; \ = 0 \text{ if } i \neq j.$$

**Figure 8.**

Semi-variogram for the Landsat data (———), fitted variograms for 3rd order stationary CAR (- - - -) and for 3rd order intrinsic CAR (— — —).



We will call $\psi^2$ a nugget parameter.

For the lattice case, we could write the underlying process as

$$X(\mathbf{t}) = \varepsilon(\mathbf{t}) + \eta(\mathbf{t}), \tag{6.1.2}$$

with $\text{Cov}(\varepsilon_i, \varepsilon_j) = \sigma^2\rho(\mathbf{t}_i - \mathbf{t}_j)$, $\rho(0) = 1$, $\text{Var}[\varepsilon(\mathbf{t})] = \psi^2$, and $\text{Cov}\{\varepsilon(\mathbf{t}), \eta(\mathbf{t})\} = 0$.

Hence (6.1.2) is an errors–in–variable model, where $\varepsilon(\mathbf{t})$ is uncontrollable error and $\eta(\mathbf{t})$ is measurement error. Representation (6.1.2) does not make sense for a continuous process; but (6.1.1) is still meaningful where the nugget parameter $\psi^2$ represents the error, depending upon the choice of the sampling frame (*i. e.*, the sites $\mathbf{t}_i$). The point to bear in mind here is that for continuous processes, the "white noise" term $\eta(\mathbf{t})$ in (6.1.2) is not meaningful. Note that if $\rho(\mathbf{h};\theta)$ is the autocorrelation function of the process, where $\theta$ denotes the range parameter (see Section 2.2), then white noise can appear with $\theta \to 0$ or $\sigma^2 \to 0$.

We have described the model with three parameters, namely nugget, range and sill $(\psi^2 + \sigma^2)$. We can easily incorporate the trend into the model. Also, we can extend these results in order to incorporate anisotropy; Brewer and Mead (1986) have used, in addition to the nugget,

$$\sigma(\mathbf{h}) = \exp[-(\mathbf{h}'\mathbf{G}^{-1}\mathbf{h})^{\beta/2}],\ 0 \le \beta \le 2,$$

where $\mathbf{G}$ is a symmetric matrix. This scheme models geometric anisotropy. Note that $\sigma(\mathbf{h})$ can be easily seen to be a covariance function, since it can be related to the characteristic function of a multivariate distribution (see Mardia, 1986). Brewer and Mead (1986) provide an intuitively plausible method to estimate the parameters, and it remains to be seen how far the exact MLEs succeed in capturing a similar behaviour.

We now describe a computational procedure.

### 6.2. A profile likelihood

Extending the profile likelihood approach of Section 3.3 is worthwhile for numerical work. Let us now take the vector $\boldsymbol{\theta}$ with just three parameters, and write

$$\sigma^2,\ \delta = \psi^2/\sigma^2,\ \text{and}\ \theta,$$

where we now assume that $\theta$ is a correlation parameter in $\mathbf{P}(\theta)$. For a given $\theta$ we can use the spectral decomposition of $\mathbf{P}(\theta)$, so that

$$\mathbf{P}(\theta) = \mathbf{C}(\theta)'\boldsymbol{\Lambda}(\theta)\mathbf{C}(\theta), \tag{6.2.1}$$

where $\mathbf{C}(\theta)$ is an orthogonal matrix, and

$$\boldsymbol{\Lambda}(\theta) = \text{diag}\{\lambda_1(\theta), \lambda_2(\theta), \ldots, \lambda_n(\theta)\}.$$

As before, suppose the "trend" is $\mathbf{F}\boldsymbol{\beta}$. Then the log–likelihood from equation (3.1.1) is simply

$$\ell = -(n/2)\log(\sigma^2) - (1/2)\sum_{i=1}^{n}\log[\delta + \lambda_i(\theta)] - \sum_{i=1}^{n} u_i^2(\theta, \boldsymbol{\beta})/\{(2\sigma^2)[\delta + \lambda_i(\theta)]\}, \tag{6.2.2}$$

where

$$\mathbf{u}(\theta, \boldsymbol{\beta}) = \mathbf{C}(\mathbf{X} - \mathbf{F}\boldsymbol{\beta}). \tag{6.2.3}$$

It can be shown, as in Mardia (1980), that

$$\hat{\sigma}^2(\delta, \theta, \boldsymbol{\beta}) = \{\sum_{i=1}^{n} u_i^2[\delta + \lambda_i(\theta)]^2\} / \sum_{i=1}^{n} [\delta + \lambda_i(\theta)]^{-2}. \tag{6.2.4}$$

The ML equations for $\delta$ and $\boldsymbol{\beta}$ do not depend on $\sigma^2$, and are

$$\hat{\boldsymbol{\beta}}(\theta, \delta) = [\mathbf{F}'\mathbf{C}(\boldsymbol{\Lambda} + \delta\mathbf{I})^{-1}\mathbf{C}\mathbf{F}]^{-1}\mathbf{C}'(\boldsymbol{\Lambda} + \delta\mathbf{I})^{-1}\mathbf{C}\mathbf{X}, \tag{6.2.5}$$

and

$$(1/n)\Big[\sum_{i=1}^{n} u_i(\theta, \hat{\boldsymbol{\beta}})^2[\hat{\delta} + \lambda_i(\theta)]^{-1}\Big]\Big[\sum_{i=1}^{n} [\hat{\delta} + \lambda_i(\theta)]^{-2}\Big] = \sum_{i=1}^{n} u_i(\theta, \hat{\boldsymbol{\beta}})^2[\hat{\delta} + \lambda_i(\theta)]^{-2}. \tag{6.2.6}$$

Now in substituting $\hat{\boldsymbol{\beta}}$ from (6.2.5), for a given $\theta$, a solution can be obtained in terms of $\delta(\theta)$. In turn, we substitute $\delta$ into (6.2.5) to obtain $\hat{\boldsymbol{\beta}}(\theta, \hat{\delta})$, and then from (6.2.4) we get $\hat{\sigma}^2 = \sigma^2(\hat{\delta}, \theta, \hat{\boldsymbol{\beta}})$. Thus the profile likelihood with respect to $\theta$ can be obtained from (6.2.2). Hence, the ML estimate of $\theta$ can be derived, which in turn gives the MLEs of $\delta$, $\boldsymbol{\beta}$, and $\sigma^2$. Alternatively we could work as we did in Section 3.3, so that the profile likelihood is a function of $\theta$ and $\delta$, since $\delta$ is not eliminated. It also should be noted that for IRF-k, equation (6.2.5) does not exist. Therefore, following Section 4.2, we must first solve equations similar to (6.2.6) for $\delta$, and then we can obtain $\hat{\sigma}^2(\theta, \delta)$ from an equation similar to (6.2.4). Next, from an equation similar to (6.2.2) we obtain the profile likelihood of $\theta$, then $\hat{\theta}$, and finally the MLEs of $\sigma^2$ and $\delta$.

*Example.* Figure 9 shows bauxite grade values in two dimensions for a region of Southern France (Marechal and Serra, 1970). The empirical, combined semi–variogram given in Figure 10 shows that there is a possible nugget. There is a peak at about $|\mathbf{h}| = 7$, which reflects the depression in the centre of the data. We fitted a power scheme with constant mean, nugget $\psi^2$, variance $\sigma^2$, and range $\alpha$. The profile log–likelihood for $\alpha$ and $\delta$ is shown in Figure 11 and appears to be quite flat. Figure 10 also shows the fitted semi–variogram. It should be borne in mind that the contours are not equi–spaced. The accompanying parameter estimates are found to be

$$\hat{\mu} = 14.31, \ \hat{\delta} = 0.278, \ (\hat{\psi}^2 = 24.51), \ \hat{\sigma}^2 = 88.18, \ \hat{\alpha} = 8.12, \ \text{and} \ \log(\ell) = -119.97.$$

Figure 12 shows contours from the ML predictor that clearly captures the depression in the data.

### 6.3. The CAR model case

For large scale data the above computational method is not feasible. We again could use the Whittle approximation, which in turn relies on the torus approximation (Besag, 1977b). Let us write $\boldsymbol{\Sigma}_0 = \sigma^2\mathbf{P}$. Let

$$\lambda_j = 1/\sum_{h} \phi_h \cos(2\pi \mathbf{j}'\mathbf{h}/n);$$

Figure 9.

Spatial distribution of Bauxite grades (Marechal & Serra, 1970).

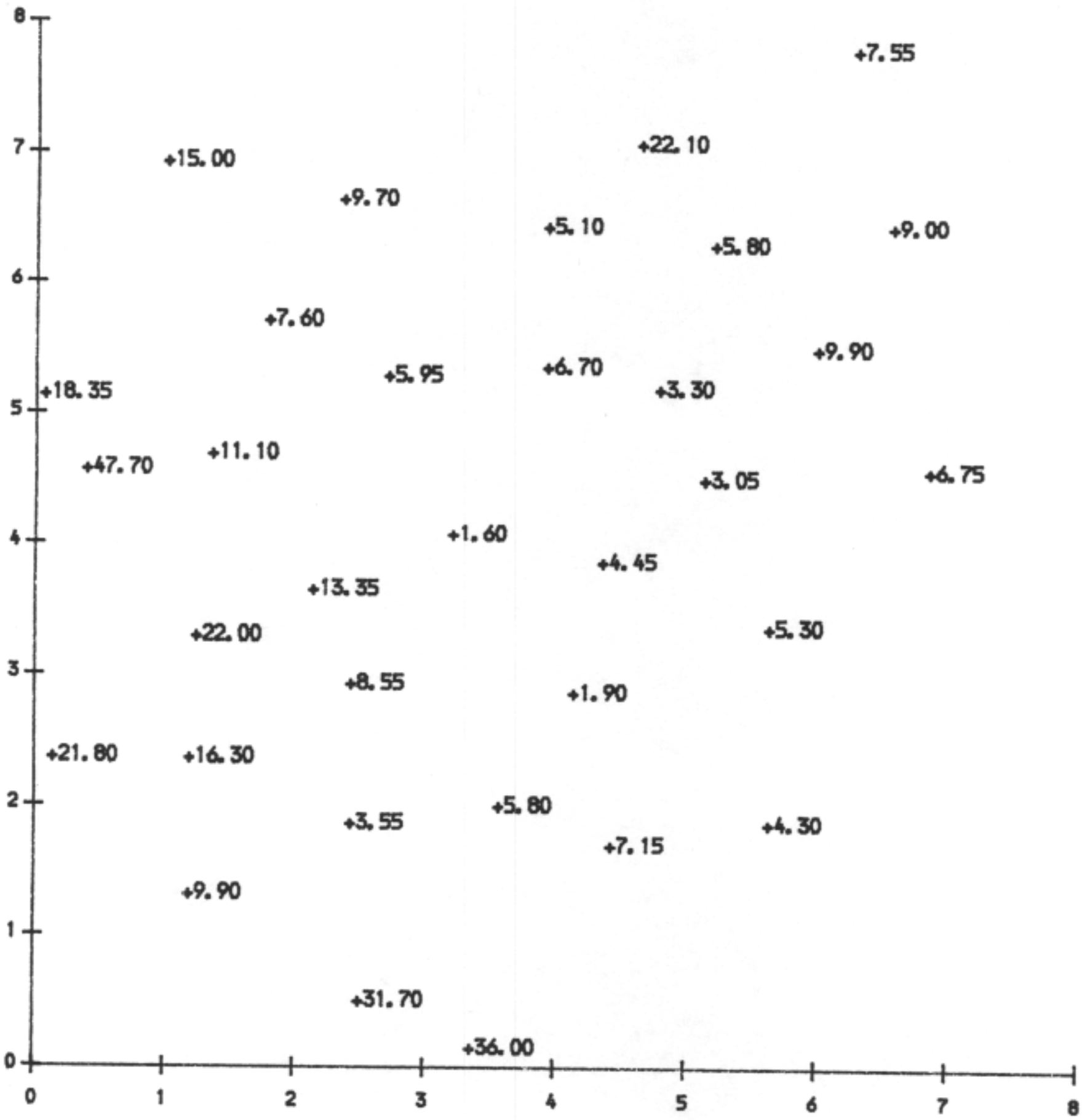


**Figure 10.**

A fitted isotropic variogram to the Bauxite data reflecting a nugget effect.

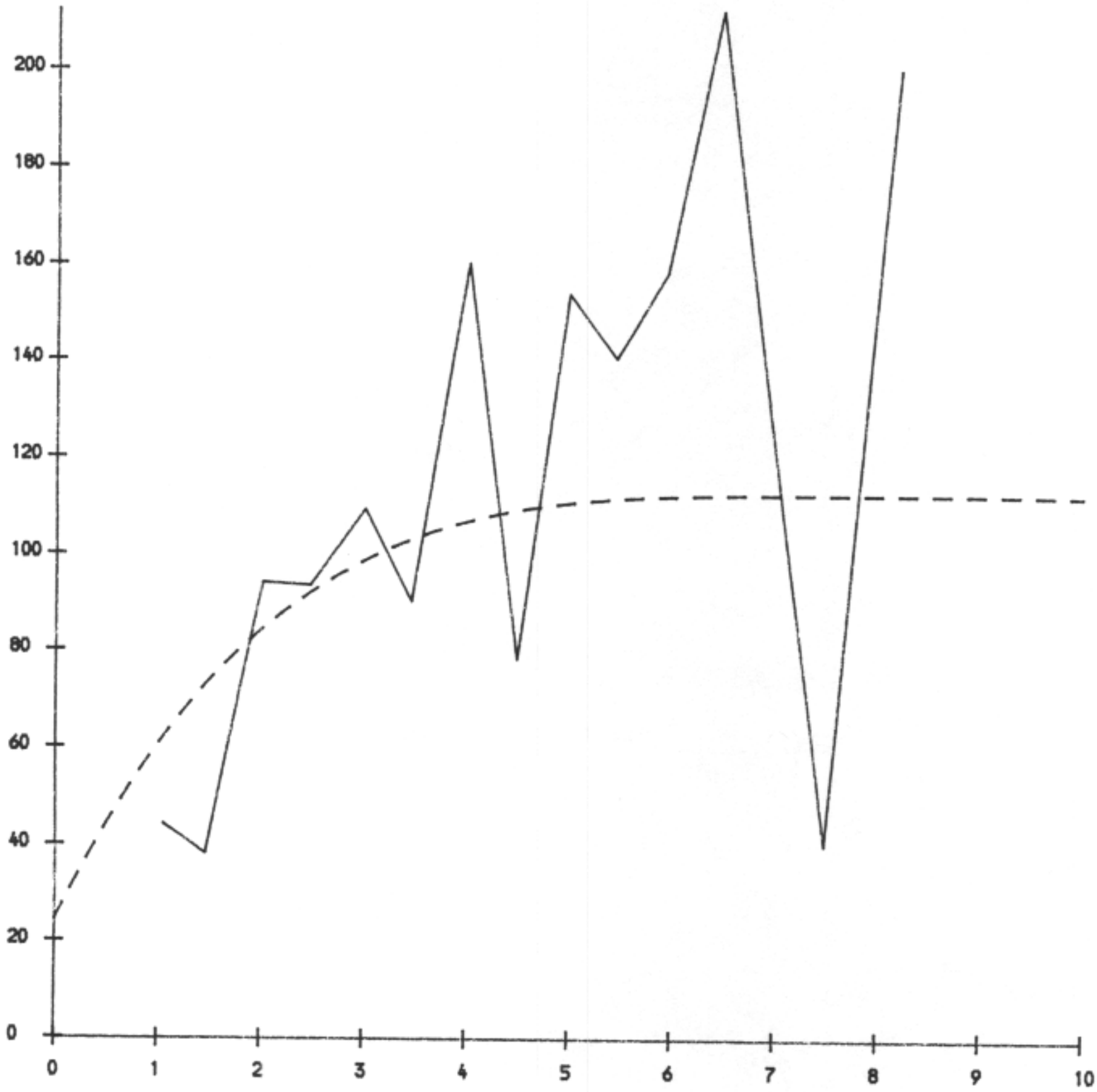


**Figure 11.**

Profile likelihood with nugget effect for the Bauxite data. [Contours:

$$1 = 123.0,\ 2 = -122.0(-0.5), \ldots, 5 = -120.50,\ 6 = -120.4(-0.1), \ldots, 8 = -120.20,$$

$$9 = -120.09(-0.02), \ldots, 13 = -120.01,\ 14 = -120.00(-0.01), \ldots, 17 = -119.97.]$$

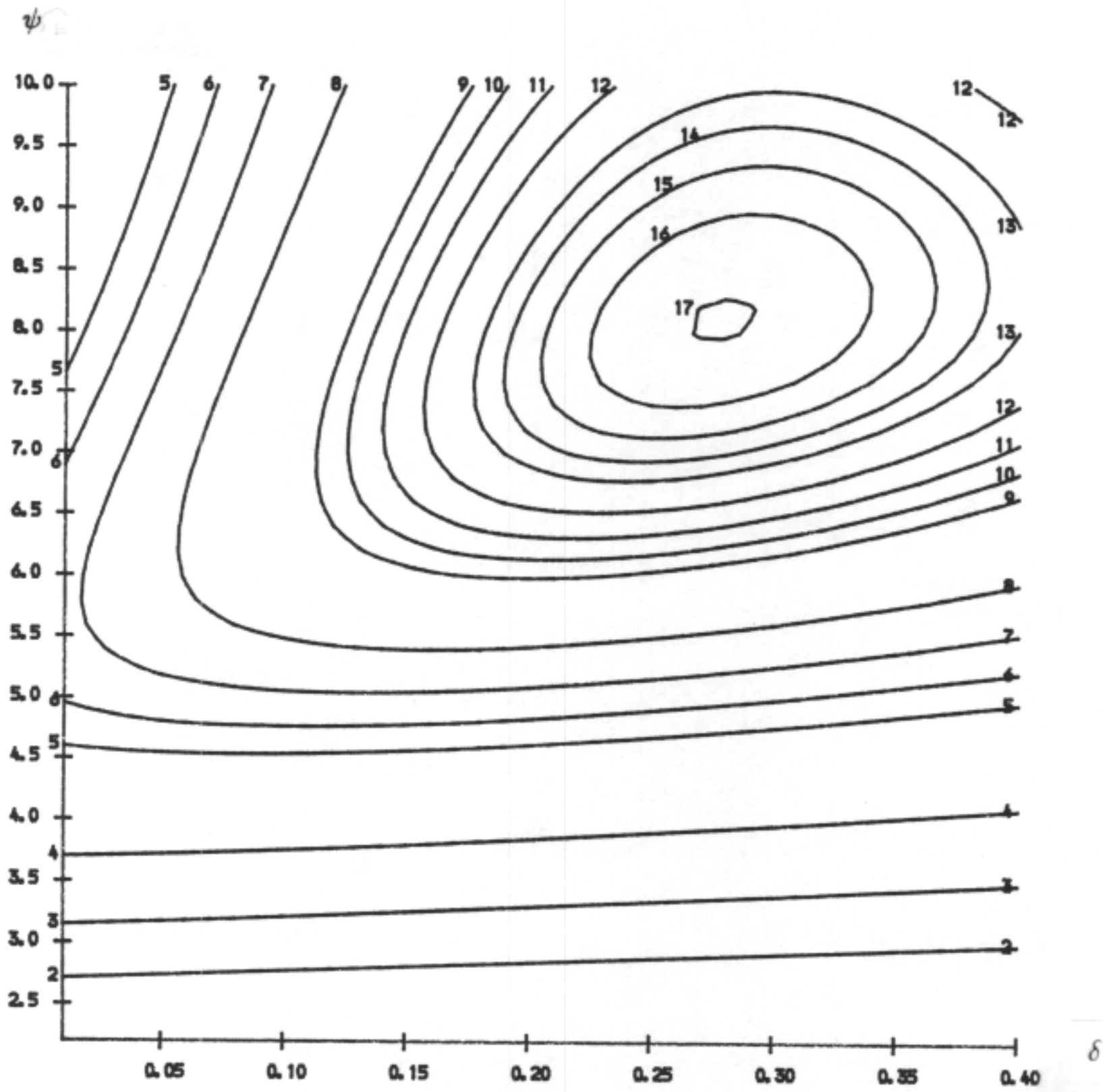
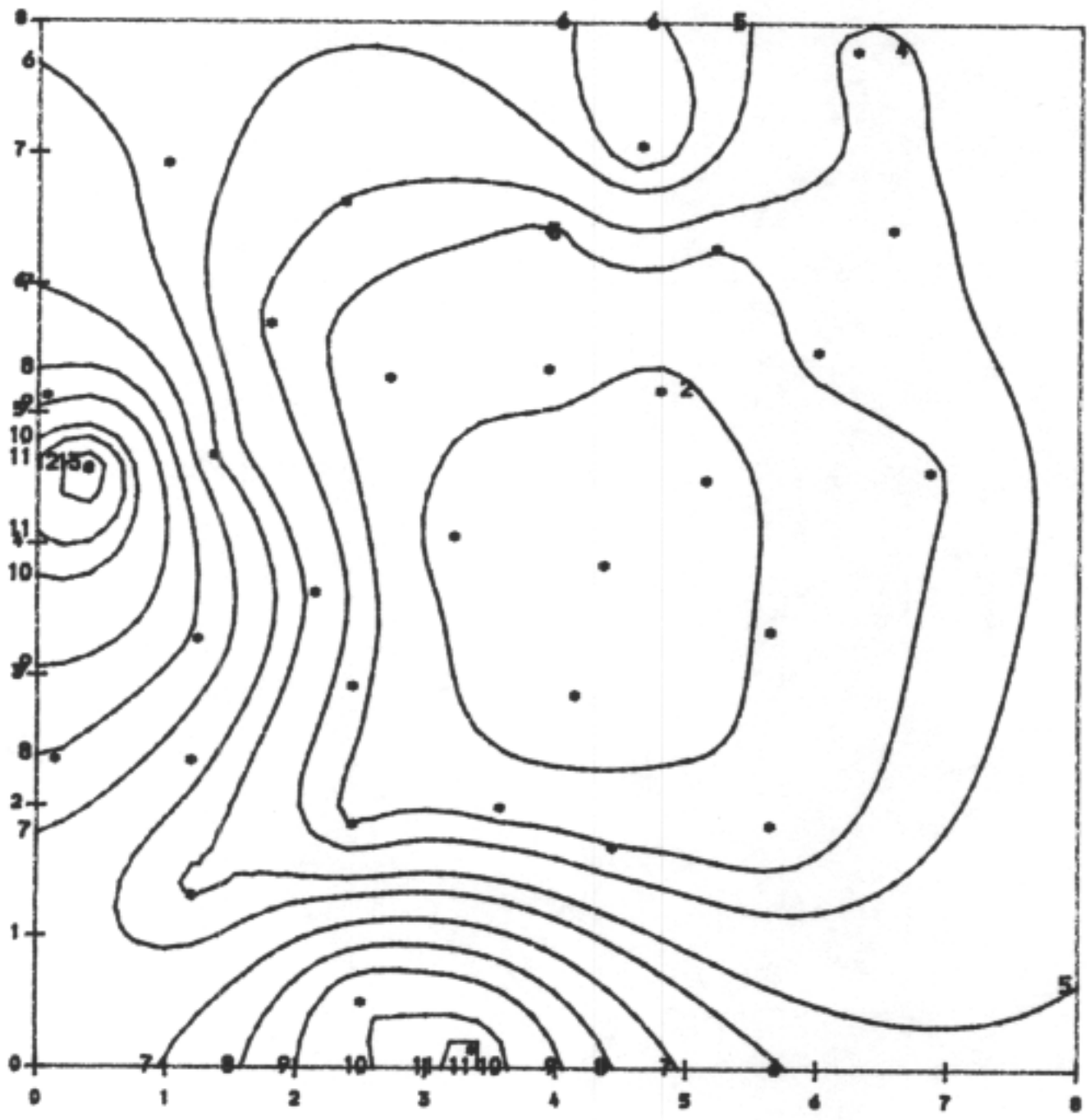


**Figure 12.**

Contour from the maximum likelihood predictor for the Bauxite data under stationary model with power scheme. [Contours: $1 = 3(3), \ldots, 3 = 8,\ 4 = 10,\ 5 = 13,\ 6 = 15(3), \ldots, 9 = 24,\ 10 = 28,\ 11 = 30,\ 12 = 33,\ 13 = 35(3), \ldots, 15 = 41$.]



then we can find an orthogonal matrix $\mathbf{W}$ (as in Section 5.3 for the torus case), given by

$$\boldsymbol{\Sigma}_0 = \mathbf{W}'\boldsymbol{\Lambda}\mathbf{W}, \quad \boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n),$$

where $\mathbf{W}$ does not depend on the parameters $\phi_h$. For the zero mean case,

$$|\boldsymbol{\Sigma}| = \Pi(\psi^2 + \lambda_j), \ \mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x} = \Sigma(\psi^2 + \lambda_j)^{-1} I(2\pi j/n),$$

where $I(\cdot)$ is the periodogram given by (5.4.14). However, we no longer have $\hat{\sigma}_h = C(\mathbf{h})$, although, as before, the ML equations can be written from the general equations. An exception is the first–order C–CAR model (Besag, 1978). Besag and Kempton (1986) have considered a regression case on the line with nugget effect, with the error being IRF–0. In general, we could add a nugget term to (4.3.1).

*Example.* Mercer and Hall (1911) have given the results of a uniformity trial on the wheat plots on a 20–by–25 lattice. Following Besag (1977b), we fit the first–order Gaussian CAR model with nugget parameter, mean and variance. Let us write

$$E(X_{ij}|\text{rest}) = \mu(1 - 2\theta_1 - 2\theta_2) + \theta_1(x_{i-1,j} + x_{i+1,j}) + \theta_2(x_{i,j-1} + x_{i,j+1}),$$

and

$$\operatorname{Var}(X_{ij}|\text{rest}) = \tau^2.$$

The accompanying estimation results are

$$\hat{\mu} = 3.95, \ \hat{\tau}^2 = 0.033, \ \hat{\delta} = 2.108, \ \hat{\theta}_1 = 0.4758, \ \text{and} \ \hat{\theta}_2 = 0.0203,$$

where $\hat{\theta}_1$ and $\hat{\theta}_2$ have SEs of 0.01, and have a high negative correlation of $-0.97$. This latter correlation reflects the relationship $|\theta_1| + |\theta_2| < 1/2$. It follows from Section 2.5 that $\hat{\psi}^2 = 0.0696$ and $\hat{\sigma}^2 = 0.1335$. Figure 13 gives the semi–variograms of the data in four directions, with fitted semi–variogram plots that indicate a nugget effect as well as anisotropy. Also, $\hat{\theta}_1 + \hat{\theta}_2 = 0.4961$ indicates that there is a trend; the plots also indicate ripples for the semi–variograms especially around $|\mathbf{h}| = 3$. In fact, the data have a periodic trend, as is revealed by plotting the empirical correlation function, or more clearly by graphing the empirical spectral density (see McBratney and Webster, 1983).

## 6.4. Asymptotics

Let us write

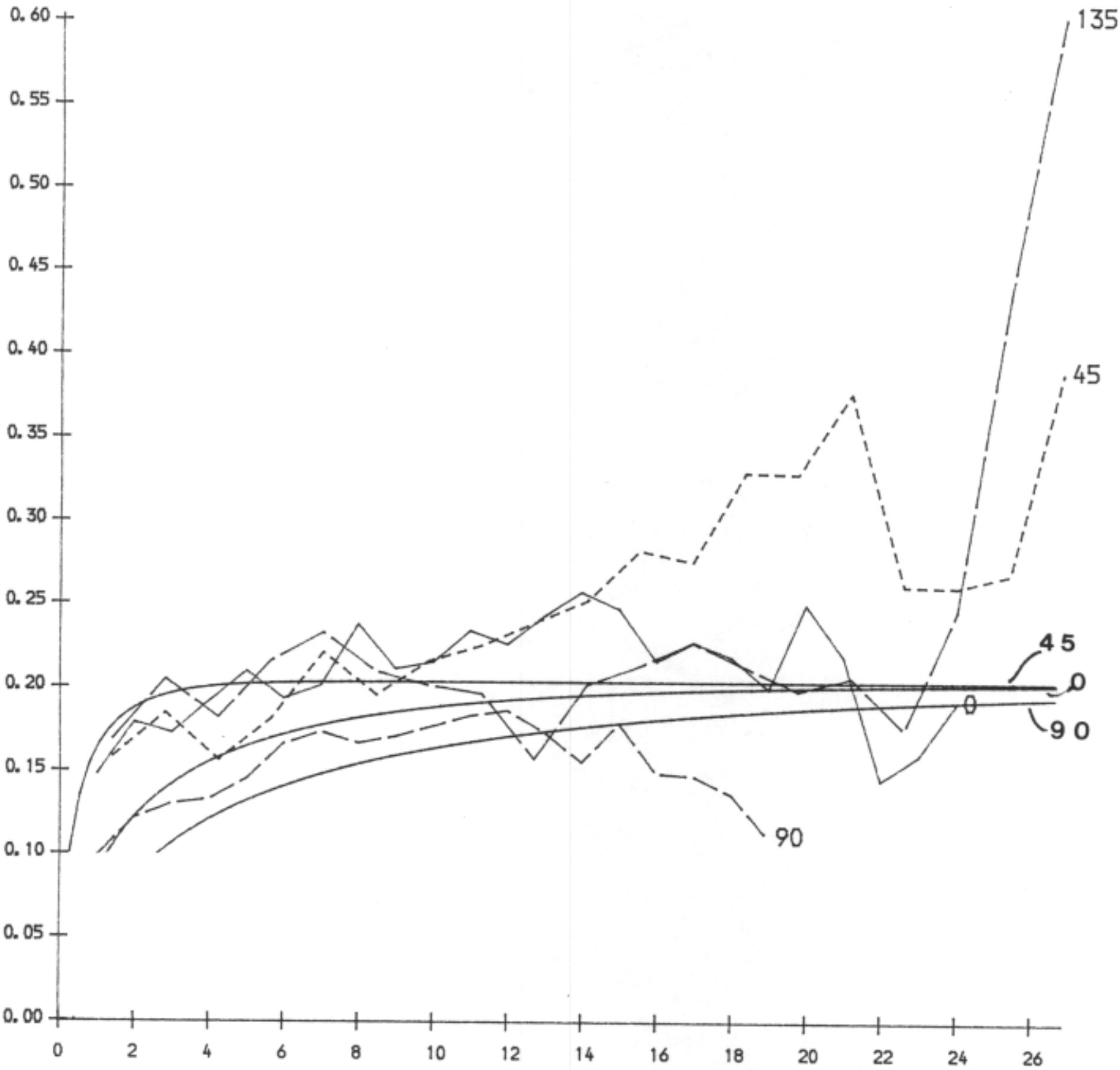
$$\sigma(\mathbf{h}; \boldsymbol{\theta}) = \sigma_1(\mathbf{h}; \boldsymbol{\theta}_2) + \theta_1\delta(\mathbf{h}), \ \theta_1 \geq 0, \tag{6.4.1}$$

where $\theta_1$ is the nugget parameter, and $\delta(\mathbf{h}) = 1$ if $\mathbf{h} = \mathbf{0}$; $\delta(\mathbf{h}) = 0$ if $\mathbf{h} \neq \mathbf{0}$. Since $\theta_1 = 0$ lies on the boundary of the parameter space, the asymptotics are not straightforward. For such cases in the independent and identically distributed variables, see Moran (1971) and Chant (1974).

Let $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ be positive definite for all values of $\boldsymbol{\theta}$, where we restrict $\theta_1 \geq 0$. Although $\sigma(\mathbf{h}; \boldsymbol{\theta})$ is not positive definite for all $\theta_1 < 0$, it is the case that $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ will remain positive definite for $\theta_1 < 0$ but near enough to zero. Let $\hat{\theta}_1^\star$ and $\hat{\theta}_1$ be the MLEs of $\theta_1$ for the unrestricted case and the restricted case, respectively. Following Mardia and Marshall (1984), it seems plausible that the likelihood will be quadratic in $\theta_1$ as $h \to \infty$. Hence $\hat{\theta}_1^\star$ will be

**Figure 13.**

For Mercer and Hall data, the semi–variogram in the four principal directions and the semi–variogram from fitted first order CAR.



asymptotically normal, and $\hat{\theta}_1$ will be asymptotically equivalent to $\text{MAX}(0, \hat{\theta}_1^*)$. Therefore, asymptotically $\hat{\theta}_1$ will have a censored normal distribution at $\hat{\theta}_1 = 0$. Hence to test

$$H_0 : \theta_1 \geq 0 \text{ versus } H_1 : \theta_1 \text{ unrestricted},$$

it can be shown for unknown $\boldsymbol{\theta}_2$ for large $n$ that, with probability $1/2$,

$$-2\log_e(\lambda) = \begin{cases} 0 & \text{if } \hat{\theta}_1 \leq 0, \\ \hat{\theta}_1/\text{Var}(\hat{\theta}_1) & \text{if } \hat{\theta}_1 > 0, \text{ distributed as } \chi_1^2, \end{cases}$$

where $\lambda$ is the likelihood ratio statistic. Hence to test $\theta_1 = 0$, one should use

$$\hat{\theta}_1^2/\text{Var}(\hat{\theta}_1) > \chi_1^2(2\alpha), \tag{6.4.2}$$

where $\text{Var}(\hat{\theta}_1)$ is computed at $\theta_1 = 0$ as described in the subsequent discussion. (It should be noted that the level of significance gets doubled.)

Recall from (3.2.2) that the $(i,j)$-th element of the information matrix $\mathbf{A}$ for $\boldsymbol{\theta}$ is $a_{ij} = (1/2)\text{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_j)$. If $\mathbf{A}$ is partitioned for $(\theta_1, \boldsymbol{\theta})$, then

$$\text{Var}(\theta_1)_{\text{ at } \theta_1=0} \sim (a_{11} - \mathbf{a}_{21}'\mathbf{A}_{22}^{-1}\mathbf{a}_{21})_{\theta_1=0}^{-1}. \tag{6.4.3}$$

Some simplification of $\mathbf{A}$ can be achieved with

$$\boldsymbol{\Sigma} = \theta_1\mathbf{I} + \theta_2\mathbf{P}(\boldsymbol{\theta}), \tag{6.4.4}$$

since we have $\boldsymbol{\Sigma}_1 = \mathbf{I}$, $\boldsymbol{\Sigma}_2 = \mathbf{P}(\boldsymbol{\theta})$, and $\boldsymbol{\Sigma}_i = \theta_2\mathbf{P}_i(\boldsymbol{\theta})$ for $i > 2$. Hence,

$$\begin{aligned} a_{11} &= \theta_2^{-2}\text{tr}(\mathbf{P}^{-2}), \\ a_{22} &= n\theta_2^{-2}, \\ a_{12} &= \theta_2^{-1}\text{tr}(\mathbf{P}^{-1}), \text{ and} \\ a_{ij} &= \theta_2^{-1}\text{tr}(\mathbf{P}^{-1}\mathbf{P}_i\mathbf{P}^{-1}\mathbf{P}_j), \ i,j = 3,4,\ldots,n. \end{aligned} \tag{6.4.5}$$

Under (6.4.4), consider on the line $\rho(\theta)$ with the correlation function

$$\rho(h;\theta) = \theta^{|h|}, \ |\theta| < 1.$$

Under $\theta_1 = 0$, we find that

$$\begin{aligned} a_{11} &= \frac{1}{2}\theta_2^{-2}(1-\theta^2)^{-2}[n + 2(2n-3)\theta^2 + (n-2)\theta^4] \\ a_{12} &= \frac{1}{2}\theta_2^{-2}(1-\theta^2)^{-1}\{2 + (n-2)(1+\theta^2)\}, \\ a_{22} &= \frac{1}{2}n\theta_2^{-2}. \end{aligned}$$

Hence asymptotically $\text{Var}(\hat{\theta}_1)$ under $\theta_1 = 0$ from (6.4.3) is

$$\theta^2/\{n\theta_2^2(1-\theta^2)^2\}.$$

Hence, the variance is well defined. For $\theta = 0$, the variance as expected, is zero.

## 7. Some extensions

### 7.1. Multivariate spatial model

Let $\mathbf{X}$ be a matrix of $n \times m$ observations, where the $i$-th row $(\mathbf{X})_i$ is the $m$–variate observation at the $i$–th site, $i = 1, 2, \ldots, n$. We can write

$$E(\mathbf{X})_i = \mathbf{f}(\mathbf{t}_i, \boldsymbol{\beta}),\ \boldsymbol{\beta} \in R^q,\ \text{and}\ \mathrm{Cov}\{(\mathbf{X})_i,\ (\mathbf{X})_j\} = \rho(\mathbf{t}_i, \mathbf{t}_j; \boldsymbol{\theta}),\ \boldsymbol{\theta} \in R^p.$$

We assume that

$$\Sigma\Sigma\alpha_i\alpha_j\rho(\mathbf{t}_i, \mathbf{t}_j; \boldsymbol{\theta})$$

is a p.d. matrix for all $\alpha_1, \alpha_2, \ldots, \alpha_s$, and that vectors $\mathbf{t}_i$ and $\mathbf{t}_j$ are contained in $R^d$. Writing $\mathbf{X}$ as a stacked vector, denoted by $\mathrm{Vec}(\mathbf{X})$, yields

$$\mathrm{Cov}[\mathrm{Vec}(\mathbf{X})] = \boldsymbol{\Sigma}(\boldsymbol{\theta}),$$

where $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is an $mn$–by–$mn$ matrix. Assuming $\mathrm{Vec}(\mathbf{X})$ to be normally distributed, then both the likelihood of $\mathbf{X}$ and its ML equations can be written as before.

Of special interest here is the linear model, where

$$E(\mathbf{X}) = \mathbf{F}\boldsymbol{\beta}.$$

The regression coefficient $\boldsymbol{\beta}$ is now a matrix (*e. g.*, for $d = 2$ with a quadratic trend, $\boldsymbol{\beta}$ is a 6–by–$m$ matrix and $\mathbf{F}$ is an $n$–by–6 matrix). We should note that $\mathbf{F}$ is the same as that for the univariate case of $m = 1$. We can use a "factorized model" for the covariance matrix (Mardia, 1984), such that

$$\mathrm{Cov}[(\mathbf{X})_i, (\mathbf{X})_j] = \rho(\mathbf{i} - \mathbf{j}; \boldsymbol{\theta})\boldsymbol{\Lambda},$$

with $\rho(0) = 1$, so that $\mathrm{Var}[(\mathbf{X})_i] = \boldsymbol{\Lambda}$ for all $i$. Here $\boldsymbol{\Lambda}$ is an $m$–by–$m$ symmetric matrix. Thus

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Gamma} \otimes \boldsymbol{\Lambda}.$$

Hence, the log–likelihood is simply

$$\text{constant} - (n/2)\log_e|\boldsymbol{\Lambda}| - (m/2)\log_e|\boldsymbol{\Gamma}| - (1/2)\mathrm{tr}[(\mathbf{X} - \mathbf{F}\boldsymbol{\beta})'\boldsymbol{\Gamma}^{-1}(\mathbf{X} - \mathbf{F}\boldsymbol{\beta})\boldsymbol{\Lambda}^{-1}].$$

We can now obtain the ML equation for $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and $\boldsymbol{\Lambda}$ for a given $\mathbf{F}$ and $\rho(\cdot)$. The profile likelihood for $\boldsymbol{\theta}$ simply maximizes

$$-(n/2)\log|\hat{\boldsymbol{\Lambda}}(\boldsymbol{\theta})| - (m/2)|\hat{\boldsymbol{\Gamma}}(\boldsymbol{\theta})|,$$

where

$$\hat{\boldsymbol{\Lambda}}(\boldsymbol{\theta}) = (1/n)[\mathbf{X} - \mathbf{F}\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})]'\boldsymbol{\Gamma}(\boldsymbol{\theta})^{-1}[\mathbf{X} - \mathbf{F}\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})],$$

and

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = [\mathbf{F}'\boldsymbol{\Gamma}(\boldsymbol{\theta})^{-1}\mathbf{F}]^{-1}\mathbf{F}'\boldsymbol{\Gamma}(\boldsymbol{\theta})^{-1}\mathbf{X}.$$

For some further details see Mardia (1984). For $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ known, the prediction problem is called co–kriging. If $\boldsymbol{\Lambda}$ is known and $\boldsymbol{\theta}$ is a scalar, then the profile is univariate and can be easily plotted. For a multivariate CAR model, see Mardia (1988).

### 7.2. Regularized process

Let $A_i$ denote the $i$-th plot and let $|A_i|$ be its area. We assume that univariate observations $X^{\star}_{A_i}$ are taken on these blocks. Then for the model

$$X(\mathbf{t}) = \mathbf{f}'(\mathbf{t})\boldsymbol{\beta} + \varepsilon(\mathbf{t}),$$

we have

$$E(X^{\star}_{A_i}) = \frac{1}{|A_i|}\left[\int_{A_i} \mathbf{f}'(\mathbf{t})d\mathbf{t}\right]\boldsymbol{\beta}, \quad \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \frac{1}{(|A_i||A_j|)}\int_{A_i}\int_{A_j} \sigma(\mathbf{s}, \mathbf{t}; \boldsymbol{\theta})d\mathbf{s}\, d\mathbf{t},$$

and $\sigma(.,\ .,\ \boldsymbol{\theta})$ is the covariance function of $\boldsymbol{\theta}$.

Usually $A_i = A$. The estimator proceeds as before, but in practice one uses $X^{\star}_{A_i} = X(\mathbf{t}_i)$ if $\mathbf{t}_i$ is the centre of $A_i$, especially when the $A_i$s are small. Thus we are back to the usual model.

### 7.3. Applications to design

Let

$$\mathbf{X} = \mathbf{F}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{F}$ is a design matrix and $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}$. Consider the $n$–by–$r$ design matrix $\mathbf{F}$ for treatments, with

$$(\mathbf{F})_{ij} = 1 \text{ if the j-th treatment is applied to the i-th plot;}$$
$$= 0 \text{ otherwise,}$$

$i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, r$, so that $\mathbf{F}'\mathbf{F} = r\mathbf{I}$. Consider the basic CAR model

$$\boldsymbol{\Sigma}^{-1} = \mathbf{I} - \theta\mathbf{N},$$

where $(\mathbf{N})_{ij} = 1$ if $i$ and $j$ are neighbours; $= 0$ otherwise, and $\theta$ is some unknown parameter.

We can use results reported in Section 2 to estimate $\theta$ and $\boldsymbol{\beta}$. Note that from (5.1.4)

$$\hat{\boldsymbol{\beta}} = (1/r)[\mathbf{F}'\mathbf{X} - \theta\mathbf{F}'\mathbf{N}(\mathbf{X} - \mathbf{F}\hat{\boldsymbol{\beta}})].$$

An alternative method is to use a Papadakis estimator of regression parameters $\boldsymbol{\beta}$ defined by (see Martin, 1982)

$$\hat{\boldsymbol{\beta}}_P = (\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'(\mathbf{X} - \hat{\beta}\mathbf{Y}),$$

where here $\hat{\boldsymbol{\beta}} = \mathbf{Y}'\mathbf{Q}\mathbf{X}/\mathbf{Y}'\mathbf{Q}\mathbf{Y}$, with $\mathbf{Y} = c\mathbf{N}\mathbf{Q}\mathbf{X}$, $c$ = a constant depending upon the layout (*e. g.*, circle, torus, or as such), and

$$\mathbf{Q} = \mathbf{I} - \mathbf{F}'(\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}.$$

For the circle, $c = 1/2$, and $\mathbf{N} = \mathbf{W} + \mathbf{W}^{-1}$, where

$$\mathbf{W} = \begin{cases} 1 & \text{along the leading upper diagonal;} \\ 1 & \text{in the lower left corner; and} \\ 0 & \text{otherwise.} \end{cases}$$

### 7.4. Missing values in lattice data

Let us assume that some data on a lattice are missing, but suppose that $\boldsymbol{\Sigma}^{-1}$ is known explicitly (*e. g.*, for a CAR model). Then one strategy is to estimate the missing values by ML prediction (with the auto–regression coefficient matrix explicitly evaluated or an approximation used), and obtain $\hat{\beta}$ and $\hat{\theta}$ on the lattice (see Martin, 1984). This procedure should not make any difference in the estimates of $\hat{\beta}$ and $\hat{\theta}$.

Also the loss of information through missing values on $\beta$ and $\theta$ can be computed as

$$I_{\beta,\theta} \text{ whole data} - I_{\beta,\theta} \text{ observed data.}$$

There is no difficulty in interpreting this measure for a single parameter; but, for higher dimensions, a determinant must be evaluated to measure the loss. Such important cases are studied in Haining *et al.* (1989).

A simple example of this situation is when the density is (5.6.5). We can maximize the density with respect to the missing values. For example, if two neighbouring values $x_i$ and $x_j$ are missing, with the first–order neighbourhood in 2–dimensions, then the respective estimates of $x_i$ and $x_j$ are simply

$$(4\tilde{x}_i + \tilde{x}_j)/5 \text{ and } (\tilde{x}_i + 4\tilde{x}_j)/5,$$

respectively, where now $\tilde{x}_i$ is the mean of the three observed neighbours of the $i$-th site.

## 8. Discussion

We have described mainly the ML method of estimation for the spatial linear model in two forms, DR and CAR. For the CAR model, there are some interesting features. For the T–CAR, the MLEs of $\boldsymbol{\theta}$ are obtained through the moment estimator of the covariance function. Another key feature is that the matrix of the eigenvectors for the covariance matrix does not involve $\boldsymbol{\theta}$. The same comment applies to the first–order C–CAR, even when the nugget parameter is present. However, the M–CAR is the realistic model, and the MLE from the approximate ML equations, $\hat{\sigma}_h = C_k(h)$ of Section 5.5, have better properties.

We have not examined alternative estimators, such as the minimum quadratic unbiased estimators for $\theta$ (see Kitanidis, 1985; Marshall and Mardia, 1985; Stein, 1986), but these estimates are closely related to the MLEs for the intrinsic model (see Stein, 1987). Also, we have not examined any exploratory techniques for spatial data (see, for example, Cressie, 1986).

Consequently, great care should be taken in formulating a model, especially since there can be some identifiability problems. For example, when the stationary model versus a trend is used, the stationary model will tend to estimate the long–term correlations if there is a trend. Plotting the data, empirical semi–variogram surface or semi–variogram in different directions, periodogram, and so forth, can be revealing. Further, the computation of the ML estimates requires care due to possible local maxima, and whenever possible a profile likelihood should be examined.

## 9. References


Besag, J. E. (1974) Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society*, **36B**: 192-236.

Besag, J. E. (1975) Statistical analysis of non–lattice data. *Statistician*, **24**: pp.79-195.

Besag, J. E. (1977a) Efficiency of pseudo–likelihood estimation for sample Gaussian fields. *Biometrika*, **64**: 616-8.

Besag, J. E. (1977b) Errors in–variables estimation for Gaussian lattice scheme. *Journal of the Royal Statistical Society*, **39B**: 73-78.

Besag, J. E. (1978) Discussion to Nearest neighbour models in the analysis of field experiments. *Journal of the Royal Statistical Society*, **40B**: 165-166.

Besag, J. E. (1981) On a system of two–dimensional recurrence equations. *Journal of the Royal Statistical Society*, **43B**: 302-309.

Besag, J. E. (1989) Towards Bayesian image analysis, to appear.

Besag, J. E. and P. A. P. Moran. (1975) On the estimation and testing of spatial interaction in Gaussian lattice processes. *Biometrika*, **62**: 555-562

Besag, J. E. and R. Kempton. (1986) Statistical analysis of field experiments using neighbouring plots. *Biometrics*, **42**: 321-351.

Brewer, A. C. and R. Mead. (1986) Continuous second order models of spatial variation with application to the efficiency of crop experiments (with discussion). *Journal of the Royal Statistical Society*, **149A**: 314-348.

Chant, D. (1974) On asymptotic tests of composite hypotheses in nonstandard conditions. *Biometrika*, **61**: 291-298.

Cliff, A. D. and J. K. Ord. (1981) *Spatial Processes*. London: Pion.

Cressie, N. (1986) Kriging nonstationary data. *Journal of the American Statistical Association*, **81**: 625-634.

Dahlhaus, R., and H. R. Künsch. (1987) Edge effects and efficient parameter estimation for stationary random fields. *Biometrika*, **74**: 877-82.

Davis, J. C. (1973) *Statistics and Data Analysis in Geology*. New York: Wiley.

Delfiner, P. (1976) Linear estimation of non–stationary spatial phenomena, in *Advances in Geostatistics in the Mining Industry*, edited by M. Guarascio, M. David, and C. Huijbregts, pp. 49-68. Dordrecht: Reidel.

Griffith, D. A. (1988) *Advanced Spatial Statistics*. Dordrecht: Kluwer.

Guyon, X. (1982) Parametric estimation for a stationary process on a d–dimensional lattice. *Biometrika*, **69**: 95-105.

Haining, R., D. A. Griffith, and R. Bennett. (1989) Maximum likelihood estimation with missing spatial data and with an application to remotely sensed data. *Communications in Statistics*, **18**: 1875-1894.

Harville, D. A. (1977) Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, **72**: 320-340.

Hocking, R. R. (1985) *The Analysis of Linear Models*. New York: Brooks/Cole.

Kalbfleisch, D. and D. A. Sprott. (1970) Applications of likelihood methods to models in-

volving large numbers of parameters. *Journal of the Royal Statistical Society*, **32B**: 175-194.

Kent, J. T. and K. V. Mardia. (1988) Spatial classification using fuzzy membership models. *Transactions of the IEEE/PAMI*, **10**: 659-671.

Kitanidis, P. K. (1983) Statistical estimation of polynomial generalized covariance functions and hydrologic applications. *Water Resources Research*, **19**: 909-921.

Kitanidis, P. K. (1985) Minimum–variance unbiased quadratic estimation of covariances of regionalised variables. *Mathematical Geology*, **17**: 195-208.

Kitanidis, P. K. (1987) Parametric estimation of covariance of regionalised variables. *Water Resources Bulletin of the American Water Resources Association*, **23**: 557-567.

Kitanidis, P. K. and R. W. Lane. (1985) Maximum likelihood parameter estimation of hydrologic spatial processes by the Gauss–Newton methods. *Journal of Hydrology*, **79**: 53-71.

Künsch, H. R. (1981) Thermodynamics and statistical analysis of Gaussian random fields. *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **98**: 407-421.

Künsch, H. R. (1983) Approximations to the maximum likelihood equations for some Gaussian random fields. *Scandinavian Journal of Statistics*, **10**: 239-246.

Künsch, H. R. (1987) Intrinsic autoregressions and related models on the two–dimensional lattice. *Biometrika*, **74**: 517-24.

Mardia, K. V. (1980) Some statistical inference problems, in Kriging II: Theory, Proceedings of the 26th International Geology Congress, pp. 113-131. Sciences de la Terre: "Advances in Automatic Processing and Mathematical Models in Geology," Series "Informatique Geologie," # 15.

Mardia, K. V. (1984) Spatial discrimination and classification maps. *Communications in Statistics*, **13**: 2181-2197.

Mardia, K. V. (1986) Discussion to the paper by Brewer and Mead. *Journal of the Royal Statistical Society*, **149A**: 341-342.

Mardia, K. V. (1988) Multi–dimensional multivariate Gaussian Markov random fields with application to image processing. *Journal of Multivariate Analysis*, **24**: 265-284.

Mardia, K. V. (1989) Markov models and Bayesian methods in image analysis. *Journal of Applied Statistics*, **16**: 125-130.

Mardia, K. V., J. T. Kent, and J. M. Bibby. (1989) *Multivariate Analysis*, 7th printing with corrections. New York: Academic Press.

Mardia, K. V. and R. J. Marshall. (1984) Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, **71**,: 135-46.

Mardia, K. V. and A. J. Watkins. (1989) On multimodality of the likelihood for the spatial linear model. *Biometrika*, **76**: 289-295.

Marechal, A. and J. Serra. (1970) Random kriging, in Geostatistics: A Colloquium, edited by D. Merriam, pp. 91-112. New York: Plenum Press.

Marshall, R. J. and K. V. Mardia. (1985) Minimum norm quadratic estimation of components of spatial covariance. *Mathematical Geology*, **17**: 517-525.

Martin, R. J. (1982) Some aspects of experimental design and analysis when errors are

correlated. *Biometrika*, **69**: 597-612.

Martin, R. J. (1984) Exact maximum likelihood for incomplete data from a correlated Gaussian process. *Communication in Statistics*, **13**: 1275-1288.

Martin, R. J. (1987) Some comments on correction techniques for boundary effects and missing value techniques. *Geographical Analysis* **19**: 273-282.

Matheron, G. (1971) The Theory of Regionalized Variables and Its Applications. Fountainbleau: Les Cahiers du Morphologie Mathematique, Fasc. No. 5.

McBratney, A. B. and R. Webster. (1981) Detection of ridge and furrow pattern by spectral analysis of crop yield. *International Statistical Review*, **49**: 45-52.

Mercer, W. B. and A. D. Hall. (1911) The experimental error of field trials. *Journal of Agricultural Science*, **4**: 107-132.

Moran, P. A. P. (1971) Maximum likelihood estimation in nonstandard conditions. *Proceedings of the Cambridge Philosophical Society*, **70**: 441-450.

Patterson, H. D. and R. Thompson. (1974) Maximum likelihood estimation of components of variance, in Proceedings of the 8th International Biometrics Conference, edited by L. Corsten and T. Postelnicu, pp. 197-208. Bucharest: Academy of the Socialist Republic of Rumania.

Plackett, R. L. (1960) *Principles of Regression Analysis*. Oxford: Clarendon Press.

Ripley, B. D. (1988) *Statistical Inference for Spatial Processes*. Cambridge: Cambridge University Press.

Stein, M. L. (1986) A modification of the minimum norm quadratic estimation of a generalised covariance function for use with large data sets. *Mathematical Geology*, **18**: 625-633.

Stein, M. L. (1987) Minimum norm quadratic estimation of spatial variograms. *Journal of the American Statistical Association*, **82**: 765-772.

Stein, M. L. (1988) Asymptotically efficient prediction of a random field with a misspecified covariance function. *Annals of Statistics*, **16**: 55-63.

Tunnicliffe–Wilson, G. (1989) On the use of marginal likelihood in time series estimation. *Journal of the Royal Statistical Society*, **51B**: 15-27.

Vecchia, A. V. (1988) Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society*, **50B**: 297-312.

Warnes, J. J. and B. D. Ripley. (1987) Problems with likelihood estimation of covariance functions of spatial Gaussian processes. *Biometrika*, **74**: 640-42.

Watkins, A. J. and K. V. Mardia. (1989) Some problems in spatial inference, to appear.

Whittle, D. (1954) On stationary processes in the plane. *Biometrika*, **41**: 434-449.

# DISCUSSION

## "Maximum likelihood estimators for spatial models"

## by Kanti V. Mardia

As an intuitively suggested extension of one–dimensional time series analysis, spatial analysis in two dimensions calls itself to attention; however, as is often the case, the introduction of an additional dimension gives rise to a series of new problems.

The first is the operational *specification* of a computable model; as is the case with spectral analysis, the presence of a trend should be explicitly recognized and taken care of. Various possibilities are suggested by Mardia: direct representation, conditional autoregression, and simultaneous autoregression. As is usual in spatial statistics (and spatial econometrics should take this more often into account), boundary problems are taken up, which can be done in various manners (torus–solution, free boundaries).

The next problem is that of *estimating* the parameters of one of the above–mentioned specifications. Maximum likelihood is a good candidate (for spatial econometrics one is referred to Paelinck and Klaassen (1979, Ch. 3) but beset with a certain number of difficulties. Is there a single global maximum (hence the suggestions of plotting the profile likelihood)? How does one compute it in the many parameters case? Furthermore, do the resulting estimators have acceptable asymptotic properties, if small sample imperfections (*e. g.*, bias) are present, as is often the case when so–called Whittle–type approximations are used? The above remarks also are applicable to intrinsic processes. A useful addition to specification and estimation is the study of errors in the variables (measurement errors), which all too often are present in empirical exercises.

Finally *testing* is derived from asymptotic normality.

The examples given by Mardia show the operational character of the methods developed and advocated; they put into light one of the difficulties besetting spatial data, to wit anisotropy (together with the trend already mentioned above). The analysis of topographic and landsat data is revealing from those points of view.

In his conclusion the author rightly draws the reader's attention to a number of key points, which may be summarized as follows:

1. the possible presence of identifiability problems (neglecting trend terms can lead to erroneously estimating long–range correlations);
2. the necessity to explicitly include anisotropy (this matches a problem in spatial econometrics, that of asymmetry of the relations postulated); and,
3. the need to tackle second–order conditions, or at least to graphically inspect the likelihood function.

### References

Paelinck, J., and L. Klaassen. (1979) *Spatial Econometrics*. Farnborough: Saxon House.

*J. H. P. Paelinck*, Erasmus University

## *PREAMBLE*

*True eloquence consists in saying all that is necessary,*
*and nothing but what is necessary.*

La Rouchefoucauld

*Any scholar who, like myself, has experienced pure mathematics will recognize in this paper the illumination of formal mathematics that so often is missing in applied statistics–definitions, theorems, proofs, lemmas, and analytics. If you have never encountered these constructions before in an advanced mathematics context, then you may find this paper difficult reading; but, it certainly is eloquent! The purpose of this paper is twofold, namely to derive (1) those formulae needed to compute the exact distributions of the Moran and Geary spatial autocorrelation indices under an assumption of normality, and (2) expressions for their asymptotic mean and variance under an assumption of non-normality. Mardia's reaction to the formalism employed here is that the results are theoretically impressive, but not yet helpful to practitioners. Perhaps Sen has been a bit too concise. On the other hand, Mardia suggests that this paper should generate further research of an applied nature.*

The Editor

# Distribution of Spatial Correlation Statistics

Ashish Sen*

*School of Urban Planning and Policy and the Urban Transportation Center, University of Illinois at Chicago, Chicago, IL 60680, U. S. A.*

**Overview:** This paper contains a number of results on the distribution of Moran and Geary statistics. Two key ones are (1) formulæ to compute the exact distributions of these statistics when observations are normally distributed, and (2) expressions for their asymptotic mean and variance, when the observations are not normal. The results are sufficiently general that they may be applied to a wide range of situations. However, in order to be somewhat specific, the presentation assumes that the spatial correlation statistics are being applied to linear least squares residuals.

## 1. Introduction

Two statistics that may be applied on regression residuals $e_1, \cdots, e_n$ to test for the existence of spatial correlation are of the form

$$c \sum_{ij} w_{ij} e_i e_j / s^2 \tag{1}$$

and

$$c \sum w_{ij} (e_i - e_j)^2 / s^2 \tag{2}$$
$$ij$$

where $c$ is a suitable constant, $s^2 = (n-k-1)^{-1} \sum_{i=1}^{n} e_i^2$ is the usual unbiased estimate of the variance of the regression error term when there are $k$ independent variables, and $w_{ij}$ is some measure of inverse distance. For example, we could have $w_{ij} = 1$ when the $i$th and $j$th observations are from contiguous zones and $w_{ij} = 0$ otherwise; or $w_{ij} = d_{ij}^{-2}$ where $d_{ij}$ is the distance between the locations where observations $i$ and $j$ were taken. The expressions (1) and (2) are obvious generalizations of statistics previously given by Moran and Geary, respectively (see Cliff and Ord, 1981).

In Section 3 we present exact distributions for these statistics under the hypothesis of no spatial correlation and under the assumption of normality of regression errors. While the formulæ are somewhat complicated and depend on the matrix of $w_{ij}$'s, they can be programmed so that for any given situation, relevant portions of tables can be obtained from a computer. They also can be used to obtain exact tail probabilities to help identify suitable approximate methods for obtaining such probabilities. In this context, it should be mentioned that computations using these exact formulæ are certainly less time consuming than Monte Carlo methods.

If the errors are not normal, the only recourse available is to invoke large sample theory and use the fact that under certain mild conditions (1) and (2) are asymptotically normal.

* I would like to thank Prof. Tony Smith, Regional Science Department, University of Pennsylvania, Prof. Muni Srivastava, Statistics Department, University of Toronto, and the discussant for many valuable comments and suggestions on an earlier draft of this paper. I would also like to express my gratitude to Ms. Marilyn Engwall for the diagrams. (School of Urban Planning and Policy and the Urban Transportation Center, University of Illinois at Chicago.)

However, we still need means and variances of the statistics, and the asymptotic expressions for these (in, say Sen, 1976) are a bit too crude. In Section 4 we present formulæ for such means and variances. It may be mentioned in passing that when the observations are non-normal, and sample sizes small, the analyst may wish to use rank-test equivalents of (1) and (2) — see Sen and Soot (1977) — or obtain critical points of (1) and (2) by permutation methods (see Cliff and Ord, 1981).

The next section is devoted to some preliminaries and to notation. Although they are known, for the sake of completeness, we also obtain means and variances of the statistics under normality.

## 2. Preliminaries

Let $\Omega_1, \ldots, \Omega_n$ be $n$ regions and assume that for each $\Omega_i$ we have an observation $y_i$ on a (dependent) variable and values $x_{i1}, \ldots, x_{ik}$ of $k$ independent variables. Then a linear regression model is written in the form

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3}$$

where $\mathbf{y} = (y_1, \ldots, y_n)'$,

$$X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix}$$

$\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$ is the vector of the error terms, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)'$ is the vector of parameters and a prime denotes matrix transpose. The least squares estimate $\mathbf{b}$ of $\boldsymbol{\beta}$ is (assuming $X$ is non-singular)

$$(b_0, \ldots, b_k)' = \mathbf{b} = (X'X)^{-1}X'\mathbf{y} \tag{4}$$

and the residuals are, therefore,

$$(e_1, \ldots, e_n)' = \mathbf{e} = \mathbf{y} - X\mathbf{b} = \mathbf{y} - X(X'X)^{-1}X'\mathbf{y} = M\mathbf{y} \tag{5}$$

where $M = I - H$ and $H = X(X'X)^{-1}X'$. It follows that $\mathbf{e}$ also could have been written as $\mathbf{e} = MX\boldsymbol{\beta} + M\boldsymbol{\epsilon} = M\boldsymbol{\epsilon}$ and, moreover, it may be verified that $M$ and $H$ are idempotent (*i. e.*, $M^2 = M$ and $H^2 = H$).

In order to be assured that the linear combination $\boldsymbol{\ell}'\mathbf{b}$ of $b_i$'s is a best (*i. e.*, minimum variance) unbiased linear estimate of $\boldsymbol{\ell}'\boldsymbol{\beta}$, three conditions — called Gauss–Markov conditions — must be met. These are

$$E(\epsilon_i) = 0 \tag{6}$$

$$E(\epsilon_i^2) = \sigma^2 (\text{a constant}) \tag{7}$$

$$E(\epsilon_i \epsilon_j) = 0 \tag{8}$$

for all $i$ and $j$. In matrix notation, the Gauss–Markov conditions become

$$E(\boldsymbol{\epsilon}) = \mathbf{0}, \quad E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma^2 I.$$

If, in addition, $\boldsymbol{\epsilon}$ is normally distributed we write $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I)$, *i. e.*, $\boldsymbol{\epsilon}$ is normally distributed with mean $\mathbf{0}$ and covariance matrix $\sigma^2 I$. From (3) it follows that if $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I)$, then $\mathbf{y} \sim N(X\boldsymbol{\beta}, \sigma^2 I)$.

When (8) holds we shall also say (since the observations are taken over space) that the $\epsilon_i$'s are spatially uncorrelated. On the other hand, if (8) does not hold and particularly if $E(\epsilon_i \epsilon_j) = \rho_{ij}$, where $|\rho_{ij}|$ declines with the spatial separation between $\Omega_i$ and $\Omega_j$, we say the $\epsilon_i$'s are spatially correlated (see Cliff and Ord, 1981, for more on this subject). As is well known, the presence of spatial correlation does not bias the estimate $\mathbf{b}$, but the covariance matrix of $\mathbf{b}$ and any quantity that depends on it (the t, the F and, indeed, most statistics used for tests) are seriously affected.

The seriousness can be seen as follows. If $\boldsymbol{\epsilon}$ has covariance matrix $\sigma^2\Omega$, then the variance of $\boldsymbol{\ell}'\mathbf{b}$ is of the form $\sigma^2 \mathbf{c}'\Omega\mathbf{c}$ with $\mathbf{c}' = \boldsymbol{\ell}'(X'X)^{-1}X'$. This contains $n(n-1)$ terms involving non–diagonal elements of $\Omega$. Therefore, even if each such element is small, their combined effect *can be* considerable. Even worse is the fact that, when we use ordinary least squares, computer packages typically compute estimates of variance under the assumption that Gauss–Markov Conditions hold, *i. e.*, $\Omega = I$. Therefore, unaccounted for non–diagonal elements can substantially affect any inferences we reach.

As mentioned in Section 1, the statistics (1) and (2) may be used to test for the existence of spatial correlation. To simplify matters we shall write both (1) and (2) in the form

$$c\mathbf{e}'Z\mathbf{e}/s^2 \tag{9}$$

which is obviously appropriate since numerators of both are quadratic forms.

Since $\mathbf{e} = M\boldsymbol{\epsilon}$, it follows that $\mathbf{e}'Z\mathbf{e} = \boldsymbol{\epsilon}'M'ZM\boldsymbol{\epsilon} = \boldsymbol{\epsilon}'B\boldsymbol{\epsilon}$ where $B = M'ZM$. Since the matrix $B$ is symmetric there exists an orthogonal matrix $\Gamma$ such that $B = \Gamma D_\lambda \Gamma'$ where $D_\lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ and $\lambda_i$'s are the eigenvalues of $B$. Therefore, writing $\Gamma\boldsymbol{\epsilon} = \mathbf{u}$, we have

$$\mathbf{e}'Z\mathbf{e} = \boldsymbol{\epsilon}'B\boldsymbol{\epsilon} = \boldsymbol{\epsilon}'\Gamma D_\lambda \Gamma'\boldsymbol{\epsilon} = \mathbf{u}'D_\lambda\mathbf{u} = \sum_{i=1}^{n} \lambda_i u_i^2. \tag{10}$$

Now consider the denominator of (9):

$$s^2 = p^{-1}\mathbf{e}'\mathbf{e}$$

where $p = n - k - 1$. Since $\mathbf{e} = M\boldsymbol{\epsilon}$ and $M$ is idempotent we may write $s^2$ as

$$p^{-1}\boldsymbol{\epsilon}'M'M\boldsymbol{\epsilon} = p^{-1}\boldsymbol{\epsilon}'M\boldsymbol{\epsilon}.$$

Obviously $BM = MB$ and hence the same matrix $\Gamma$ that diagonalized $B$ also diagonalizes $M$ (see Bellman, 1960, p. 56). Hence

$$B = \Gamma D_\lambda \Gamma', \text{ and } M = \Gamma D_\xi \Gamma' \tag{11}$$

where $D_\xi = \text{diag}(\xi_1, \ldots, \xi_n)$. Since $M$ is idempotent, its eigenvalues $\xi_i$ are either one or zero and since the rank of $M$ is $p = n - k - 1$, exactly $p$ of the eigenvalues are ones. Therefore, we may write

$$c\mathbf{e}'Z\mathbf{e}/s^2 = c\boldsymbol{\epsilon}'B\boldsymbol{\epsilon}/p^{-1}\boldsymbol{\epsilon}'M\boldsymbol{\epsilon} = c\sum_{i=1}^{n}\lambda_i u_i^2 / p^{-1}\sum_{i=1}^{p} u_i^2 = cP, \tag{12}$$

where $P = U/V$, $U = \sum_{i=1}^{n} \lambda_i u_i^2$ and $V = p^{-1} \sum_{i=1}^{p} u_i^2$.

It is a property of normal distributions (see Rao, 1973; Srivastava and Khatri, 1979) that if $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I)$, then $\mathbf{u} \sim N(0, \sigma^2 I)$. From (11), and the fact that the $\xi_i$'s are either one or zero (and therefore $D_\xi^2 = D_\xi$), we get

$$\boldsymbol{\epsilon}' M \boldsymbol{\epsilon} = \boldsymbol{\epsilon}' \Gamma D_\xi \Gamma' \boldsymbol{\epsilon} = \mathbf{u}' D_\xi \mathbf{u} = \mathbf{u}' D_\xi' D_\xi \mathbf{u}.$$

Also, because $D_\xi$ is diagonal with elements either 0 or 1, $p$ of the components of $D_\xi \mathbf{u} = \mathbf{u}^\star$ are identical to those in $\mathbf{u}$ and the remainder are zeros. Now

$$\begin{aligned} \boldsymbol{\epsilon}' B \boldsymbol{\epsilon} = \boldsymbol{\epsilon}' M Z M \boldsymbol{\epsilon} &= \boldsymbol{\epsilon}' \Gamma D_\xi \Gamma' Z \Gamma D_\xi \Gamma' \boldsymbol{\epsilon} = \mathbf{u}' D_\xi \Gamma' Z \Gamma D_\xi \mathbf{u} \\ &= \mathbf{u}^\star (\Gamma' Z \Gamma) \mathbf{u}^\star = (\mathbf{u}^\star)' Z^\star \mathbf{u}^\star \end{aligned}$$

where $Z^\star = \Gamma' Z \Gamma$, showing that the same $p$ $u_i$'s are in the numerator of (12) as are in its denominator. Because of this and because $\mathbf{u}$ is normal, it follows that $P$ and $V$ are independent (Cliff and Ord, 1981, p.43, Theorem 1, Pitman, 1937; the original result is due to R. A. Fisher). An important consequence of this is that

$$E(P^s) E(V^s) = E(U^s) \tag{13}$$

**Theorem 2.1.**

When $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I)$, *i. e.*, when $\mathbf{y} \sim N(X\beta, \sigma^2 I)$, the mean and variance of (9) are

$$E(\mathbf{e}' Z \mathbf{e} / s^2) = \mathrm{tr}[B] \tag{14}$$

and

$$\mathrm{var}(\mathbf{e}' Z \mathbf{e} / s^2) = 2(n-k+1)^{-1} \{(n-k-1)\mathrm{tr}[B^2] - (\mathrm{tr}[B])^2\} \tag{15}$$

where var( $\cdot$ ) stands for the 'the variance of', and $B = M'ZM$.

As mentioned earlier, this result is known, e.g., see Ripley (1981, p. 100) or Brandsma and Ketellapper (1979).

Proof of Theorem 2.1:

From (10)

$$E(\mathbf{e}' Z \mathbf{e}) = E(\sum_{i=1}^{n} \lambda_i u_i^2) = \sigma^2 \sum_{i=1}^{n} \lambda_i = \sigma^2 \mathrm{tr}[D_\lambda] = \sigma^2 \mathrm{tr}[B] \tag{16}$$

since $\mathrm{tr}[B] = \mathrm{tr}[\Gamma' D_\lambda \Gamma] = \mathrm{tr}[D_\lambda \Gamma \Gamma'] = \mathrm{tr}[D_\lambda]$. Also, since $u_i$'s are independent normal with mean 0 and variance $\sigma^2$,

$$\begin{aligned} E(\mathbf{e}' Z \mathbf{e})^2 &= E(\sum_{i=1}^{n} \lambda_i u_i^2)^2 \\ &= E[\sum_{i=1}^{n} \lambda_i^2 u_i^4 + \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \lambda_i \lambda_j u_i^2 u_j^2] \\ &= \mu_4 \sum_{i=1}^{n} \lambda_i^2 + \sigma^4 \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \lambda_i \lambda_j \end{aligned} \tag{17}$$

where $\mu_4 = E(u_i^4) = 3\sigma^4$. Hence

$$\begin{aligned}\text{var}(\mathbf{e}'Z\mathbf{e}) &= E(\mathbf{e}'Z\mathbf{e})^2 - [E(\mathbf{e}'Z\mathbf{e})]^2 \\ &= \mu_4 \sum_{i=1}^{n} \lambda_i^2 + \sigma^4 \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \lambda_i \lambda_j - \sigma^4 (\sum_{i=1}^{n} \lambda_i)^2,\end{aligned}$$

and since

$$(\sum_{i=1}^{n} \lambda_i)^2 = \sum_{i=1}^{n} \lambda_i^2 + \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \lambda_i \lambda_i$$

it follows that

$$\text{var}(\mathbf{e}'Z\mathbf{e}) = (\mu_4 - \sigma^4) \sum_{i=1}^{n} \lambda_i^2 = 2\sigma^4 \sum_{i=1}^{n} \lambda_i^2 = 2\sigma^4 \text{tr}[B^2]. \tag{18}$$

From (18) and (16), we have, alternatively

$$E(\mathbf{e}'Z\mathbf{e})^2 = 2\sigma^4 \text{tr}[B^2] + \sigma^4 \text{tr}[B])^2. \tag{19}$$

Now replacing $B$ by $M$ in (15) and (19) we get

$$E(s^2) = p^{-1} E(\boldsymbol{\epsilon}' M \boldsymbol{\epsilon}) = p^{-1} \sigma^2 \text{tr}[M] = \sigma^2 \tag{20}$$

and

$$E(s^2)^2 = p^{-2} E(\boldsymbol{\epsilon}' M \boldsymbol{\epsilon})^2 = p^{-2} \{2\sigma^4 \text{tr}[M] + \sigma^4 (\text{tr}[M])^2\} = \sigma^4 (2p^{-1} + 1). \tag{21}$$

It is now simple to verify that (14) follows from (13), (16) and (20). To verify (15), notice that from (13), (19) and (21)

$$E(P^2) = (2\text{tr}[B^2] + (\text{tr}[B])^2)/(1 + 2p^{-1}).$$

Therefore, from (14)

$$\begin{aligned}\text{var}(P) &= E(P^2) - (E(P))^2 \\ &= 2(1 + 2p^{-1})^{-1} \text{tr}[B^2] + (\text{tr}[B])^2 ((1 + 2p^{-1})^{-1} - 1) \\ &= 2(p + 2)^{-1} \{p\text{tr}[B^2] - (\text{tr}[B])^2\}.\end{aligned}$$

This proves the theorem.

## 3. Exact Distribution of $P$ under Normality

Since $P$ is a ratio of quadratic forms of normal variables, there are several methods available for computing its distribution function (cdf) under the hypothesis of no spatial correlation, although perhaps not as many as one would expect. In this section we describe versions of the two key ones. Sections 3.1 and 3.2 deal with the case where the $\lambda_i$'s are distinct, while Section 3.3 is concerned with the $\lambda_i$'s having common values. Section 3.1 is devoted to the proof of a theorem on the distribution of $U$, which then is used in Section 3.2 to provide formulæ for the cdf of $P$. Notice that without loss of generality we can set $\sigma = 1$ and assume that $\mathbf{u} \sim N(0, I)$.

### 3.1. Distribution of $U$ when $\lambda_i$'s are distinct

**Theorem 3.1**

Without loss of generality we can ignore zero valued $\lambda_i$'s. Let $\mu_1 > \ldots > \mu_{n_1}$ be the negative valued $\lambda_i$'s (if any) and $\nu_1 < \ldots < \nu_{n_2}$ be the positive valued $\lambda$'s with $n = n_1 + n_2$. Define

$$\int_{(+)} = \sum_{k=1}^{\lfloor (n_2+1)/2 \rfloor} (-1)^{k+1} \int_{\nu_{2k-1}^{-1}}^{\nu_{2k}^{-1}}$$

and

$$\int_{(-)} = \sum_{k=1}^{\lfloor (n_1+1)/2 \rfloor} (-1)^{k+1} \int_{\mu_{2k-1}^{-1}}^{\mu_{2k}^{-1}}$$

where $\lfloor \psi \rfloor$ is the integer part of $\psi$, $\mu_{n_1+1} = -\infty$ and $\nu_{n_2+1} = \infty$. Further define

$$D(\lambda) = \left[ - \prod_{1=1}^{n} (1 - \lambda\lambda_i) \right]^{-1/2}.$$

Then if the $\mathbf{u} \sim N(0, I)$, the cdf $F(z)$ of $U$ is

$$F(z) = \begin{cases} 1 - \pi^{-1} \int_{(+)} \lambda^{-1} D(\lambda) \exp(-\lambda z/2)\, d\lambda & \text{for } z \geq 0 \\ \pi^{-1} \int_{(-)} \lambda^{-1} D(\lambda) \exp(-\lambda z/2)\, d\lambda & \text{for } z < 0. \end{cases} \tag{22}$$

Notice that $\int_{(+)}$ and $\int(-)$ depend on $n_1$, $n_2$ and the $\lambda_i$'s. However, no confusion need occur if we note that whenever these symbols arise, we simply act as if they were exactly equivalent to their definitions above. This theorem is similar to one in Smirnov (1937) and also may be obtained from Plackett (1960, pp. 20-22). However, Smirnov made a mistake in signs which carried over into several subsequent papers (some of which did not refer to Smirnov!). The mistake is pointed out in the proof of the theorem which is given after the following lemma.

**Lemma 3.1**

The absolute value of each of the integrals that comprise either $\int_{(+)} \lambda^{-1} D(\lambda)\, d\lambda$ or $\int(-)\lambda^{-1} D(\lambda)\, d\lambda$ is less than a constant $M > 0$.

Proof:

Let $a$ be a number between $\nu_{2k-1}^{-1}$ and $\nu_{2k}^{-1}$. For $\nu_{2k-1}^{-1} \leq \lambda \leq a$,

$$|\lambda^{-1}| \prod_{\substack{i=1 \\ i \neq 2k-1}}^{n} |1 - \lambda\lambda_i|^{-1/2} < M_1$$

for some $M_1 > 0$. Hence

$$\int_{\nu_{2k-1}^{-1}}^{a} \lambda^{-1} D(\lambda)\, d\lambda < M_1 \int_{\nu_{2k-1}^{-1}}^{a} |1 - \lambda_{2k-1}\lambda|^{-1/2}\, d\lambda$$
$$= 2M_1 \lambda_{2k-1}^{-1/2} (a - \lambda_{2k-1}^{-1})^{1/2} < M_2(\text{say}).$$

Similarly it can be shown that if $2k \leq n_2$, $\int_a^{\nu_{2k}^{-1}} \lambda^{-1} D(\lambda) d\lambda$ is finite. When $2k = n_2 + 1$, *i. e.*, when $\nu_{2k} = \infty$, let $a - \nu_{n_2}^{-1} = c$ and

$$\int_a^{\infty} \lambda^{-1} D(\lambda) d\lambda < \left( \prod_{i=1}^{n} \lambda_i^{-1/2} \right) \int_c^{\infty} x^{-(n/2+1)} dx$$

which also is finite. Similarly, each of the components of $\int_{(-)} \lambda^{-1} D(\lambda) d\lambda$ may be shown to be less than some number $M$.

Proof of Theorem 3.1

Since each $u_i \sim N(0,1)$, it follows that $u_i^2$ has a chi–square distribution with 1 degree of freedom. Hence (Rao, 1973, p.167), its characteristic function is $(1 - 2\imath t)^{-1/2}$ where $\imath = \sqrt{-1}$. Therefore, the characteristic function of $\lambda_i u_i^2$ is $(1 - 2\imath t\lambda_i)^{-1/2}$ (Rao, 1973, p. 100). Since the $u_i$'s are independent, the characteristic function of $U = \sum_{i=1}^{n} \lambda_i u_i^2$ is (Rao, 1973, p. 104)

$$\Phi_U(t) = \prod_{i=1}^{n} (1 - 2\imath t\lambda_i)^{-1/2}. \tag{23}$$

The distribution function $F(z)$ of $U$ can be found from (23). From the general inversion theorem for characteristic functions (Wilks, 1962, p. 252), we have

$$\begin{aligned} F(z) - F(0) &= (2\pi)^{-1} \int_{-\infty}^{\infty} \Phi_U(t)[1 - \exp(\imath tz)](\imath t)^{-1} \, dt \\ &= (2\pi)^{-1} \int_{-\imath\infty}^{\imath\infty} \lambda^{-1} D(\lambda)[1 - \exp(-\lambda z/2)] \, d\lambda \end{aligned} \tag{24}$$

on making the substitution $\lambda = 2\imath t$. Hence for two different values $z$ and $z'$,

$$F(z) - F(z') = (2\pi)^{-1} \int_{-\imath\infty}^{\imath\infty} \lambda^{-1} D(\lambda)[\exp(-\lambda z'/2) - \exp(-\lambda z/2)] \, d\lambda. \tag{25}$$

Consider first the case where $z$ and $z'$ are non–negative. Call the integrand in (25) $g$ and consider the non–negative half plane of $\lambda$, *i. e.*, the half plane containing complex values of $\lambda$ with non–negative real parts. On this half plane $[D(\lambda)]^{-1}$ can take zero values only on the positive real half axis. Therefore, $g$ is analytic over the entire half plane from which the slits $S_k$ (see Exhibits 1.a and 1.b) have been removed and by Cauchy's Theorem (see, for example, Nehari, 1961 or Hille, 1959),

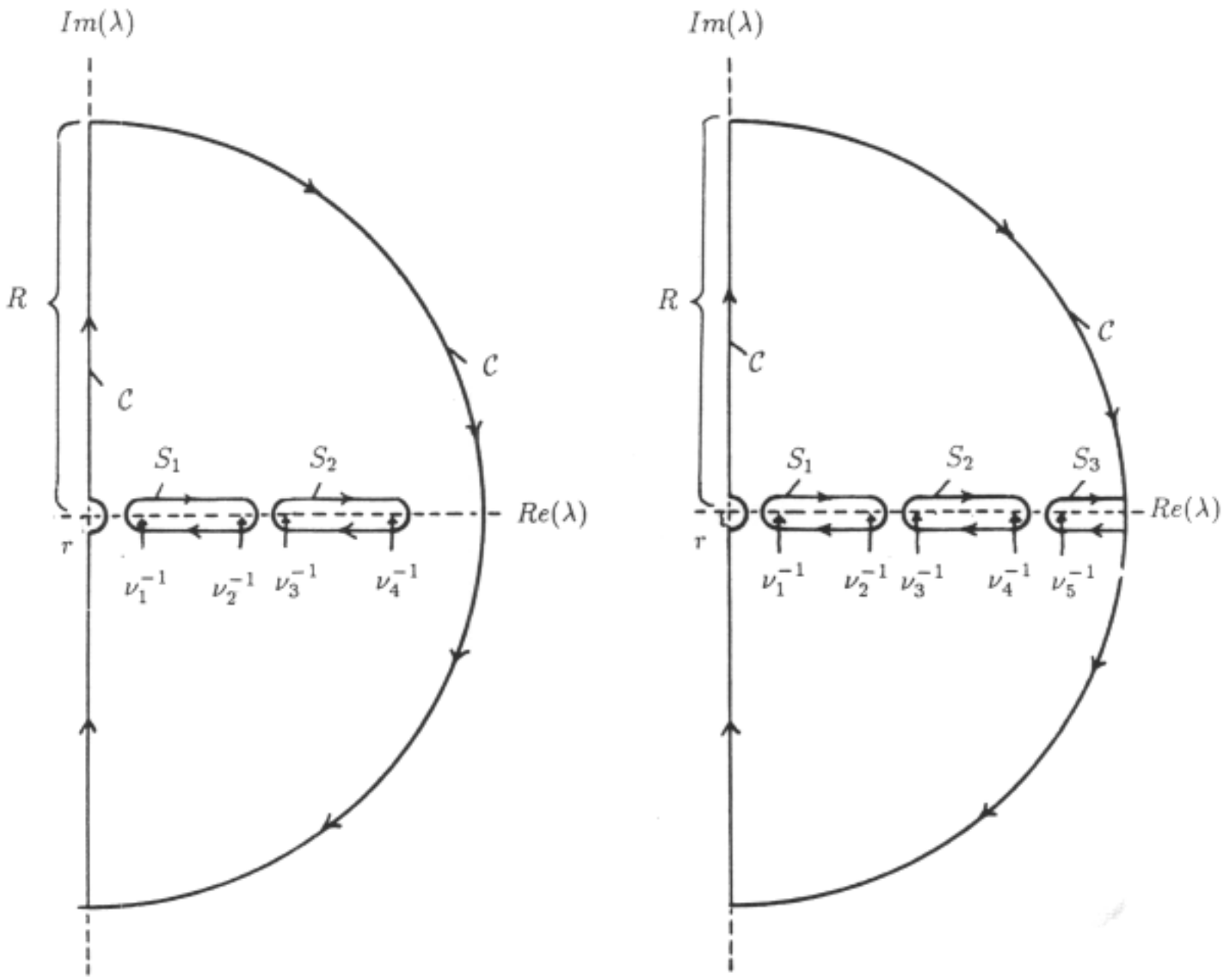
$$\int_C g + \sum_k \int_{\partial S_k} g = 0 \tag{26}$$

where the integral $\int_C$ is taken over the contour $C$ illustrated in Exhibits 1.a and 1.b, and $\int_{\partial S_k}$ is the integral over the boundary $\partial S_k$ of the slit $S_k$. When $R \to \infty$, the integral of $g$ goes to zero over the part of $C$ that is semicircular with radius $R$. Hence,

$$\lim_{R \to \infty} \int_C g = \int_{-\imath\infty}^{\imath\infty} g.$$

**Exhibit 1. a, left; b, right.**

Contour $C$ and slits $S_k$.




Therefore, from (26),

$$\int_{-i\infty}^{i\infty} g = -\sum_k \int_{\partial S_k} g. \tag{27}$$

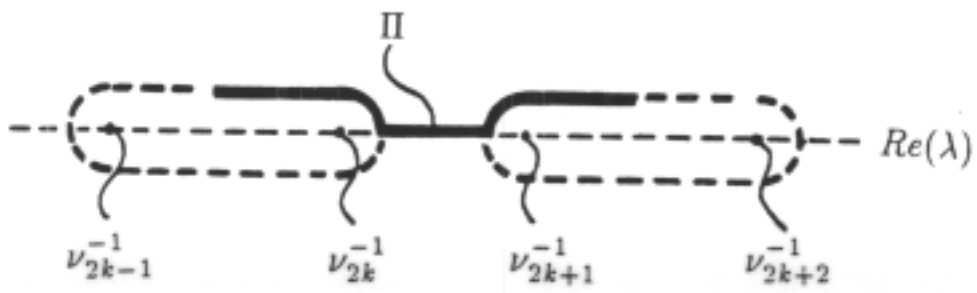
For any branch of $D(\lambda)$, as $\lambda$ moves from the lower boundary of any of the slits $S_k$ around the point $\nu_{2k-1}^{-1}$ to the upper boundary, $\arg(1-\nu_{2k-1}\lambda)$ changes by $2\pi$ while all other $\arg(1-\nu_i\lambda)$'s remain the same. Hence $\arg[D(\lambda)]$ changes by $\pi$, and consequently $D(\lambda)$ and $g$ change sign. Moreover, if $\lambda$ follows the path $\Pi$ (see Exhibit 2) from the upper boundary of $S_k$ around the points $\nu_{2k}^{-1}$ and $\nu_{2k+1}$ to the upper boundary of $S_{k+1}$, both $\arg(1-\nu_{2k}\lambda)$ and $\arg(1-\nu_{2k+1}\lambda)$ change by $\pi$, while other $\arg(1-\nu_i\lambda)$'s remain the same. Hence, again $d(\lambda)$ changes sign (this is the discussion Smirnov missed. I found the mistake when I used Smirnov's formula and obtained values of a distribution function which were greater than one). The value of $g$ over the semicircular ends of the narrow slits $S_k$ of width $2r$ is of the order $r^{-1/2}$. Therefore, the integral of $g$ over one of these semicircles goes to 0

as $r \to 0$. Consequently, as $r \to 0$,

$$F(z) - F(z') = \pm\pi^{-1} \int_{(+)} \lambda^{-1} D(\lambda)[\exp(-\lambda z'/2) - \exp(-\lambda z/2)]\, d\lambda. \tag{28}$$

**Exhibit 2.**

Path $\pi$.




As $z' \to \infty$, $F(z') \to 1$ and $\int_{(+)} \lambda^{-1} D(\lambda)\exp(-\lambda z'/2) \to 0$. Therefore

$$F(z) = 1 \pm \pi^{-1} \int_{(+)} \lambda^{-1} D(\lambda)\exp(-\lambda z/2)\, d\lambda. \tag{29}$$

We choose that branch of $D(\lambda)$ which replaces '$\pm$' by '$-$' because, as we now show, $\int_{(+)} \lambda^{-1} D(\lambda)\exp(-\lambda z/2)\, d\lambda > 0$.

For $n_2 \leq 2$, $\int_{(+)} g$ is obviously positive. If $n_2 \geq 3$, then by Lemma 3.1

$$\int_{(+)} g > \exp[-z/(2\nu_2)] \int_{\nu_1^{-1}}^{\nu_2^{-1}} \lambda^{-1} D(\lambda)\, d\lambda - (n_2/2) M \exp[(-z/(2\nu_3)]$$

which is positive when $z$ is large enough. Since $\int_{(+)} g$ is obviously continuous, if it were negative for some value of $z$, it would be zero for some other value, $z'' < \infty$. Then for either branch of $D(\lambda)$, $F(z'')$ would be 1, which is impossible.

This proves the part of (22) for $z \geq 0$. The proof for $z \leq 0$ is quite similar; only now we would consider a mirror image (in the imaginary axis) of the contour $C$ and slits $S_1', S_2', \ldots$ enclosing pairs of points $\mu_i^{-1}$. This proves the theorem.

If the reader is at all disturbed by the prospect of discontinuity of $F(z)$ at $z = 0$, he needs to notice the following: consider a region in the $\lambda$ plane, enclosed by a large circular contour of radius $R$, from which slits $S_k$ and $S_k'$ have been removed as has been a small circular hole of radius $r$ around $\lambda = 0$ (to account for the $\lambda^{-1}$ in $g$). By the residue theorem (see Nehari, 1961, or Hille, 1959) the integral around the small hole is $2\pi i \lim_{\lambda \to 0} \lambda g = 1$. Notice also that as we move from the upper boundary of $S_1'$ around the upper part of the circumference of the small circular hole to the upper boundary of $S_k$, $\arg(g)$ changes first

by $\pi/2$, then by $\pi$ and finally by $\pi/2$ for a total of $2\pi$. Therefore, $\int_{\partial S_1'}$ and $\int_{\partial S_1}$ have the same sign. Hence it may be shown that

$$\int_{(-)} \lambda^{-1} D(\lambda)\, d\lambda + \int (+)\lambda^{-1} D(\lambda)\, d\lambda - 1 = 0 \tag{30}$$

which establishes continuity.

### 3.2. Distribution of $P$ when $\lambda_i$'s are Distinct

In order to get the distribution of $P$ from Theorem 3.1, we can proceed in at least two different ways. One is the approach taken by Imhoff (1961). Let $G(z)$ be the distribution of $P$. Then

$$G(z) = \Pr(P \le z) = \Pr(U \le zV) = \Pr(U - zV \le 0) = \Pr(\sum_{i=1}^{n} [\lambda_i - z] u_i^2 \le 0)$$

Thus, for each $z$, $G(z)$ is the same as $F(0)$ where $F(z)$ is as in Theorem 3.1 with the difference that $\lambda_i$'s are replaced by $[\lambda_i - z]$. One computational advantage of this approach is that $\exp[-z\lambda]$, the repeated computation of which can be quite time consuming, becomes 1. Several alternative expressions and approximations also have been found for such $F(0)$'s — see Koertz and Abrahamse (1969, Ch. 5), Imhoff (1961), L'Esperance *et al.* (1976), White (1978).

Another approach is more classical. We present it as a theorem:

**Theorem 3.2**

Define

$$\psi(z, \lambda) = \begin{cases} 1 - \lambda z & \text{when } 0 \le \lambda z \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Then, in the same notation as in Theorem 3.1, the distribution $G(z)$ of $P$ is

$$G(z) = \begin{cases} 1 - \pi^{-1} \int_{(+)} \lambda^{-1} D(\lambda) [\psi(z, \lambda/p)]^{\nu - 1}\, d\lambda & \text{for } z \ge 0 \\ \pi^{-1} \int_{(-)} \lambda^{-1} D(\lambda) [\psi(z, \lambda/p)]^{\nu - 1}\, d\lambda & \text{for } z < 0, \end{cases} \tag{31}$$

where $2\nu = p$ and $\nu$ is not necessarily an integer.

Before using (31) in numerical computations, it is desirable to plot the function $\psi(z, \lambda/p)$ for some representative values of the arguments. This would help in determining time saving limits of integration. In this context, it might also be mentioned that since numerical integration is equivalent to computing the dot–product of (rather long) vectors, modern compilers can perform the operation rather efficiently.

Notice that as $p \to \infty$, (31) becomes (22), as indeed should have been expected. Koopmans (1942, see also von Neumann, 1942, and Plackett, 1960) obtained the density function corresponding to (31) in terms of a complex integral from which (31) can be derived in a manner analogous to the proof of Theorem 3.1. Mulholland (1970) also has obtained a result similar to Theorem 3.2.

Proof of Theorem 3.2:

The density function (p.d.f.) of $U$ can easily be obtained from (22) by differentiation with respect to $z$. It is

$$f(z) = \begin{cases} +(2\pi)^{-1} \int_{(+)} D(\lambda)\exp\,(-\lambda z)\, d\lambda & \text{for } z \geq 0 \\ -(2\pi)^{-1} \int_{(-)} D(\lambda)\exp(-\lambda z)\, d\lambda. & \text{for } z < 0. \end{cases}$$

Therefore, the $s$th moment of $U$ is

$$E(U^s) = (2\pi)^{-1} \left[ \int_0^\infty \int_{(+)} D(\lambda) z^s \exp(-\lambda z)\, d\lambda dz \right.$$
$$\left. - \int_{-\infty}^0 \int_{(-)} D(\lambda) z^s \exp(-\lambda z)\, d\lambda dz. \right. \tag{32}$$

The Gamma function $\Gamma(s+1)$ is defined as $\int_0^\infty x^s \exp(-x)\, dx$. Therefore,

$$\int_0^\infty z^s \exp(-\lambda z/2) dx = (2/\lambda)^{s+1} \Gamma(s+1).$$

The integrability of the relevant functions in (32) is easily established. Hence we can use Fubini's theorem to exchange the order of integration. Therefore, the first integral within square brackets in (32) can be written as

$$\Gamma(s+1) \int_{(+)} D(\lambda)(2/\lambda)^{s+1}\, d\lambda.$$

After carrying out a similar exercise with the second integral we get

$$E(U^s) = (2\pi)^{-1}\Gamma(s+1) \cdot \left[ \int_{(+)} D(\lambda)(2/\lambda)^{s+1}\, d\lambda + (-1)^s \int_{(--)} D(\lambda)(2/\lambda)^{s+1}\, d\lambda \right].$$

where $\int_{(--)}$ is obtained from $\int_{(-)}$ by replacing each $\nu_i^{-1}$ by $-\nu_i^{-1}$.

Since $pV$ has a chi–square distribution with $p$ degrees of freedom, it can be shown that

$$E[(pV)^s] = 2^s \Gamma(s+\nu)/\Gamma(\nu).$$

Hence, by (13), it follows that

$$E[(pP)^s] = \pi^{-1}[\Gamma(s+1)\Gamma(\nu)/\Gamma(\nu+s)]$$
$$\cdot \left[ \int_{(+)} D(\lambda)\lambda^{-s-1}\, d\lambda + (-1)^s \int_{(-)} D(\lambda)\lambda^{-s-1}\, d\lambda \right]. \tag{33}$$

Now consider the function

$$h(z,\lambda) = \pi^{-1}(\nu-1)D(\lambda)[\psi(z,\lambda)]^{\nu-2}.$$

Since it can easily be verified that $z^r h(z,\lambda)$ is integrable over the appropriate region, we can use Fubini's theorem to invert order of integration and get

$$\begin{aligned}\int_0^\infty z^s \int_{(+)} h(z,\lambda)\,d\lambda dz &= \pi^{-1}(\nu-1)\int_{(+)} D(\lambda)\int_0^{\lambda^{-1}} z^s(1-\lambda z)^{\nu-2}\,dz d\lambda \\ &= \pi^{-1}(\nu-1)\int_{(+)} D(\lambda)\lambda^{-s-1}\int_0^1 (\lambda z)^s(1-\lambda z)^{\nu-1}\,d(\lambda z)d\lambda \\ &= \pi^{-1}[\gamma(s+1)\Gamma(\nu)/\Gamma(\nu+s)]\int_{(+)} D(\lambda)\lambda^{-s-1}\,d\lambda, \qquad (34)\end{aligned}$$

on noting that $\int_0^1 x^{a-1}(1-x)^{b-1}\,dx$ is the beta–function $\Gamma(a)\Gamma(b)/\Gamma(a+b)$ (see Wilks, 1962, p. 174). Therefore, (34) is the first term on the right side of (33). A similar result can be derived for the second term. Hence,

$$E[(pP)^s] = \int_0^\infty z^s \int_{(+)} h(z,\lambda)\,d\lambda dz - \int_{-\infty}^0 z^s \int_{(-)} h(z,\lambda)\,d\lambda dz. \qquad (35)$$

Therefore, by the uniqueness theorem for moments of bounded random variables, it follows that the p.d.f. of $pP$ is $\int_{(+)} h(z,\lambda)\,d\lambda$ when $z \geq 0$, and when $z < 0$ it is $\int_{(-)} h(z,\lambda)\,d\lambda$. A straightforward change of variable yields the p.d.f. of $P$ as

$$\begin{cases} +(2\pi\nu)^{-1}(\nu-1)\int_{(+)} D(\lambda)[\psi(z,\lambda/p)]^{\nu-2}\,d\lambda \text{ for } z \geq 0 \\ -(2\pi\nu)^{-1}(\nu-1)\int_{(-)} D(\lambda)[\psi(z,\lambda/p)]^{\nu-2}\,d\lambda \text{ for } z < 0. \end{cases} \qquad (36)$$

To complete the theorem and obtain the distribution function all we need do is integrate the density function (36) from 0 to $z$. When $z < 0$, it is easy to see that this integral [which is simply the integral of the lower expression in (36)] is the lower expression in (31). When $z \geq 0$, integration of the upper expression in (36) from 0 to $z$, followed by the use of (30), yields the upper part of (31) as the required distribution function.

### 3.3. Distribution of $P$ when Values of $\lambda_i$'s are Repeated

In most practical applications the $\lambda_i$'s are distinct, particularly when the regions considered are irregular — as census tracts, states and Zip–code zones usually are. If a pair (or two) of the $\lambda_i$'s become alike, there is perhaps not much lost by adding a small number to one and subtracting it from the other in order to make them distinct. However, this recourse is not too satisfying if a large number of pairs of $\lambda_i$'s are the same. This can happen when observations are taken over a regular lattice or over regions bounded by a uniform grid (*e.g.*, quarter sections). In this section we first consider the case when

$$U = \sum_{\ell=1}^{2}\sum_{i=1}^{m} \lambda_i u_{i\ell}^2, \qquad (37)$$

$\lambda_i$'s are distinct and $n = 2m$ is necessarily even. Then we shall consider the case when some $\lambda_i$'s are singletons while others come in pairs. Finally, we shall briefly examine the general case when $\lambda_i$'s may be repeated arbitrarily often.

As before, let $\mu_1, \ldots, \mu_{m_1}$ be the negative $\lambda_i$'s and let $\nu_1, \ldots, \nu_{m_2}$ be the positive $\lambda_i$'s. Let

$$B_{k_{(+)}} = \prod_{i=1}^{m_1} (1 - \mu_i/\nu_k)^{-1} \prod_{\substack{i=1 \\ i \neq k}}^{n_2} (1 - \nu_i/\nu_k)^{-1}$$

$$B_{k_{(-)}} = \prod_{\substack{i=1 \\ i \neq k}}^{m_1} (1 - \mu_i/\mu_k)^{-1} \prod_{\substack{i=1 \\ i \neq k}}^{n_2} (1 - \nu_i/\mu_k)^{-1}.$$

Then, for $z \geq 0$, considering the same semicircular contour $C$ that we did in the proof of Theorem 3.1, and for $z < 0$, considering the mirror image of $C$ in the imaginary axis, and applying the residue theorem, it may be shown that the distribution function of $U$ is

$$F(z) = \begin{cases} 1 - \sum_k^{n_2} B_{k_{(+)}} \exp(-z\nu_k/2) & \text{for } z \geq 0 \\ \sum_k^{n_1} B_{k(-)} \exp(-z\mu_k/2) & \text{for } z < 0. \end{cases} \tag{38}$$

The result is very well known, having been given by several authors (see Box, 1954).

Using (38) and following steps similar to those in the proof of Theorem 3.2, it may be shown that the distribution of $P$, when $U$ has the form (37), is

$$F(z) = \begin{cases} 1 - \sum_k^{n_2} B_{k(+)} (\psi[z, (p\nu_k)^{-1}])^{\nu - 1} & \text{for } z \geq 0 \\ \sum_k^{n_1} B_{k(-)} (\psi[z, (p\mu_k)^{-1}])^{\nu - 1} & \text{for } z < 0 \end{cases} \tag{39}$$

where $\psi$ is the same function we defined in Theorem 3.2. Obviously, since no integration is involved, (38) is easier to use than Theorem 3.1.

When some $\lambda_i$'s are singletons while others come in pairs, we can write $U = U_1 + U_2$, where

$$U_1 = \sum_{i=1}^{q} \lambda_{i(1)} u_i^2 \tag{40}$$

with distinct $\lambda_{i(1)}$'s and $U_2$ is as in (37). The p.d.f. $f_1$ of $U_1$ can be found from Theorem 3.1, and $f_2$, the p.d.f. of $U_2$, can be found from (38). The p.d.f. $f(z)$ of $U$ can then be found by convolution:

$$f(z) = \int_{-\infty}^{\infty} f_1(z - y) f_2(y)\, dy.$$

Since $z$ enters both $f_1(z)$ and $f_2(z)$ only as an argument of the exponential function, such a convolution is easy to obtain analytically and is roughly of the same form as $F(z)$ in Theorem 3.1. The distribution of the corresponding $P$ can be found using the same means as in the proof of Theorem 3.2, and is numerically no more difficult to compute than (31).

It is unlikely that, in practice, we would encounter values of $\lambda_i$ repeated more than three times, and, therefore, the discussion above should cover most practical situations. However, for completeness, we briefly outline a treatment for the most general case where $U = U_1 + U_2$, with $U_1$ as in (40) and $U_2$ containing coefficients the values of which are repeated an even number of times. The distribution of $U_2$ can be obtained using the residue theorem. It will have the same general form as (38), but $B_k$'s will now be polynomials and not constants.

The distribution of $U$ again can be found analytically by convolution, although now it will contain $\Gamma$–functions (because $B_k$ are polynomials). That of $P$ can be obtained using the same method as in the proof of Theorem 3.2, but because of the $\Gamma$–functions in the convolution, it would contain $\beta$–functions.

## 4. Asymptotic Mean and Variance

If we do not assume that $\boldsymbol{\epsilon}$ is normal, then in general we cannot say much about $\mathbf{u}$. Consequently we must proceed differently and indeed the exact mean and variance of (9) are not available. However, we can obtain the mean and variance of the numerator of (9) and obtain asymptotic results by letting the denominator $s^2 \to \sigma^2$. These results will be sharper than the ones in Sen (1976).

To obtain the mean of the numerator $c\mathbf{e}'Z\mathbf{e}$ we can proceed directly. As before we can write $\mathbf{e}'Z\mathbf{e} = \boldsymbol{\epsilon}'MZM\boldsymbol{\epsilon} = \boldsymbol{\epsilon}'B\boldsymbol{\epsilon}$, and, since we are interested in the distribution under the hypothesis, $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2 I$. Hence

$$\begin{aligned} E(\boldsymbol{\epsilon}'B\boldsymbol{\epsilon}) &= E(\text{tr}[\boldsymbol{\epsilon}'B\boldsymbol{\epsilon}]) = E(\text{tr}[B\boldsymbol{\epsilon}\boldsymbol{\epsilon}']) = \text{tr}[E(B\boldsymbol{\epsilon}\boldsymbol{\epsilon}')] \\ &= \text{tr}[BE(\boldsymbol{\epsilon}\boldsymbol{\epsilon}')] = \text{tr}[B(\sigma^2 I)] = \sigma^2 \text{tr}[B] \end{aligned} \tag{41}$$

as before. Hence $E(c\mathbf{e}'Z\mathbf{e}) = c\sigma^2\text{tr}(B)$ and $E(P) = \text{tr}(B)$.

In order to obtain the variance, we first compute

$$E(\boldsymbol{\epsilon}'B\boldsymbol{\epsilon})^2 = E(\boldsymbol{\epsilon}'B\boldsymbol{\epsilon}\boldsymbol{\epsilon}'B\boldsymbol{\epsilon}) = E(\text{tr}[\boldsymbol{\epsilon}B\boldsymbol{\epsilon}\boldsymbol{\epsilon}'B\boldsymbol{\epsilon}]) = E(\text{tr}[B\boldsymbol{\epsilon}\boldsymbol{\epsilon}'B\boldsymbol{\epsilon}\boldsymbol{\epsilon}']) = \text{tr}[BC] \tag{42}$$

where

$$C = E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}'B\boldsymbol{\epsilon}\boldsymbol{\epsilon}'). \tag{43}$$

To compute (43) we note that

$$E(\boldsymbol{\epsilon}_i^4) = \mu_4 \quad \text{and} \quad E(\boldsymbol{\epsilon}_i^2\boldsymbol{\epsilon}_j^2) = \sigma^4 \text{ when } i \neq j, \tag{44}$$

and all other expectations of the products of four $\boldsymbol{\epsilon}_i$'s are zero, *i. e.*,

$$\begin{aligned} &E(\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}, 3_j) = E(\boldsymbol{\epsilon}_i)E(\boldsymbol{\epsilon}_j^3) = 0 \text{ for } i \neq j, \\ &E(\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_j^2\boldsymbol{\epsilon}_k) = 0 \text{ for } i \neq j, i \neq k, \\ &E(\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_j\boldsymbol{\epsilon}_k\boldsymbol{\epsilon}_{\boldsymbol{\ell}}) = 0 \text{ for } i \neq j, k, \boldsymbol{\ell}. \end{aligned} \tag{45}$$

Now set $\boldsymbol{\epsilon}\boldsymbol{\epsilon}'B = A = (a_{i\boldsymbol{\ell}})$ (i.e, A is the matrix with elements $a_{i\boldsymbol{\ell}}$). Obviously $a_{i\boldsymbol{\ell}} = \boldsymbol{\epsilon}_i \sum_{k=1}^n \boldsymbol{\epsilon}_k b_{k\boldsymbol{\ell}}$. Since the $(\boldsymbol{\ell}, j)$th element of $\boldsymbol{\epsilon}\boldsymbol{\epsilon}'$ is $\boldsymbol{\epsilon}_{\boldsymbol{\ell}}\boldsymbol{\epsilon}_j$, if we set $C = (c_{ij})$,

$$c_{ij} = \sum_{\boldsymbol{\ell}} c_{\boldsymbol{\ell}}(i,j) \text{ where } c_{\boldsymbol{\ell}}(i,j) = E[\boldsymbol{\epsilon}_i(\sum_{k=1}^n \boldsymbol{\epsilon}_k b_{k\boldsymbol{\ell}})\boldsymbol{\epsilon}_{\boldsymbol{\ell}}\boldsymbol{\epsilon}_j].$$

From (45) we see that the terms of $c_{\boldsymbol{\ell}}(i,j)$ are zeros unless

$$i = j \text{ and } k = \boldsymbol{\ell} \tag{46}$$

$$i = \boldsymbol{\ell} \text{ and } k = j \tag{47}$$

$$i = k \text{ and } \boldsymbol{\ell} = j. \tag{48}$$

First consider $i = j$. Then the only non-zero term in $c_\ell(i,i)$ occurs when $k = \ell$. Therefore, if $i \neq \ell$, $c_\ell(i,i) = b_{\ell\ell}\sigma^4$, and if $i = \ell$, $c_i(i,i) = b_{ii}\mu_4$. It follows that

$$c_{ii} = \sum_{\ell=1}^{n} b_{\ell\ell}\sigma^4 + b_{ii}(\mu_4 - \sigma^4) = \sigma^4 \mathrm{tr}[B] + b_{ii}(\mu_4 - \sigma^4).$$

If $i \neq j$, then $c_\ell(i,j) \neq 0$ only if either $i = \ell$ or $j = \ell$. In the former case $c_i(i,j) = \sigma^4 b_{ji}$ and in the latter case $c_j(i,j) = \sigma^4 b_{ij}$. Since B is symmetric, this yields

$$c_{ij} = 2\sigma^4 b_{ij}.$$

Thus

$$\begin{aligned} C = \sigma^4 &\begin{pmatrix} \mathrm{tr}[B] & 2b_{12} & \dots & 2b_{1n} \\ 2b_{21} & \mathrm{tr}[B] & \dots & 2b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 2b_{n1} & 2b_{n2} & \dots & \mathrm{tr}[B] \end{pmatrix} \\ &+ (\mu_4 - \sigma^4)\,\mathrm{diag}(b_{11}, b_{22}, \dots, b_{\mathrm{nn}}) \\ &= 2\sigma^4 B + (\mu_4 - 3\sigma^4)\,\mathrm{diag}(b_{11}, b_{22}, \dots, b_{\mathrm{nn}}) + \sigma^4(\mathrm{tr}[B])I_n. \end{aligned}$$

From (41) and (42)

$$\mathrm{var}(\mathbf{e}'Z\mathbf{e}) = \mathrm{tr}[BC] - \sigma^4(\mathrm{tr}[B])^2 = 2\sigma^4\mathrm{tr}[B^2] + (\mu_4 - 3\sigma^4)\sum_{i=1}^{n} b_{ii}^2. \tag{49}$$

It has been shown (Sen, 1976) that $\mathbf{e}'Z\mathbf{e} = \boldsymbol{\epsilon}'B\boldsymbol{\epsilon}$ is asymptotically normal when $b_{ii} \to 0$ and some other mild conditions mentioned in that paper are met. In most practical situations where a small set of observations are not unduly influential these conditions will always be met. Then, of course, $(\mu_4 - 3\sigma^4) \to 0$. Since $s^2 \to \sigma^2$ almost surely, the asymptotic normality of $e'Ze/s^2$ follows from Slutsky's Theorem (see Rao, 1971, p. 122). Hence the asymptotic variance of $P$ is $2\mathrm{tr}[B^2]$.

## 5. References


Bellman, R. A. (1962) *Introduction to Matrix Analysis*, McGraw–Hill, New York.

Box, G. E. P. (1954) Some Theorems on Quadratic Forms Applied in the Study of Variance Problems, I. Effect of Inequality of Variance in the one–way Classification, *Annals of Mathematical Statistics*, **26** 464-477.

Brandsma, A. S. and R. H. Ketellapper (1979) Further Evidence on Alternative Procedures for Testing Spatial Correlation Amongst Regression Disturbances, in *Exploratory and Explanatory Statistical Analysis of Spatial Data*, C.P.A. Bartels and R.H. Ketellaper, Eds. Martinus Nijhoff, Boston, 113-136.

Cliff, A. D. and J. K. Ord. (1981) *Spatial Processes*, Pion, London.

Hille, E. (1959) *Analytic Function Theory, Vol. 1*, Ginn and Co., Boston.

Imhoff, J. P. (1961) Computing the Distribution of Quadratic Forms in Normal Variables, *Biometrika* **48** 419-426.

Koertz, J. and A. P. J. Abrahamse (1969) *On the Theory and Applications of the General Linear Model*, Rotterdam University Press.

Koopmans, T. C. (1942) Serial Correlation and Normal Quadratic Forms in Normal Variables, *Annals of Mathematical Statistics*, **13** 14-33.

L'Esperance, W. L., D. Chall and D. Taylor (1976) An Algorithm for Determining the Distribution Function of the Durbin-Watson Statistic, *Econometrica*, **44** 1325-1346.

Mullholland, H. P. (1970) On Singularities of Sampling Distributions, and Particular for Ratios of Quadratic Forms, *Biometrika*, **57** 155-174.

Nehari, Z. (1961) *Complex Analysis*, Allyn and Bacon, Boston.

Pitman, E. J. G. (1937) The 'Closest' Estimates of Statistical Parameters, *Proceedings, Cambridge Philosophical Society*, **33** 212.

Plackett, R. L. (1960) *Principles of Regression Analysis*, Clarendon Press, Oxford.

Rao, C. R. (1973) *Linear Statistical Inference and its Applications*, Wiley.

Ripley, B. D. (1981) *Spatial Statistics*, Wiley.

Sen, A. (1976) Large Sample–Size Distribution of Statistics Used in Testing for Spatial Correlation, *Geographical Analysis*, **9** 175-184.

Sen, A. and S. Soot. (1977) Rank Tests for Spatial Correlation, *Environment and Planning A*, **9** 897-903.

Smirnov, N. V. (1937) O Raspredelenie $\omega^2$ kriteriia Mizesa, *Matematicheskii Sbornik, 2* **44** 973-993. A French summary is given in Smirnoff, N. V. (1936) Sur la Distribution de $\omega^2$ (Criterium de M. R. v. Mises), *Com. Rend. Acad. Sci. (Paris)* **202** 449-452.

Srivastava, M. S. and C. G. Khatri (1979) *An Introduction to Multivariate Statistics*, North–Holland.

Von Neumann, J. (1941) Distribution of the Ratio of the Mean Square Successive Difference to the Variance, *Annals of Mathematical Statistics*, **12** 367-395.

White, K. J. (1978) A General Computer Program for Econometric Models — SHAZAM, *Econometrica*, **46** 151–159.

Wilks, S. S. (1962) *Mathematical Statistics*, Wiley.

## DISCUSSION

### "Distribution of spatial correlation statistics"

### by Ashish Sen

The paper looks into the difficult problem of obtaining the exact distribution of spatial correlation statistics P for normal residuals under the null hypothesis of spatial independence. Also, the first two moments of P are obtained together with the first two asymptotic moments for non–normal variables. Theoretically the results are impressive, but I believe that substantial further work is required to make them applicable in practice. My main comments are as follows:

(i) How will the actual formulæ for the distribution function help with numerical examples? Cannot the same work be achieved in practice by simulating the percentage points under the null hypothesis? [I am not clear of the precise overlap of Theorem 3.2 with Mulholland (1970)].

(ii) The mean and variance of the statistic P under the normality assumption are given in Section 2. How will these help in approximating the distribution function? Should one take higher moments and use a Beta distribution, for example? The work of Jones (1987) and the references therein are relevant to obtain the moments.

(iii) The asymptotic variance in Section 4 also must depend on $\mathrm{cov}(\mathbf{e}'\mathbf{Ze}, s^2)$. This value can be obtained in the same way as $\mathrm{var}(\mathbf{e}'\mathbf{Ze})$ in Section 4. Using the asymptotic variance for the ratio through Taylor series expansion one can obtain an improved approximation to the variance of P for random samples from non–normal variables. The use of the permutation approach of Box and Watson (1962) and Mardia (1970) will be a step forward. Note that for $b_{ii} \to 0$ asymptotically, the variance of P under the normal and non–normal cases is equivalent. Thus, at least for the first two moments, the effect of non–normality seems to be minimal! Does this result require the assumption that the kurtosis is negligible? A few simulation studies should prove useful.

Now coming to a point of detail, I might mention that Theorem 2.1 follows trivially from a well–known result (see Mardia, Kent and Bibby, 1979, p. 95, Exercise 3.4.21) that if $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$, then

$$\mathrm{var}(\mathbf{X}'\mathbf{AX}) = 2\mathrm{tr}(\mathbf{A\Sigma})^2,$$
$$\mathrm{cov}(\mathbf{X}'\mathbf{AX}, \mathbf{X}'\mathbf{BX}) = 2\mathrm{tr}(\mathbf{A\Sigma B\Sigma}).$$

Here $\mathbf{e} \sim N(0, \mathbf{M})$ where $\mathbf{M}$ is idempotent so that

$$\mathrm{cov}(\mathbf{e}'\mathbf{Ze}, \mathbf{e}'\mathbf{e}) = 2\sigma^4\mathrm{tr}(\mathbf{ZM}) = 2\sigma^4\mathrm{tr}(\mathbf{M}'\mathbf{ZM}) = 2\sigma^4\mathrm{tr}(\mathbf{B})$$
$$\mathrm{var}(\mathbf{e}'\mathbf{Ze}) = 2\mathrm{tr}(\mathbf{ZM})^2 = 2\sigma^4\mathrm{tr}(\mathbf{B})^2,$$
$$\mathrm{var}(\mathbf{e}'\mathbf{e}) = 2\sigma^4 p.$$

Further, $P = \mathbf{e}'\mathbf{Ze}/\mathbf{e}'\mathbf{e}$ and $V = \mathbf{e}'\mathbf{Ze}$, are independent under normality of $\mathbf{e}$ because $P$ is a scale free quantity. Hence $\mathrm{var}(P)$ can be written down.

In conclusion, I am sure that Dr. Sen's paper will generate further research work with the aim of some specific recommendations for the practitioners.

## References

Box, G. E. S., and G. S. Watson. (1962) Robustness to nonnormality of regression tests. *Biometrika*, **49**, 93-106 (correction, *Biometrika*, **52**, 669).

Jones, M. C. (1987) On moments of ratios of quadratic forms in normal variables. *Statistics & Probability Letters*, **6**, 129-136.

Mardia, K. V. (1979) The effect of nonnormality on some multivariate tests and robustness to nonnormality in the linear model. *Biometrika*, **58**, 105-121.

Mardia, K. V., J. T. Kent, and J. M. Bibby. (1979) *Multivariate Analysis*. New York: Academic Press.

*Kanti V. Mardia*, University of Leeds

## *PREAMBLE*

*Inspiration and genius — one and the same.*

Victor Hugo

*This paper, in an ingenious way, mingles the themes of information content of data and the latent spatial dependence of geo–referenced data. To date most of the spatial statistics literature has been devoted to model specification issues, in order to handle or accommodate spatial dependence. But rather than adjusting computations in order for standard statistical tables to serve as proper references in a conventional way, Richardson exploits the redundant information aspect of geo–referenced data so that standard statistical tables can be used in a novel way. Accordingly, the purpose of this paper is to present a modified test of association based upon the correlation coefficient, and then evaluate its performance. One question Sen asks is whether or not these two approaches are equivalent? Richardson's presentation should inspire much subsequent research on this equivalence topic, as well as various extensions of the modification to other classical statistics. The notion of degrees of freedom as an index of information contained in data is employed in the implementation of Richardson's modification, which essentially filters out the redundant information seemingly quiescent in geo–referenced data by determining the correct degrees of freedom index; radical changes in this index can occur, as is indicated by tabulated results. But, except for extremely small degrees of freedom, the magnitude of the t–statistic does not change very much for a given probability level across its degrees of freedom range. This fact suggests that spatial dependence may make little difference to the drawing of inferences, other than in cases that are close to a selected critical value; fortunately other statistics, such as chi square, may better demonstrate Richardson's point. Another prominent drawback Sen notes is that this approach fails to acknowledge the impact spatial dependence has on important statistical properties of estimators, such as efficiency. Nevertheless, the elegance of Richardson's approach should open new areas of research in spatial statistics.*

The Editor

# Some Remarks on the Testing of Association Between Spatial Processes

Sylvia Richardson *

*INSERM U.170 - 16 Avenue Paul Vaillant Couturier, 94807 Villejuif Cedex, France*

**Overview:** This paper is concerned with the problem of testing for association between spatially defined variables exhibiting spatial autocorrelation. This problem occurs commonly in geography with variables typically defined as area averages of dichotomous or continuous variables recorded at the individual level. Classical statistical methods, such as correlation or regression analysis, are not directly applicable to the situation of spatially autocorrelated variables. A modified test of association based on the correlation coefficient is reviewed and results of its performance (Type I error, power, robustness to departure from normality) as well as that of a nonparametric measure of association, Tjøstheim's index, are given. This comparison shows the weak performance of the nonparametric index of association. For the case of several variables, the modified test can be extended to assess the significance of the partial correlation coefficient. Alternatively, regression analysis with the spatially parametrised error variance–covariance matrix can be performed. The two approaches are discussed and compared on some examples of geographical epidemiology concerning the relationship between lung cancer mortality and industrial factors.

## 1. Introduction

The problem of testing association between spatially defined variables occurs commonly in science and its applications. Standard examples are found in the fields of geography and regional science, for instance when relating consumption of agricultural output to road accessibility (Cliff and Ord, 1981). Upton and Fingleton (1985) discuss cases in ecology concerned with the relationships between, for example, plant species or island flora and locational or environmental characteristics. Examples in sociology and political science are described by Doreian (1981), with reference to voting behaviour and socio–economic or political factors. In epidemiology, etiological clues to environmental risk–factors are sometimes sought through their joint analysis with disease incidence or mortality maps (Doll, 1984; Armstrong and Doll, 1975).

As seen in these studies, the data consist of variables observed at different locations that can be considered as observations of stochastic processes exhibiting typically some spatial dependence (*i. e.*, dependence between variables indexed by nearby points). This dependence or *spatial autocorrelation* can arise in different ways. It might be an intrinsic characteristic of the process itself, such as the existence of interactions between the sites of a diffusion or of a contagious phenomenon. Spatial autocorrelation might also result indirectly from the influence on the variable considered of other related factors varying continuously

* Chantal Guihenneuc and Virginie Lasserre of the Laboratoire de Statistiques Médicales de l'Université de Paris V provided invaluable computing assistance. Evelyne Przybilski is warmly thanked for her secretarial assistance. This work was supported by a EURATOM contract # BI6-126F.

through space. Such is the case for variables defined as area averages of dichotomous variables recorded at the individual level. Even if there is *a priori* no direct inter–individual influence on a particular dichotomous variable, spatial autocorrelation between aggregated values will often be observed since spatially close individuals will often share some common influencing characteristics. The epidemiological examples that will be discussed in the final section of this paper belong to this category.

Classical statistical methods based on correlation or regression analysis are not directly applicable to the situation of spatially autocorrelated variables. The consequences of neglecting existing autocorrelation in regression analysis have been pointed out by Johnston (1972) for time series and Cliff and Ord (1981) for spatial series.

The problem of dealing with spatial autocorrelation in testing for association may be tackled in various ways. Time series methods proposed by Haugh (1976) such as the prewhitening of each series before analysis of the cross correlations could be extended to the spatial domain by considering appropriate filtering. Using this approach, Pierce (1977) found only weak evidence of a relationship between pairs of economic time series that were traditionally considered as related. In a subsequent paper, Geweke (1981) compares several tests of independence between stationary time series, in particular Haugh's test and an F test on the regression parameters of a mixed regressive–autoregressive model between the two series. He concludes that in many cases the proportion of Type II errors of Haugh's test is larger than that of F tests of the regression coefficients. For spatial series, no comparable study has been done to date although the prewhitening of series has been discussed by Griffith (1980). It is likely that a similar conclusion to Geweke would hold and that tests of independence or estimates of regression coefficients based on the original series will be more efficient than those based on their residuals after prewhitening.

Alternatively standard measures of association such as the correlation coefficient can be adapted to take account of autocorrelation. This approach has been developed by Clifford, Richardson and Hémon (1989) (CRH 89) and will be reviewed in Section 2.

New indices of association can also be proposed. A non–parametric index of association was developed by Tjøstheim (1978). A comparative study (Type I error and power) of this index and of the aforementioned modified test of the correlation coefficient will be presented in Section 3. Section 4 will complement this discussion by studying the robustness of the modified test of the correlation coefficient for departures from normality. Finally, classical regression analysis can be extended specifically to take into account the spatial structure of the data by modelling the variance–covariance error matrix. In Section 5 this approach is reviewed and a modified test of partial correlation, which is an extension to the case of several variables of the modified test of simple correlation, is presented. In addition the results given by different choices of models for the variance–covariance error matrix are compared with those given by the modified test of partial correlation on some examples of geographical epidemiology.

## 2. Modified test of association

We are interested in data sets that consist of a set $\mathbf{A}$ of $\mathbf{N}$ locations numbered from $\mathbf{1}$ to $\mathbf{N}$ and a set of pairs of observations $\{(\mathbf{X}_\alpha, \mathbf{Y}_\alpha), \alpha \varepsilon \mathbf{A}\}$, where each pair is indexed by its location. We shall use the following notation:

$$\bar{\mathbf{X}} = N^{-1}(\sum X_\alpha)$$
$$\mathbf{s_{XY}} = N^{-1} \sum (X_\alpha - \bar{X})(Y_\alpha - \bar{Y})$$
$$\mathbf{s_{X^2}} = N^{-1} \sum (X_\alpha - \bar{X})^2$$

similar expression can be written for $\bar{\mathbf{Y}}$ and $\mathbf{s_{Y^2}}$.

Modified tests of association based either on $\mathbf{s_{XY}}$, the empirical covariance between pairs of observations $\{(\mathbf{X}_\alpha, \mathbf{Y}_\alpha), \alpha \varepsilon \mathbf{A}\}$, or based on $\mathbf{r_{XY}}$, the corresponding empirical correlation coefficient, have been proposed by Clifford, Richardson and Hémon (1989). These tests rely on an estimation of the variance of $\mathbf{s_{XY}}$ and $\mathbf{r_{XY}}$ that takes the internal autocorrelations into account.

In the classical case where the elements of $\mathbf{Y}$ are normal i.i.d. random variables conditional on $\mathbf{X}$ or vice–versa, then $\mathbf{r_{XY}}$ has the standard null distribution with p.d.f.

$$\mathbf{f_N(r) = (1 - r^2)^{1/2(N-4)} / B(1/2, 1/2(N-2)), \qquad |r| \leq 1}$$

where $\mathbf{B}$ is the beta function.

The expectation of $\mathbf{r_{XY}}$ is zero and its variance is equal to $\mathbf{(N-1)^{-1}}$. Critical values of $\mathbf{r_{XY}}$ are usually obtained from $t$–tables since $\mathbf{(N-2)^{1/2} r / (1 - r^2)^{1/2}}$ has a $t$–distribution with $\mathbf{N-2}$ degrees of freedom under these assumptions. We shall refer to this statistic as $t_{\mathbf{N-2}}$. This is also the $t$–statistic that is calculated in testing the significance of the linear regression either of $\mathbf{Y}$ on $\mathbf{X}$, or of $\mathbf{X}$ on $\mathbf{Y}$.

### 2.1. The variance of $\mathbf{r_{XY}}$

Suppose now that $\mathbf{X}$ and $\mathbf{Y}$ are independent but that both $\mathbf{X}$ and $\mathbf{Y}$ are multivariate normal vectors with constant means and variance–covariance matrices $\mathbf{\Sigma_X}$ and $\mathbf{\Sigma_Y}$ respectively. The variance of $\mathbf{r_{XY}}$ is inflated by positive autocorrelation. This was shown asymptotically in the time–series context by Bartlett (1935) and in the spatial context by Richardson and Hémon (1981). This needs to be taken into account in the testing method.

It can be shown that, to the first order, the variance of $\mathbf{r_{XY}}$, $\sigma_{\mathbf{r}}^{\mathbf{2}}$ is:

$$\sigma_{\mathbf{r}}^{\mathbf{2}} = \frac{\mathbf{var(s_{XY})}}{\mathbf{E(s_{X^2}) E(s_{Y^2})}}, \tag{1}$$

and that this approximation is exact in some special cases. The variance of $\mathbf{s_{XY}}$ can be evaluated if some hypotheses on the spatial structure of $\mathbf{\Sigma_X}$ and $\mathbf{\Sigma_Y}$ are imposed. We suppose that pairs in $\mathbf{A \times A}$ can be divided into strata $\mathbf{S_0, S_1, \ldots S_K}$ such that the covariances within strata remain constant, *i. e.*,

$$\mathbf{cov}(\mathbf{X}_\alpha, \mathbf{X}_\beta) = \mathbf{C_X(k)} \quad \text{if} \quad (\alpha, \beta) \varepsilon \mathbf{S_k},$$

and

$$\mathbf{cov}(\mathbf{Y}_\alpha, \mathbf{Y}_\beta) = \mathbf{C_Y(k)} \quad \text{if} \quad (\alpha, \beta)\,\varepsilon\, \mathbf{S_k},$$

with $S_0 = \{(\alpha, \alpha),\ \alpha\,\varepsilon\, A\}$. This formulation is flexible enough to permit non-isotropy or other aspects of inhomogeneity to be taken into account.

An estimate of the variance of $\mathbf{s_{XY}}$ is then derived:

$$\mathbf{N^{-2}} \sum_{\mathbf{k}} \mathbf{N_k \hat{C}_X(k) \hat{C}_Y(k)}, \tag{2}$$

$\mathbf{N_k}$ is the number of pairs in strata $\mathbf{S_k}$ and $\mathbf{\hat{C}_X(k)}$ [respectively $\mathbf{\hat{C}_Y(k)}$] is the estimated autocovariance:

$$\mathbf{\hat{C}_X(k)} = \sum_{\mathbf{S_k}} \mathbf{(X_\alpha - \bar{X})(X_\beta - \bar{X})/N_k}$$

Thus the estimate (2) takes into account the autocorrelation of both $\mathbf{X}$ and $\mathbf{Y}$, and leads to an estimate of the variance of $\mathbf{r_{XY}}$.

$$\mathbf{\hat{\sigma}_r^2} = \sum \mathbf{[N_k \hat{C}_X(k) \hat{C}_Y(k)]/[N^2 s_{X2} s_{Y2}]}. \tag{3}$$

In a simulation study reported in CRH 89, mutually independent autoregressive processes $\mathbf{X}$ and $\mathbf{Y}$ were generated on a 12-by-12 lattice. This allowed a comparison to be made between the asymptotic variance of $\mathbf{r_{XY}}$, the first order approximation (1), the average estimated $\mathbf{\hat{\sigma}_r^2}$ given by (3) and the empirical variance of $\mathbf{r_{XY}}$ over 4000 trials. The results are illustrated in Figure 1 and clearly show that the asymptotic value is far too large for highly autocorrelated processes. Note that in time series analysis a first assessment of the cross-correlogram is traditionally done *via* a similar asymptotic expression. For moderate autocorrelation there is little difference between the empirical and the average estimated variance given by (3). For high autocorrelation in each processes the estimated variance is consistently too low.

### 2.2. Modified tests

A modified $t$-test $(\mathbf{t_{\widehat{M}-2}})$ was proposed based on an estimated *effective sample size* $\mathbf{\widehat{M}}$, $\mathbf{\widehat{M} = 1 + \hat{\sigma}_r^{-2}}$, that rejects the null hypothesis of no association when:

$$\mathbf{|(\widehat{M} - 2)^{1/2} r (1 - r^2)^{-1/2}| > t^{\alpha}_{\widehat{M}-2}} \tag{4}$$

where $\mathbf{t^{\alpha}_{\widehat{M}-2}}$ is the critical value of the $t$-statistic with $\mathbf{\widehat{M} - 2}$ d.f. The quantity $\mathbf{\widehat{M}}$ takes into account the spatial autocorrelation in the variables $\mathbf{X}$ and $\mathbf{Y}$ and is typically less than $\mathbf{N}$ for positively autocorrelated processes.

Equivalently a standardised covariance can be used:

$$\mathbf{W = N s_{XY} [\sum N_k \hat{C}_X(k) \hat{C}_Y(k)]^{-1/2}}$$

and tested as a standard normal relying upon central limit theorems for spatially dependent variables (Bolthausen, 1982; Guyon and Richardson, 1984).

**Figure 1.**

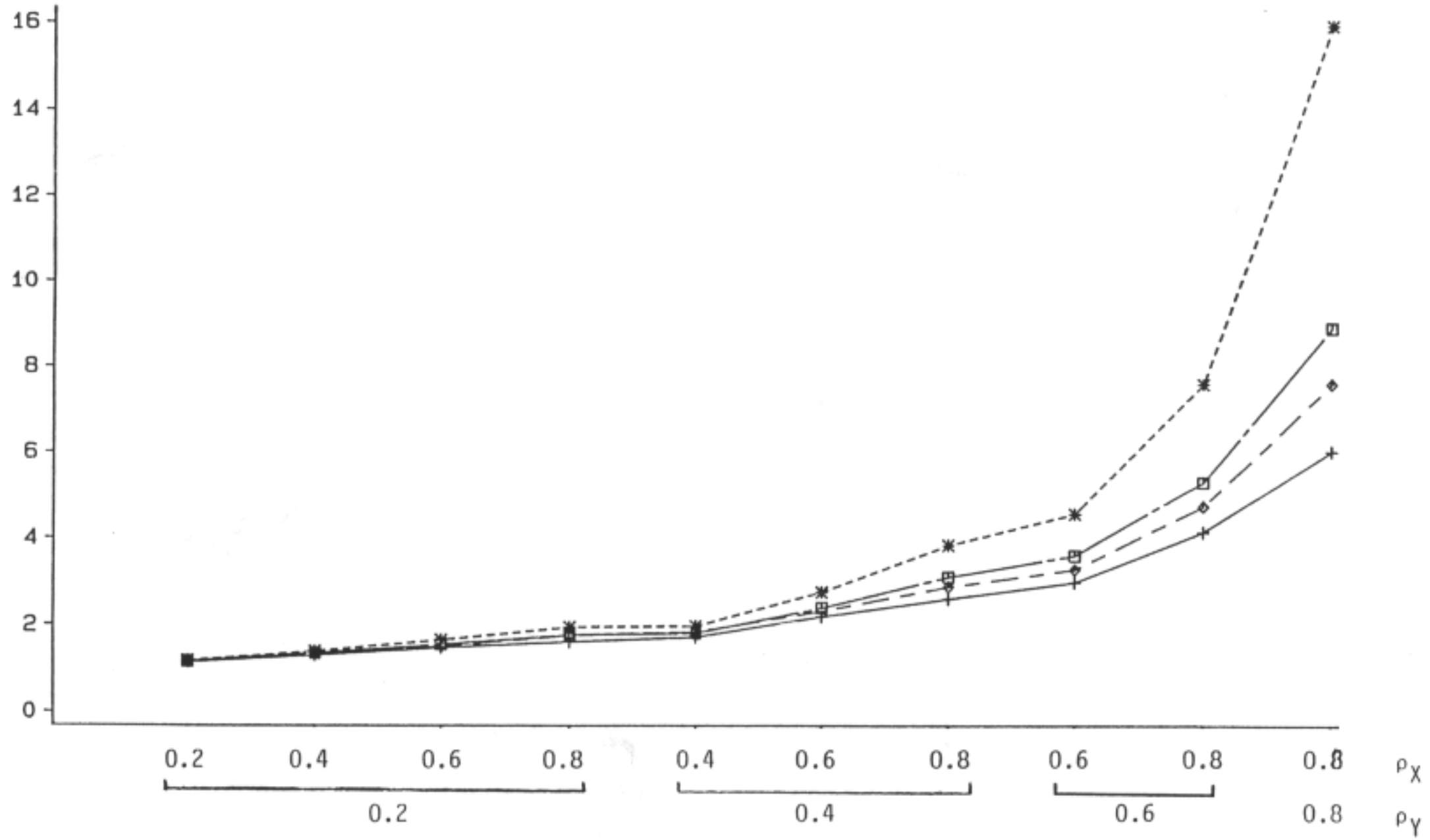



Figure 1 : The variance of r for two mutually independant simultaneous autoregressive processes on a 12x12 lattice : comparison of the asymptotic value (*----------*), the first order approximation (1) (□– - - - □), the empirical variance of r over 4000 simulations (◊_____ ◊) the average $v_r$ of $\hat{\sigma}^2_r$ given by (3) (+ _____+).
The values on the abscissa are $\rho_X(1)$ and $\rho_Y(1)$, the nearest neighbour autocorrelations for X and Y.

The Type I errors and the power of the modified tests were investigated in several simulation studies and were shown to be satisfactory (CRH 89, Richardson and Clifford, 1988). For small spatial domains, the Type I errors of the $t_{\widehat{M}-2}$ test are closer to their nominal value than those of the **W** test. Otherwise the two tests give equivalent results. Their performance will again be illustrated in the following section, which is concerned with a study of the performance of non–parametric tests in cases of spatial autocorrelation.

## 3. Comparison of non–parametric tests of association and the modified tests

A classical non–parametric measure of association between two variables is Spearman's rank correlation $r_S$ which evaluates a correlation coefficient between the ranks of $X_\alpha$ and $Y_\alpha$. It can be calculated as

$$r_S = 1 - \frac{6}{n(n^2-1)} \sum_{\alpha=1}^{N} d_\alpha^2,$$

where $d_\alpha$ is the difference between the ranks of $X_\alpha$ and $Y_\alpha$.

A non–parametric spatial index of association was proposed by Tjøstheim (1978) and generalised by Hubert and Golledge (1982). It is based on the sum of the "distances" between locations of similar ranks for the variables **X** and **Y**, with a suitably chosen distance function. The values $\{X_\alpha, \alpha = 1, \ldots, N\}$ and $\{Y_\alpha, \alpha = 1, \ldots, N\}$ are first ranked. Let $X(i)$ and $Y(i)$ denote respectively their $i^{th}$ rank and let $\{k_X(i), l_X(i)\}$ and $\{k_Y(i), l_Y(i)\}$ denote the respective coordinates of $X(i)$ and $Y(i)$. Supposing that the coordinate system has been centered, Tjøstheim's index **A** can be written as:

$$A = \frac{\sum_{i=1}^{N} \{k_X(i)k_Y(i) + l_X(i)l_Y(i)\}}{\sum_{i=1}^{N} \{k_X(i)^2 + l_X(i)^2\}}.$$

Note that the denominator of **A** is a constant of the coordinate system and does not depend on **X** (or **Y**).

To evaluate the moments of $r_S$ or **A** under the null distribution of no association between **X** and **Y**, it is assumed in both cases that the **N**! permutations of the values of one of the variables, the other staying fixed, are equally likely. For autocorrelated variables **X** and **Y**, this is no longer true and the existing autocorrelations are perturbed by the permutations. As in the case of the Pearson correlation coefficient $r_{XY}$, the variance of $r_S$ and **A** are increased by positive autocorrelation whereas the variances classically used to test under the null hypothesis are assumed to be fixed. This leads to over–significant tests in the case of positive autocorrelation. A Monte Carlo simulation was carried out in order to evaluate this and to compare the power of the different tests.

*Simulation model*

In order to stay close to the examples that will be analysed in the last section, a simulation was carried out on an irregular grid of points defined by the administrative centres of the French *département* ($N = 82$). Spatial dependence was introduced directly on the variance–covariance matrix of the multivariate normal distributions by considering a disc model where the covariance between 2 points is defined as being proportional to the intersection area of 2 discs centered on those points (see §5.3 for a precise definition). The shape of the covariance

function of this model shows a fairly linear decrease with distance. This shape is similar to the observed variograms; *i.e.*, the plot of

$$\mathbf{N_k^{-1}} \sum_{(\alpha,\gamma)\,\varepsilon\,\mathbf{S_k}} (\mathbf{X}_\alpha - \mathbf{X}_\gamma)^2$$

against the average distance between locations in $\mathbf{S_k}$, of a number of variables that will be considered in our examples. The parameter of the disc model is chosen so that the autocorrelation for a distance of 40 km between points is equal to 0.2, 0.4, 0.6 or 0.8. We denote this autocorrelation by $\rho(\mathbf{1})$ indexed by the name of the variable.

For each chosen value of $\rho_{\mathbf{X}}(\mathbf{1})$ an $\mathbf{N}$–by–$\mathbf{N}$ matrix $\mathbf{\Sigma_X}$ is generated, with diagonal elements equal to 1, following the disc model. $\mathbf{\Sigma_X}$ is then triangularised, $\mathbf{\Sigma_X = LL^t}$, and a realisation of $\mathbf{X}$ distributed as $\mathbf{N(0, \Sigma_X)}$ is obtained by first generating a vector of $\mathbf{N}$ i.i.d. $\mathbf{N(0,1)}$ and then pre–multiplying this vector by $L$.

The distances between the centres of *départements* were partitioned into 15 classes of 50 km intervals each. This gives 15 strata $\mathbf{S_1, \ldots, S_{15}}$; the stratum $\mathbf{S_0} = \{(\alpha, \alpha), \alpha\,\varepsilon\,\mathbf{A}\}$. These strata are used in the calculation of $\mathbf{t_{\widehat{M}-2}}$ as defined in §2 formula (4).

### 3.1. Type I error for $\mathbf{r_S}$ and A

Exhibit 1 gives the observed percentages of Type I errors together with their 95% confidence limits for 4 statistics: $\mathbf{t_{\widehat{M}-2}}$, $\mathbf{t_{N-2}}$, $\mathbf{r_S}$ and $\mathbf{A}$, testing at a 5% nominal level the association between $\mathbf{X}$ and $\mathbf{Y}$ under the null hypothesis of independence between $\mathbf{X}$ and $\mathbf{Y}$. Five hundred simulations were done for each combination of $\rho_{\mathbf{X}}(\mathbf{1})$ and $\rho_{\mathbf{Y}}(\mathbf{1})$. The performance of the modified $\mathbf{t_{\widehat{M}-2}}$ statistic is satisfactory as the value 5% belongs to all the confidence intervals and there is no systematic variation with increasing autocorrelation. On the other hand, the Type I errors of the non–adjusted test $\mathbf{t_{N-2}}$ and $\mathbf{r_S}$ are clearly increasing with increasing autocorrelation, reaching values around 30% instead of the nominal 5% in the most highly autocorrelated case. This result had already been observed (CRH 89) and here we see also that, not unexpectedly, the behaviours of $\mathbf{t_{N-2}}$ and $\mathbf{r_S}$ are quite similar since they both use $\mathbf{N-2}$ degrees of freedom.

For Tjøstheim's index $\mathbf{A}$, the influence of the autocorrelation on the significance level is less important. The difference with the nominal 5% level is only clearly apparent in the most highly autocorrelated case ($0.8 \times 0.8$) with Type I errors nearly tripled. We checked the empirical variance of $\mathbf{A}$ in this case and found it to be around 1.6 times its theoretical value. Hence as for $\mathbf{r_S}$, Tjøstheim's index $\mathbf{A}$ leads to over–significant tests but only in the presence of high autocorrelation in both the variables.

### 3.2. Comparison of the powers of the modified $\mathbf{t_{\widehat{M}-2}}$ test and of the index A

The discussion of the comparative performances of Spearman's $\mathbf{r_S}$ and Tjøstheim's index $\mathbf{A}$ in terms of a general cross–product statistic between a measure of spatial proximity and a measure of non–spatial proximity was initiated by Glick (1982), and taken up by Hubert and Golledge (1982) and Upton and Fingleton (1985). This formulation is helpful in highlighting the contrast between the index $\mathbf{A}$ and $\mathbf{r_S}$. The index $\mathbf{A}$ uses a relatively sophisticated measure of spatial proximity, the distances between locations of similar ranks, but a $0-1$ classification of non–spatial proximity in terms of identical ranks. On the other hand, $\mathbf{r_S}$

Exhibit 1.

**Proportion of type I errors and 95 % confidence limits for the modified $t_{\hat{M}-2}$ test, the standard $t_{N-2}$ test, Spearman's $r_S$ and Tjøstheim's index A, for a nominal 5 % test level in the case of two mutually independent processes generated by a disc model on a network of 82 points.**

| $\rho_Y(1)$ | $\rho_X(1)$ | 0.0 | 0.2 | 0.4 | 0.6 | 0.8 |
|---|---|---|---|---|---|---|
| 0.0 | $t_{\hat{M}-2}$ | 3.4 ± 1.6 | | | | |
| | $t_{N-2}$ | 3.2 ± 1.5 | | | | |
| | $r_S$ | 4.4 ± 1.8 | | | | |
| | A | 4.0 ± 1.7 | | | | |
| 0.2 | $t_{\hat{M}-2}$ | 6.0 ± 2.1 | 6.4 ± 2.1 | | | |
| | $t_{N-2}$ | 6.2 ± 2.1 | 7.4 ± 2.3 | | | |
| | $r_S$ | 6.2 ± 2.1 | 5.8 ± 2.1 | | | |
| | A | 4.6 ± 1.8 | 4.8 ± 1.9 | | | |
| 0.4 | $t_{\hat{M}-2}$ | 4.6 ± 1.8 | 5.8 ± 2.1 | 4.4 ± 1.8 | | |
| | $t_{N-2}$ | 5.6 ± 2.0 | 7.0 ± 2.2 | 8.0 ± 2.4 | | |
| | $r_S$ | 6.0 ± 2.1 | 7.0 ± 2.2 | 8.2 ± 2.4 | | |
| | A | 5.4 ± 2.0 | 6.2 ± 2.1 | 5.4 ± 2.0 | | |
| 0.6 | $t_{\hat{M}-2}$ | 5.6 ± 2.0 | 4.4 ± 1.8 | 5.0 ± 1.9 | 4.8 ± 1.9 | |
| | $t_{N-2}$ | 5.8 ± 2.1 | 5.8 ± 2.1 | 9.8 ± 2.6 | 13.6 ± 3.0 | |
| | $r_S$ | 5.0 ± 1.9 | 6.2 ± 2.1 | 9.0 ± 2.5 | 16.8 ± 3.3 | |
| | A | 4.6 ± 1.8 | 5.6 ± 2.0 | 4.2 ± 1.8 | 6.4 ± 2.1 | |
| 0.8 | $t_{\hat{M}-2}$ | 4.4 ± 1.8 | 5.0 ± 1.9 | 4.2 ± 1.8 | 5.2 ± 1.9 | 3.0 ± 1.5 |
| | $t_{N-2}$ | 5.2 ± 1.9 | 7.8 ± 2.4 | 11.4 ± 2.8 | 20.2 ± 3.5 | 35.6 ± 4.2 |
| | $r_S$ | 4.6 ± 1.8 | 9.4 ± 2.6 | 12.0 ± 2.8 | 18.8 ± 3.4 | 34.4 ± 4.2 |
| | A | 5.0 ± 1.9 | 4.2 ± 1.8 | 8.8 ± 2.5 | 6.4 ± 2.1 | 14.0 ± 3.0 |

The parameters $\rho_X(1)$ and  of the disc models for X and Y respectively are equal to the autocorrelations at 40 km. 500 simulations were carried out in each cell.

only considers a $0-1$ measure of spatial proximity (identical locations or not) and a less crude measure of non–spatial proximity, the square of the differences between the ranks. Extension of the index **A** to a statistic that equally involves both types of proximity has been proposed by Hubert and Golledge.

We now discuss the results of a Monte Carlo study aimed at comparing the power of the modified $\mathbf{t}_{\widehat{\mathbf{M}-\mathbf{2}}}$ test and of the index **A** under two sets of alternative hypotheses:

$\mathbf{H_1}$: **Y** and **X** are related by a linear relationship with autocorrelated error term,

$\mathbf{H_2}$: **Y** and **X** are related by a local permutation with a random disturbance,

for different levels of autocorrelation in **X**. The comparison is limited to these two statistics because their significance levels can be controlled and fixed at around 5% (except for **A** in one case of very high autocorrelation). Comparison with $\mathbf{r_S}$ is consequently not possible except in unautocorrelated cases, which are of little interest here. Two contrasting alternative hypotheses were chosen: $\mathbf{H_1}$ in the classical framework of linear regression particularly suited to the modified $\mathbf{t}_{\widehat{\mathbf{M}-\mathbf{2}}}$ test of the correlation coefficient and $\mathbf{H_2}$ as a local rearrangement particularly adapted to the index **A**.

*3.2.1. Observed power of* $\mathbf{t}_{\widehat{\mathbf{M}-\mathbf{2}}}$ *and* **A** *under* $\mathbf{H_1}$

The alternative hypothesis of linear regression between **Y** and **X** was defined as:

$$\mathbf{H_1}: \mathbf{Y} = \mathbf{aX} + \mathbf{W}, \quad \mathbf{X} \sim \mathbf{N}(\mu_{\mathbf{X}}, \mathbf{\Sigma_X}), \quad \mathbf{W} \sim \mathbf{N}(\mu_{\mathbf{W}}, \mathbf{\Sigma_W})$$

and **X** and **W** independent. It is difficult to calculate theoretically the power of the modified test $\mathbf{t}_{\widehat{\mathbf{M}-\mathbf{2}}}$ or that of **A** because their distribution under $\mathbf{H_1}$ is not precisely known. Their power can be assessed by simulations.

Two independent spatially autocorrelated processes **X** and **W** were generated on the grid of the administrative centres of French *départements* as Gaussian variables with a disc model for their autocovariance. Without loss of generality $\sigma^2_{\mathbf{X}} = \sigma^2_{\mathbf{W}}$ was chosen and hence the correlation $\rho_{\mathbf{XY}}$ between **X** and **Y** was only dependent on the parameter **A**. Five hundred trials were carried out for several levels of autocorrelation in **X** and **W** and for the values $\rho_{\mathbf{XY}} = 0.2$ and $0.4$. The grid contained $\mathbf{N = 82}$ points. Results for higher values of $\rho_{\mathbf{XY}}$ are not reported because the power of the $\mathbf{t}_{\widehat{\mathbf{M}-\mathbf{2}}}$ was very close to 1. In these simulations a 5% nominal level was chosen. The results are presented in Exhibit 2.

With respect to $\mathbf{t}_{\widehat{\mathbf{M}-\mathbf{2}}}$, one can see that the observed power decreases when there is strong autocorrelation in the **X** or the **W** variable. This is to be expected since the variance of $\mathbf{r_{XY}}$ increases and the effective degrees of freedom diminish with increasing autocorrelation. In Richardson and Clifford (1988), the power of $\mathbf{t}_{\widehat{\mathbf{M}-\mathbf{2}}}$ was shown to be quite close to a reference value given by the power of the classical $t$–test based on a number of observations compatible with the observed variance of $\mathbf{r_{XY}}$. The observed power of the **W** statistic is not shown because it is nearly identical to that of $\mathbf{t}_{\widehat{\mathbf{M}-\mathbf{2}}}$. A theoretical approximation of the power of **W** was also given in this earlier paper.

On the other hand, the power of Tjøstheim's index **A** under $\mathbf{H_1}$ is very low. Even in the case $\rho_{\mathbf{XY}} = 0.4$ where the power of $\mathbf{t}_{\widehat{\mathbf{M}-\mathbf{2}}}$ is in most cases over 80%, that of the index **A** does not exceed its significance level. Indeed it is around 5% in most cases and only reaches

**Exhibit 2.**

(continued)

**Observed power and 95 % confidence intervals for the modified $t_{\hat{M}\text{-}2}$ test and Tjostheim's index A under an alternative hypothesis of linear model : Y = aX + W where both X and W follow a disc model, X and W independent and of equal variance and a is chosen so that the correlation $\rho_{XY}$ between X and Y takes the value 0.2 or 0.4.**

| | | $\rho_X(1) = 0$ | | $\rho_X(1) = 0.2$ | | $\rho_X(1) = 0.4$ | | $\rho_X(1) = 0.6$ | | $\rho_X(1) = 0.8$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\rho_W(1)$ | | $\rho_{XY} = 0.2$ | $\rho_{XY} = 0.4$ | $\rho_{XY} = 0.2$ | $\rho_{XY} = 0.4$ | $\rho_{XY} = 0.2$ | $\rho_{XY} = 0.4$ | $\rho_{XY} = 0.2$ | $\rho_{XY} = 0.4$ | $\rho_{XY} = 0.2$ | $\rho_{XY} = 0.4$ |
| 0.4 | $t_{\hat{M}\text{-}2}$ | **44.6** [40.2-49.0] | **95.4** [93.6 - 97.2] | **36.8** [32.6-41.0] | **93.0** [90.8-95.2] | **37.6** [33.4-41.8] | **91.0** [88.5-93.5] | **34.4** [30.2-38.6] | **88.6** [85.8-91.4] | **21.4** [17.8-25.0] | **79.2** [75.6-82.8] |
| | **A** | **5.4** [3.4-7.4] | **6.0** [3.9 - 8.1] | **6.4** [4.3-8.5] | **5.6** [3.6- 7.6] | **5.6** [3.6 - 7.6] | **7.4** [5.1 - 9.7] | **5.2** [3.3 - 7.1] | **5.4** [3.4 - 7.4] | **7.6** [5.3 - 9.9] | **11.0** [8.3 - 13.7] |
| 0.6 | $t_{\hat{M}\text{-}2}$ | **48.4** [44.0-52.8] | **96.4** [94.8-98.0] | **36.4** [32.2-40.6] | **94.8** [92.9-96.7] | **36.2** [32.0-40.4] | **93.0** [90.8-95.2] | **28.0** [24.1-31.9] | **81.0** [77.6-84.4] | **20.0** [16.5-23.5] | **56.4** [52.1-60.7] |
| | **A** | **3.8** [2.1-5.5] | **3.8** [2.1 - 5.5] | **4.4** [2.6-6.2] | **4.4** [2.6- 6.2] | **4.6** [2.8 - 6.4] | **5.0** [3.1 - 6.9] | **4.4** [2.6- 6.2] | **6.0** [3.9 - 8.1] | **9.2** [6.7 - 11.7] | **11.0** [8.3 - 13.7] |
| 0.8 | $t_{\hat{M}\text{-}2}$ | **52.4** [48.0-56.8] | **96.8** [95.3 - 98.3] | **48.4** [44.0-52.8] | **93.4** [91.2-95.6] | **42.4** [38.1-46.7] | **89.4** [86.7-92.1] | **28.6** [24.6-32.6] | **76** [72.3-79.7] | **12.0** [9.2 - 14.8] | **41.8** [37.5-46.1] |
| | **A** | **5.0** [3.1-6.9] | **5.4** [3.4 - 7.4] | **6.4** [4.3-8.5] | **4.2** [2.4- 6.0] | **7.2** [ 4.9- 9.5] | **4.8** [2.9- 6.7] | **4.6** [2.8 - 6.4] | **7.4** [5.1 - 9.7] | **13.0** [10.1-15.9] | **15.4** [12.2-18.6] |

The parameters $\rho_X(1)$ and $\rho_W(1)$ of the disc models for X and W are equal to the autocorrelation at 40 km. 500 simulations were carried out in each cell.

**Exhibit 2.**

**Observed power and 95 % confidence intervals for the modified $t_{\hat{M}\cdot 2}$ test and Tjøstheim's index A under an alternative hypothesis of linear model : Y = aX + W where both X and W follow a disc model, X and W independent and of equal variance and a is chosen so that the correlation $\rho_{XY}$ between X and Y takes the value 0.2 or 0.4.**

| | | $\rho_X(1) = 0$ | | $\rho_X(1) = 0.2$ | | $\rho_X(1) = 0.4$ | | $\rho_X(1) = 0.6$ | | $\rho_X(1) = 0.8$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\rho_W(1)$ | | $\rho_{XY} = 0.2$ | $\rho_{XY} = 0.4$ | $\rho_{XY} = 0.2$ | $\rho_{XY} = 0.4$ | $\rho_{XY} = 0.2$ | $\rho_{XY} = 0.4$ | $\rho_{XY} = 0.2$ | $\rho_{XY} = 0.4$ | $\rho_{XY} = 0.2$ | $\rho_{XY} = 0.4$ |
| 0.0 | $t_{\hat{M}\cdot 2}$ | **42.4** [38.1-46.7] | **96.6** [95 - 98.2] | **44.2** [39.8-48.6] | **96.0** [94.3-97.7] | **43.2** [38.9-47.5] | **96.6** [95.0-98.2] | **39.2** [34.9-43.5] | **92.2** [89.8-94.6] | **35.8** [31.6 - 40.0] | **87.2** [84.3-90.1] |
| | **A** | **5.8** [3.8-7.8] | **5.6** [3.6 - 7.6] | **7.4** [5.1-9.7] | **6.0** [3.9- 8.1] | **5.8** [3.8 - 7.8] | **5.0** [3.1 - 6.9] | **4.8** [2.9-6.7] | **6.8** [4.6 - 9.0] | **7.6** [5.3 - 9.9] | **10.0** [7.4 - 12.6] |
| 0.2 | $t_{\hat{M}\cdot 2}$ | **42.4** [38.1-46.7] | **97.4** [96.0 - 98.8] | **42.2** [37.9-46.5] | **95.6** [93.8-97.4] | **39.0** [34.7-43.3] | **94.0** [91.9-96.1] | **34.6** [30.4-38.8] | **9.2** [89.6-94.4] | **30.4** [26.4 - 34.4] | **83.6** [80.4-86.8] |
| | **A** | **4.8** [2.9-6.7] | **5.4** [3.4 - 7.4] | **6.2** [4.1-8.3] | **6.8** [4.6- 9.0] | **3.8** [2.1 - 5.5] | **5.6** [3.6 - 7.6] | **4.2** [2.4 - 6.0] | **5.4** [3.4 - 7.4] | **6.2** [4.1 - 8.3] | **9.4** [6.8 - 12.0] |

The parameters $\rho_X(1)$ and $\rho_W(1)$ of the disc models for X and W are equal to the autocorrelation at 40 km. 500 simulations were carried out in each cell.

10 to 15% in cases of high autocorrelation in $\mathbf{X}$, where results reported in Exhibit 1 show that its significance level also increases. This is the reason for the apparent increase of power of the index $\mathbf{A}$ with autocorrelation that has also been noted by Glick (1982), although he did not relate it to Type I errors.

Hence, as expected from the construction of the two statistics $\mathbf{t}_{\widehat{\mathbf{M}}-\mathbf{2}}$ and $\mathbf{A}$, $\mathbf{t}_{\widehat{\mathbf{M}}-\mathbf{2}}$ has good power in the regression framework whilst the index $\mathbf{A}$ is clearly unable to recognize that type of association if there is any random error.

*3.2.2. Observed power of* $\mathbf{t}_{\widehat{\mathbf{M}}-\mathbf{2}}$ *and* $\mathbf{A}$ *under* $\mathbf{H_2}$

The second alternative hypothesis was constructed as follows. The 82 *départements* were separated into 16 groups of contiguous *départements* (composed of from 4 to 7 *départements* to balance their total surface area). The process $\mathbf{X}$ was generated with unit variance following a disc model, and then its values were randomly permuted within each of the 16 groups. This created a variable $\widetilde{\mathbf{X}}$ deduced from $\mathbf{X}$ by a local permutation. The variable $\mathbf{Y}$ was then defined as: $\mathbf{Y} = \widetilde{\mathbf{X}} + \mathbf{dW}$, where $\mathbf{W}$ are i.i.d. $\mathbf{N(0,1)}$ variables and $\mathbf{0 \leq d \leq 1}$. Five hundred simulations were carried out for several values of $\rho_{\mathbf{X}}(\mathbf{1})$ and $\mathbf{d}$. The results are presented in Exhibit 3. Several observations can be made:

(a) When there is no autocorrelation, the variable $\widetilde{\mathbf{X}}$ is considered as independent of $\mathbf{X}$ by the statistic $\mathbf{t}_{\widehat{\mathbf{M}}-\mathbf{2}}$, which only looks at association at the same site. The index $\mathbf{A}$ has maximum power when $\mathbf{d = 0}$, but its power decreases dramatically as soon as there is a random disturbance, even of small variance. For instance, it reaches 1/4 of its value when the variance of $\mathbf{dW}$ is $\mathbf{(1/4)^2}$ of the variance of $\widetilde{\mathbf{X}}$.

(b) When $\mathbf{X}$ is autocorrelated, then the power of $\mathbf{t}_{\widehat{\mathbf{M}}-\mathbf{2}}$ is substantially increased, being not too far from that of $\mathbf{A}$ when $\mathbf{d = 0}$ and $\rho_{\mathbf{X}}(\mathbf{1}) = 0.8$ and always higher than $\mathbf{A}$ even when $\mathbf{d}$ is small.

In conclusion, it is apparent that even under an alternative hypothesis well adapted to the index $\mathbf{A}$, the power of $\mathbf{A}$ is weak as soon as a more realistic situation including random disturbances is analysed. This considerably limits the interest of using the index $\mathbf{A}$ to detect a spatial shift. On the other hand, when there was autocorrelation in the $\mathbf{X}$ variable, the statistic $\mathbf{t}_{\widehat{\mathbf{M}}-\mathbf{2}}$ had a reasonable power to detect a local spatial rearrangement. One could conjecture that an extension of the $\mathbf{t}_{\widehat{\mathbf{M}}-\mathbf{2}}$ to *spatially lagged cross-correlations* (*i. e.*, between $\mathbf{X}_\alpha$ and $\mathbf{Y}_\beta$, $\beta$ in a neighbourhood of $\alpha$) would prove more powerful.

## 4. Robustness of the modified tests of association

The construction of the modified tests as well as the study of their performance was done under Gaussian hypotheses. In this section we investigate some aspects of the robustness of these tests to departure from normality. Three types of departure from normality will be considered, namely

(a) truncated Gaussian variables

(b) lognormal variables

(c) mixtures of Gaussian variables.

For each of these cases the Type I errors of $\mathbf{t}_{\widehat{\mathbf{M}}-\mathbf{2}}$, $\mathbf{t}_{\mathbf{N}-\mathbf{2}}$, and $\mathbf{A}$ are evaluated by Monte Carlo simulations with 500 trials and at a 5% nominal level of significance. In (*a*) mutually

**Exhibit 3.**

**Observed power and 95 % confidence limits for the modified $t_{\hat{M}-2}$ test and Tjøstheim's index A under an alternative hypothesis of linear model : $Y = \tilde{X} + dW$ (see §3.2.2) where X is generated by a disc model and W are i.i.d N(0,1).**

| | $\rho_X(1)$ | 0.0 | 0.2 | 0.4 | 0.6 | 0.8 |
|---|---|---|---|---|---|---|
| | $\hat{C}_Y(k)$ | - 0.02 | 0.01 | 0.05 | 0.13 | 0.33 |
| d=0.0 | $t_{\hat{M}-2}$ | 5.6 ± 2.0 | 6.6 ± 2.2 | 10.6 ± 2.7 | 33.2 ± 4.1 | 79.2 ± 3.6 |
| | A | 100 | 100 | 100 | 100 | 100 |
| | $\hat{C}_Y(k)$ | - 0.02 | 0.01 | 0.04 | 0.12 | 0.32 |
| d=0.125 | $t_{\hat{M}-2}$ | 5.2 ± 1.9 | 6.0 ± 2.1 | 10.6 ± 2.7 | 32.4 ± 4.1 | 80.2 ± 3.5 |
| | A | 52.2 ± 4.4 | 54.6 ± 4.4 | 54.8 ± 4.4 | 61.0 ± 4.3 | 73.8 ± 3.9 |
| | $\hat{C}_Y(k)$ | - 0.02 | 0.01 | 0.04 | 0.10 | 0.30 |
| d=0.25 | $t_{\hat{M}-2}$ | 5.8 ± 2.0 | 5.6 ± 2.0 | 10.8 ± 2.7 | 32.8 ± 4.1 | 78.4 ± 3.6 |
| | A | 25.4 ± 3.8 | 22.6 ± 3.7 | 26.8 ± 3.9 | 26.6 ± 3.9 | 48.8 ± 4.4 |
| | $\hat{C}_Y(k)$ | - 0.02 | 0.0 | 0.03 | 0.09 | 0.24 |
| d=0.5 | $t_{\hat{M}-2}$ | 5.0 ± 1.9 | 6.2 ± 2.1 | 9.8 ± 2.6 | 28.2 ± 3.9 | 72.6 ± 3.9 |
| | A | 10.0 ± 2.6 | 12.2 ± 2.9 | 10.4 ± 2.7 | 16.8 ± 3.3 | 38.4 ± 4.3 |
| | $\hat{C}_Y(k)$ | - 0.02 | - 0.01 | 0.01 | 0.05 | 0.14 |
| d =1 | $t_{\hat{M}-2}$ | 5.8 ± 2.0 | 5.6 ± 2.0 | 8.6 ± 2.5 | 18.2 ± 3.4 | 51.2 ± 4.4 |
| | A | 4.6 ± 1.8 | 7.2 ± 2.3 | 7.8 ± 2.4 | 9.4 ± 2.6 | 22.0 ± 3.6 |

The parameters $\rho_X(1)$ and $\rho_W(1)$ of the disc models for X and W are equal to the autocorrelation at 40 km and $\hat{C}_Y(1)$ is the overage observed autocorrelation of Y in the first strata $S_1$ defined in §3. 500 simulations were carried out in each cell.

independent variables $\mathbf{X}$ and $\mathbf{Y}$ were generated following a disc model, an then their values were set equal to the chosen truncation limits if they exceeded these limits. In (b) mutually independent autocorrelated variables $\mathbf{U}$ and $\mathbf{V}$ were first generated by a disc model based upon i.i.d. $\mathbf{N(0, \ln(2))}$ variables [rather than $\mathbf{N(0,1)}$], and then $\mathbf{X}$ and $\mathbf{Y}$ were defined as follows:

$$\mathbf{X} = [\ \mathbf{\exp(U)/\sqrt{2}}] - \mathbf{1},$$

and

$$\mathbf{Y} = [\ \mathbf{\exp(V)/\sqrt{2}}] - \mathbf{1}.$$

This definition implies that $\mathbf{X}$ and $\mathbf{Y}$ follow centered lognormal distributions, with unit variance. The correlation between $\mathbf{X}_\alpha$ and $\mathbf{X}_\beta$ is equal to $2^{\rho_U(\alpha,\beta)} - 1$, with

$$\rho_{\mathbf{U}}(\alpha, \beta) = \mathbf{E}(\mathbf{U}_\alpha \mathbf{U}_\beta).$$

In (c) mutually independent autocorrelated variables $\mathbf{U}$ and $\mathbf{V}$ were first generated by a disc model, and then the variable $\mathbf{X}$ was defined as equaling $\mathbf{U}$ with probability $\gamma$ and $\mathbf{bU}$ with probability $\mathbf{1-\gamma}$, $\mathbf{0 \le \gamma \le 1}$; $\mathbf{Y}$ was defined similarly with respect to $\mathbf{V}$. The variance of $\mathbf{X}$ thus was equal to $[\gamma + \mathbf{b^2}(\mathbf{1-\gamma})]$, and the correlation between $\mathbf{X}_\alpha$ and $\mathbf{X}_\beta$ was equal to:

$$\frac{[\gamma + \mathbf{b}(\mathbf{1-\gamma})]^{\mathbf{2}}}{[\gamma + \mathbf{b^2}(\mathbf{1-\gamma})]} \cdot \rho_{\mathbf{U}}(\alpha, \beta).$$

Results are presented in Exhibit 4 for several symmetric truncation levels, in Exhibit 5 for the lognormal case, and in Exhibit 6 for several mixture coefficients $\gamma$ and $b = 3$. The observed Type I errors are, on the whole, not much different from those appearing in Exhibit 1. For $\mathbf{t}_{\widehat{\mathbf{M}}-\mathbf{2}}$ the significance levels are not altered in the lognormal case (b), but there is a slight tendency toward over–conservative levels in cases of higher autocorrelations with some Type I errors being below 5% in cases (a) or (c). The results are nearly identical for W. Non–symmetric truncations also were tried and the results were found to be similar.

In case (c), mixtures with $b = 6$, 9 or 12 also were simulated for highly autocorrelated $\mathbf{U}$ and $\mathbf{V}$ processes (see Exhibit 7). The tendency of $\mathbf{t}_{\widehat{\mathbf{M}}-\mathbf{2}}$ to be over–conservative again was apparent, with Type I errors around 2% in nearly all instances. Note that when the coefficient $\mathbf{b}$ is increased, the internal autocorrelation is decreased, with this balancing effect rendering comparisons awkward; other combinations of $\mathbf{b}$ and $\gamma$ appear to be less informative. For $\mathbf{t}_{\mathbf{N}-\mathbf{2}}$ and the index $\mathbf{A}$, the results are comparable to those presented in Exhibit 1, with high autocorrelation resulting in inflated significance levels.

In conclusion, the $\mathbf{t}_{\widehat{\mathbf{M}}-\mathbf{2}}$ test is shown to be quite robust to small departures from normality in terms of its significance level, with a tendency to give over–conservative test results in situations of high autocorrelation. It would be interesting to develop a family of permutation tests based upon a restricted set of permutations that would preserve some aspect of the spatial structure and to compare the performance of those tests to that of the modified $\mathbf{W}$ and $\mathbf{t}_{\widehat{\mathbf{M}}-\mathbf{2}}$ tests.

**Exhibit 4.**

**Proportion of type I errors and 95 % confidence limits for the modified $t_{\hat{M}-2}$ test and Tjøstheim's index A, for a nominal 5 % test level in the case of two mutually independent processes generated by a disc model on a network of 82 points and truncated at chosen limits.**

| Truncations limits | | $\rho_Y(1) = \rho_X(1)$ | | |
|---|---|---|---|---|
| | | 0.0 | 0.4 | 0.8 |
| [- 1.64 ; 1.64] | $t_{\hat{M}-2}$ | 4.8 ± 1.9 | 5.8 ± 2.0 | 3.2 ± 1.5 |
| | $t_{N-2}$ | 4.2 ± 1.8 | 8.2 ± 2.4 | 36.4 ± 4.2 |
| | A | 5.4 ± 2.0 | 6.4 ± 2.1 | 8.8 ± 2.5 |
| [- 1.28 ; 1.28] | $t_{\hat{M}-2}$ | 4.4 ±1.8 | 5.0 ± 1.9 | 3.0 ± 1.5 |
| | $t_{N-2}$ | 4.6 ± 1.8 | 8.2 ± 2.4 | 35.6 ± 4.2 |
| | A | 8.4 ± 2.4 | 8.6 ± 2.5 | 10.8 ± 2.7 |
| [- 1.04 ; 1.04] | $t_{\hat{M}-2}$ | 4.6 ±1.8 | 5.4 ± 2.0 | 3.4 ± 1.6 |
| | $t_{N-2}$ | 4.2 ± 1.8 | 8.0 ± 2.4 | 34.0 ± 4.2 |
| | A | 7.2 ± 2.3 | 8.8 ± 2.5 | 7.8 ± 2.4 |
| [- 0.84 ; 0.84] | $t_{\hat{M}-2}$ | 4.6 ±1.8 | 4.6 ± 1.8 | 3.6 ± 1.6 |
| | $t_{N-2}$ | 4.6 ± 1.8 | 7.8 ± 2.4 | 31.6 ± 4.1 |
| | A | 10.2 ± 2.7 | 7.8 ± 2.4 | 10.0 ± 2.6 |

The parameters $\rho_X(1)$ and $\rho_Y(1)$ of the disc models for X and Y are equal to the autocorrelations at 40 km. 500 simulations were carried out in each cell.

Exhibit 5.

**Proportion of type I errors and 95 % confidence intervals for the modified $t_{\hat{M}\text{-}2}$ test, the standard $t_{N\text{-}2}$ test and Tjøstheim's index A, for a nominal 5 % test level in the case of two mutually independent lognormal processes X and Y generated from processes U and V following a disc model on a network of 82 points.**

| $\rho_V(1) = \rho_U(1)$ | 0.0 | 0.2 | 0.4 | 0.6 | 0.8 |
|---|---|---|---|---|---|
| $2\rho(1) - 1$ | 0 | 0.15 | 0.32 | 0.52 | 0.74 |
| $t_{\hat{M}\text{-}2}$ | **5.2** [3.3 - 7.1] | **5.6** [3.6 - 7.6] | **6.2** [4.1 - 8.3] | **5.8** [3.8 - 7.8] | **5.2** [3.3 - 7.1] |
| $t_{N\text{-}2}$ | **5.0** [3.1 - 6.9] | **6.0** [3.9 - 8.1] | **7.2** [4.9 - 9.5] | **11.6** [8.8 - 14.4] | **29.4** [25.4 - 33.4] |
| A | **4.4** [2.6 - 6.2] | **6.2** [4.1 - 8.3] | **5.8** [3.8 - 7.8] | **6.0** [3.9 - 8.1] | **11.2** [8.4 - 14] |

The parameters $\rho_U(1)$ and $\rho_V(1)$ of the disc models for U and V respectively are equal to the autocorrelations at 40 km. The resulting autocorrelation $\rho_X(1)$ the process in X is equal to $2\rho(1) - 1$ and similarly to Y. 500 simulations were carried out in each cell.

Exhibit 6.

**Proportion of type I errors and 95 % confidence intervals for the modified $t_{\hat{M}\text{-}2}$ test, the standard $t_{N\text{-}2}$ test and Tjøstheim's index A, for a nominal 5 % test level in the case of two mutually independent processes X and Y, each generated from a mixture of disc processes U with probability $\gamma$ and 3U with probability 1-$\gamma$.**

| | $\rho_U(1)=\rho_V(1)$ | 0.0 | 0.2 | 0.4 | 0.6 | 0.8 |
|---|---|---|---|---|---|---|
| | $\rho_X(1)=\rho_Y(1)$ | 0.0 | 0.16 | 0.32 | 0.48 | 0.64 |
| $\gamma$=0.9 | $t_{\hat{M}\text{-}2}$ | **3.2** [1.7 - 4.5] | **3.4** [1.8 - 5.0] | **2.2** [0.9 - 3.5] | **3.0** [1.5 - 4.5] | **2.2** [0.9 - 3.5] |
| | $t_{N\text{-}2}$ | **3.6** [2.0 - 5.2] | **4.0** [2.3 - 5.7] | **6.0** [3.9 - 8.1] | **10.2** [7.5 - 12.9] | **34.2** [30.0 - 38.4] |
| | A | **5.8** [3.8 - 7.8] | **5.4** [3.4 - 7.4] | **3.4** [1.8 - 5.0] | **5.0** [3.1 - 6.9] | **10.8** [8.1 - 13.5] |
| | $\rho_X(1)=\rho_Y(1)$ | 0.0 | 0.15 | 0.30 | 0.45 | 0.60 |
| $\gamma$=0.8 | $t_{\hat{M}\text{-}2}$ | **4.8** [2.9 - 6.7] | **3.6** [2.0 - 5.2] | **4.2** [2.4 - 6.0] | **2.8** [1.4 - 4.2] | **3.4** [1.8 - 5.0] |
| | $t_{N\text{-}2}$ | **5.0** [3.1 - 6.9] | **4.8** [2.9 - 6.7] | **6.8** [4.6 - 9.0] | **13.4** [10.4 - 16.4] | **36.8** [32.6 - 41.0] |
| | A | **6.4** [4.3 - 8.5] | **4.2** [2.4 - 6.0] | **4.4** [2.6 - 6.2] | **5.4** [3.4 - 7.4] | **11.4** [8.6 - 14.2] |
| | $\rho_X(1)=\rho_Y(1)$ | 0.0 | 0.15 | 0.30 | 0.45 | 0.60 |
| $\gamma$=0.7 | $t_{\hat{M}\text{-}2}$ | **4.6** [2.8 - 6.4] | **3.8** [2.1 - 5.5] | **4.2** [2.4 - 6.0] | **5.0** [3.1 - 6.9] | **3.4** [1.8 - 5.0] |
| | $t_{N\text{-}2}$ | **5.0** [3.1 - 6.9] | **4.4** [2.6 - 6.2] | **6.8** [4.6 - 9.0] | **14.6** [11.5 - 17.7] | **35.6** [31.4 - 39.8] |
| | A | **7.4** [5.1 - 9.7] | **2.8** [1.4 - 4.2] | **5.0** [3.1 - 6.9] | **5.4** [3.4 - 7.4] | **11.8** [9.0 - 14.6] |

The parameters $\rho_U(1)$ and $\rho_V(1)$ of the disc models for U and V respectively are equal to the autocorrelation at 40 km.

The resulting autocorrelation $\rho_X(1)$ in the processes X is equal to $(9\text{-}8\gamma)^{-1}\ (3 - 2\gamma)^2\ \rho_U(1)$ and similarly for $\rho_Y(1)$. 500 simulations were carried out in each cell.

Exhibit 7.

**Proportion of type I errors and 95 % confidence intervals for the modified $t_{\hat{M}-2}$ test, the standard $t_{N-2}$ test and Tjøstheim's index A, for a nominal 5 % test level in the case of two mutually independent processes X and Y each generated from a mixture of disc processes U with probability $\gamma$ and bU with probability $1-\gamma$.**

| | | $\rho_U(1) = \rho_V(1) = 0.8$ | | | |
|---|---|---|---|---|---|
| | b | 3 | 6 | 9 | 12 |
| | $\rho_X(1) = \rho_Y(1)$ | 0.77 | 0.65 | 0.52 | 0.41 |
| $\gamma$=0.99 | $t_{\hat{M}-2}$ | **1.8** [0.6 - 3.0] | **1.8** [0.6 - 3.0] | **1.8** [0.6 - 3.0] | **1.8** [0.6 - 3.0] |
| | $t_{N-2}$ | **32.8** [28.7 - 36.9] | **32.8** [28.7 - 36.9] | **32.8** [28.7 - 36.9] | **32.6** [28.5 - 36.7] |
| | A | **11.0** [8.3 - 13.7] | **11.2** [8.4 - 14] | **10.8** [8.1 - 13.5] | **10.6** [7.9 - 13.3] |
| | $\rho_X(1) = \rho_Y(1)$ | 0.69 | 0.45 | 0.31 | 0.24 |
| $\gamma$=0.95 | $t_{\hat{M}-2}$ | **2.0** [0.8 - 3.2] | **1.8** [0.6 - 3.0] | **1.8** [0.6 - 3.0] | **2.2** [0.9 - 3.5] |
| | $t_{N-2}$ | **35.2** [31.0 - 39.4] | **33.6** [29.5 - 37.7] | **33.4** [29.3 - 37.5] | **34.0** [29.8 - 38.2] |
| | A | **9.0** [6.5 - 11.5] | **14.2** [11.1 - 17.3] | **9.6** [7.0 - 12.2] | **11.0** [8.3 - 13.7] |
| | $\rho_X(1) = \rho_Y(1)$ | 0.64 | 0.4 | 0.29 | 0.23 |
| $\gamma$ =0.9 | $t_{\hat{M}-2}$ | **2.2** [0.9 - 3.5] | **2.8** [1.4 - 4.2] | **3.2** [1.7 - 4.7] | **3.8** [2.1 - 5.5] |
| | $t_{N-2}$ | **34.2** [30.0 - 38.4] | **36.2** [32.0 - 40.4] | **34.2** [30.0 - 38.4] | **33.0** [28.9 - 37.1] |
| | A | **10.8** [8.1 - 13.5] | **11.0** [8.3 - 13.7] | **9.8** [7.2 - 12.4] | **9.6** [7.0 - 12.2] |

The parameters $\rho_U(1)$ and $\rho_V(1)$ of the disc models for U and V respectively are equal to the autocorrelation at 40 km.

The resulting autocorrelation $\rho_X(1)$ in the processes X is equal to :

$[\gamma + b^2(1-\gamma)]^{-1} [\gamma + b(1-\gamma)]^2 \cdot \rho_U(1)$ and similarly for $\rho_Y(1)$. 500 simulations were carried out in each cell.

## 5. Partial correlations and multiple regressions

In the Gaussian framework the modified tests of association can be extended to test partial correlations. Alternatively, regression models can be specified and estimated, in which case spatial autocorrelation should be included in the variance–covariance matrix of the residuals, if necessary. We have thought it interesting to compare the results of these two different approaches on some examples. For this purpose we shall first describe how the modified $\mathbf{t}_{\widehat{\mathbf{M}}-\mathbf{2}}$ test is extended and indicate some results on its performance (§5.1); then we shall discuss the regression approach (§5.2); finally, we shall give some comparative results based upon examples concerning the relationship between male lung cancer mortality, smoking and industrial exposure analysed at the geographical level of the French *départements* (§5.3).

### 5.1. Extension of the modified $\mathbf{t}_{\widehat{\mathbf{M}}-\mathbf{2}}$ statistic to the testing of partial correlations.

This extension was detailed in Richardson (1989) and we shall now briefly describe its main features. For the sake of clarity the method is going to be described for testing the association between two variables $(\mathbf{Y}_\alpha, \mathbf{Z}_\alpha)$ adjusted for a third one, $\mathbf{X}_\alpha, \alpha \varepsilon \mathbf{A}$. Its generalisation to any number of adjustment variables is straightforward.

We suppose that the **3N** vector $\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \\ \mathbf{Z} \end{pmatrix}$ follows a multivariate normal distribution. Then the joint distribution of $(\mathbf{Y}, \mathbf{Z})$ conditional on $\mathbf{X}$ is also multivariate normal. Therefore, we can test the following hypothesis: $\rho_{\mathbf{YZ}\,.\,\mathbf{X}} = \mathbf{0}$, where $\rho_{\mathbf{YZ}\,.\,\mathbf{X}}$ is the partial correlation between $\mathbf{Y}$ and $\mathbf{Z}$ conditional on $\mathbf{X}$, by testing that the correlation between the residuals of the regressions of $\mathbf{Y}$ on $\mathbf{X}$ and of $\mathbf{Z}$ on $\mathbf{X}$ is zero. Hence, the method outlined in §2.1 and 2.2 equations (1) through (4) can be extended to test a partial correlation coefficient.

In practice this leads to using the modified $\mathbf{t}_{\widehat{\mathbf{M}}-\mathbf{2}}$ statistic on the residuals of the linear regression of $\mathbf{Y}$ on $\mathbf{X}$ and of $\mathbf{Z}$ on $\mathbf{X}$, respectively. These residuals are estimated by ordinary least squares (OLS), since the OLS regression estimates are unbiased.

In summary, the following steps are followed:

Step 1: regress $\mathbf{Y}$ on $\mathbf{X}$ by OLS, giving estimated residuals $\widehat{\mathbf{U}}$,

Step 2: regress $\mathbf{Z}$ on $\mathbf{X}$ by OLS, giving estimated residuals $\widehat{\mathbf{V}}$,

Step 3: test the correlation coefficient between $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$ using the modified test statistic $\mathbf{t}_{\widehat{\mathbf{M}}-\mathbf{2}}$ given in formula (4), §2.2 with

$$\widehat{\mathbf{M}} = \widehat{[\mathbf{var}(\mathbf{r}_{\widehat{\mathbf{U}},\widehat{\mathbf{V}}})]^{-1}} + \mathbf{1}.$$

Thus, the degrees of freedom are adjusted with respect to the spatial autocorrelation in the conditional distributions.

Simulation results reported in Richardson (1989) indicate that the performance of this extended $\mathbf{t}_{\widehat{\mathbf{M}}-\mathbf{2}}$ test is satisfactory with regard both to significance levels and to power. Indeed, its observed power was close to that of a classical test of partial correlation based upon a number of observations compatible with the observed variance of $\mathbf{r}_{\mathbf{YZ}\,.\,\mathbf{X}}$, the empirical partial correlation coefficient.

### 5.2. Regression with a spatially parametrised variance–covariance error matrix.

In the regression models involving spatially distributed variables, the spatial structure can be taken into account by allowing spatial autocorrelation in the error variable; in other words, considering models of the form (where the subscript indicates the dimension of the matrix):

$$\mathbf{Y_{n\times 1}} = \mathbf{X_{n\times p}}\,\beta_{\mathbf{p}\times 1} + \mathbf{U_{n\times 1}}, \tag{5}$$

with $\mathbf{U}$ following a multinormal distribution $\mathbf{N(0, \Sigma_U)}$.

In the classical framework of independent errors distributed as $\mathbf{N(0, \sigma^2 I)}$, tests of a regression coefficient $\beta_i$, $\mathbf{1 \le i \le p}$, and tests of the partial correlation between $\mathbf{Y}$ and $\mathbf{X_i}$ conditional on $\{\mathbf{X_j}, \mathbf{j \neq i}\}$ are equivalent. In the more general framework of (5), tests of the coefficients $\beta_i$ are done *conditionally* on an estimated structure for $\mathbf{\Sigma_U}$, since the matrix $\mathbf{\Sigma_U}$ is only known for theoretical cases.

Different approaches to the modelling of $\mathbf{\Sigma_U}$ have been suggested. They basically follow three lines:

(i) assuming a specific parametric model for $\mathbf{U}$,

(ii) assuming a known functional form for $\mathbf{\Sigma_U}$, and

(iii) direct estimation of $\mathbf{\Sigma_U}$.

Estimation in the framework of (i) was first discussed by Ord (1975), who considered a specific autoregressive model for $\mathbf{U}$ of the form:

$$\mathbf{U = cWU} + \varepsilon, \qquad \varepsilon \sim \mathbf{N(0, \sigma^2 I)}, \tag{6}$$

where $\mathbf{W}$ is a known matrix of weights representing contiguity between the spatial locations. Examples of the fitting of models (5) and (6) are given in Bodson and Peeters (1975), Cliff and Ord (1981), Doreian (1981), Bivand (1984), and Haining (1987). The regression model defined by equations (5) and (6) is analogous in the spatial context to the commonly used time series method of ARMA modelling of residual errors in regression, which has been extensively used (see Glasbey, 1988, for a recent discussion of this topic).

For the alternative approach (ii), varied functional forms for $\mathbf{\Sigma_U}$ have been considered. Ripley (1981) discusses classes of spatial covariance functions that ensure $\mathbf{\Sigma_U}$ always is non–negative definite. These include the family of functions first introduced by Whittle (1954) in which the covariance at $\mathbf{r}$ is proportional to $\mathbf{r^v K_v(ar)}$, $\mathbf{v > 0}$, where $\mathbf{r}$ denotes the distance between points in $\mathbf{R^2}$, and $\mathbf{K_v}$ are the modified Bessel functions of the second kind. Setting $\mathbf{v = 1/2}$ gives an exponential correlation function depending upon only one parameter. Cook and Pocock (1983), in their study of the association between water hardness and cardiovascular deaths, used a more general form of the exponential decline function that depended upon two coefficients.

Including the possibility of anisotropy, Vecchia (1988) considered a general form for a rational spectral density function of a two–dimensional process, whose covariance can be expressed in terms of derivatives of the Bessel function $\mathbf{K_0(ar)}$. This family of covariances overlaps partially with those introduced by Whittle.

There is another interesting family of functions for $\mathbf{\Sigma_U}$ in which the covariance between two points is defined as being proportional to the intersection area of two discs of common

radius centered on those points. This family, often referred to as the disc model, was selected for simulating non–lattice autocorrelated processes in our work on the modified tests.

Other functional forms for $\mathbf{\Sigma_U}$ that do not necessarily ensure that $\mathbf{\Sigma_U}$ is non–negative definite also have been tried. This is the case for the quadratic distance function proposed by Agterberg (1984) and used by Haining (1987) for modelling the autocorrelation in trend surface analysis.

The choice of parameterisation of $\mathbf{\Sigma_U}$ is sometimes made by plotting the variogram of the OLS residuals. Care has to be taken when interpreting the variogram since it is sensitive to the number of pairs of data points used to estimate the empirical covariance at a particular distance. The number of pairs will vary with the distance, typically increasing at first. Ripley (1981) advises the use of cross–validation by successive deletion of data points to assess the fit of the model chosen for the covariance function.

Direct estimation of $\mathbf{\Sigma_U}$ in approach (iii) was advocated by Arora and Brown (1977) in a case where spatial data at different time intervals were available. In defining $\mathbf{\Sigma_U}$, Haining also has considered using direct estimates of the residual autocorrelation or a finite number of spatial lags (defined through a contiguity matrix), and zero elsewhere. Here, again, the resulting covariance matrix is not necessarily non–negative definite.

Once the model for $\mathbf{\Sigma_U}$ has been specified, estimation can be carried out in a Gaussian framework by maximum likelihood (ML) techniques, in cases (i) and (ii), when $\mathbf{\Sigma_U}$ is non–negative definite, or through some iterative method based upon generalised least squares, similar to that first proposed by Cochrane and Orcutt for time series (1949). Mardia and Marshall (1984) have studied the asymptotic properties of the maximum likelihood estimators (MLE) for model (5) in the cases of (i) and (ii). Assuming that $\mathbf{Y}$ is a Gaussian process, they give conditions that ensure the consistency and the asymptotic normality of $(\widehat{\beta},\widehat{\theta})$, the MLEs of $(\beta,\theta)$, $\theta$ being the vector of parameters of $\mathbf{\Sigma_U}$. Conditional upon $\widehat{\theta}$, the variance–covariance matrix for $\widehat{\beta}$ can be calculated, and hence tests of the regression coefficients can be performed. Because of the conditionality involved, the standard errors of $\widehat{\beta}$ might be underestimated, which leads some authors to consider a Bayesian approach. Hepple (1979) carried out a Bayesian analysis of the model defined by equations (5) and (6), assuming a simple linear regression with a constant $\beta_1$, a slope $\beta_2$, and uniform diffuse priors for $\beta = (\beta_1, \beta_2)$, $\mathbf{log}(\sigma)$, and $\rho$. He shows how the conditional posterior distribution of $\beta_2$ is sensitive to values of $\rho$, and he derives the bivariate posterior distribution for $\beta_2$ and $\rho$, with mode corresponding to MLEs of $\beta_2$ and $\rho$.

When the data set contains a large number of points, MLE becomes computationally very heavy, and is fraught with difficulties. Mardia and Marshall (1984) used the Fisher scoring technique. A study by Warnes and Ripley (1987) showed that this method usually converges only to the nearest local maximum, and in some cases did not converge at all. Furthermore, they found cases of multimodal profile likelihoods that renders hazardous the search for a maximum. Moreover, Ripley (1988) reports some simulation results where the global maximum found is well away from the true value. In a recent paper, Vecchia (1988) proposed carrying out estimation, within the class of covariance functions defined earlier in this paper, with the help of successive *approximate likelihood functions* that are much easier to handle. He applied his method successfully to simulated data sets and to water level data where he wanted to estimate a trend.

Trend fitting, or regression when $\mathbf{X}$ is a matrix of polynomial powers and cross–products of geographic coordinates, is indeed an area where the estimation of model (5) has been much used. As pointed out by Haining (1987), residual autocorrelation can arise either from false specification of the order of the trend, or from local scale effects arising from spatial processes that operate at an intermediate scale between the regional trend and the local residuals. Haining analysed aerial survey data of marine pollution using three approaches: model (6) fitted via ML techniques, Agterberg's quadratic covariance function, and a direct estimation of $\mathbf{\Sigma_U}$ based upon six spatial lags. He encountered some problems of convergence in the last two approaches, due possibly to the matrix $\mathbf{\Sigma_U}$ not satisfying the non–negative definite property. When comparison was possible, he found some moderate differences between the resulting trend estimates given by the three methods. He also commented that the direct estimation of $\mathbf{\Sigma_U}$ is the least satisfactory.

Haining's results seem to indicate that regression estimates can be quite sensitive to the choice of model for $\mathbf{\Sigma_U}$. This is an interesting problem, and we have chosen to investigate it in a different manner by restricting the comparisons to models for $\mathbf{U}$ and $\mathbf{\Sigma_U}$ that satisfy the non–negative definiteness property, and by using the same estimation technique (namely ML) in all cases. Finally, the same examples will be analysed by means of the modified $\mathbf{t_{\widehat{M}-2}}$ test for partial correlations.

### 5.3. Comparison of regression results obtained by different modelling of the variance–covariance error matrix and by modified $\mathbf{t_{\widehat{M}-2}}$ tests

The examples analysed in this section concern the relationship between male lung cancer, some industrial factors, and smoking. Among the different cancer sites, male lung cancer has been frequently associated with industrial exposure (Pastorino *et al.*, 1984; Benhamou *et al.*, 1988), and it has been estimated that 15% of all lung cancer arising in men in the U. S. A. could be due to occupational risk factors (Doll and Peto, 1981). This figure has been the subject of recent debates (Simonato et al, 1988) and several authors have tried to estimate it using data from different case–control studies (Vineis *et al.*, 1988; Ronco *et al.*, 1988). Hence it is particularly interesting to test the link between male lung cancer and industrial exposure at a geographical level. If the results that emerge are consistent with those of case–controls or cohort studies, it becomes possible to calculate an estimation of attributable risk due to occupational factors based on comprehensive geographical census data rather than data from individual epidemiological studies. A study of the biases associated with estimation from ecological (grouped) data was done in Richardson, Stücker and Hémon (1987). In view of the strong link between smoking and lung cancer, some adjustment based on a measure of smoking consumption is also needed. We present some results on four branches of industry: metal, general engineering, mining and textile works.

#### *5.3.1. The data*

Male lung cancer mortality rate has been standardised over the age 35-74 and over a 2 year period, 1968-1969. The data were provided by the French National Institute for Health and Medical Research (INSERM) at the level of the French *départements*. Cigarette sales data were compiled by the French Nationalised Tobacco Company (SEITA). To take into account the time lag between smoking and the onset of a lung pathology, cigarette sales per inhabitant were recorded in 1953. Demographic data on the percentage of employed males

in the metal industry, in general engineering, in the textile industry and in mining were taken from the 1962 census (INSEE). After the grouping of the *départements* around Paris into one area and the exclusion of four others owing to the poor quality of the data, $\mathbf{N = 82}$ locations were retained indexed by the coordinates of the administrative centres of the *départements*.

*5.3.2. Spatial models for the variance–covariance error matrix*

Four models for the variance–covariance error matrix $\mathbf{\Sigma_U}$ in the regression model:

$$\mathbf{Y = Xb + U}$$

[cf. (5) §3.2] were chosen. The first two involve only one shape parameter for the autocorrelation.

(a) $\quad \mathbf{U = cWU} + \varepsilon, \qquad \varepsilon \sim \mathbf{N(0, \sigma^2 I)},$

$\mathbf{W}$ is defined as an $(82 \times 82)$ $0-1$ contiguity matrix on the French *départements* normalised such that each row sum is equal to unity, and with non–zero $(\mathbf{i,j})$ element if *départements* $\mathbf{i}$ and $\mathbf{j}$ have a common border length. In this case the covariance matrix $\mathbf{\Sigma_U}$ is equal to $\sigma^2(\mathbf{I - cW})^{-1}(\mathbf{I - cW^t})^{-1}$.

(b) $\quad \mathbf{\Sigma_U}$ follows a disc model.

Letting $\mathbf{d_{ij}}$ be the distance between locations of *départements* $\mathbf{i}$ and $\mathbf{j}$, the $(\mathbf{i,j})$ element of $\mathbf{\Sigma_U}$ is then given by $\sigma^2 \mathbf{f_a}(\mathbf{d_{ij}})$ with:

$$\mathbf{f_a}(\mathbf{r}) = \frac{\mathbf{2}}{\pi}[\cos^{-1}(\frac{\mathbf{r}}{\mathbf{2a}}) - \frac{\mathbf{r}}{\mathbf{2a}}(1 - \frac{\mathbf{r^2}}{\mathbf{4a^2}})^{1/2}] \quad r \leq 2a,$$
$$\mathbf{f_a}(\mathbf{r}) = 0 \quad r > 2a.$$

This covariance function exhibits a fairly linear decrease with increasing distance (cf. Ripley, 1981 p. 56) with the value for $\mathbf{f_a'(0)}$, the slope of the tangent at zero, equal to $\mathbf{-2(a\pi)^{-1}}$. To illustrate its behaviour using a linear approximation for small $\mathbf{r}$, we can say that if the autocorrelation is equal to 0.8 for a distance $\mathbf{r} = 40$ km [implying $\mathbf{f_a'(0)} = 0.005$] then the autocorrelation will become zero after a distance of $\mathbf{2a = 2 \times 2(\pi \times 0.005)^{-1} = 255}$ km.

Next we consider classes of covariances with two shape parameters for the autocorrelation.

(c) $\quad \mathbf{\Sigma_U}$ follows an exponential model.

The $(\mathbf{i,j})$ element of $\mathbf{\Sigma_U}$ is given by $\sigma^2 \gamma \mathbf{e}^{-\lambda \mathbf{d_{ij}}}$. This model was used by Cook and Pocock (1983) after inspection of the variogram of the OLS. residuals of their regression model.

(d) $\quad \mathbf{\Sigma_U}$ follows a Bessel model.

The $(\mathbf{i,j})$ element of $\mathbf{\Sigma_U}$ is given by $\sigma^2 \mathbf{g_{v,a}}(\mathbf{r})$ with

$$\mathbf{g_{v,a}}(\mathbf{r}) = \frac{\mathbf{1}}{\mathbf{2^{v-1}\Gamma(v)}}(\mathbf{ar})^{\mathbf{v}}\mathbf{K_v}(\mathbf{ar}), \quad \mathbf{v > 0}, \quad \mathbf{a > 0}.$$

All models were fitted by ML. The numerical maximisation was performed with a safeguarded quadratic interpolation method for models (a) and (b), simplified by the use of

eigenvalues in case (a) following the remark made by Ord (1975). For models (c) and (d) a quasi–Newton method with finite difference gradient was used. Starting points were provided after visual comparison between the shapes of the theoretical correlograms and those calculated from the OLS residuals. New starting points were tried when the program indicated convergence problems. Clearly the possibility of having encountered only local maxima cannot be discarded. The only lengthy maximisation was the one for model (d). The parameter v was restricted to lie in the range $[0.1 - 2.9]$ after a visual inspection. All maximisations were performed on a Compaq 386 microcomputer using IMSL library routines (1987) UVMIF (1–parameter) or BCONF (2–parameter).

*5.3.3. Relative efficiency of OLS with regard to generalised least squares (GLS)*

In a recent paper, Krämer and Donninger (1987) have given some results on the relative efficiency of OLS with regard to GLS for the autoregressive model (a). They define the relative efficiency **e** by the quotient of the traces of the covariance matrices for GLS and OLS respectively:

$$\mathbf{e} = (\mathbf{X^t \Sigma_U^{-1} X})^{-1} / [(\mathbf{X^t X})^{-1} \mathbf{X^t \Sigma_U X} (\mathbf{X^t X})^{-1}].$$

They show that for $\mathbf{\Sigma_U} = (\mathbf{I} - \mathbf{cW})^{-1}(\mathbf{I} - \mathbf{cW^t})^{-1}$, the limit of **e** is unity as **c** tends to its maximal value of **1** provided the regressors **X** include a constant term. This result is of limited practical significance since, as they point out, for intermediate values of **c**, the loss of efficiency can be substantial. We have computed this ratio for all the fitted regression models in order to pursue this problem further.

*5.3.4. Regression results*

*Metal industry and general engineering workers*

There is broad agreement amongst the four models in finding a statistically significant link between lung cancer rates and these two industries, both with and without adjustment for cigarette sales (Exhibits 8 and 9). Nevertheless, we observe some amount of variation among the models concerning the regression slopes or their significance levels, with closer similarity within the 1–parameter or the 2–parameter models. The inclusion of cigarette sales in the regression clearly improves the fit in all models with higher ML values. Even after this adjustment, there is still a substantial amount of autocorrelation among the residuals as shown in the estimated autocorrelation at 40 km. For both industries the disc model gives lower estimates of this residual autocorrelation and consequently higher t values. The fit of the exponential and the Bessel model are very close, with slightly higher likelihoods for the exponential model. From visual inspection of the observed and estimated correlograms, the Bessel model seems to give the closer fit and the 1–parameter disc model is clearly not flexible enough.

*Mining industry*

A non–significant association between male lung cancer rates and the percent of workers employed by the mining industry is found in all four models (Exhibit 10). The slopes are decreased after adjustment for cigarette sales and there is close agreement on residual autocorrelation and on the values of the $t$–statistics between all four models.

## Exhibit 8.

**Linear regression between lung cancer mortality rates for men and metal industry workers with and without adjustment on cigarette sales : standard, modified $t_{\hat{M}\text{-}2}$ tests and results for different parametrisations of the error variance-covariance matrix.**

| Metal industry workers | no adjustment | | Adjustment on cigarette sales | |
|---|---|---|---|---|
| Standard test : β | $29.1 \times 10^{-4}$ | | $17.3 \times 10^{-4}$ | |
| t (p value) | 7.1 | ($< 10^{-9}$) | 5.27 | ($10^{-6}$) |
| $t_{\hat{M}\text{-}2}$ | 3.00 | ($9.5 \times 10^{-3}$) | 3.48 | ($1.4 \times 10^{-3}$) |
| | $\hat{M} = 16$ | | $\hat{M} = 37$ | |

**Autoregressive model (a)**

| Metal industry workers | no adjustment | | Adjustment on cigarette sales | |
|---|---|---|---|---|
| β | $18.9 \times 10^{-4}$ | | $12.6 \times 10^{-4}$ | |
| t (p value) | 4.34 | ($4 \times 10^{-5}$) | 3.70 | ($3 \times 10^{-4}$) |
| log M.L. | 1443.5 | | 1490.0 | |
| model parameter : $\hat{c}$ | 0.574 | | 0.379 | |
| $e^+$ | 0.54 | | 0.82 | |

**Disc model (b)**

| Metal industry workers | no adjustment | | Adjustment on cigarette sales | |
|---|---|---|---|---|
| β | $18.7 \times 10^{-4}$ | | $15.5 \times 10^{-4}$ | |
| t (p value) | 4.22 | ($6 \times 10^{-5}$) | 4.69 | ($10^{-5}$) |
| log M.L. | 1431.1 | | 1487.4 | |
| model parameter : $2\hat{a}$ | 100.4 | | 65.8 | |
| autocorrelation ρ (1)* | 0.493 | | 0.227 | |
| $e^+$ | 0.66 | | 0.834 | |

**Exponential model (c)**

| Metal industry workers | no adjustment | | Adjustment on cigarette sales | |
|---|---|---|---|---|
| β | $15.7 \times 10^{-4}$ | | $10.1 \times 10^{-4}$ | |
| t (p value) | 3.51 | ($7 \times 10^{-4}$) | 2.90 | ($4.8 \times 10^{-3}$) |
| log M.L. | 1445.4 | | 1491.6 | |
| model parameters : $\hat{\gamma},\hat{\lambda}$ | 0.725 , $0.35 \times 10^{-2}$ | | 0.569 , $0.79 \times 10^{-2}$ | |
| autocorrelation ρ(1)* | 0.629 | | 0.414 | |
| $e^+$ | 0.28 | | 0.58 | |

**Bessel model (d)**

| Metal industry workers | no adjustment | | Adjustment on cigarette sales | |
|---|---|---|---|---|
| β | $17.1 \times 10^{-4}$ | | $10.9 \times 10^{-4}$ | |
| t (p value) | 3.8 | ($2.8 \times 10^{-4}$) | 3.17 | ($2.2 \times 10^{-3}$) |
| log M.L. | 1442.3 | | 1491.3 | |
| model parameters : $\hat{\nu},\hat{a}$ | 0.304 , $0.98 \times 10^{-2}$ | | 0.211 , 0.01 | |
| autocorrelation ρ(1)* | 0.499 | | 0.378 | |
| $e^+$ | 0.50 | | 0.69 | |

* autocorrelation calculated using estimated model parameters at a distance of 40 km.
\+ relative efficiency as defined in §5.3.3.

## Exhibit 9.

**Linear regression between lung cancer mortality rates for men and general engineering workers with and without adjustment on cigarette sales : standard, modified $t_{\hat{M}\text{-}2}$ tests and results for different parametrisations of the error variance-covariance matrix.**

| General engineering workers | no adjustment | | Adjustment on cigarette sales | |
|---|---|---|---|---|
| Standard test : β | $32.4 \times 10^{-4}$ | | $18.9 \times 10^{-4}$ | |
| t (p value) | 4.23 | $(6 \times 10^{-5})$ | 3.54 | $(7 \times 10^{-4})$ |
| $t_{\hat{M}\text{-}2}$ (p value) | 2.72 | $(1.03 \times 10^{-2})$ | 2.88 | $(5.7 \times 10^{-3})$ |
| | $\hat{M} = 35$ | | $\hat{M} = 55$ | |

### Autoregressive model (a)

| General engineering workers | no adjustment | | Adjustment on cigarette sales | |
|---|---|---|---|---|
| β | $25.7 \times 10^{-4}$ | | $16.7 \times 10^{-4}$ | |
| t (p value) | 4.77 | $(0.8 \times 10^{-5})$ | 3.67 | $(4 \times 10^{-4})$ |
| log M.L. | 1449.0 | | 1492.5 | |
| model parameter : $\hat{c}$ | 0.736 | | 0.586 | |
| $e^+$ | 0.31 | | 0.33 | |

### Disc model (b)

| General engineering workers | no adjustment | | Adjustment on cigarette sales | |
|---|---|---|---|---|
| β | $21.6 \times 10^{-4}$ | | $18.4 \times 10^{-4}$ | |
| t (p value) | 4.28 | $(5 \times 10^{-5})$ | 3.96 | $(2 \times 10^{-4})$ |
| log M.L. | 1443.2 | | 1485.2 | |
| model parameter : $2\hat{a}$ | 190.1 | | 85.4 | |
| autocorrelation ρ (1)* | 0.732 | | 0.403 | |
| $e^+$ | 0.21 | | 0.73 | |

### Exponential model (c)

| General engineering workers | no adjustment | | Adjustment on cigarette sales | |
|---|---|---|---|---|
| β | $21.8 \times 10^{-4}$ | | $14.23 \times 10^{-4}$ | |
| t (p value) | 4.0 | $(0.4 \times 10^{-3})$ | 3.24 | $(1.7 \times 10^{-3})$ |
| log M.L. | 1449.9 | | 1495.5 | |
| model parameters : $\hat{\gamma},\hat{\lambda}$ | 0.884 , $0.34 \times 10^{-2}$ | | 0.787 , $0.49 \times 10^{-2}$ | |
| autocorrelation ρ(1)* | 0.772 | | 0.647 | |
| $e^+$ | 0.22 | | 0.37 | |

### Bessel model (d)

| General engineering workers | no adjustment | | Adjustment on cigarette sales | |
|---|---|---|---|---|
| β | $22.3 \times 10^{-4}$ | | $15.1 \times 10^{-4}$ | |
| t (p value) | 4.05 | $(1.2 \times 10^{-4})$ | 3.36 | $(1.2 \times 10^{-3})$ |
| log M.L. | 1448.7 | | 1493.6 | |
| model parameters : $\hat{\nu},\hat{a}$ | 0.381 , $0.61 \times 10^{-2}$ | | 0.417 , 0.013 | |
| autocorrelation ρ(1)* | 0.695 | | 0.537 | |
| $e^+$ | 0.28 | | 0.52 | |

* autocorrelation calculated using estimated model parameters at a distance of 40 km.
+ relative efficiency as defined in §5.3.3.

## Exhibit 10.

**Linear regression between lung cancer mortality rates for men and mining industry workers with and without adjustment on cigarette sales : standard, modified $t_{\hat{M}\text{-}2}$ tests and results for different parametrisations of the error variance-covariance matrix.**

| Mining industry workers | no adjustment | Adjustment on cigarette sales |
|---|---|---|
| Standard test : β | $19.9 \times 10^{-4}$ | $9.8 \times 10^{-4}$ |
| t (p value) | 3.16 ($2.2 \times 10^{-3}$) | 2.14 ($3.5 \times 10^{-2}$) |
| $t_{\hat{M}\text{-}2}$ (p value) | 2.37 ($2.2 \times 10^{-2}$) | 2.23 ($2.8 \times 10^{-2}$) |
| | $\hat{M} = 47$ | $\hat{M} = 89$ |

**Autoregressive model (a)**

| Mining industry workers | no adjustment | Adjustment on cigarette sales |
|---|---|---|
| β | $5.2 \times 10^{-4}$ | $3.0 \times 10^{-4}$ |
| t (p value) | 1.03 (0.31) | 0.78 (0.43) |
| log M.L. | 1429.6 | 1480.3 |
| model parameter : $\hat{c}$ | 0.687 | 0.539 |
| e+ | 0.42 | 0.66 |

**Disc model (b)**

| Mining industry workers | no adjustment | Adjustment on cigarette sales |
|---|---|---|
| β | $4.2 \times 10^{-4}$ | $1.8 \times 10^{-4}$ |
| t (p value) | 0.95 (0.34) | 0.55 (0.58) |
| log M.L. | 1425.7 | 1478.5 |
| model parameter : $2\hat{a}$ | 240.1 | 184.1 |
| autocorrelation ρ (1)* | 0.788 | 0.723 |
| e+ | 0.21 | 0.31 |

**Exponential model (c)**

| Mining industry workers | no adjustment | Adjustment on cigarette sales |
|---|---|---|
| β | $4.5 \times 10^{-4}$ | $2.4 \times 10^{-4}$ |
| t (p value) | 0.94 (0.35) | 0.66 (0.51) |
| log M.L. | 1435.9 | 1485.7 |
| model parameters : $\hat{\gamma},\hat{\lambda}$ | 0.852 , $0.32 \times 10^{-2}$ | 0.743 , $0.50 \times 10^{-2}$ |
| autocorrelation ρ(1)* | 0.749 | 0.609 |
| e+ | 0.35 | 0.60 |

**Bessel model (d)**

| Mining industry workers | no adjustment | Adjustment on cigarette sales |
|---|---|---|
| β | $4.6 \times 10^{-4}$ | $2.9 \times 10^{-4}$ |
| t (p value) | 0.97 (0.33) | 0.80 (0.43) |
| log M.L. | 1434.2 | 1484.4 |
| model parameters : $\hat{v},\hat{a}$ | 0.358 , $0.63 \times 10^{-2}$ | 0.311 , $0.92 \times 10^{-2}$ |
| autocorrelation ρ(1)* | 0.663 | 0.524 |
| e+ | 0.411 | 0.65 |

* autocorrelation calculated using estimated model parameters at a distance of 40 km.
+ relative efficiency as defined in §5.3.3.

*Textile industry*

There is a discrepancy between the results given by the disc model that finds a borderline association after adjustment for cigarette sales, and other models that find a non–significant association with the textile industry (Exhibit 11). Note that for all models, the inclusion of cigarette sales has led to a higher $t$–value for the regression coefficient of mining. In one case (exponential model without adjustment for cigarettes), the maximisation encountered numerical problems and the maximum found might be only local.

Amongst the investigated associations with industrial exposure we thus found overall general agreement between the results given by different parametrisations of the variance–covariance error matrix. Nevertheless, significance levels did vary by a non–negligible ratio and in one case this variation could lead to a different interpretation of the results. Hence, it is important, when using this approach, to check the coherence of the results for at least two different models of $\mathbf{\Sigma_U}$. A further note of caution is also warranted when one compares the results given by the autoregressive model (a) for two choices of $\mathbf{W}$: a non–standardised $(0-1)$ matrix (results not shown) and the same matrix with row sums equal to 1. Some discrepancies arise even though the matrices use the same definition of "contiguity". With the non–standardised $\mathbf{W}$ matrix we found, for instance, a significant association with mining ($t = 2.02$, $p = 0.046$) before adjustment for cigarette sales and a borderline association ($t = 1.69$, $p = 0.095$) after adjustment. Overall the choice of standardising $\mathbf{W}$ led to higher ML values and to closer results with the other models than when using the non–standardised version. Finally one needs to recall that contrary to the time series case, it is difficult to ensure that a global maximum has always be found.

*Relative efficiency*

All the relative efficiencies calculated were far from unity even though a constant term was always included in the regression. This reinforces the necessity of modelling the spatial structure of the residuals. The relative efficiency of the exponential model was always lower than that of the Bessel model, pointing again to the exponential model as having a slightly better performance than the other models in our examples.

*5.3.5. Comparison between the regression results and the modified* $\mathbf{t_{\widehat{M}-2}}$ *statistic results*

For the mining industry and general engineering workers, the $\mathbf{t_{\widehat{M}-2}}$ test agrees overall with the results of the regressions by finding a statistically significant link. The agreement is closer once adjustment for cigarette sales has been performed, with levels of significance of the same order of magnitude as those given by the exponential or Bessel models for both industries. The agreement is poorer before adjustment when there is a higher residual autocorrelation with values of $\mathbf{t_{\widehat{M}-2}}$ lower than those given by the models. Recall that the adjustment of the d.f. carried out by $\mathbf{t_{\widehat{M}-2}}$ uses directly the autocorrelations $\mathbf{\hat{C}_X(k)}$ and $\mathbf{\hat{C}_Y(k)}$ for the $\mathbf{k}$ strata and consequently is adapted to the whole correlograms, whereas the model approach centres its estimations mainly on the first few classes of higher autocorrelation. In cases of lower autocorrelation levels, the first strata give most of the contribution to the modified d.f., and hence agreement might be expected to be closer between the two approaches.

For the mining industry there is a divergence of results since the modified $\mathbf{t_{\widehat{M}-2}}$ statistic finds a significant link at the 5% level whereas none of the four models do. Recall that a

## Exhibit 11.

**Linear regression between lung cancer mortality rates for men and textile industry workers with and without adjustment on cigarette sales : standard, modified $t_{\hat{M}-2}$ tests and results for different parametrisations of the error variance-covariance matrix.**

| Textile industry workers | no adjustment | | Adjustment on cigarette sales | |
|---|---|---|---|---|
| Standard test : $\beta$ | $18.3 \times 10^{-4}$ | | $11.2 \times 10^{-4}$ | |
| t (p value) | 2.57 | $(1.2 \times 10^{-2})$ | 2.28 | $(2.5 \times 10^{-2})$ |
| $t_{\hat{M}-2}$ (p value) | 1.52 | (0.14) | 1.89 | $(6.4 \times 10^{-2})$ |
| | $\hat{M} = 30$ | | $\hat{M} = 57$ | |

### Autoregressive model (a)

| Textile industry workers | no adjustment | | Adjustment on cigarette sales | |
|---|---|---|---|---|
| $\beta$ | $0.6 \times 10^{-4}$ | | $4.2 \times 10^{-4}$ | |
| t (p value) | 0.1 | (0.9) | 0.96 | (0.37) |
| log M.L. | 1428.6 | | 1480.6 | |
| model parameter : $\hat{c}$ | 0.709 | | 0.543 | |
| $e^+$ | 0.36 | | 0.66 | |

### Disc model (b)

| Textile industry workers | no adjustment | | Adjustment on cigarette sales | |
|---|---|---|---|---|
| $\beta$ | $1.8 \times 10^{-4}$ | | $9.1 \times 10^{-4}$ | |
| t (p value) | 0.33 | (0.74) | 2.06 | (0.042) |
| log M.L. | 1426.4 | | 1474.6 | |
| model parameter : $2\hat{a}$ | 189.8 | | 83.4 | |
| autocorrelation $\rho$ (1)* | 0.732 | | 0.39 | |
| $e^+$ | 0.24 | | 0.72 | |

### Exponential model (c)

| Textile industry workers | no adjustment | | Adjustment on cigarette sales | |
|---|---|---|---|---|
| $\beta$ | $-0.5 \times 10^{-4}$ | | $3.28 \times 10^{-4}$ | |
| t (p value) | - 0.093 | (0.92) | 0.80 | (0.43) |
| log M.L. | 1435.0 | | 1485.9 | |
| model parameters : $\hat{\gamma},\hat{\lambda}$ | 0.871 , $0.31 \times 10^{-2}$ | | 0.755 , $0.52 \times 10^{-2}$ | |
| autocorrelation $\rho(1)$* | 0.771 | | 0.613 | |
| $e^+$ | 0.21 | | 0.42 | |

### Bessel model (d)

| Textile industry workers | no adjustment | | Adjustment on cigarette sales | |
|---|---|---|---|---|
| $\beta$ | $1.8 \times 10^{-4}$ | | $4.6 \times 10^{-4}$ | |
| t (p value) | 0.32 | (0.75) | 1.09 | (0.28) |
| log M.L. | 1431.4 | | 1484.4 | |
| model parameters : $\hat{\nu},\hat{a}$ | 0.471 , 0.01 | | 0.353 , 0.011 | |
| autocorrelation $\rho(1)$* | 0.649 | | 0.518 | |
| $e^+$ | 0.35 | | 0.59 | |

* autocorrelation calculated using estimated model parameters at a distance of 40 km.
+ relative efficiency as defined in §5.3.3.

comparable significant link was also obtained when a non–standardised **W** matrix was used in model (a). This example was also investigated after removal of a linear trend component (see Richardson, 1989), and after this further adjustment the $\mathbf{t}_{\widehat{\mathbf{M}-\mathbf{2}}}$ statistic becomes clearly non–significant. This points to the existence of a simple gradient–like structure for the residuals of mining, cigarette sales and lung cancer that is taken into account by the regression models, but not by the modified $\mathbf{t}_{\widehat{\mathbf{M}-\mathbf{2}}}$ test (without trend removal).

For the textile industry, the $\mathbf{t}_{\widehat{\mathbf{M}-\mathbf{2}}}$ finds a non–significant link, agreeing qualitatively with the results of three of the models. Nevertheless, the significance levels are lower than those given by these three models and not far from the borderline significance given by the disc model. Hence the results given by the $\mathbf{t}_{\widehat{\mathbf{M}-\mathbf{2}}}$ statistic are in this case intermediate between the disc model and the other three models.

In conclusion, except for mining, the results given by the modified $\mathbf{t}_{\widehat{\mathbf{M}-\mathbf{2}}}$ statistic would lead to the same conclusions as those given by the regression models. We note however that the values of the $\mathbf{t}_{\widehat{\mathbf{M}-\mathbf{2}}}$ statistic tend to be more moderate in their adjustment for spatial autocorrelation than those obtained by the regressions with parametrised variance–covariance error matrix, which find the link either still strongly significant or not at all. This is probably due to the more flexible nature of the adjustment for spatial structure carried out by $\mathbf{t}_{\widehat{\mathbf{M}-\mathbf{2}}}$. It is also clearly apparent that the use of standard regressions with i.i.d. errors would lead to quite erroneous conclusions.

## 6. Concluding remarks

In this paper we have compared different approaches to the testing of association between spatially autocorrelated variables.

One approach consists of adjusting classical tests on simple or partial correlation coefficients to take into account spatial autocorrelation. The proposed modified tests that we developed were shown to have satisfactory Type I error and power. These methods do not require the identification of a particular parametric model for the type of spatial autocorrelation and they only involve straightforward calculations that can be done on small computers. As these tests are derived in the Gaussian framework, a study of the robustness of their performance to departure from normality was also carried out. There was no measurable effect on Type I errors except in highly autocorrelated cases where the tests became too conservative.

Alternatively new measures of association could be employed such as Tjøstheim's non–parametric index **A**. In a Monte Carlo study we found an expected increase in Type I errors of the index **A** in cases of high autocorrelation. Furthermore its observed power was lower than that of the modified test in most cases, even under an alternative hypothesis of local spatial permutation provided that a random error term is included.

Finally, the results of the modified tests were compared to those obtained by generalised regression with different spatial models for the variance–covariance error matrix on four examples of geographical epidemiology. There was general agreement between the two approaches except in one example. The correspondence between the modified test and general regressions was greater in the two cases where there was also close agreement between the results of the different spatial parametrisations. The modified tests can thus be seen as a first step in the testing of association, to be carried out before any parametric modelling

that probably involves some degree of arbitrariness and some computational difficulties.

## 7. References


Agterberg, F. (1984) Trend surface analysis, in *Spatial Statistics and Models*, edited by G. Gaile and C. Willmot, pp. 147-171. Dordrecht: Reidel.

Armstrong, B., and R. Doll. (1975) Environmental factors and cancer incidence and mortality in different countries, with special reference to dietary practices. *International Journal of Cancer*, **15**, 617-631.

Arora, S., and M. Brown. (1977) Alternative approaches to spatial autocorrelation: an improvement over current practice. *International Regional Science Review*, **2**, (1), 67-78.

Bartlett, M. (1935) Some aspects of the time–correlation problem in regard to tests of significance. *Journal of the Royal Statistical Society*, **98**, 536-543.

Benhamou, S., E. Benhamou, and R. Flamant. (1988) Occupational risk factors of lung cancer in a French case–control study. *British Journal of Industrial Medicine*, **45**, 231-233.

Bivand, R. (1984) Regression modeling with spatial dependence: an application of some class selection and estimation methods. *Geographical Analysis*, **16**, (1), 25-37.

Bodson, P., and D. Peeters. (1975) Estimation of the coefficients of a linear regression in the presence of spatial autocorrelation. An application to a Belgian labour–demand function. *Environment and Planning A*, **7**, 456-472.

Bolthausen, E. (1982) On the central limit theorem for stationary random fields. *Annals of Probability*, **10**, 1047-1050.

Cliff, A., and J. Ord. (1981) *Spatial Processes: Models and Applications*. Pion: London.

Clifford, P., S. Richardson, and D. Hémon. (1989) Assessing the significance of the correlation between two spatial processes. *Biometrics*, **45**, (1), 123-134.

Cochrane, D., and G. Orcutt. (1949) Applications of least–square regressions to relationships containing autocorrelated error terms. *Journal of the American Statistical Association*, **44**, 32- 61.

Cook, D., and S. Pocock. (1983) Multiple regression in geographic mortality studies with allowance for spatially correlated errors. *Biometrics*, **39**, 361-371.

Doll, R. (ed.). (1984) The geography of disease, *British Medical Bulletin*. Published for the British Council by Churchill Livingstone.

Doll, R., and P. Peto. (1981) The causes of cancer. *Journal of the National Cancer Institute*, **66**, 1191-1308.

Doreian, P. (1981) Estimating linear models with spatially distributed data, in *Sociological Methodology*, edited by S. Leinhardt, pp. 359-388. San Francisco: Jossey–Bass.

Geweke, J. (1981) A comparison of tests of the independence of two covariance–stationary time series. *Journal of the American Statistical Association*, **76**, (374), 363-373.

Glasbey, C. (1988) Examples of regression with serially correlated errors. *The Statistician*, **37**, 277-291.

Glick, B. (1982) A spatial rank order correlation measure. *Geographical Analysis*, **14**, (2), 177-181.

Griffith, D. (1980) Towards a theory of spatial statistics. *Geographical Analysis*, **12**, (4),

325-339.

Guyon, X, and S. Richardson. (1984) Vitesse de convergence du théorème de la limite centrale pour des champs faiblement dépendants. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, **66**, 297-317.

Haining, P. (1987) Trend surface models with regional and local scales of variation with an application to aerial survey data. *Technometrics*, **29**, 461-469.

Haugh, L. (1976) Checking the independence of two covariance– stationary time series: a univariate residual cross–correlation approach. *Journal of the American Statistical Association*, **71**, 378-385.

Hepple, L. (1979) Bayesian analysis of the linear model with spatial dependence, in *Exploratory and Explanatory Statistical Analysis of Spatial Data*, edited by C. Bartels and R. Ketellapper, pp. 179-199 . Boston: Martinus Nijhoff.

Hubert, L., and R. Golledge. (1982) Measuring association between spatially defined variables: Tjøstheim's index and some extensions. *Geographical Analysis*, **14**, (3), 273-278.

Johnston, J. (1972) *Econometric Methods*, 2nd edition. New York: McGraw-Hill.

Krämer, W., and C. Donninger. (1987) Spatial autocorrelation among errors and the relative efficiency of OLS in the linear regression model. *Journal of the American Statistical Association*, **82**, (398), 577-579.

Mardia, K., and R. Marshall. (1984) Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, **71**, 135-146.

Ord, K. (1975) Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, **70**, 120-126.

Pastorino, U., F. Berrino, A. Gervasio, V. Pesenti, E. Riboli, and P. Crosignomi. (1984) Proportion of lung cancers due to occupational exposure. *Journal of the National Cancer Institute*, **33**, 231-237.

Pierce, D. (1977) Relationships and the lack thereof between economic time series, with special reference to money and interest rates. *Journal of the American Statistical Association*, **72**, (357), 11-26.

Richardson, S., and D. Hémon. (1982) On the variance of the sample correlation between two independent lattice processes. *Journal of Applied Probability*, **18**, 943-948.

Richardson, S., and P. Clifford. Testing association between spatial processes. Presented at the A.M.S. conference on *Spatial Statistics and Imaging*, Bowdoin College (1988).

Richardson, S. (1990) A method for testing the significance of geographical correlations with application to industrial lung cancer in France. *Statistics in Medicine*, **9**, 515-528.

Richardson, S., I. Stücker, and D. Hémon. (1987) Comparison of relative risks obtained in ecological and individual studies: some methodological considerations. *International Journal of Epidemiology*, **16**, 111-120.

Ripley, B. (1981) *Spatial Statistics*. New York: Wiley.

Ripley, B. (1988) *Statistical Inference for Spatial Processes*. Cambridge: Cambridge University Press.

Ronco, G., G. Ciccone, D. Mirabelli, B. Troia, and P. Vineis. (1988) Occupation and lung cancer in two industrialized areas of Northern Italy. *International Journal of Cancer*, **41**,

354-358.

Simonato, L., P. Vineis, and A. Fletcher. (1988) Estimates of the proportion of lung cancer attributable to occupational exposure. *Carcinogenesis*, **9**, 1159-1165.

Tjøstheim, D. (1978) A measure of association for spatial variables. *Biometrika*, **65**, (1), 109-114.

Upton, G., and B. Fingleton. (1983) *Spatial Data Analysis by Example*. Chichester: Wiley.

Vecchia, A. (1988) Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society*, B **50**, (2), 297-312.

Vineis, P., T. Thomas, R. Hayes, W. Blot, T. Mason, L. Williams Pickle, P. Correa, E. Fontham, and J. Schoenberg. (1988) Proportion of lung cancers in males, due to occupation, in different areas of the USA. *International Journal of Cancer*, **42**, (6), 851-856.

Warnes, J., and B. Ripley. (1987) Problems with likelihood estimation of covariance functions of spatial Gaussian processes. *Biometrika*, **74**, (3), 640-642.

Whittle, P. (1954) On stationary processes in the plane. *Biometrika*, **41**, 434-449.

# DISCUSSION

## "Some remarks on the testing of association between spatial processes"

## by Sylvia Richardson

The problem of deciding if spatially distributed variables are associated is one the analyst of spatial data frequently encounters. Richardson has considered three commonly used (and one somewhat less common used) measures in the study of associations between pairs of variables. When one or all of the variables are spatially correlated, the distributions of the measures can be substantially affected. Richardson's paper proposes a method for correcting for these effects and shows by Monte Carlo means that the corrected statistics behave rather well.

Since the theoretical discussion in this paper deals mainly with the conditional distribution of these statistics given all but one of the variables, I, too, will focus on this situation. This is the familiar regression model where one considers the independent variables as given. Under spatial correlation, regression estimates are affected in at least two ways. One is the loss of efficiency, which affects parameter estimates as well as measures of association. The other, perhaps more serious effect is that the distribution of any statistic under the hypothesis of independence is often quite far from that under spatial correlation.

There have been some efforts to address both problems simultaneously using generalized least squares. In such an approach the covariance matrix $\mathbf{\Sigma}$ needs to be estimated, and given that usually there are far more elements of such a matrix than observations, some restrictions need to be placed on this matrix. Richardson's assumption that elements of $\mathbf{Cov(U)}$ be given by a step function on $\mathbf{A} \times \mathbf{A}$ is such an assumption, and, if the steps are large enough, can form the basis on which to estimate $\mathbf{\Sigma}$. Such an estimate would fall under the general rubric of EGLS estimation (Judge, *et al.*, 1985). Several methods exist for obtaining estimates of $\mathbf{\Sigma}$ and thereby the parameter estimates and measures of association (including Rao's MINQUE theory — see Rao, 1973). The difficulty of this approach in many applications is that the estimate of $\mathbf{\Sigma}$ is not positive definite. In that case it is not always clear what needs to be done, although a sometimes satisfactory way out is to delete some observations.

Richardson has taken a different approach, addressing only the second problem. She uses the traditional measures, but corrects them for spatial correlation. The method, though ingenious, is not entirely new. Approximating the distribution of beta–distributed random variables by a beta distribution with parameters chosen on the basis of the first few parameters has been used before (*e. g.*, Theil and Nagar, 1961). The formulæ are easy to compute and the Monte Carlo comparisons show the measures to be functioning well.

If this commentator has any significant questions, they are with the Monte Carlo comparisons. Richardson has compared modified statistics largely with statistics that totally ignore spatial correlation. I would like to see comparisons with EGLS versions of the statistics. If her approach performs as well as or better than these, then her contribution would have been most valuable.

## References

Judge, G., W. Griffiths, R. Hill, H. Lutkepohl, and T-C. Lee. (1985) *The Theory and Practice of Econometrics*, 2nd ed. New York: Wiley.

Rao, C. (1973) *Linear Statistical Inference and its Applications*. New York: Wiley.

Theil, H., and A. Nagar. (1961) Testing the independence of regression disturbances. *Journal of the American Statistical Association*, **56**, 793-806.

*Ashish Sen*, University of Illinois/Chicago

## *PREAMBLE*

*A man should first direct himself in the way he should go.*
*Only then should he instruct others.*

Buddha

*This maxim is being exercised here by Upton, who reports on results stemming from problems encountered in studying the geographic distribution of voting changes over time. Statistical analysis of data can tell us a great deal about the nature of reality, although as yet it does not fully disclose a deep understanding of geo-referenced data. To this end, the future of spatial statistics in our age of sophisticated computing often is hotly debated, as can be seen in the earlier commentary made by Ripley on Ord, and by Ord on Griffith, for example, in this volume. Upton's paper follows the emerging tradition of computer-intensive statistical analysis that is so characteristic these days of a computer-rich scientific research environment. The purpose of this paper is to outline the problem of interpolating regional values, graphically representing regional data, and using regional data to make inferences about individual data (known as the Fallacy of Division in Logic, and more popularly known as the ecological fallacy). Upton continually gives instructions based upon his empirical experiences. Griffith's reaction to this paper is that Upton addresses two questions that deal with unresolved issues in spatial statistics, while addressing a third that has received considerable attention in the cartography literature. He further notes that again, as both Martin and Richardson, in her commentary on Martin's paper, point out, a fuller dialogue is needed between quantitative geographers and professional statisticians in order to erase such unawareness gaps.*

The Editor

# Information from Regional Data

Graham J. G. Upton

*Department of Mathematics, University of Essex, Colchester, Essex, CO4 3SQ, England*

**Overview:** This paper is concerned with three problems relating to regional data, namely, interpolation of regional values, pictorial representation of regional data, and the use of regional data to make inferences about individual data. In Section 2, regional values are treated as point values, with their regionality being recognized by giving particular weight to the values for neighbouring regions. A general weight function of the form $w = \text{(Population)}^i\,\text{(Area)}^j / \text{(Distance)}^k$ is used, where the distance concerned is that between the point representing the region of interest and the point representing the target region. Optimal values appear to be $i = 0$, $j = 1$, and $k = 3$, for first and second neighbours, with zero weights for more distant regions. A region may be small in size but have great importance—for example, Greater London in the context of population. This familiarly leads to the production of cartograms. Section 3 introduces two computer–based approaches to the production of cartograms, one based on the representation of regions by an appropriate number of points, and the other, due to my colleague, Dr. D. Fremlin, based on treating regions as polygons. Both methods lead to revised maps in which regions have areas proportional to importance. Section 4 is concerned with deducing individual values from regional values, and outlines a novel approach due to the Danish political scientist, Dr. S. Thomsen. Thomsen's method relies on the assumption that a cross–tabulation of two variables of interest may be regarded as a discretization of a bivariate normal distribution, with the underlying variables having values dependent on a common set of unmeasured (latent) variables. In the example considered, the method proves outstandingly accurate in retrieving information at the individual level from aggregate data.

## 1. Introduction

Motivation for the present work stems from problems experienced in carrying out a study of voting changes, which occurred between 1983 and 1987, in English constituencies (Upton, 1989). This study was concerned with disaggregating these voting changes into a number of components. Components of especial interest to political scientists concern the effect on the fortunes of a party of fielding a female candidate as opposed to a male candidate, and the extent of the personal allegiance that a sitting Member of Parliament might gain. The study suggested that, in a constituency in which 50,000 voted, the effect of fielding a female candidate resulted in a loss of between 125 and 250 votes, and that the advantage of incumbency was worth between 400 and 500 votes. Despite the small size of these effects, there were four constituencies (out of 512) in which the outcome might have been different if the candidates had been of the opposite sex.

In attempting to estimate such small effects, it is evident that one must reduce the "noise" in the data so far as is practicable by controlling for the variation between the characteristics of the constituencies. Some control would be possible using covariates such as the social class breakdown of the constituencies, or a breakdown of housing types. However, in the earlier paper I took the view that these should be regarded as genuine local effects, and instead

attempted to control for both tactical voting (using the 1987 voting profile) and geographical variation.

In the 1960s and 1970s, the changes between elections in allegiance towards the competing parties were remarkably homogeneous; a national swing towards the Conservative party, say, would be as apparent in the changes experienced in a Northumbrian constituency as in a Cornish constituency. However, in the 1980s, marked regional variations had become apparent, with the South of England moving away from the Labour party, so that it became commonplace to talk of a "North–South divide." Cozens and Swaddle (1987) have remarked that "[the election of] 1987 witnesses the emergence of fully–fledged regional patterns in Britain" (see also Johnston, 1985). Of course, while there was no sharp line separating North and South, the presence of this "divide" did imply that a change in the allegiance of a Northumbrian constituency between 1983 and 1987 was no longer a good guide to changes in the South.

In order to correct for geographical variations across the country, in my *Electoral Studies* paper I considered each constituency as being represented by a single point on the two–dimensional plane. The value of the voting change to be expected in a constituency then was taken to be that estimated by a weighted average of the voting changes experienced in all the remaining English constituencies, and this estimated change was subtracted from that actually experienced, leaving behind a residual in which the gender and incumbency effects would play a more prominent role. Clearly, different weighting procedures will lead to different estimated changes, and I chose weights based upon an inverse power of the distances between the notional point positions of the constituencies. The inverse power chosen was that which produced estimates most closely resembling the changes actually experienced, resulting in the use of an inverse square law.

Having estimated the residual local effects, a geographical element was still visible, implying that the weighting procedure adopted was sub–optimal. Therefore, a major theme of this present contribution is the comparison of alternative weighting schemes. For the most part we shall use data referring to the states of the United States, since these areal units form a more tractable data set. We shall consider several different sets of data, and will provide some general recommendations concerning a likely optimal procedure for interpolating regional data.

Having estimated the magnitudes of local effects, by whatever method, it is natural to consider presenting these estimates in the form of a map. However, the majority of the population live in conurbations, and this is reflected in a clustering of the positions of administrative groupings (*e. g.*, the English parliamentary constituencies). Using standard geographical co–ordinates to represent these areal units results in a map that is difficult to interpret, since the details of the fluctuations in local effects within the conurbations are almost invisible, unless the entire map is made unacceptably large.

Consequently, a second theme is the construction of scaling procedures that give equal prominence to each constituency. With constituencies represented as points, this problem implies a rearrangement of the points so that they are approximately uniformly distributed. With constituencies represented as polygons, achieving this goal implies a rescaling of the polygons, preserving their connectivities between regions, so that regions of equal population occupy the same area on the revised map. This latter approach, which results in the production of a cartogram, is not confined to a treatment of population, and can be easily

extended to give a "fair" representation of other aspects of regions. The work reported here is that of my colleague, Dr. Fremlin, and I am grateful to him for allowing me to report it in advance of publication.

The final theme of this paper is the so–called "ecological fallacy." This concept is concerned with the extent to which the actions of individuals can be deduced from knowledge of their aggregate behaviour. Here I report on the work of the Danish political scientist, Professor Thomsen, who has devised a method for estimating the body of a transition table from knowledge of its margins.

## 2. Spatial interpolation

### 2.1. Alternative approaches

Assume that we have $n$ irregularly shaped regions whose positions are known, and that we wish to estimate the value of some variable, $y$, for a particular region, $R$, using the known value of $y$ for at least one other region.

The motivation here is that the value of $y$ for region $R$ is either unknown or is suspected of having been influenced by some local factors. In the first case we have a genuine missing data problem, while in the second case (which prompted this study) it is the residual local effect that we are effectively estimating.

A number of general approaches to this problem could be considered. If we have information for all (or most) of the regions, on a vector of other relevant variables, say $x$, then we probably would do best to model the variation in $y$ across the regions using standard multiple regression techniques for some function of $x$. In order to account for spatial effects, we might include location as a parameter in the model, or we might allow for a spatially correlated error structure (cf., Upton and Fingleton, 1985, Ch. 5). Two cases now arise, depending upon whether or not the $x$ values for region $R$ are known.

If the $x$ values for region $R$ are known, then substitution of these values into the regression model, using the values of the parameters estimated from the data from the other regions, will result in an estimate of the $y$ value for region $R$ whose reliability can be gauged from the associated standard errors.

If the $x$ values for region $R$ are not known, then a possible procedure would be as follows. First, use some interpolation procedure to estimate the $x$ values for region $R$, then estimate the $y$ value for region $R$ using the fitted relation between $y$ and $x$ as described above. This approach appears to be novel, its properties are unknown, and it remains a topic for further research.

If there is no information on any relevant $x$–variables, then we must base our estimate on the available $y$–values, and the problem becomes one of interpolation in two dimensions.

### 2.2. Areal interpolation

Although there is a very large amount of literature on spatial interpolation, very little of it refers to the interpolation of data of the regional form being considered here. The special characteristic of the present data is that the $y$ values refer to aggregate (or average) values for mapped regions whose boundaries are known. From the map we can obtain several pieces of information, which include the area of each region, the proportions of the edge of each region that border every other region, and the contiguities that exist between regions.

Interpolation could be based on any of these discernible properties of the regions on the map. To these we add population (if that information is available), since, if the variable $y$ has any relation to human attributes, then it seems plausible to suppose that if $R$ has two equi–sized neighbours (see Figure 1*a*), then that region having the greater population would have the greater bearing on the $y$ value for $R$.

Kennedy and Tobler (1983) suggested that interpolation should be based solely on the information provided by contiguities between neighbouring regions, with weights being proportional to the total lengths of contiguous edge. However, this procedure requires the storage of a non–trivial amount of information, and also can lead to difficulties, as illustrated in Figure 1*b*. This figure shows Region $R$ bordered by two regions of equal size, $S$ and $T$. Logic suggests that each should be equally weighted, yet the Kennedy–Tobler procedure would result in region $S$ being given much greater weight as a consequence of the jagged boundary between $R$ and $S$.

"Jagged" boundaries result naturally when two regions are bounded by a natural feature, such as a river, while "straight" boundaries result from the use of lines of longitude or latitude. Accurate measurement will be difficult in the case of river boundaries. Further, if the river broadens into a lake, then a decision has to be made as to whether or not the two regions really have a common border at that point (after all, regions separated by the sea—the ultimate lake—would not be regarded as bordering one another).

Tobler and Kennedy (1985) describe the application of their procedure to the interpolation of values on a regular pixel mesh, and to interpolation on an irregular resel network (the states of the United States). They also extend their procedure to include second neighbours, giving the necessary formulae for the pixel mesh, and an illustration of the results for the resel case. One should note that their Figure 4 is a fine example of misleading statistics! This figure shows a scatter diagram of fitted $(\hat{y})$ and actual $(y)$ values together with the regression line of $\hat{y}$ on $y$. This line is $Y = 39.3 + 0.52y$ (there is a typographical error on the figure) and leads to a multiple correlation $R^2$ value of 0.72. However, the true predictor is $\hat{y}$, not $Y$, and the relevant line is the 45 degree line $\hat{y} = y$, which has a much smaller $R^2$ value. The authors' claim that the value of 0.72 represents "an average success rate of 72%" is meaningless.

Judging by the recent review of Lam (1983), the Kennedy–Tobler procedures are the only ones that have been suggested for the direct interpolation of one regional $y$ value from its companions. With one exception, all other methods discussed by Lam in her review paper are of the "overlay" type in which the $y$ values from one geographical subdivision, the "target" zones, are computed from those for some alternative geographical subdivision, the "source" zones.

The exception is a procedure due to Tobler (1979) that allows one to interpolate sub–regional values from aggregated regional values. A brief description follows, and a simple worked example is provided by Lam (1983).

When a $y$ value is quoted for a region and a different $y$ value is quoted for its neighbour, then, if we assume homogeneity within the region, this implies that there is a sharp discontinuity along the edges of the regions. However, when one crosses the border from one region to its neighbour, one does not necessarily expect to immediately perceive that fact—only rarely is there a clear demarcation line in terms of ground cover, population density, social

**Figure 1.**

The effects of boundary and population on regional interpolation.

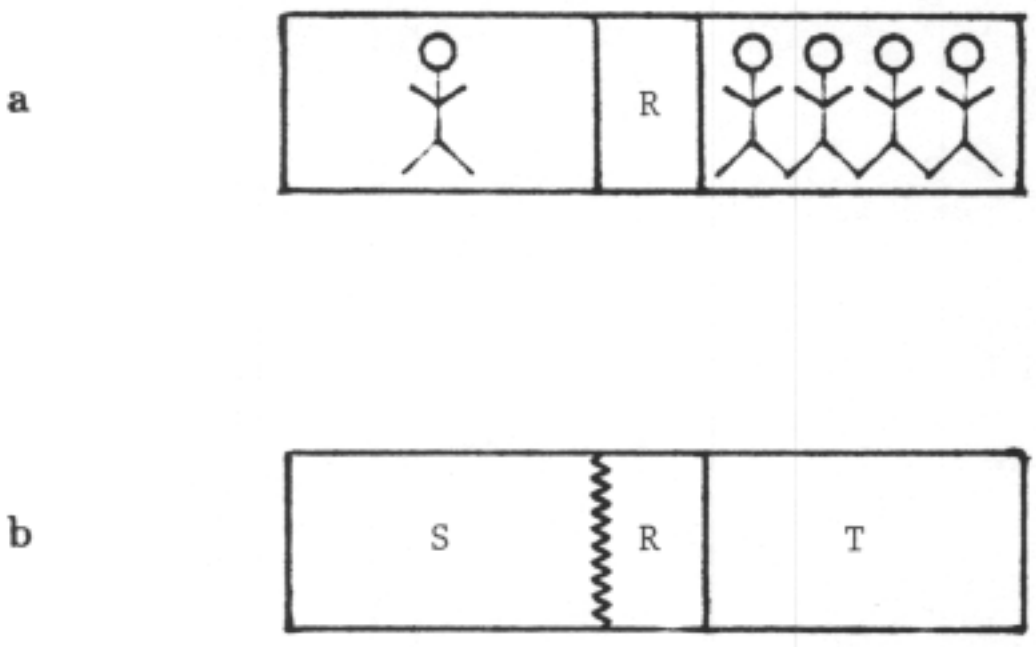
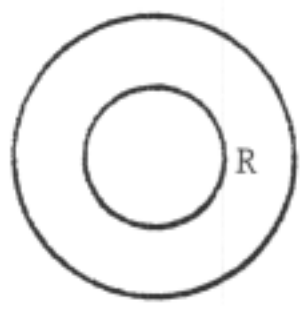



**Figure 2.**

A pathological example of a problem with the centre of gravity.



class split, or whatever. Tobler suggested a possible procedure for introducing a smooth lack of homogeneity into regional data so that these edge discontinuities are removed.

In order to illustrate Tobler's idea, consider the example of population density. For a region $R$, of area $A$ and population density $\rho$, the total population is clearly $A\rho$. Suppose that the point $(x, y)$ lies within $R$. Under the homogeneity assumption, the density at $(x, y)$ is $\rho$ for all points within $R$, thus leading to a plateau of density with edge–discontinuities. Tobler outlines a procedure for replacing the constant density $\rho$ with a position–dependent density $D(x, y)$, which is such that

(i) $\iint D(x, y)dx\,dy = A\rho$ ,

(ii) there is no discontinuity along the edges of neighbouring regions.

Because of the double integration involved, this procedure is said to be volume preserving, or *pycnophylactic*.

### 2.3. Point interpolation

We now move from these area–based procedures to alternatives in which the $y$ value for a region is assigned to some central point. Tobler and Kennedy (1985) assert that their procedure is "far superior to the use of arbitrary points ... to represent geographic areas," but they cited no evidence to support that statement. Upton (1985) showed that the estimates of state population density obtained using points were at least as good as those obtained by Kennedy and Tobler (1983), and further evidence of this is presented later in this section. The Tobler–Kennedy statement probably derives from the critique by Porter (1958) of the use of arbitrary points in the construction of isopleths.

A major problem that arises when representing an area by a "central" point is the choice of location for that point. As Figure 2 illustrates for a pathological case, there may be no point lying within the region that could reasonably be described as its centre! Moreover, even for a circular region the choice of location for the representative point is not clear cut. Suppose, for example, that this circular region consists of an uninhabited upper semi–circle containing forest, and a lower semi–circle representing a densely populated urban area. If our interpolation is concerned with some variable that is independent of population and afforestation, such as rainfall, then the centre of the circle would be a suitable choice as the representative point. If we are concerned with some property of forests, then we should choose a location in the upper semi–circle, while if we are concerned with people, or people's attitudes, then a location in the lower semi–circle would be appropriate.

In this study, these niceties have been ignored and the same central points have been used throughout. In the case of the contiguous regions of the United States, the "centre" has been taken to be the co–ordinates reported for that region in a standard atlas. In the case of the English constituencies, the "centre" has been taken to be the approximate co–ordinates of the population centre of gravity.

The huge literature on spatial interpolation using point values is spread over many disciplines, but almost exclusively deals with the case where a $y$ value refers to an actual point location. There are two broad approaches to the interpolation of data of this type. One approach is to postulate the existence of some underlying continuum and fit a so–called trend surface. The principal problem with this approach is that the order of this surface is not known. An underlying first–order surface (a plane) is unlikely to be the case,

and once we move to higher–order surfaces we may achieve rather improbable estimated surfaces, leading to implausible estimates that lie outside the feasible range for $y$ (Crain, 1970). One alternative here is to fit "local" surfaces subject to edge constraints that obviate discontinuities; this approach can be very expensive in terms of computational effort (see, for example, Haining *et al.*, 1984).

The second approach is to use, as an estimate, a simple weighted average of the available $y$–values. One major disadvantage here is that there are a very large number of plausible alternative weighting schemes. A second disadvantage is that this approach is innately conservative, in the sense that every estimated value will lie within the range of the observed $y$–values so that maxima will always be under–estimated and minima will always be over–estimated. Nevertheless, this approach has the great virtue of simplicity, and it is for that reason that we choose this approach for further study. In the context of genuine point values, Lam (1981) notes that this approach almost always leads to reasonable results.

Upton (1985) used inter–regional distances alone as the basis for weights. However, this is not really plausible as Figure 3*a* demonstrates. Regions $S$ and $T$ have centres that are equi–distant from the centre of $R$, but it would seem strange not to give more weight to $S$ than to $T$.

Further evidence of the need to take area into account is provided by Figure 3*b*, which shows region $S$ divided into two sub–regions, $S_1$ and $S_2$, that have centres equi–distant from the centre of $R$. Suppose that the value of $y$ is uniform throughout $S$. Then before subdivision we have one $y$ value at a distance $d$ from the centre of $R$, and after subdivision we have two such values—yet nothing has changed but our imposition of a dividing line through $S$. If we weight by the area of $S$, then a single $y$–value is replaced by two half–weight $y$–values, and sanity is preserved. Note that the Kennedy–Tobler procedure also deals effectively with this problem.

Clearly area cannot be the sole ingredient of the weight function, since we can expect positive spatial correlation to prevail such that nearby $y$–values will have more relevance than more distant $y$–values. We can measure distance in many different ways (for example travel times), but the two methods considered here are in terms either of "flat–earth" distance or of contiguities.

In the case of the English constituencies, the co–ordinates used were based upon the standard National Grid, while in the case of the United States the distances were calculated in the following way. Let $(x_r, y_r)$ and $(x_s, y_s)$ be the co–ordinates (latitude, longitude) of the centres of regions $R$ and $S$; then $d_{rs}$, the distance between these two centres, is defined as

$$d_{rs}^2 = (1.6) \times (x_r - x_s)^2 + (y_r - y_s)^2, \tag{1}$$

where the quantity 1.6 is a rough correction factor adjusting for the fact that one degree of longitude is not the same length as a degree of latitude. The reference to a flat–earth distance is a reflection that the distances calculated are not great–circle distances, and in view of this restriction nothing more sophisticated than equation (1) (which also ignores variations in distance with longitude) seems worthwhile.

The most obvious way to include distance in the weight function is to use some inverse power of distance, $d^{-k}$, although other functions of distance such as $\exp(-ad)$, $\exp(-ad^2)$ and $\exp(-ad^2)/(b + d^2)$ (see Ripley, 1981, Ch. 4) have been suggested. In these latter

**Figure 3.**

The relevance of area to spatial interpolation.

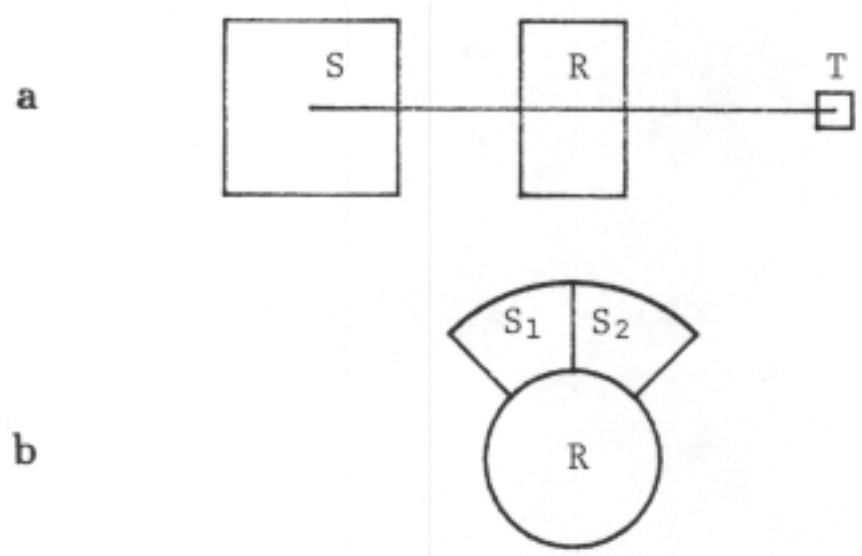
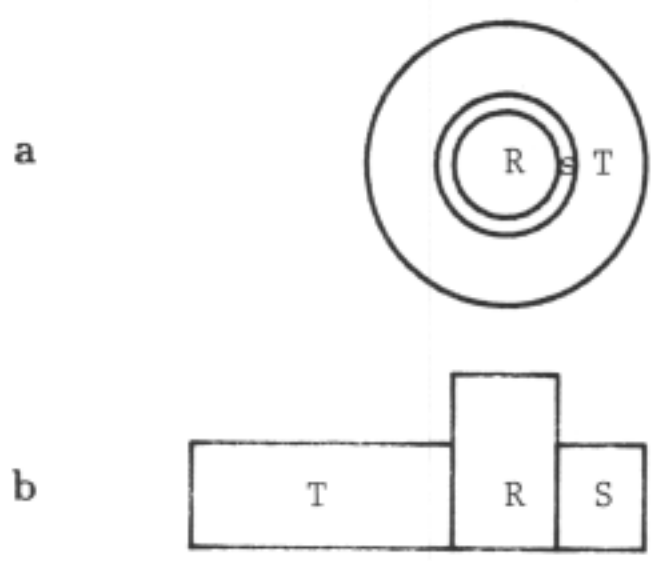



**Figure 4.**

Problems with weights that use edge contact only.



functions the choice of values for the arbitrary constants $a$ and $b$ will be dependent upon the unit of measurement (Stetzer, 1982). Therefore, in this present study the simple $d^{-k}$ weighting is used with the sensitivity of the estimates to the choice of $k$ being investigated, and recommendations being made concerning the optimal value of $k$.

The method suggested by Kennedy and Tobler (1983) implies that remoteness of one region from another could be measured using contiguity rather than distance *per se*, implying estimation using a weighted average of the $y$ values exhibited by first neighbours. Clearly one should not expect to perceive sharply differing local characteristics simply by crossing over a border—any changes are sure to be gradual.

First neighbours alone, however, could be misleading, as the admittedly far–fetched Figure 4*a* illustrates. The region $R$ has only one first neighbour, $S$, but an analyst surely would want to take rather more notice of region $T$ than of region $S$.

Figure 4*b* illustrates another situation. Here regions $S$ and $T$ border $R$ and account for the same proportions of the border of $R$. Region $T$ is very much bigger than region $S$, but in this case one would not really want to give it that much more prominence because its own centre of gravity is so distant from $R$.

Taken together, Figures 4*a* and 4*b* suggest that it would probably be unwise to confine attention only to first neighbours, and that distance still has a role to play. Accordingly, attention now will turn to investigating a number of alternative weight functions of the form

$$w = (\text{Population})^{i} (\text{Area})^{j} / (\text{Distance})^{k}, \tag{2}$$

with $i = 0$ or $1$, $j = 0$ or $1$, and $k$ taking on values in the range 0 to 6 inclusive. These weight functions are used in conjunction with all of the available data, with just the first neighbours, or with both first and second neighbours.

### 2.4. The United States data sets

Although this study was motivated by data from English elections, most attention will be paid to selected, very much smaller data sets relating to the contiguous states of the United States. These data sets are listed in Appendix A, together with the areas of the states and their co–ordinates as given in the gazetteer of the 1973 edition of *The Mitchell Beazley Concise Atlas of the Earth*. The first data set (I) is that used by Kennedy and Tobler (1983), and subsequently by Upton (1985), in their earlier works on spatial interpolation. This data set refers to the population densities of the forty–eight contiguous states (at an unspecified date). The second data set (II) has been the subject of some discussion in the context of the plotting of geographical information. These data were first illustrated using an unclassed choropleth map by Gale and Halperin (1982), and subsequently using rectangle charts by Cleveland and McGill (1984) and Dunn (1987). These data refer to the murder rates in 1978, expressed as murders per 100,000 inhabitants. Both of these data sets omit any specific information concerning the District of Columbia.

The remaining five data sets include information on the District of Columbia, and were chosen so as to provide a spectrum of applications. They were gleaned from the Statistical Abstract of the United States 1986 (table given in brackets) and are as follows:

(IIIa,b) the 1982 mortality rates of infants aged less than 1 (numbers per 1000) (Appendix V),

(IVa,b) the proportions of whites in the 1980 resident population (Table 32),

(Va,b) the population change between 1970 and 1980 expressed as a percentage of the 1970 population (Table 13),

(VIa,b) the percentage of voters supporting the Republican candidate for president in 1984 (Table 412), and

(VII) the numbers of food stamp recipients per 1000 population in 1984 (Appendix V).

The first four of these data sets contain one extreme outlier, the District of Columbia (D. C.), which is a very small densely populated area in which the vast majority of the residents are black. This areal unit has very high infant mortality and has experienced a net exodus in recent years. Because of these features we can expect that all the interpolation procedures will fare much worse when the Washington, D. C., data are included. For this reason these first four *Statistical Abstract* data sets were analysed twice, once (a) excluding the D. C. data, and once (b) including it.

### 2.5. Results for the United States data sets

In order to assess the merits of the various interpolation procedures, the true value, $v_r$ for each region was compared with the interpolated value, $v_r^*$. A summary of the overall effectiveness of a procedure is given by evaluating

$$D = \sum (v_r - v_r^*)^2, \tag{3}$$

with the sum being over either the 48 contiguous states or over the 49 regions including the District of Columbia, as appropriate.

Since the actual magnitudes of the items in the various data sets differ considerably from each other, it is necessary to place each set of rival $D$ values on a comparable footing. To this end, the smallest $D$ value in each collection was set equal to 100, and the remaining $D$ values in each collection were scaled accordingly. The resulting values are summarised in Table 1.

The first four columns of Table 1 refer to the particular form of weight function, namely equation (2), that has been used. The upper half of the table refers to cases where $k > 0$, implying that an inverse power of distance is included in the weight function. In the lower half of the table $k = 0$, implying that no explicit account is taken of inter–regional distances. Within each half of the table, three variants of the weight function are applied to all the data, to first and second neighbours only, or to first neighbours only. The variants have $(i, j) = (0, 0)$, $(1, 0)$ or $(0, 1)$. A fourth variant $(i, j) = (1, 1)$ gave values intermediate between the last two cases.

Thus Row 10 of Table 1 uses the grand mean of all the remaining values as an estimate of $v_r$, with no account being taken of either their distance from region $R$ or their area or population size. This naturally leads to very poor estimates, which is evident in the table, where one can see, for example, that for the population density data set (Set I) the value of $D$ was 4.2 times greater than the best value that was obtained. The scaled $D$ values given in this row are (approximately) scaled versions of the variances of the various $y$ values. A

TABLE 1
COMPARISON OF EFFICIENCIES OF 18 INTERPOLATION PROCEDURES ON 11 DATA SETS
(Each data set indexed at 100 for the overall optimal procedure)

| Distance | Population | Area | Neighbours | Data Set I | II | IIIa | IIIb | IVa | IVb | Va | Vb | VIa | VIb | VII | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| * | | | All | 114 | 102 | 109 | 138 | 105 | 156 | 124 | 126 | 105 | 123 | 110 | 119 |
| * | * | | All | 123 | 108 | 110 | 131 | 128 | 151 | 144 | 140 | 120 | 116 | 100 | 125 |
| * | | * | All | 102 | 102 | 103 | 109 | 110 | 103 | 129 | 121 | 106 | 100 | 100 | 108 |
| * | | | 1st + 2nd | 112 | 100 | 101 | 100 | 105 | 135 | 110 | 110 | 105 | 122 | 116 | 111 |
| * | * | | 1st + 2nd | 118 | 107 | 106 | 113 | 127 | 132 | 134 | 125 | 121 | 112 | 104 | 116 |
| * | | * | 1st + 2nd | 100 | 101 | 100 | 105 | 109 | 102 | 118 | 113 | 107 | 100 | 108 | 106 |
| * | | | 1st | 120 | 108 | 120 | 115 | 100 | 105 | 100 | 100 | 107 | 118 | 123 | 111 |
| * | * | | 1st | 119 | 109 | 114 | 113 | 120 | 114 | 130 | 125 | 125 | 114 | 110 | 117 |
| * | | * | 1st | 111 | 108 | 112 | 108 | 106 | 100 | 109 | 109 | 100 | 101 | 114 | 107 |
| | | | All | 420 | 287 | 221 | 169 | 264 | 202 | 217 | 205 | 136 | 140 | 175 | 222 |
| | * | | All | 435 | 325 | 224 | 166 | 293 | 204 | 224 | 208 | 147 | 138 | 174 | 231 |
| | | * | All | 459 | 302 | 229 | 176 | 264 | 199 | 242 | 232 | 141 | 145 | 177 | 232 |
| | | | 1st + 2nd | 221 | 133 | 102 | 100 | 130 | 142 | 119 | 116 | 118 | 123 | 129 | 130 |
| | * | | 1st + 2nd | 253 | 207 | 128 | 114 | 204 | 168 | 134 | 127 | 128 | 123 | 148 | 158 |
| | | * | 1st + 2nd | 221 | 155 | 109 | 105 | 142 | 145 | 124 | 121 | 115 | 122 | 148 | 128 |
| | | | 1st | 165 | 131 | 122 | 115 | 111 | 106 | 100 | 100 | 107 | 119 | 129 | 119 |
| | * | | 1st | 211 | 153 | 129 | 113 | 155 | 128 | 130 | 126 | 125 | 125 | 139 | 139 |
| | | * | 1st | 153 | 154 | 125 | 108 | 130 | 122 | 109 | 109 | 100 | 122 | 143 | 125 |

NOTE. * denotes information that has been used.

large value in this row indicates that the best procedure for that data set was particularly effective in dealing with the observed spatial variations in the data values.

In the case of the procedures that use a distance component, $d^{-k}$, the actual power chosen varies from weight function to weight function and from data set to data set. All values of $k$ between 0 and 6, in steps of 0.25, were considered, and the reported (scaled) $D$ values are those that correspond to the optimal choice of $k$. The variability of $D$ with changing $k$ within a data set, and the variability in the optimal values of $k$ from data set to data set are discussed later.

Table 1 shows clearly that procedures that incorporate a distance component almost always do very much better than those from which this variable is omitted. Generally speaking, when there is no distance component in the weight function, the smaller $D$ values result from confining attention to the immediate neighbours of a region, while taking averages over all the remaining regions is markedly disastrous in the case of the population density and murder data sets.

When there is a distance component included, the picture is less clear as to which regions should be given non–zero weights. In one case (food stamp recipients—set VII) it is preferable to use information from all the regions, in four cases using both first and second neighbours is best (Sets I, II, IIIa, IIIb), in five cases (Sets IVa, IVb, Va, Vb, VIa) using only the first neighbours is best, and in the remaining case (Set VIb) using only the first neighbours is marginally inferior to the other alternatives. In contrast, using population as a component of the weighting procedure does not work well, even in the case of Sets VIa and VIb, which relate to opinion data (proportions supporting the Republicans).

Using a simple inverse distance weighting in order to combine information from all other constituencies (Row 1 of Table 1), which was the method suggested by Upton (1985) and used in Upton (1989), never provided optimal results with these data sets and occasionally led to seriously incorrect estimates. This technique performed least well for the data sets containing information on the District of Columbia (the sets marked with a suffix b), and the contrast with those excluding that information highlights the sensitivity of this procedure to the presence of an outlier.

The Kennedy–Tobler estimates for the population density data (Set I), which were based on edge–weighted first neighbours, with no distance effect, lead to a scaled index of 143, substantially less than the unweighted first neighbour estimate (165), but clearly far from optimal.

Although there is no especial reason for expecting the same weighting procedure to be optimal for every data set, from Table 1 there do appear to be some generalisations that can be made. The index having the lowest overall average is that involving a weighting function of the form Area $/d^k$. There is little to choose between the use of this weighting function applied to first neighbours only (average index 107), to first and second neighbours (average index 106), or to all regions (average index 108). Each of these produced the optimal index for two of the eleven data sets. However, figures given for the distance weighted index values are those for the optimal choice of $k$. Since this optimal value varies from data set to data set, some investigation of the sensitivity of these estimates to the choice of $k$ is required.

Table 2 illustrates the dependence on $k$ of interpolation procedures that use weights given by $w =$ Area $/d^k$. This table shows that the average indices using first neighbours are much less dependent on the choice of $k$ than are the averages based on greater numbers of regions. However, there is a trade–off between remoteness of the observations from the target locality and quantity of observations. For the 49 contiguous regions of the United States, the number of first neighbours varies between 1 and 8, with a mean of 4.5, while the total number of first and second neighbours varies between 3 and 24 with a mean of 11.9.

Combining findings from Tables 1 and 2 suggests that the following five procedures give very comparable results:

(i) the unweighted average of first neighbours,

(ii) the weighted average of first neighbours, with weights given by $w =$ Area $/d^2$,

(iii) the weighted average of first neighbours, with weights given by $w =$ Area$/d^3$,

(iv) the weighted average of first and second neighbours, with weights given by $w =$ Area$/d^3$, and

(v) the weighted average over all regions, using weights given by $w =$ Area $/d^3$.

The simplest of these five procedures is clearly (i), while procedure (iv) has the smallest average index, and also the smallest maximum index for the eleven data sets so far considered. However, all of the above testing has been done on data obeying the contiguities of the United States, and it is clearly sensible to look at data relating to some other region for confirmation of these findings.

TABLE 2
THE EFFECT OF THE PARAMETER k IN $d^{-k}$ WEIGHTED AREA-BASED ESTIMATORS
(Each data set indexed at 100 for the overall optimal procedure)

| | | Data Set | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Procedure | Value of k | I | II | IIIa | IIIb | IVa | IVb | Va | Vb | VIa | VIb | VII | Average |
| All | 0 | 459 | 302 | 229 | 176 | 264 | 199 | 242 | 232 | 141 | 145 | 177 | 233 |
| Other | 1 | 352 | 228 | 158 | 136 | 204 | 167 | 158 | 151 | 119 | 126 | 136 | 176 |
| Regions | 2 | 200 | 144 | 113 | 112 | 140 | 122 | 129 | 121 | 107 | 104 | 108 | 127 |
| | 3 | 132 | 108 | 103 | 112 | 112 | 103 | 132 | 124 | 107 | 101 | 100 | 112 |
| | 4 | 109 | 102 | 108 | 143 | 112 | 127 | 138 | 133 | 111 | 126 | 108 | 120 |
| | 5 | 103 | 104 | 119 | 172 | 120 | 159 | 143 | 142 | 115 | 151 | 119 | 132 |
| | | 22 | 16 | 12 | 10 | 14 | 12 | 9 | 9 | 10 | 11 | 12 | Optimal value of 4k |
| First | 0 | 221 | 155 | 109 | 105 | 142 | 145 | 124 | 121 | 115 | 122 | 148 | 137 |
| and | 1 | 186 | 131 | 103 | 105 | 127 | 130 | 118 | 113 | 110 | 112 | 127 | 124 |
| Second | 2 | 146 | 111 | 100 | 107 | 114 | 110 | 123 | 115 | 107 | 102 | 113 | 113 |
| Neighbours | 3 | 117 | 102 | 102 | 112 | 109 | 102 | 131 | 123 | 108 | 101 | 108 | 110 |
| | 4 | 103 | 102 | 109 | 144 | 113 | 128 | 138 | 133 | 112 | 126 | 113 | 120 |
| | 5 | 100 | 105 | 119 | 173 | 122 | 159 | 144 | 142 | 116 | 152 | 121 | 132 |
| | | 20 | 14 | 9 | 2 | 10 | 12 | 4 | 5 | 9 | 11 | 12 | Optimal value of 4k |
| First | 0 | 153 | 154 | 125 | 108 | 130 | 122 | 109 | 109 | 100 | 122 | 143 | 125 |
| Neighbours | 1 | 146 | 133 | 117 | 110 | 116 | 110 | 116 | 112 | 102 | 108 | 125 | 117 |
| Only | 2 | 134 | 117 | 112 | 112 | 107 | 103 | 126 | 118 | 107 | 101 | 115 | 114 |
| | 3 | 121 | 110 | 113 | 118 | 107 | 101 | 135 | 126 | 112 | 104 | 114 | 115 |
| | 4 | 113 | 109 | 119 | 150 | 113 | 129 | 141 | 136 | 116 | 130 | 121 | 125 |
| | 5 | 111 | 110 | 128 | 177 | 122 | 160 | 146 | 144 | 120 | 154 | 128 | 136 |
| | | 20 | 15 | 10 | 0 | 10 | 11 | 0 | 0 | 0 | 9 | 11 | Optimal value of 4k |

### 2.6. The English constituency voting data

Consider data concerning voting profiles for the 523 English constituencies for the General Elections of 1979, 1983 and 1987. Further, consider only the shares of the vote obtained by the three main parties, with these shares being scaled so as to sum to one for each constituency at each election.

These election data are used in two different ways. First, the change in the percentage of the vote obtained by a party between two successive elections is analyzed. Three parties and two inter–election periods gives rise to six possible data sets, and the values reported refer to the aggregate fit over all six sets. The second treatment is of the actual percentages themselves, and here results refer to the aggregate fit over all nine party/election combinations.

Results reported in Table 3 for the two data set groups are divided into two sections, one incorporating an area component in the weight function and one omitting this component. A convenient source of information about constituency areas is a publication of the Office of Population Censuses and Surveys (1981). Results are given for weights involving inverse distance powers between 0 and 4 inclusive. These results are indexed so that 100 represents the best fit found for that data set.

The findings reported here confirm those for the United States data. If there is no knowledge about the constituency areas, then the simple mean of first neighbours works

very well for both data set groups. If area measures are used in the weighting function, then the minimax choice is $w = \text{(Area)}/d^3$ for first and second neighbours.

A noticeable feature of the calculations that does not show up in Table 3 is the difference in the computational effort between using all of the data and a restricted set of neighbours. For these 523 English constituencies the number of first neighbours varies between 1 and 15, with an average of 5.5, while the combined total of first and second neighbours varies between 2 and 45 with an average of 18.2. The English constituencies have larger average numbers of neighbours than the states of the United States because a smaller proportion of the regions are situated on a boundary.

TABLE 3
ALTERNATIVE $d^{-k}$ INTERPOLATION PROCEDURES
APPLIED TO VOTING DATA FOR 523 ENGLISH CONSTITUENCIES

| Data Set | Procedure | With Area Component | | | | | Without Area Component | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | k: | 0 | 1 | 2 | 3 | 4 | 0 | 1 | 2 | 3 | 4 |
| Inter- | All | 136 | 124 | 106 | 104 | 114 | 123 | 107 | 101 | 110 | 120 |
| election | 1st + 2nd | 120 | 110 | 104 | 108 | 117 | 102 | 100 | 105 | 115 | 123 |
| change | 1st | 116 | 111 | 112 | 119 | 125 | 104 | 107 | 115 | 124 | 131 |
| Party | All | 298 | 245 | 153 | 114 | 115 | 301 | 182 | 144 | 138 | 140 |
| percentages | 1st + 2nd | 179 | 142 | 116 | 108 | 110 | 115 | 107 | 108 | 115 | 124 |
| | 1st | 132 | 118 | 109 | 109 | 112 | 100 | 100 | 106 | 114 | 121 |

### 2.7. Stetzer's results

Stetzer (1982) was interested in a somewhat different problem, namely the optimal choice of a weights matrix for use with STARIMAR models. Stetzer was concerned with point values and simulated data having a variety of autocorrelation structures. Stetzer considered both the accuracy of parameter estimates and the accuracy of forecasts using various weight matrices. For irregular meshes, which correspond to the present situation, Stetzer considered both 0/1 weights based on the connectivities in minimum spanning trees (roughly equivalent to the use of first neighbours with no distance, population or areal weighting), and two alternative distance–decay weightings. These distance weightings took either the form $d^{-1}$ or $\exp(-cd)$, where $c$ was given some appropriate value; but, the weights were set to zero if $d > D_{\max}$, the cut–off distance. Stetzer reports results for three cut–off distances, with the smaller $D_{\max}$ value restricting attention to nearby points and the largest $D_{\max}$ value allowing information from all areal unit values to be used. Therefore, these three values correspond reasonably well to the three levels of neighbour considered in the present study.

Confining attention to the results that Stetzer obtained for the errors in forecasts, three generalizations emerged:

(i) using distance–decay weights, the largest of the $D_{\max}$ values was worst and the smallest was preferable,

(ii) the $d^{-1}$ weights were preferable to the $\exp(-cd)$ weights, for Stetzer's choice of $c$, and

(iii) the simple connectivity weights were often much inferior to the distance–decay

weights.

Although the contexts are distinctly different, these findings are reassuringly similar to those being reported in the present analysis.

### 2.8. Recommendations

If information concerning the areas of the various regions is not available, then apparently the best procedure is to take *the simple average of the immediate (first) neighbours*. However, disregarding the area of a region seems somewhat illogical, since if a homogeneous first–neighbour of a region $R$ were to suddenly disintegrate into $m$ pieces, all bordering the region $R$, then the weight of that region would increase by a factor of $m$. Hence, despite the undoubted success of the simple first–neighbour estimate, an area–based estimate seems more reliable. With large numbers of regions there is a huge computational gain from confining attention to just a few constituencies, though one is reluctant to ever use just one constituency as an estimator of another. For these reasons the preferable procedure would seem to be to use *first and second neighbours, with weighting (Area)* $/d^3$.

Figure 5 displays results for this procedure applied to the population density data. Since population densities have a highly skewed distribution, with most being small (smallest is Wyoming at 3.4) and a few being very large (largest is New Jersey at 953.1), the data have been plotted on a log–log scale, so that all the states receive comparable prominence. This figure may be contrasted with Figure 4 of Tobler and Kennedy (1985).

A perfect estimation procedure would result in all of the points in the diagram lying on the 45 degree line shown. Given the logarithmic scaling, the greatest vertical deviations from the line correspond to the greatest *multiplicative* discrepancies. These are for Nevada (observed 4.4, predicted 102.7) and California (observed 127.6, predicted 6.4). Although neighbours, these states are separated by mountains, and evidently any interpolation scheme that fails to take proper account of barriers of this kind is doomed to failure. The greatest absolute error is that for the state of New Jersey (observed 953.1, predicted 346.4), though as noted before, every interpolation scheme based on weighted averages is certain to underestimate the maximum and overestimate the minimum.

## 3. Redrawing the map

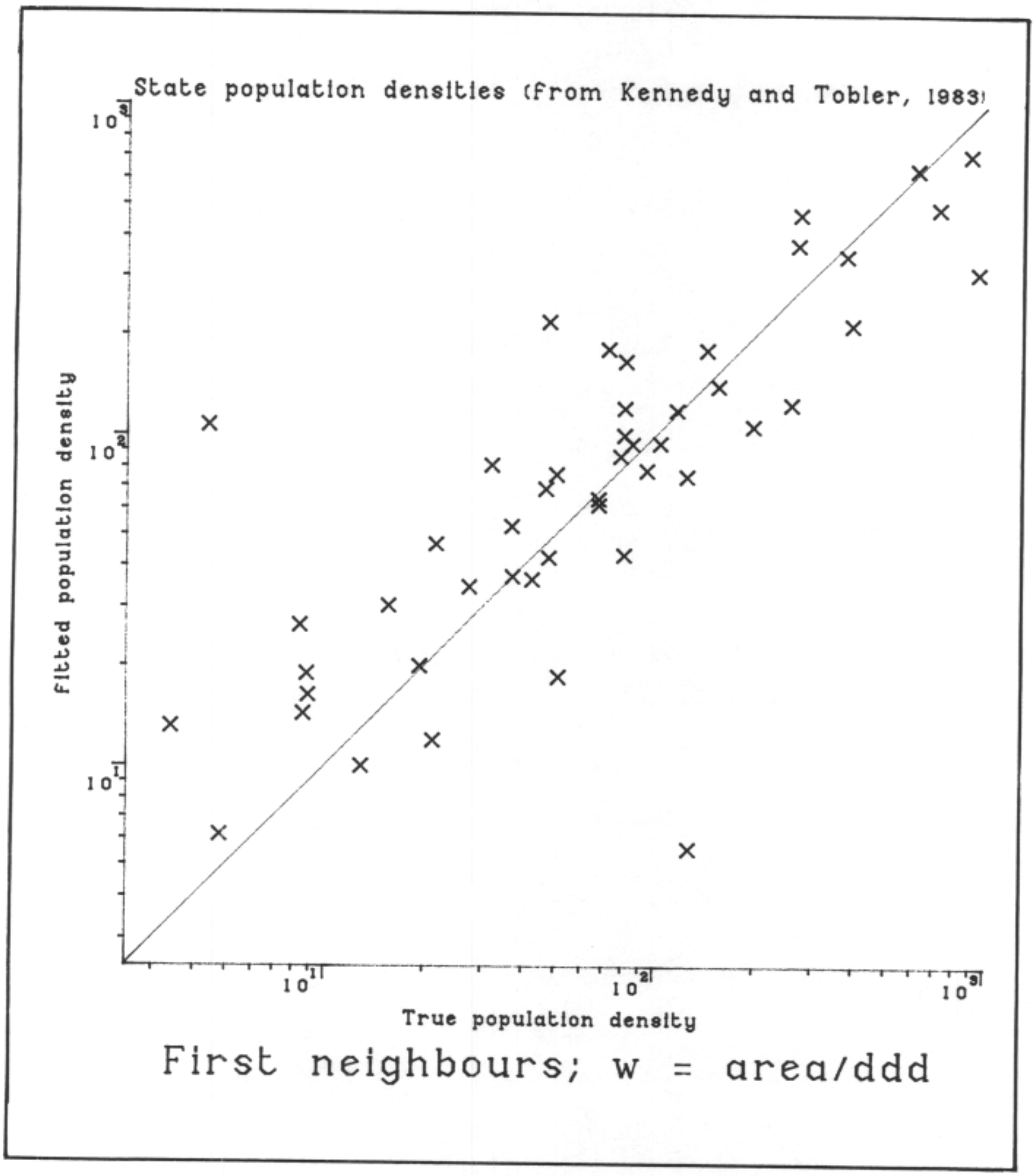
As mentioned in the introduction, the motivation for this section came from a need to coherently display information concerning English constituencies on a single map. One standard procedure for illustrating quantitative geographic information is the use of a choropleth map, described as the "lazy man's map" by Evans and Jones (1981), who discuss various alternatives to it. More recent alternatives involve the use of rectangle charts (Cleveland and McGill, 1984; Dunn, 1987), though these do not address the problems of the visibility of small but important locations.

In this section, two computational approaches will be considered, one resulting in a redistribution of point locations, and the other, due to my colleague, Dr. Fremlin, resulting in a direct redrafting of a map to produce a cartogram. An alternative approach, when there are only small numbers of locations, involves multidimensional scaling using contiguities to specify structure (see, *e. g.*, Gatrell, 1981); this technique is not feasible with present programs when the numbers of locations are much in excess of 100.

The point relocation procedure will be illustrated in the context of the British parlia-

**Figure 5.**

Scatter diagram, using a log–log scale, of observed and fitted population density values.



mentary constituencies that motivated this study; Fremlin's procedure will be illustrated in the context of the United States.

### 3.1. Point locations

Representing each of the 631 contiguous British constituencies (including the Western Isles constituency, but omitting the Northern Ireland constituencies and the remote Orkney and Shetland constituency) by a point located at the approximate centre of gravity of its population, leads to the map displayed in Figure 6. The choice of the point location was subjective and based upon a scrutiny of the relevant constituency map.

In Figure 6, the constituencies have been subdivided into 11 groups following the classification employed by Waller (1983). A breakdown of the allocation of counties to groups is given in Appendix B. The most obvious features of this map are the concentrations of constituencies in Greater London, Birmingham, Liverpool and Manchester, Tyneside, Edinburgh and Glasgow, and south Wales. If one wants to display information about these urban constituencies, then this map will have to be enormous if Central London is to be visible! The more usual solution is to have magnified maps of the conurbations displayed adjacent to a main map, with the conurbation information on these inset maps and the rural information on the main map. This procedure works well unless one is interested in the urban–rural interface, or in national trends rather than local details. It was the latter requirement that motivated this study.

The objective now is to rearrange the points in Figure 6 in such a way that all are equally visible, and so that both the relative and absolute geographic positionings of the constituencies are preserved (*e. g.*, London constituencies remain adjacent to each other and continue to be located towards the South–East corner of the revised map).

Consider, for simplicity, a rectangle of land within England. If the point density within the rectangle is uniform, then it follows that the density of the $x$–coordinates will be uniform, and that the same will be true for the $y$–coordinates. This suggests that if a rectangle contains a non–uniform arrangement of points, with non–uniform spreads of coordinates, then a simple rescaling of the axes might suffice to correct matters. However, Figure 7 demonstrates that this is not the case.

Figure 7 shows a rectangle containing eight points whose $x$–coordinates are equi–spaced, and whose $y$–coordinates are equi–spaced, yet two of these points have considerably greater visibility than the rest. Study of the marginal coordinate distributions provides no information concerning the clusters present in the pattern. Although this is an extreme case, it will be observed that similar features appear in the real data of Figure 6, with London and Glasgow being in opposite corners of the "British rectangle."

Nevertheless, the concept of coordinate scaling underlies the method actually adopted, which involves a scaling procedure in which $x$–coordinates and $y$–coordinates are alternately and iteratively adjusted (one should note that the resulting solution is not unique). Ideally the number of points is arranged to be a power of 2 and, in this case, at each iteration all the data have either their $x$–coordinate or their $y$–coordinate adjusted, with the number of data points being scaled in precisely the same way and being halved each time.

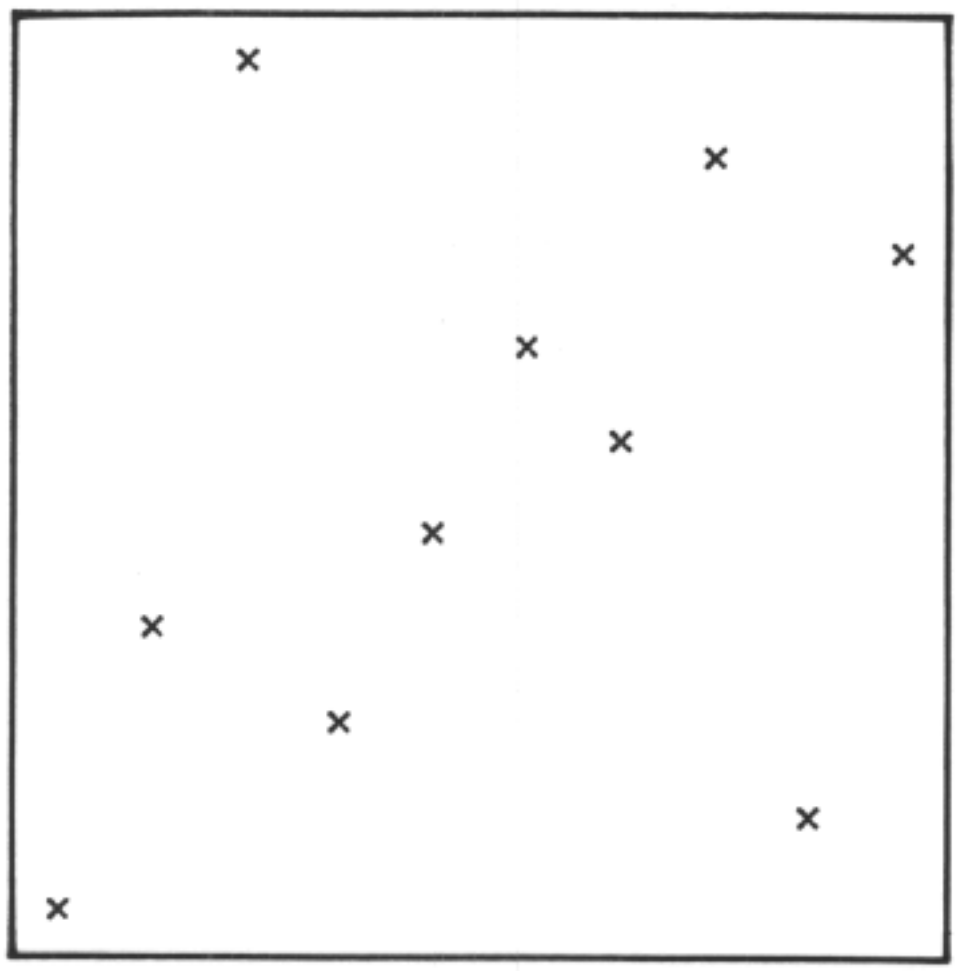
In Figure 8*a* the study area is a rectangle with width $w$ and height $h$. For convenience suppose the origin of the axes is at the bottom left corner. Let $M_x$ be the median $x$–

**Figure 6.**

Location of 631 British constituencies. (Symbols refer to different regions—see Appendix B.)

**Figure 7.**

Ten points, non–uniformly spread, but with equi–spaced $x-$ and $y-$coordinates.



coordinate of the $N$ points. Ideally, as remarked above, $N$ is arranged to be a power of 2; but this may prove impractical, and so a procedure for general $N$ will be described. If $N$ is even, with $N = 2n$, then $M_x$ is the average of the $n$th and $(n+1)$–th ordered $x$–coordinates. If $N$ is odd, with $N = 2n + 1$, then $M_x$ is equal to the $(n + 1)$–th ordered $x$–coordinate. Let $R_x = 2M_x/w$. If the points were distributed evenly over the rectangle then $R_x$ would be equal to 1. If $R_x$ is less than 1 then this indicates clustering toward the left of the rectangle, while $R_x$ greater than 1 indicates clustering toward the right. Denote the group of points having the n smallest $x$–coordinates as group $A$, and the remainder as group $B$. Let $x_A$ denote the $x$–coordinate of a point in group $A$ and $x_B$ denote the $x$–coordinate of a point in group $B$. Revised coordinates $x^*_A$ and $x^*_B$ may be calculated using the formulæ

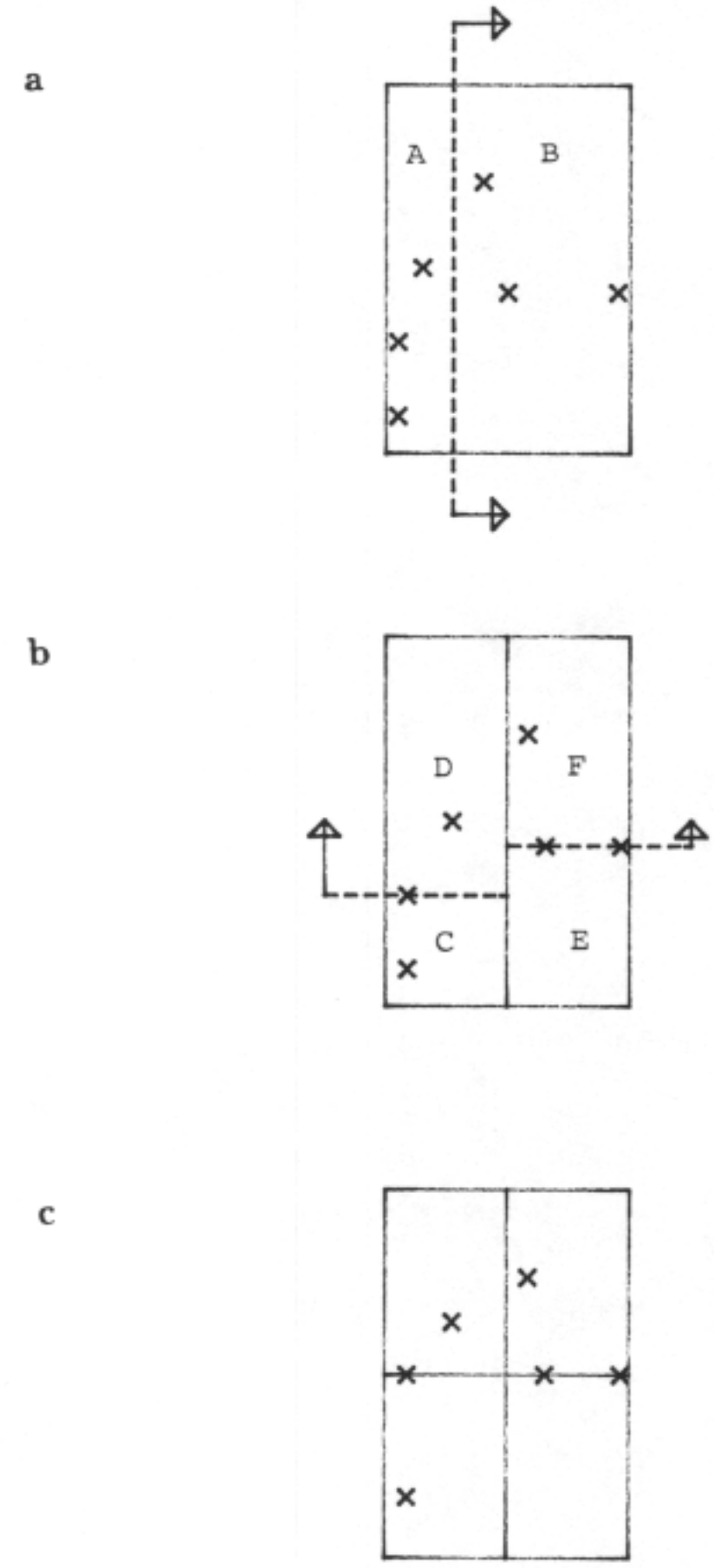
$$x^*_A = x_A/R_x, \quad x^*_B = [w + (2x_B - wR_x)/(2 - R_x)]/2. \tag{4}$$

These computations result in a set of revised $x$–coordinates having median $w/2$, so that in general there are equal numbers of points in the left– and right–hand sides of the rectangle. In the case of $N$ being odd there also will be one point having a revised $x$–coordinate equal to $w/2$, and this $x$–coordinate will remain unchanged throughout subsequent iterations. The adjustments are illustrated in Figure 8*a*, and lead to the revised $x$–coordinates shown in Figure 8*b*.

This first iteration concludes with a scaling of the $y$–coordinates. In this scaling the

**Figure 8.**

One complete iteration of the scaling procedure.



two groups $A$ and $B$ are scaled *separately*. If $N$ is even, then each of $A$ and $B$ contains $N/2 = n$ points. If $N$ is odd, then the point that has the median $x$–coordinate is assigned to either $A$ or $B$, the assignment being based on the location of its nearest neighbour. If its nearest neighbour belongs to group $A$, then this point also belongs to that group; otherwise this point is assigned to group $B$.

Now the $y$–coordinates of group $A$ may be scaled by dividing the points in that group into subgroups, say $C$ and $D$, with the $y$–coordinates of subgroup $C$ being smaller than those of subgroup $D$. The median $y$–coordinate, $M_y$, is calculated in the manner described earlier and, taking $R_y = M_y/h$, the revised $y$–coordinates are given by

$$y_C^\star = y_C/R_y, \quad y_D^\star = [h + (2y_D - hR_y)/(2 - R_y)]/2, \tag{5}$$

which are an exact equivalent of the previous $x$–coordinate equations given by (4) above.

Next, turning attention to group $B$, one may form subgroups $E$ and $F$, calculate the values of $M_y$ and $R_y$ for group $B$ (which will, in general, be different than those for group $A$), and revise the $y$–coordinates for group $B$ using equation (5), with the new $R_y$ value and with the labels $C$ and $D$ being replaced by $E$ and $F$. This procedure is illustrated in Figure 8*b*, and leads to the revised positions shown in Figure 8*c*. This step completes one cycle of the iterative procedure. If any point has been assigned the median $y$–coordinate value $(h/2)$, then this coordinate will be unchanged through subsequent iterations.

At the start of the next cycle there are four separate groups, namely $C$, $D$, $E$ and $F$, each of which may be treated as a separate entity. The scaling formulæ (4) and (5) will require adjustment in order to take account of which quarter of the original square is being treated; nevertheless, the general procedure remains the same. At the end of this cycle there will be eight groups to consider: each successive cycle will result in a doubling of the number of groups to be considered, while the number of points in these groups will be roughly halved on each occasion. The reduction is not exactly a halving because of the presence of an increasing number of points whose coordinates have been fixed by virtue of their occupying a median position. The iterations continue until each subregion contains only a single point that is relocated at the centre of that subregion. One should note that in the case where $N$ is a power of 2, only at the last stage will the final coordinates of a point be determined.

This foregoing description of the procedure assumes that one is working with a rectangular study region. However, if one simply encloses the point locations of Figure 6 within a rectangle and then applies the procedure, the outcome would be an unrecognizable "rectangularized" Britain! In order to avoid this problem the artificial rectangle must be filled up around the existing land constituencies with a lattice of "sea constituencies." These synthetic areal units are positioned on a grid having approximately the same density of points as the land constituencies, so that the general shape of the land is preserved. Small random increments may be added to the lattice point positions so as to avoid problems with multiple tied coordinates; and, these fictitious constituencies may be given identifiers that will prevent them from being plotted when the map finally is redrawn. The revised map given in Figure 9 involved a supplement of 764 sea constituencies to the 631 original constituencies (the supplementary number is large because the land mass of Britain is inclined relative to the National grid, resulting in a wide rectangle—a different coordinate system could result in a closer fitting encompassing rectangle).

That the algorithm has been reasonably successful in giving each constituency approximately equal prominence is immediately apparent. Studies of the patterns of regional symbols confirm that this procedure does not scatter constituencies unduly, and preserves their general relative positions. However, a curious feature of the procedure becomes apparent when Figure 9 is held at eye level and is viewed from the bottom of the page. Apparently there are a number of nearly empty "channels" running through the point pattern. These channels are a consequence of the splitting procedure and arise in the following manner. Consider the initial split of the 1395 (land and sea) constituencies. Precisely one has its $x$–coordinate assigned to the median value $w/2$. At the next iteration there are four groups, two of size 348 and two of size 349. The latter two cause precisely two constituencies to have their $x$–coordinates fixed (at $w/4$ or $3w/4$). At the next iteration there are 16 groups each of size 87, so that four constituencies have their $x$–coordinates fixed at each of $w/8$, $3w/8$, $5w/8$ and $7w/8$. Successive iterations lead to steadily increasing numbers of $x$–coordinates being fixed, with, in general, each subdivision of $w$ being allocated to a greater number of constituencies. However, none of the subsequent allocations are to $w/2, w/4, 3w/4, w/8, \ldots$, and so, inevitably, these channels appear in the data.

Therefore, the channels apparent in Figure 9 are a direct consequence of the fact that the number of points being rearranged (1395) is not a power of 2. With, for example, 1024 points, there will be no premature allocations to any of $w/2, w/4, 3w/4, \ldots$. Instead, in the final stage the points will be allocated to $w/2048, 3w/2048, 5w/2048, \ldots$, and thus all of them will lie on an evenly spaced grid.

Figures 10*a* and 10*b* show the results of confining attention to the 523 English constituencies and surrounding them by 501 "sea constituencies." The figures differ in that in Figure 10*a*, in each iteration, the first adjustment was made to the $x$–coordinates, and the second to the $y$–coordinates, whereas in Figure 10*b*, in each iteration, the first adjustment was made to the $y$–coordinates, and the second to the $x$–coordinates. Examination of the figures reveals that while most of the regions remain as compact entities, there are a few instances in which some constituencies have lost contact with the bulk of their region. Figure 10*b* is noticeably worse in this respect.

In addition to the arbitrary order of coordinate adjustment, there are many other arbitrary elements, such as the precise position of the enveloping rectangle and the choice of positions for the phantom sea constituencies; there is no unique end product.

### 3.2. From points to areas

After all the point positions have been adjusted to their final positions, one then could redraw the constituency boundaries. However, in view of the arbitrary nature of the transformations employed, no attempt will be made here to restore the contiguities present in the original data. But, by connecting each constituency to all its original neighbours, using the revised coordinates, an idea of the distortions that have been induced may be attained. Figure 11 shows results for the English constituencies, using the revised coordinates illustrated in Figure 10*a*.

If all the constituencies were in their same relative positions, then one would see a network of short lines with very few crossovers. Figure 11 shows that this holds true for Greater London and the North West, but, not unexpectedly, the vast expansions of these regions have resulted in considerable stress around their peripheries. The initial binary

**Figure 9.**

Map of Britain showing constituencies in revised positions (symbols correspond to regions).

**Figure 10a.**

Revised English constituency positions with first revision performed on $x$ co–ordinates.

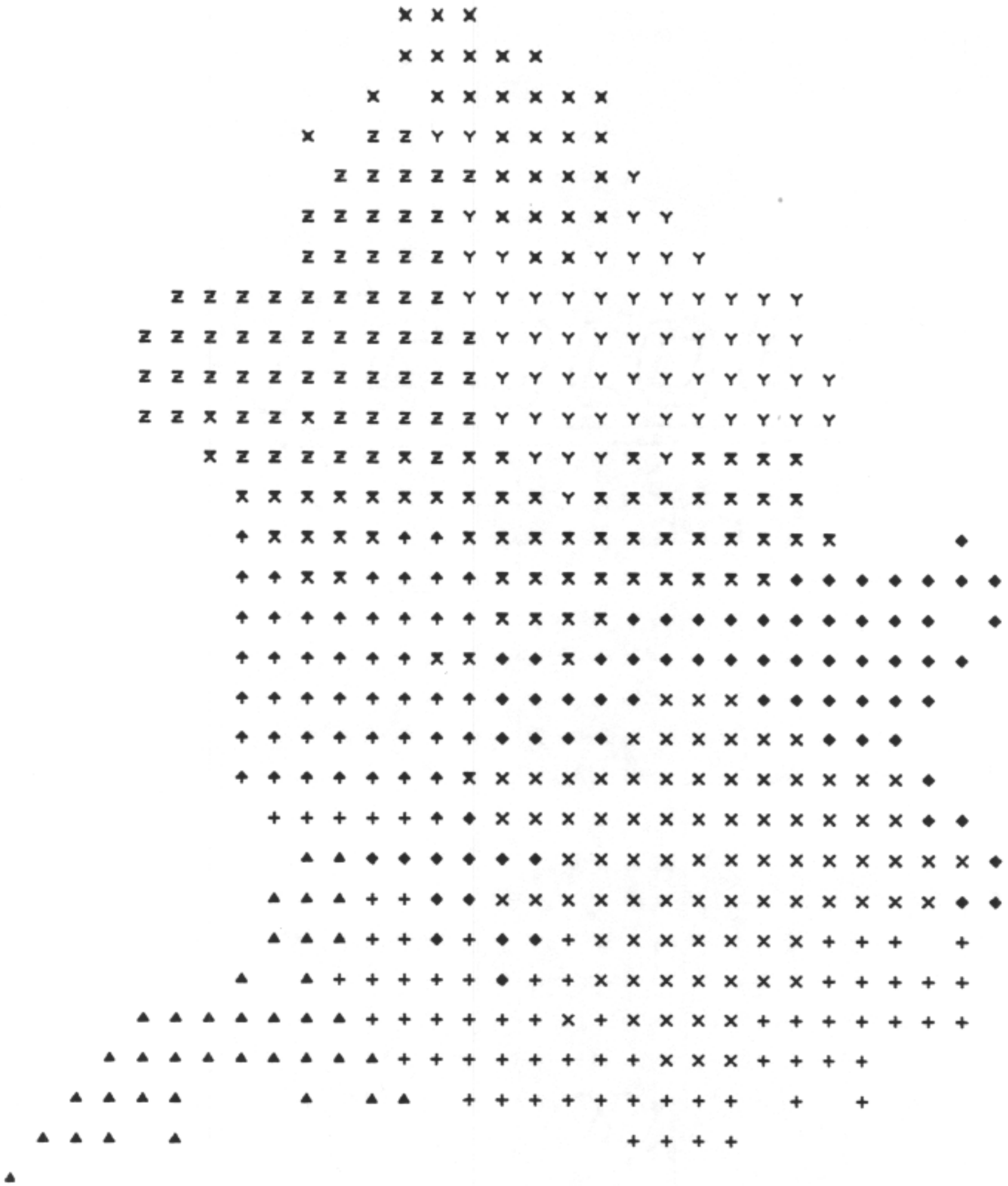
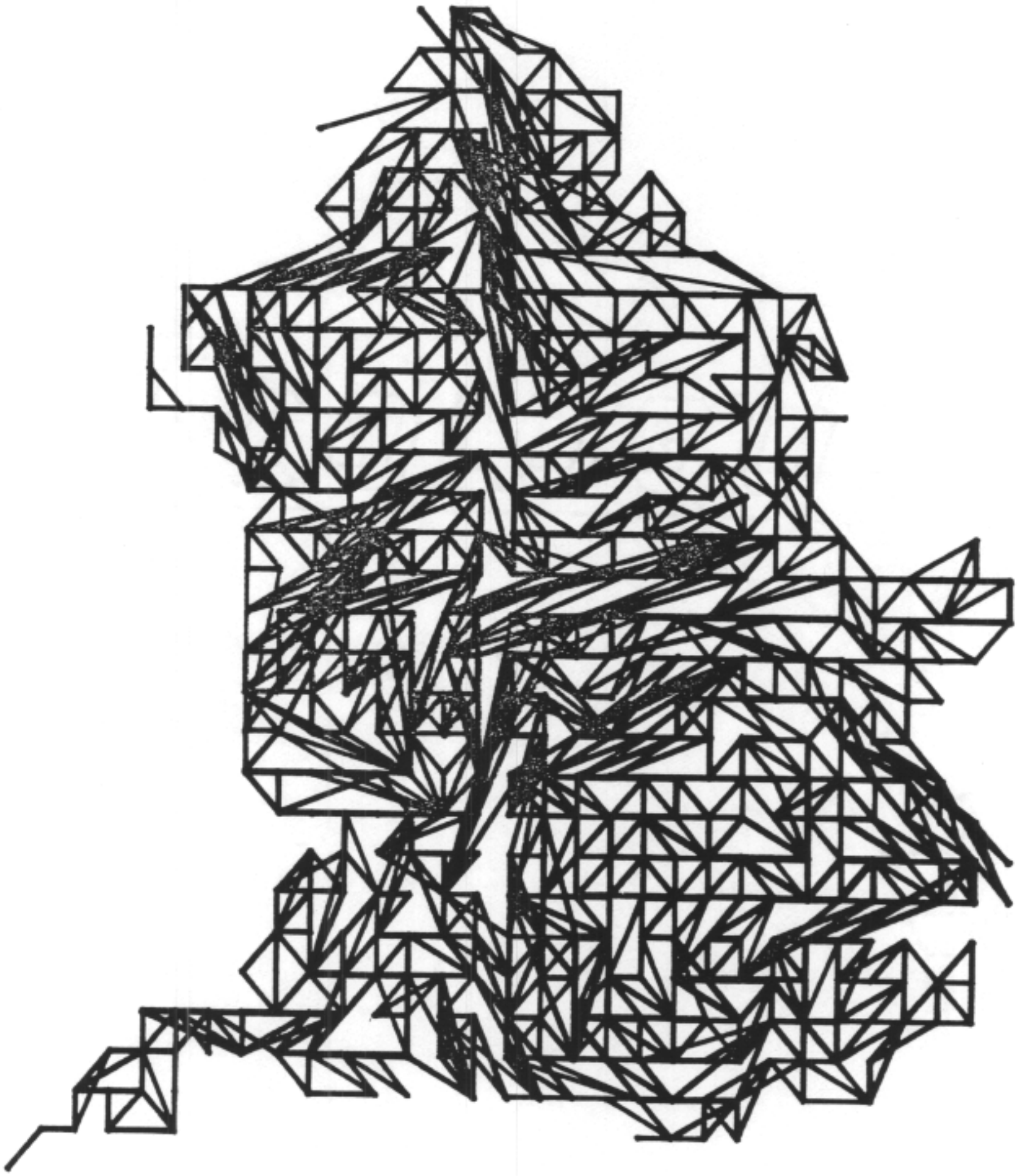


**Figure 10b.**

Revised English constituency positions with first revision performed on $y$ co–ordinates.

**Figure 11.**

Map of revised nearest neighbour connectivities showing the distortions induced by scaling.



splits impart particularly noticeable displacements, which are evidenced not only by the long diagonal lines, but also by the lack of horizontal lines in the central column, as well as the lack of vertical lines in the central row. Those constituencies most affected fall to the east of an expanding Birmingham and lie in the west of Yorkshire, which are being squeezed by the simultaneous expansions of Tyneside in the North region and Merseyside in the North West.

In order to quantify the stress, a list was made, for each English constituency, of its ten nearest neighbours. After each iteration a count was made of the number of these nearest neighbours that remained among the ten nearest constituencies according to the revised positions. The results are summarised in Table 4.

Table 4 shows that in the final map transformation arrangement (Figure 10*a*) all constituencies retained at least one of their original ten nearest neighbours in their revised ten, with only two retaining all ten original neighbours. Not unexpectedly, the greatest changes are experienced during the first few iterations. Constituencies affected most adversely are listed in this table, and interestingly these are not persistent, which indicates that subsequent iterations can "repair the damage" done by previous iterations. Bosworth, Sutton Coldfield, and Bromsgrove are constituencies in central England in the neighbourhood of Birmingham that become displaced because of the simultaneous expansions of the Birmingham and Greater London conurbations. Skipton and Ripon, Bishop Auckland, and Keighley become displaced because of the simultaneous expansions of the Tyneside and west Yorkshire conurbations. In terms of actual first neighbours, Bishop Auckland retains 3 out of 5, Bromsgrove 4 out of 10, and Keighley 2 out of 4; the figures in Table 4 are pessimistic with respect to these constituencies, since each is physically close to a multi–constituency urban centre.

TABLE 4
DISPLACEMENT OF THE TEN NEAREST NEIGHBOURS OF 523 ENGLISH CONSTITUENCIES AFTER EACH ITERATION OF THE ADJUSTMENT PROCEDURE

| After | Number Still in Nearest 10 | | | | | | | | | | | Mean | Worst |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Iteration | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Number | Cases |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 523 | 10.0 | * |
| 1 | 0 | 0 | 1 | 3 | 2 | 6 | 9 | 14 | 58 | 210 | 220 | 9.1 | * |
| 2 | 0 | 0 | 1 | 7 | 10 | 19 | 31 | 68 | 124 | 179 | 84 | 8.2 | * |
| 3 | 1 | 0 | 2 | 15 | 18 | 51 | 71 | 118 | 121 | 93 | 33 | 7.2 | Skipton, Ripon |
| 4 | 0 | 1 | 5 | 14 | 23 | 55 | 115 | 120 | 101 | 69 | 20 | 6.8 | Bosworth |
| 5 | 0 | 2 | 8 | 16 | 37 | 81 | 127 | 115 | 81 | 47 | 9 | 6.4 | Bosworth, Sutton Coldfield |
| 6 | 0 | 3 | 8 | 21 | 68 | 116 | 114 | 103 | 68 | 20 | 2 | 5.9 | Bishop Auckland, Bromsgrove, Keighley |

Figure 11 shows the stress induced when an arbitrarily oriented rectangle including arbitrarily placed "sea constituencies" was used to surround the genuine constituencies. The placement of the "sea constituencies" and the size of the rectangle used are unlikely to make much difference to the outcome; but an improved rearrangement undoubtedly could be obtained by choosing the orientation of the rectangle so as to minimize the stress induced. An untested hypothesis is that the rectangle of minimum area would be near optimal.

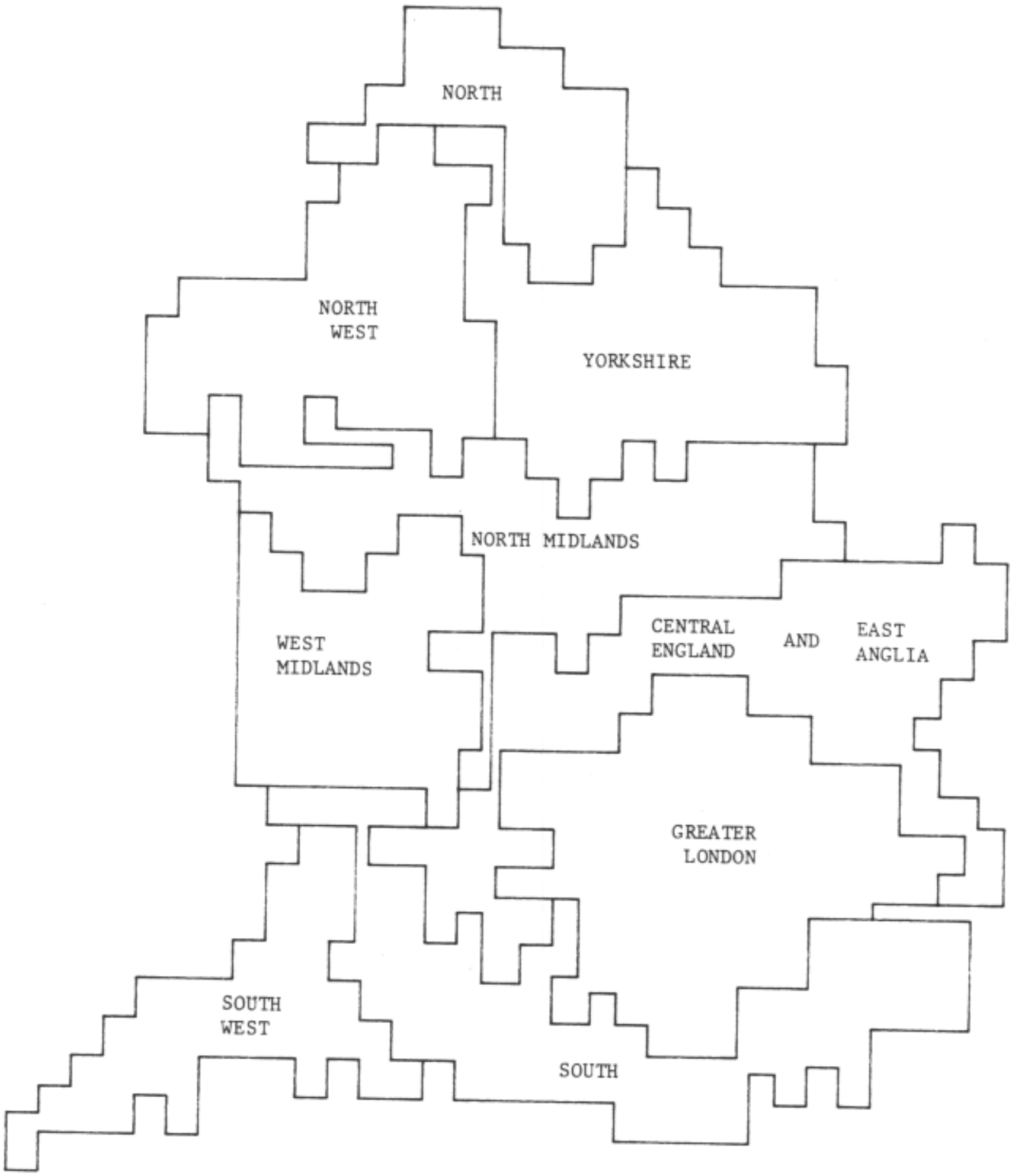
The array of symbols in Figure 10*a* is potentially distracting, so Figures 12*a* and 12*b* show the effects of rescaling in areal terms. Figure 12*a* shows the regions of England as defined by Waller (1983), while Figure 12*b* is a rendering of Figure 10*a* into a comparable

**Figure 12a.**

Map of England showing regions before constituency rearrangement.

**Figure 12b.**

Regions of England after rescaling (from Figure 10a).



form. This figure suggests that the North Midlands and Yorkshire regions have become noticeably misshapen, and have developed a number of "pseudopodia" corresponding to the few severe constituency displacements noted earlier.

Before leaving the English voting data, Figure 13 shows the application of this procedure to the study of voting behaviour within the context of the 1987 general election. The symbols $C$, $L$ and $A$ have been used to denote constituencies in which the Conservative, Labour and Alliance parties had majorities of more than 10% over their nearest rival. Marginal seats (less than a 10% lead) are indicated by one of the integers $1, 2, \ldots, 6$, depending upon the voting outcome. The main features apparent from this figure are the North–South divide, with Labour seats primarily in the North and the conurbation centres, and the clusters of marginal constituencies on the urban–rural interfaces of the West Midlands and the South-East of London. The boundaries marked are those from Figure 12*a*.

One should note that the type of data portrayed in Figure 13 can be much better displayed in colour. With colour one replaces the $C$, $L$ and $A$ symbols by different coloured dots, and the six integers by the letters $C$, $L$ and $A$ in an appropriate colour. The colour represents the party that won the seat and the letter identifies the runner–up. More quantitative displays also are possible.

### 3.3. Rescaling areas

This section reviews a very flexible procedure devised by my colleague Dr. Fremlin for the computer production of cartograms. This procedure is still under development, but seems to me to provide exciting possibilities, since it works directly with areas, and hence guarantees that existing geographical contiguities are preserved. Figure 14 shows the final output of the current version of the procedure. This figure shows all the states of the USA (together with the District of Columbia) scaled in such a way that the area of each is proportional to its population. The figure is plainly recognizable, and while the scaling obviously distorts the states, the undesirable "pseudopodia" of the previous method (see Figure 12*b*) are not present here; recognition of individual states is a simple matter, except for a few in the mountains of the west (and except for Alaska!).

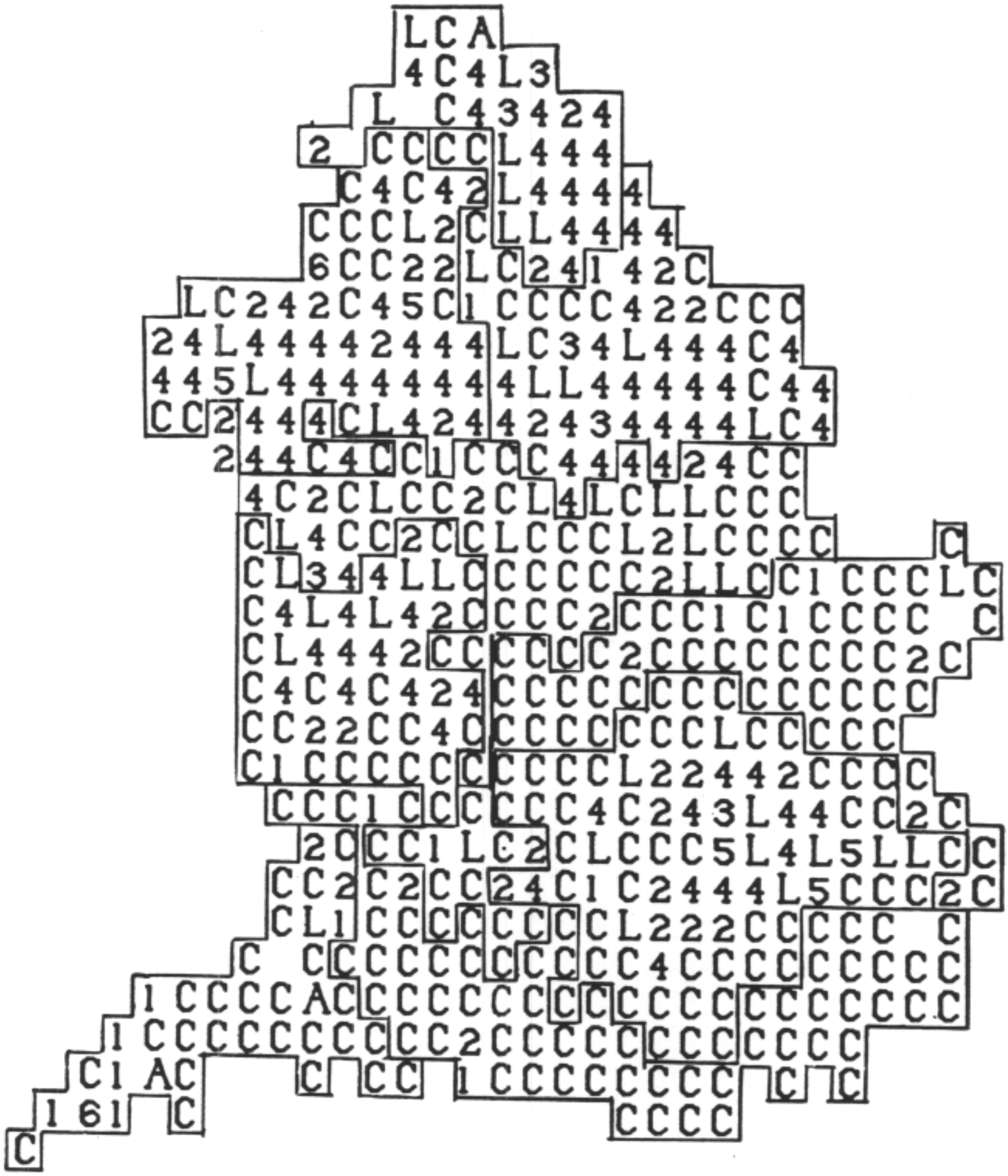
The idea of harnessing the computer to produce cartograms is not novel (*e. g.*, Tobler, 1973), but has received rather little attention in the literature. In part this may be because of the natural reluctance of the cartographer to accept the potentially inartistic output of the computer. An example of an unacceptable computer–produced cartogram is provided by Sen (1976), and the warning of Griffin (1980) is relevant: "the novelty of an automated approach may lead to intemperate haste in its utilization ...." It should be noted that the computer–produced cartogram of 1970 United States population reproduced on page 119 of Muehrcke and Muehrcke (1986) is in fact grossly inaccurate.

This is not the place for a detailed mathematical description of Fremlin's procedure, but it is possible to give a general idea of his strategy. The essential input to the program is a list of the coordinate positions of the vertices of the polygons that are used to approximate the boundaries of the states, together with the values (here population sizes) for the states. The procedure can be used equally well to display the importance of the states in terms of features other than population—all that would be required is a change in the input values.

The first stage of the iterative procedure consists of dividing the present version of the country map into a rectangular array of cells, some of which may be completely or partly

**Figure 13.**

The results for the 523 English constituencies in the 1987 election.

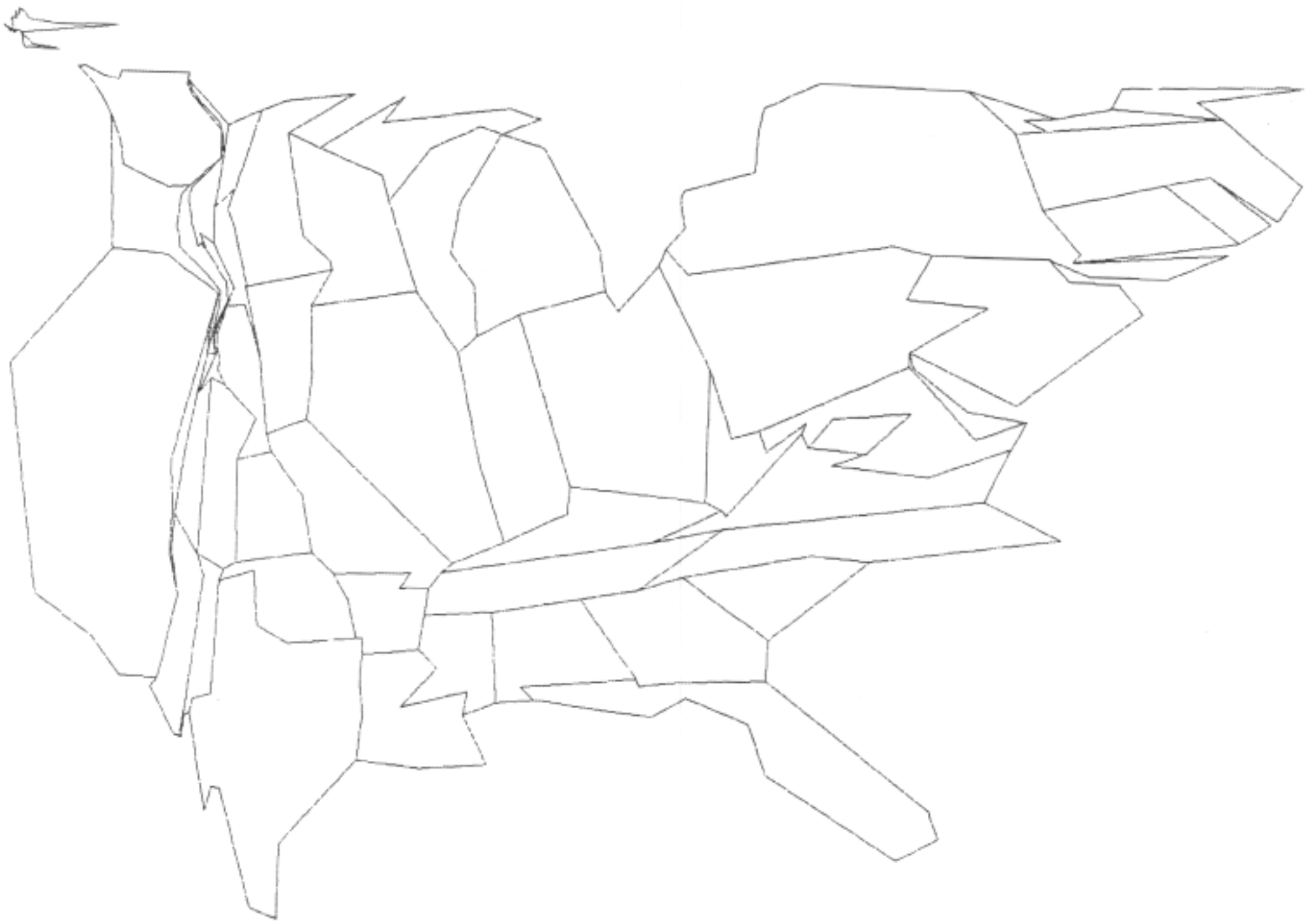


Key: C, L, A Seats having ≥10% majority for the Conservative (C), Labour (L) and Alliance (A) parties.

1-6 refer to other (marginal) seats in which the party orders are as follows: 1 CAL, 2 CLA, 3 LAC, 4 LCA, 5 ALC, 6 ACL.

**Figure 14.**

The United States (including Alaska) scaled using Fremlin's procedure and retaining contiguities.



filled with sea. For the first iteration the "present version" is the usual geographical map. The total land area and the estimated total population (assuming uniform density within a state) then is calculated separately for each column of the rectangle. From these results calculation of the average population density within that column is easy, and this average value is used to invent "sea people" to inhabit the areas of sea lying in this column, so that the overall density for the column matches the density for the land in that column.

Various improvements could be considered in the implementation of the first stage, which, in the form described above, suffers from the assumption of uniform population density within

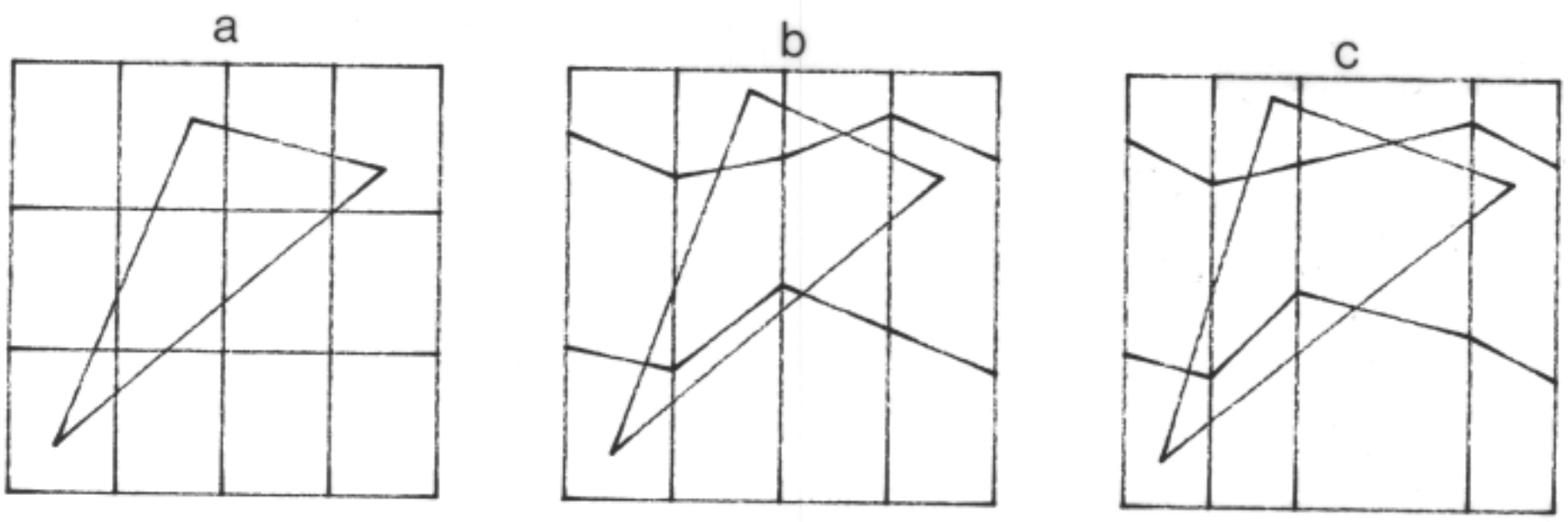
a state. If individual population figures and precise boundaries are available for some or all of the subregions of a state (*e. g.*, principal cities or counties), then this information could be used to provide improved estimates of the populations in the cells of the rectangular array. Alternatively, it might be thought appropriate to employ Tobler's (1979) pycnophylactic approach to produce a smoothed version of the underlying population density before implementing the first stage procedure.

The second stage also treats each column separately. The initially rectangular cells are now distorted into trapezia having vertical sides—in other words, the initially horizontal top and bottom of each cell are tilted in some appropriate manner. The nature of this tilting is a function of the population densities of the cells in neighbouring columns, with the aim being to make the fluctuations in density as gradual as possible.

After these adjustments have been made, the original rectangular mesh (Figure 15*a*) is changed into a mesh consisting of equi–spaced vertical lines and "wavy" horizontal lines (Figure 15*b*). A final adjustment now consists of an alteration in the column widths so as to reflect the variations in population density from column to column (Figure 15*c*). It is natural to consider altering the spacings so as to make each adjusted column have the same density. However, this adjustment turns out to create most undesirable distortions in the map, because of the large fluctuations in density that are present during the first iterations, and in practice a less extreme adjustment of the same character is used.

**Figure 15.**

The stages in an iteration of Fremlin's procedure.




The preceding step completes the first iteration, after which it is a comparatively easy matter to compute the revised locations of the vertices defining the regions. The map is now rotated through 90 degrees and the entire procedure is repeated. The iterations continue until the variation in population density from state to state has become acceptably small.

One necessary refinement to the procedure is required in order to avoid state boundaries crossing over each other, and hence resulting in figure–eight shapes! When computing the adjusted vertex locations, a check is made that this crossover has not happened. If it has happened, then this iteration is abandoned and is restarted with extra "vertex" coordinates being interpolated along the offending sides. This procedure may need to be implemented several times for any given iteration, but ultimately these undesired crossovers are eliminated.

It should be noted that, while the computer performs the necessary calculations, there are a number of variable parameters within the computer program. Moreover, the skill of the cartographer is translated into choosing values for these parameters that give the most acceptable cartogram for a given data set. One cannot expect optimal values for these parameters to exist that will apply to all data sets; in other words, there is still considerable scope for the blending of art with science.

## 4. Inferences from aggregate data

### 4.1. The "ecological fallacy"

The central problem of interest in this section is the following:

Given information concerning a number of *groups* of individuals, what can we conclude about the *individuals* separately?

All attempts to answer this type of question are coloured by knowledge of the results set out in a seminal paper by Robinson (1950), who used data on race and literacy taken from the 1930 U. S. census. These data are reproduced in two forms in Table 5.

TABLE 5
INFORMATION ON RACE AND LITERACY PROVIDED BY THE 1930 US CENSUS

**(a) A cross–classification for the entire country (figures in thousands)**

| | Black | White | Total |
|---|---|---|---|
| Illiterate | 1512 | 2406 | 3918 |
| Literate | 7780 | 85574 | 93354 |
| Total | 9292 | 87980 | 97272 |

**(b) Percentage figures by region**

| | Percent Black | Percent Illiterate |
|---|---|---|
| New England | 1.1 | 3.7 |
| Mid Atlantic | 4.0 | 3.5 |
| East North Central | 3.7 | 2.1 |
| West North Central | 2.6 | 1.4 |
| South Atlantic | 27.6 | 8.3 |
| East South Central | 27.2 | 9.6 |
| West South Central | 18.8 | 7.2 |
| Mountain | 0.9 | 4.2 |
| Pacific | 1.1 | 2.1 |

Robinson considered the question "To what extent are race and literacy interrelated?" Tables 5*a* and 5*b* both provide information about this relationship. Table 5*a* provides

information at the individual level, whereas Table 5*b* provides aggregated information. Multiplying the figures in Table 5*b* by the region populations and summing these products yields the marginal totals (9292, 87980, 3918, 93354) of Table 5*a*.

A correlation coefficient may be calculated from Table 5*a* using the four cell frequencies (1512, 2406, 7780 and 85574). There are various possible correlation formulæ depending upon how the two classifying variables are treated. If specific values (for example 0 and 1) are assigned to the two categories of each classifying variable, then the Pearson product–moment correlation coefficient is appropriate. This was the coefficient calculated by Robinson, and it proves to have the value 0.203, which suggests a rather slight connection between race and literacy.

Thomsen (1987) suggested calculating the *tetrachoric* correlation coefficient, which has value 0.747. The tetrachoric coefficient is appropriate when the two classifying variables are innately continuous variables (such as height and weight), but just happen to have been dichotomized (*e. g.*, into categories short and tall, light and heavy). When the joint distribution of the classifying variables is bivariate normal, this correlation coefficient is an unbiased estimate of the actual population correlation.

In the present case, neither of these correlation coefficients seems entirely appropriate, since although literacy can certainly be regarded as a continuous variable, it is difficult to regard color in that light. However, since the normal distribution is commonly used as an approximation for the binomial, the tetrachoric coefficient seems a better choice than the product moment coefficient used by Robinson.

For Table 5*b* Robinson calculated the correlation between percentage black and percentage illiterate, weighting these percentages by the populations of the regions concerned; this has been termed the *ecological* correlation, which may be denoted as $\rho_E$. Specifically, if region $i$ has population $n_i$ and the values of the two characteristics of interest are denoted by $p_i$ and $q_i$, then $\rho_E$ is calculated from the expression

$$\rho_E = \sum n_i(p_i - \bar{p})(q_i - \bar{q}) / \{\sum n_i(p_i - \bar{p})^2 \sum n_i(q_i - \bar{q})^2\}^{1/2}, \qquad (6)$$

where $\bar{p}$ is the mean of the $p$–values and $\bar{q}$ is the mean of the $q$–values. The nine pairs of regional values in Table 5*b* give the strikingly high correlation of 0.946, with Robinson having taken the discrepancy between 0.203 and 0.946 to imply that aggregated results might well give a highly misleading impression of individual–level relations between variables.

The high correlation obtained for Table 5*b* is clearly a consequence of the huge gap between the values for the three southern regions and those for the rest of the map. Because the gap is apparent in both columns of data, the correlation is very great. Firebaugh (1978) segregated the regions into two "super–regions" on this basis, and showed that, for explaining the variations in literacy rates, it is the interaction between race and super–region that is the dominant explanatory variable, rather than race itself. In a more detailed analysis, Hanushek *et al.* (1974) performed a multiple regression of literacy rates on race, available schooling and various other variables. They concluded that it was the schooling variable that played the dominant role: in 1930 in the southern states the proportions of those eligible for schooling that were actually enrolled in schools were lower than in other states. Since these same states contained the highest proportions of blacks, in the absence of schooling information it is race that appears to be the controlling variable.

Firebaugh, Hanushek *et al.*, and the many other scholars who have written about the problem posed by Robinson are in general agreement that, where there is a difference between the correlation in the table of individual–level data and the correlation apparent in a spatially aggregated form of that same data, this difference will be attributable to some unmeasured "latent" variable that has a marked spatial variation that matches the aggregation used. For Table 5 this would be the "available schooling" variable.

Robinson also considered the effect of a less extreme spatial aggregation, using state percentages, which resulted in an apparent correlation of 0.773 (very close to the tetrachoric value of 0.747). A detailed example of the way in which increasing spatial aggregation can lead to increasing apparent correlation is provided by Yule and Kendall (1950, Ch. 13). The fact that results vary according to the aggregation used is part of a more general problem entitled "the modifiable areal unit problem" by Openshaw and Taylor (1981). These authors regarded the general problem as follows:

> ... it would appear that geographers are doing their best to build what can only be described as intrinsically non–geographical models of geographical phenomena based on unreasonable assumptions regarding the nature of zonal data.

Closely related to these problems, and with the same essential cause—unmeasured latent variable(s)—are the so–called "nonsense" regressions and Simpson's paradox (Simpson, 1951). A typical example of the former is the strong positive relationship found between the numbers of clergy in Cuba at various decades of the century and the amount of rum consumed in Havana at those times (both increase as the population increases). Simpson's paradox is a discrete data analogue of the problems that arise when data from two distinct populations are combined—separate positive correlations then may appear to become a negative correlation.

### 4.2. Thomsen's approach

Political scientists are interested in the so–called "floating voters" whose political allegiance changes from election to election. Aggregate election results convey only the gross changes of support—if the Democrat vote increases by 5%, then (neglecting births, deaths and so forth) all one can say is that $(5+x)\%$ of the electorate moved from Republican to Democrat and $x\%$ moved from Democrat to Republican. Little is known about the size of $x$.

Expressed in terms of the layouts of Tables 5*a* and 5*b*, the constituency aggregated figures are comparable to the data of Table 5*b*, and the interest of the political scientist is in reconstructing the figures in a tabular form akin to Table 5*a*. There have been several attempts to achieve this goal, of which the most recent are those by Brown and Payne (1986), Johnston *et al.* (1982), and Thomsen (1987). These authors attended a recent workshop on ecological regression at Lund University, where their various approaches were discussed and where the consensus appeared to be that Thomsen's approach was the most promising.

Thomsen's description of his procedure is naturally couched specifically in terms of voting figures, but his rationale is not confined to voting and will be described here in general terms. In the case where each variable of interest has just two categories (as for the data of Table 5), the procedure leads to an explicit formula for the reconstruction of the individual–level data. When more than two categories are involved, an iterative computer program is required, with a version suitable for a PC being available from Professor Thomsen (Department of Political Science, University of Aarhus, Denmark).

In general terms, one is interested in quantifying the joint occurrence of the four category combinations of the two binary variables $X$ and $Y$, both of which, without loss of generality, can be assumed to take on the values 0 or 1. An implicit assumption of the procedure is that the values of both variables are governed by the values of one or more unobserved latent variables; this is termed the latent structure approach. There is no need to specify the number of latent variables, nor to give them names. The theory extends to cover any number (providing certain assumptions are met). Solely for simplicity of presentation, suppose that there are just two latent variables, denoted by $Z_1$ and $Z_2$.

Since different regions have different characteristics, one can expect that the mean values of $Z_1$ and $Z_2$ for the inhabitants of region $i$, $\mu_{1i}$ and $\mu_{2i}$ say, will differ from those in region $j$ ($\mu_{1j}$ and $\mu_{2j}$). Since $Z_1$ and $Z_2$ are unmeasured quantities, a general scale should be set in some arbitrary way, which will be done here by assuming that for each $Z$–variable the various $\mu$–values have mean 0 and variance 1. Within region $i$ individuals will have their own specific $Z$–values, which will vary about their regional means $\mu_{1i}$ and $\mu_{2i}$. Basic assumptions are that the $Z$–variables have a common variance, $\sigma^2$, that this variance is the same for each region, and that within a region their joint distribution is bivariate normal. Thus, within region $i$ the joint probability density function of the $Z$–values may be written as $\phi(\mu_{1i}, \mu_{2i}, \sigma^2)$.

Combining the assumptions about the variability of the regional means and the variability of $Z$–values within a region, and assuming that the joint distribution of regional means also is bivariate normal, the overall joint distribution of the $Z$–values across all the regions has the bivariate normal density function $\phi(0, 0, 1 + \sigma^2)$.

Since the $Z$–variables influence $X$ (and $Y$), it is reasonable to assume that, for a given individual, the probability that $X$ takes on the value 0 is some function of the values of $Z_1$ and $Z_2$ for that individual. A convenient assumption is that $P(X = 0) = \Phi(a + bz_1 + cz_2)$, where the symbol $\Phi$ refers to the distribution function of the unit normal distribution. Likewise, $P(Y = 0) = \Phi(d + ez_1 + fz_2)$.

Let the overall proportion of people for whom $X$ takes the value 0 be denoted by $p$. Given these various normality assumptions, the value of $p$ is given by the double integration of the product of $\Phi(a + bz_1 + cz_2)$ and $\phi(0, 0, 1 + \sigma^2)$ over the ranges of $Z_1$ and $Z_2$. Similar double integrations render expressions for $q$, the overall proportion of people for whom $Y$ takes the value 0, and for the joint probability $P(X = 0, Y = 0)$. Combining this pair of expressions yields the following one, for the overall correlation coefficient $\rho_A$:

$$\rho_A = (be + cf)(1 + \sigma^2)/\{[1 + (b^2 + c^2)(1 + \sigma^2)][1 + (e^2 + f^2)(1 + \sigma^2)]\}^{1/2}. \qquad (7)$$

Equation (7) is the quantity that is estimated by the tetrachoric coefficient from a complete 2–by–2 table.

At this point a recapitulation of which quantities are supposedly known and which are supposedly unknown is worthwhile. The problem posed at the beginning of this section was in effect "What can be deduced about individuals given regional values?" Given the regional values $(p_1, p_2, \ldots)$ for $P(X = 0)$, and the regional values $(q_1, q_2, \ldots)$ for $P(Y = 0)$, calculating the values of $p$ and $q$ is easy. Thus, in the race/literacy example these computations are given by the marginal totals of Table 5*a* as $3918/97272 = 0.0403$ and $9292/97272 = 0.0955$. These then are known values, as are the values $p_1, p_2, \ldots, q_1, q_2, \ldots$

given as percentages in Table 5*b*. However, the individual cell entries in Table 5*a* (*e. g.*, 1512) are supposedly unknown, and constitute what is to be estimated.

If the value of $\rho_A$ were known, then, given $p$ and $q$, one could reconstruct the cell entries. To do so exactly requires a computer program, but an approximation is provided by the formula

$$P(X=0,Y=0) = \{k - [k^2 - 8\rho_A(1+\rho_A)pq]^{1/2}\}/(4\rho_A), \tag{8}$$

where $k = 1-\rho_A+2\rho_A(p+q)$. This formula is a straightforward inversion of an approximation originally suggested by Yule (1897).

Although the value of $\rho_A$ is unknown, $\rho_E$, the ecological correlation displayed by the set of individual regional $p_i$ and $q_i$ values, can be calculated using equation (6). One also could calculate the ecological correlation between a transformation of the $p_i$ values and the same transformation of the $q_i$ values. Thomsen shows that, if a particular transformation of the values is used, then this correlation is given by the expression

$$\rho_E = (be+cf)/[(b^2+c^2)(e^2+f^2)]. \tag{9}$$

Comparing expression (7) for $\rho_A$ with expression (9) for $\rho_E$ indicates that if $1/(1+\sigma^2)$ is negligible by comparison with the sums $(b^2+c^2)$ and $(e^2+f^2)$, then $\rho_E$ will be approximately equal to $\rho_A$; otherwise it will be greater than $\rho_A$. In practice it appears that $\rho_E$ often is very close to $\rho_A$, providing that $\rho_E$ is computed from reasonably homogeneous areas—thus, for Robinson's data the value of $\rho_E$ for state data exceeds $\rho_A$ by only 0.026, whereas for the data aggregated over regions the excess correlation is appreciably greater. Thomsen himself feels that a state has too large an area to be regarded as a homogeneous unit, and hence advocates performing calculations using data at the county level.

The transformation used to obtain equation (9) was the so–called probit transformation, which is closely approximated by the simpler logit transformation, in which the proportion $p_i$ is replaced by $\log_e\{p_i/(1-p_i)\}$. An example of the affiliated calculations is given below.

### 4.3. An example

Table 6*a* shows the numbers of households in the four regions of the United States in 1980. These data have been extracted from Tables 64 and 527 of the 1981 edition of the *Statistical Abstract of the United States*. The numbers of households classified as black and the numbers in receipt of food stamps also are shown, together with the corresponding $p$ and $q$ values (thus, for example, $1632/17447 = 0.09354$). These particular characteristics were chosen because Table 527 of the Abstract also conveys the information that a total of 2409 black families were in receipt of food stamps. The intention here is to use Thomsen's procedure to obtain an estimate of this total figure from ecological data, typified by those appearing in Table 6*b*. Knowing the true value furnishes a guide to how well the procedure has performed.

Table 6*b* summarizes, for the Northeast region, the information available in Tables 32 and 203 of the 1982 edition of the Abstract. Although this information refers to persons rather than households, this reference creates no difficulty providing that, for example, the mean size of black households does not vary appreciably from state to state within a region. The (false) assumption of equal family sizes for black and non–black families is not required—this simply generates a scaling factor that cancels out in the computation of $\rho_E$.

TABLE 6
INFORMATION CONCERNING HOUSEHOLDS OF THE UNITED STATES

**(a) Numbers of households (figures in thousands) in various regions in March 1980**

| Region | Total | Black | Receiving food stamps | p | q |
|---|---|---|---|---|---|
| Northeast | 17447 | 1632 | 1359 | 0.09354 | 0.07789 |
| Midwest | 20933 | 1808 | 1251 | 0.08637 | 0.05976 |
| South | 25523 | 4125 | 2401 | 0.16162 | 0.09407 |
| West | 15205 | 840 | 900 | 0.05524 | 0.05919 |

**(b) Numbers of people (figures in thousands) in Northeast region at April 1, 1980**

| State | Total | Receiving food stamps | Others | Weighted logit | Black | Others | Weighted logit |
|---|---|---|---|---|---|---|---|
| ME | 1125 | 139 | 986 | − 65.7 | 3 | 1122 | −198.7 |
| NH | 921 | 53 | 868 | − 84.8 | 4 | 917 | −164.9 |
| VT | 511 | 46 | 465 | − 52.3 | 1 | 510 | −140.9 |
| MA | 5737 | 446 | 5291 | −187.3 | 221 | 5516 | −243.7 |
| RI | 947 | 87 | 860 | − 70.5 | 28 | 919 | −107.4 |
| CT | 3108 | 174 | 2934 | −157.5 | 217 | 2891 | −144.4 |
| NY | 17558 | 1804 | 15754 | −287.2 | 2402 | 15156 | −244.1 |
| NJ | 7365 | 600 | 6765 | −207.9 | 925 | 6440 | −166.5 |
| PA | 11864 | 1030 | 10834 | −256.3 | 1047 | 10817 | −254.4 |

**(c) Results**

| Region | Correlation Between Weighted Logits | Estimated number of black families receiving food stamps | |
|---|---|---|---|
| Northeast | 0.7236 | 450 | |
| Midwest | 0.5649 | 289 | |
| South | 0.7883 | 1298 | |
| West | 0.9087 | 372 | |
| | | 2409 | Total |
| | | 2417 | True number |
| | | 8 | Error of estimation |

One should note that if Table 6*a* had referred to people rather than households, then it could have been constructed from an extended version of Table 6*b*. In essence, therefore, one is hoping to estimate the overall number for the black/food stamp combination simply from aggregate state values. Thus this is a classic example of the ecological problem.

The calculation of the weighted logits in Table 6*b* requires explanation. Consider the 17558 households in New York state, of which 1804 received food stamps. Thus, $p_{NY} = 1804/17558$, and hence $U_{NY} = \log_e [p_{NY}/(1 - p_{NY})] = \log_e(1804/15754)$. Similarly, $q_{NY} =$

2402/17558, so that $V_{NY} = \log_e[q_{NY}/(1-q_{NY})] = \log_e(2402/15156)$. In the same way, $U$–values and $V$–values can be obtained for each of the other states in the region, and a straightforward (unweighted) correlation between the $U$–values and the $V$–values would involve sums such as $C = (U_{ME}V_{ME} + \cdots + U_{MA}V_{MA})$. However, sums calculated in this way fail to account for the precision with which the various logit values are known. Evidently more attention should be paid to states with large populations because these will give more precise estimates. It turns out that an appropriate set of weights is provided by the set of state populations $N_{ME}, N_{NH}, \ldots, N_{PA}$. Therefore the value of $C$ is replaced by the weighted version $(N_{ME}U_{ME}V_{ME} + \cdots + N_{MA}U_{MA}V_{MA})$. In view of the other sums that need to be calculated, it is convenient to write the product $N_{ME}U_{ME}V_{ME}$ as $N_{ME}^{1/2}U_{ME} \cdot N_{ME}^{1/2}V_{ME}$, and the quantities $N^{1/2}U$ and $N^{1/2}V$ define the weighted logits in Table 6*b*.

Table 6*c* summarizes the final results for this procedure. The value for $\rho_E$ for the Northeast region is 0.7236. Using this value for $\rho_A$ in equation (8), with the values of $p$ and $q$ (0.09354 and 0.07789) given in Table 6*a*, leads to the estimate $P(X=0, Y=0) = 0.02579$. Multiplying this value by the regional population, 17447, one gets an estimated 450 thousand black families in the Northeast region that received food stamps. Summing the regional estimates leads to an estimated total of 2409 thousand black families receiving stamps in the nation as a whole. This is remarkably close (a 0.33% error) to the true value of 2417 thousand; but, it would be unwise to expect that the method always would perform that well.

## 5. Conclusions

Regions provide a natural and convenient framework for summarising geographical data. Invariably government statistics are summarised in this way, using various levels of aggregation, ranging from aggregations of states, states themselves, counties, cities, wards, and postal districts.

One might have the impression that data so frequently used and so commonly cited would be transparent to the user. However, as this paper has shown, this is not the case. It is not always easy to display regional data in a manner appropriate to its use, and the map used often is a matter of judgement by the cartographer. At present interpolation of univariate regional data essentially is an empirical task. Multivariate regional data provide a minefield of potential misunderstandings.

Given the difficulties inherent with this common type of geographical data, it is a surprise to find that it has been accorded rather little attention in either the geographical or the statistical literature. The purpose of this paper has been to highlight these difficulties, and hence to stimulate further research on the interesting problems that regional data present.

## 6. References

Brown, P., and C. Payne. (1986) Aggregate data, ecological regression and voting transitions, *Journal of the American Statistical Association.* **81**, 452-460.

Cleveland, W., and R. McGill. (1984) Graphical perception: Theory, experimentation, and application to the development of graphical methods, *Journal of the American Statistical Association.* **79**, 531-554.

Cozens, P., and K. Swaddle. (1987) The British General Election of 1987, *Electoral Studies.*

**6**, 263-266.

Crain, I. (1970) Computer interpolation and contouring of two–dimensional data: a review, *Geoexploration*. **8**, 71-86.

Dunn, R. (1987) Variable–width framed rectangle charts for statistical mapping, *The American Statistician*. **41**, 153-156.

Evans, I., and K. Jones. (1981) Ratios and closed number systems, in *Quantitative Geography: A British View*, edited by (N. Wrigley and R. Bennett, pp. 123-134. London: Routledge and Kegan Paul.

Firebaugh, G. (1978) A rule for inferring individual–level relationships from aggregate data, *American Sociological Review*. **43**, 557-572.

Gale, N., and W. Halperin. (1982) A case for better graphics: the unclassed choropleth map, *The American Statistician*. **36**, 330-336.

Gatrell, A. (1981) Multidimensional scaling, in *Quantitative Geography: A British View*, edited by N. Wrigley and R. Bennett, pp. 151-163. London: Routledge and Kegan Paul.

Griffin, T. (1980) Cartographic transformation of the thematic map base, *Cartography*, **11**, 163-174.

Haining, R., D. Griffith, and R. Bennett. (1984) A statistical approach to the problem of missing spatial data using a first–order Markov model, *The Professional Geographer*. **36**, 338-345.

Hanushek, E., J. Jackson, and J. Kain. (1974) Model specification, use of aggregate data, and the ecological correlation fallacy, *Political Methodology*. **1**, 89-107.

Johnston, R. (1985) *The Geography of English Politics*. London: Croom Helm.

Johnston, R., A. Hay, and P. Taylor. (1982) Estimating the sources of spatial change in election results: a multiproportional matrix approach, *Environment and Planning A*. **14**, 951-961.

Kennedy S., and W. Tobler. (1983) Geographic interpolation, *Geographical Analysis*. **15**, 151-156.

Lam, N. (1981) The reliability problem of spatial interpolation models, *Modeling and Simulation*. **12**, 869-876.

Lam, N. (1983) Spatial interpolation methods: a review, *The American Cartographer*. **10**, 129-149.

Muehrcke, P., and J. Muehrcke. (1986) *Map Use: Reading, Analysis and Interpretation*. Madison, WI: JP Publications.

Office of Population Census and Surveys. (1981) *Census 1981: Parliamentary Constituency Monitors—1983 Boundaries*. London: Government Statistical Office.

Openshaw, S., and P. Taylor. (1981) The modifiable areal unit problem, in *Quantitative Geography: A British View*, edited by N. Wrigley and R. Bennett, pp. 60-69. London: Routledge and Kegan Paul.

Porter, P. (1958) Putting the isopleth in its place, *Proceedings*, Minnesota Academy of Science. **26**: 372-384.

Ripley, B. (1981) *Spatial Statistics*. Chichester: Wiley.

Robinson, W. (1950) Ecological correlations and the behavior of individuals, *American Sociological Review*. **15**: 351-357.

Sen, A. (1976) On a class of map transformations, *Geographical Analysis*. **8**: 23-37.

Simpson, E. (1951) The interpretation of interaction in contingency tables, *Journal of the Royal Statistical Society*. **13B**: 238-241.

*Statistical Abstract of the United States, 1981, 1982, 1986*. Washington: U. S. Department of Commerce, Bureau of the Census.

Stetzer, F. (1982) Specifying weights in spatial forecasting models: the results of some experiments, *Environment and Planning A*. **14**: 571-584.

*The Mitchell Beazley Concise Atlas of the Earth, 1973*. London: Mitchell Beazley.

Thomsen, S. (1987) *Danish Elections 1920–79: A Logit Approach to Ecological Analysis and Inference*. Aarhus: Politica.

Tobler, W. (1973) A continuous transformation used for districting, *Annals of the New York Academy of Science*. **219**: 215-220.

Tobler, W. (1979) Smooth pycnophylactic interpolation for geographical regions, *Journal of the American Statistical Association*. **74**: 121-127.

Tobler, W., and S. Kennedy. (1985) Smooth multidimensional interpolation, *Geographical Analysis*. **17**: 251-257.

Upton, G. (1985) Distance–weighted geographic interpolation. *Environment and Planning A*. **17**: 667-671.

Upton, G. (1989) The components of voting change in England 1983–1987, *Electoral Studies*. **8**: 59-74.

Upton, G., and B. Fingleton. (1985) *Spatial Data Analysis by Example*, Volume 1: Point Pattern and Quantitative Data. Chichester: Wiley.

Waller, R. (1983) *The Atlas of British Politics*. London: Croom Helm.

Yule, G. (1897) On the theory of correlation, *Journal of the Royal Statistical Society*. **60**: 812-854.

Yule, G., and M. Kendall. (1950) *An Introduction to the Theory of Statistics* (14th ed.). London: Griffin.

## APPENDIX A

Data for the United States

| State | Area | Latitude | Longitude | Data Set I | II | III | IV | V | VI | VII |
|---|---|---|---|---|---|---|---|---|---|---|
| Alabama | 131994 | 32.83 | 87.00 | 67.9 | 13.3 | 13.8 | 73.8 | 13.1 | 60.5 | 151.1 |
| Arizona | 295121 | 34.00 | 112.00 | 15.6 | 9.4 | 9.3 | 82.4 | 53.1 | 66.4 | 70.4 |
| Arkansas | 135403 | 34.83 | 93.67 | 37.0 | 9.1 | 10.1 | 82.7 | 18.9 | 60.5 | 112.4 |
| California | 406377 | 37.50 | 119.50 | 127.6 | 11.7 | 9.9 | 76.2 | 18.5 | 57.5 | 61.9 |
| Colorado | 269347 | 39.50 | 105.50 | 21.3 | 7.3 | 9.1 | 89.0 | 30.8 | 63.4 | 51.9 |
| Connecticut | 12667 | 41.75 | 72.75 | 623.6 | 4.1 | 11.1 | 90.1 | 2.5 | 60.7 | 46.6 |
| Delaware | 5023 | 39.17 | 75.50 | 276.5 | 6.7 | 14.1 | 82.1 | 8.4 | 59.8 | 63.6 |
| D. C | 164 | 38.90 | 77.02 | 999.9 | 999.9 | 21.2 | 26.9 | −15.6 | 13.7 | 122.0 |
| Florida | 140798 | 28.00 | 82.00 | 125.5 | 11.0 | 12.8 | 84.0 | 43.5 | 65.3 | 55.8 |
| Georgia | 150946 | 32.83 | 83.25 | 79.0 | 14.4 | 12.7 | 72.2 | 19.1 | 60.2 | 96.5 |
| Idaho | 214271 | 45.00 | 115.00 | 8.6 | 5.4 | 9.9 | 95.6 | 32.4 | 72.4 | 55.9 |
| Illinois | 144677 | 40.00 | 89.00 | 199.4 | 9.9 | 13.6 | 80.8 | 2.8 | 56.2 | 95.9 |
| Indiana | 93423 | 40.00 | 86.25 | 143.9 | 6.2 | 11.4 | 91.1 | 5.7 | 61.7 | 72.9 |
| Iowa | 145509 | 42.25 | 93.25 | 50.5 | 2.6 | 10.2 | 97.4 | 3.1 | 53.3 | 64.6 |
| Kansas | 212623 | 38.75 | 98.25 | 27.5 | 5.7 | 10.4 | 91.7 | 5.1 | 66.3 | 48.4 |
| Kentucky | 103139 | 37.50 | 85.25 | 81.2 | 9.0 | 12.0 | 92.3 | 13.7 | 60.0 | 149.9 |
| Louisiana | 115755 | 31.25 | 92.25 | 81.0 | 15.8 | 13.0 | 69.2 | 15.4 | 60.8 | 140.1 |
| Maine | 80587 | 45.25 | 69.25 | 32.1 | 2.7 | 9.0 | 98.7 | 13.2 | 60.8 | 96.9 |
| Maryland | 25576 | 39.00 | 76.75 | 396.6 | 8.2 | 11.9 | 74.9 | 7.5 | 52.5 | 68.5 |
| Massachusetts | 20342 | 42.25 | 71.83 | 727.0 | 3.7 | 10.1 | 93.5 | 0.8 | 51.2 | 59.2 |
| Michigan | 148080 | 44.00 | 85.00 | 156.2 | 10.6 | 12.1 | 85.0 | 4.3 | 59.2 | 112.2 |
| Minnesota | 206825 | 46.00 | 94.25 | 48.0 | 2.0 | 9.5 | 96.6 | 7.1 | 49.6 | 54.3 |
| Mississippi | 122806 | 32.83 | 89.50 | 46.9 | 12.6 | 15.4 | 64.1 | 13.7 | 61.9 | 190.1 |
| Missouri | 179257 | 38.50 | 93.50 | 67.8 | 10.4 | 11.7 | 88.4 | 5.1 | 60.0 | 74.1 |
| Montana | 378009 | 47.00 | 110.00 | 4.8 | 4.8 | 10.1 | 94.1 | 13.3 | 60.5 | 66.7 |
| Nebraska | 199274 | 41.50 | 100.00 | 19.4 | 3.0 | 10.0 | 94.9 | 5.7 | 70.6 | 54.2 |
| Nevada | 285724 | 39.00 | 117.00 | 4.4 | 15.5 | 10.2 | 87.4 | 63.8 | 65.8 | 34.0 |
| New Hampshire | 23382 | 43.58 | 71.67 | 81.7 | 1.4 | 11.0 | 98.8 | 24.8 | 68.6 | 29.7 |
| New Jersey | 19417 | 40.25 | 74.50 | 953.1 | 5.4 | 11.7 | 83.2 | 2.7 | 60.1 | 63.6 |
| New Mexico | 315471 | 34.50 | 106.00 | 8.4 | 10.2 | 11.3 | 75.1 | 28.1 | 59.7 | 107.4 |
| New York | 123180 | 43.00 | 75.00 | 381.3 | 10.3 | 12.1 | 79.5 | − 3.7 | 53.8 | 103.5 |
| North Carolina | 126992 | 35.50 | 80.00 | 104.1 | 10.8 | 13.7 | 75.8 | 15.7 | 61.9 | 75.3 |
| North Dakota | 180180 | 47.50 | 100.25 | 8.9 | 1.2 | 10.6 | 95.9 | 5.7 | 64.8 | 39.4 |
| Ohio | 106610 | 40.25 | 82.75 | 260.0 | 6.9 | 11.5 | 88.9 | 1.3 | 58.9 | 104.5 |
| Oklahoma | 178503 | 35.50 | 98.00 | 37.2 | 8.5 | 12.3 | 85.9 | 18.2 | 68.6 | 77.6 |
| Oregon | 250078 | 44.00 | 121.00 | 21.7 | 5.0 | 10.5 | 94.6 | 25.9 | 55.9 | 79.7 |
| Pennsylvania | 116709 | 40.75 | 77.50 | 272.3 | 6.2 | 11.6 | 89.8 | 0.5 | 53.3 | 89.9 |
| Rhode Island | 2743 | 41.67 | 71.50 | 902.5 | 4.0 | 10.0 | 94.7 | − 0.3 | 51.7 | 73.8 |
| South Carolina | 78528 | 34.00 | 81.00 | 85.7 | 11.5 | 16.1 | 68.8 | 20.5 | 63.6 | 112.4 |
| Tennessee | 107003 | 35.83 | 85.50 | 94.9 | 9.4 | 12.0 | 83.5 | 16.9 | 57.8 | 110.2 |
| South Dakota | 197475 | 44.25 | 100.00 | 8.8 | 1.9 | 10.2 | 92.7 | 3.7 | 63.0 | 63.7 |
| Texas | 681244 | 31.50 | 99.00 | 42.7 | 14.2 | 10.9 | 78.7 | 27.1 | 63.6 | 74.6 |
| Utah | 213390 | 39.50 | 111.50 | 12.9 | 3.7 | 11.0 | 94.7 | 37.9 | 74.5 | 44.2 |
| Vermont | 24110 | 43.83 | 72.75 | 47.9 | 3.3 | 9.3 | 99.1 | 15.0 | 57.9 | 84.9 |
| Virginia | 103230 | 37.50 | 78.75 | 116.9 | 8.8 | 12.8 | 79.1 | 14.9 | 62.3 | 64.9 |
| Washington | 172929 | 47.50 | 120.50 | 51.2 | 4.6 | 10.6 | 91.5 | 21.1 | 55.8 | 60.7 |
| West Virginia | 62709 | 38.75 | 80.50 | 72.5 | 6.8 | 11.4 | 96.2 | 11.8 | 55.1 | 135.8 |
| Wisconsin | 141508 | 44.75 | 89.50 | 81.1 | 2.5 | 9.5 | 94.4 | 6.5 | 54.2 | 74.1 |
| Wyoming | 252171 | 43.00 | 107.50 | 3.4 | 7.1 | 9.8 | 95.0 | 41.3 | 70.5 | 43.1 |

Data Set I: Population density

Data Set II: Murders (per 100,000)

Data Set III: Infant mortality (per 1000)

Data Set IV: Percent white, 1980

Data Set V: Percent increase in population

Data Set VI: Percent voting Republican

Data Set VII: Food stamp recipients (per 1000)

## APPENDIX B

The British Regions used by Waller (1983)

1. **The South West**:

Counties of Avon, Cornwall, Devon, Dorset, Somerset — 38 constituencies.

2. **The South of England**:

Counties of East Sussex, Gloucestershire, Hampshire, Isle of Wight, Kent, Surrey, West Sussex, Wiltshire — 68 constituencies.

3. **Greater London**:

84 constituencies.

4. **Central England and East Anglia**:

Counties of Bedfordshire, Buckinghamshire, Cambridgeshire, Essex, Hertfordshire, Norfolk, Oxfordshire, Suffolk — 69 constituencies.

5. **The West Midlands**:

Counties of Hereford and Worcester, Shropshire, Warwickshire, West Midlands — 47 constituencies.

6. **The North Midlands**:

Counties of Cheshire, Derbyshire, Leicestershire, Lincolnshire, Northamptonshire, Nottinghamshire, Staffordshire — 63 constituencies.

7. **The North West**:

Counties of Greater Manchester, Lancashire, Merseyside — 64 constituencies.

8. **Yorkshire**:

Counties of Cleveland, Humberside, North Yorkshire, South Yorkshire, West Yorkshire — 60 constituencies.

9. **The North of England**:

Counties of Cumbria, Durham, Northumberland, Tyne and Wear — 30 constituencies.

10. **Wales**:

38 constituencies

11. **Scotland**:

72 constituencies.

## DISCUSSION

### "Information from regional data"

### by Graham J. G. Upton

Upton addresses three interesting questions concerning spatial statistics, within the context of several empirical examples. Two of these questions deal with unresolved issues in spatial statistics, while one illustrates the lack of awareness by many non–geography spatial statisticians of developments that have occurred in the geographical sciences. This latter view is reinforced by some of the graphics research that has been undertaken at the National Center for Supercomputer Applications (University of Illinois), too, and in part can be seen as a justification for the establishment of the National Center for Geographic Information and Analysis by the National Science Foundation.

Upton first discusses the determination of appropriate weights for numerically handling spatial autoregressive structures. He reports an all too common finding in this section, namely that those weights producing estimates most closely resembling the geographic pattern in question still leave a visible geographical element in residuals. Knowing how to statistically evaluate this residual situation is quite difficult, if presently not impossible, as Cliff and Ord (1981) have noted. The reader should recognize here that Upton is somewhat in error, in that the approach he suggests is not really all that novel. Work on the missing data problem for spatial surfaces has been studied in considerable detail by Martin (1984, forthcoming) as well as Haining, Griffith, and Bennett (1984, 1989). Flowerdew and Green (1988, 1989) have explored the problem of estimating disaggregated areal unit values from aggregate values. Further, two drawbacks should be acknowledged concerning Upton's analysis. First, as fractals and cartographic line generalization research indicate, boundary length measuring is both difficult and of questionable accuracy. Second, an overlooked and very serious problem with trend surface modelling is the underlying geographic distribution of points (see Unwin and Wrigley, 1987). As a supplement, Griffith (1988, 1989) has reported some interesting findings concerning outlier diagnostics for geo–referenced data that at least supplement Upton's work, here, too.

This first section raises several troublesome questions in this reader's mind, as well. First, does ignoring the "... variation in distance with longitude ..." introduce serious measurement error? Second, is the "first neighbors only" scheme equivalent to a conditional autoregressive (CAR) model, and is the "first plus second neighbors" scheme equivalent to a simultaneous autoregressive model (SAR)? Third, does the standardization of results for Table I really make the different data sets comparable? In other words, what is the source of and structure of error being evaluated? Is the aggregate fit of the model over all six data sets achieved with covariance analysis? Finally, while Ripley (1988) argues for a CAR model, Upton ends up arguing for an SAR model; why do these two scholars differ on this point?

In the next section Upton's preoccupation with computational constraints fails to take into account their considerable relaxation by supercomputing capabilities. A major portion of this section seems like an attempt to re–invent the wheel, for Upton actually seems to be talking about continuous cartograms (see Tobler, 1963; Monmonier, 1982). From a numerical point of view, the use of "sea constituencies" is bothersome, mostly because these imaginary areal units comprise such a large proportion of the final geo–referenced data set (763 imaginary versus 631 actual, in one case, and 523 imaginary versus 501 actual in an-

other case). Griffith (1982) has discussed the problem of shape and boundary buffer zones, in terms of percentages of internal and border areal units. Perhaps one needs to determine some threshold value beyond which an approach like Upton's becomes too artificial. Certainly the impact of having such a disproportionate number of manufactured areal units on the final statistical properties of an analysis merits very close scrutiny.

The final section of Upton's paper deals with the general problem of modifiable areal units and statistical inference complications introduced by aggregation. Again the Flowerdew and Green (1988, 1989) treatment of this topic is of relevance. Arbia's (1989) recent book addresses this issue, too, in a very imaginative and systematic way. Another important but often overlooked theme where the ecological fallacy emerges is in location–allocation modelling, where an areal unit centroid is used as the geo–referenced coordinate for all items located within that unit (introducing severe measurement error in many cases). Upton discusses the tetrachoric correlation coefficient, which is merely the square root of a chi square statistic that has been divided by the sample size. But this constrained estimation approach appears elsewhere in the spatial analysis literature, and so a comparison is in order. Certainly the volume of work that has been published on entropy maximizing models is relevant here, as it follows this same strategy. Upton's reflection on the question of "What is an appropriate areal unit?" is reminiscent of the familiar geographic question of "What is a region?"; perhaps some insights can be gained by reviewing the classical literature on this topic.

All in all, Upton's paper is a treatment of three interesting issues, with continual reference to selected empirical examples. His conceptual reflections upon these empirical examples is the major strength of this paper. His lack of acknowledgement of considerable relevant geographic literature supports the contention that a fuller dialogue is needed between quantitative geographers and professional statisticians. This, indeed, is the gap that this very book seeks to help fill!

**References**

Arbia, G. (1989) *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems*. Dordrecht: Kluwer.

Cliff, A., and J. Ord. (1981) *Spatial Processes*. London: Pion.

Flowerdew, R., and M. Green. (1988) "Statistical methods for inference between incompatible zonal systems," paper presented at Initiative #1, NCGIA, UC/Santa Barbara, December 12-16.

Flowerdew, R., and M. Green. (1989) "Inference between incompatible zonal systems using the EM algorithm," paper presented to the Sixth European Colloquium on Theoretical and Quantitative Geography, Chantilly, France, September 5-9.

Griffith, D. (1982) Geometry and spatial interaction, *Annals* of the Association of American Geographers, **72**, 332-346.

Griffith, D. (1988) "Interpretation of standard influential observations regression diagnostics in the presence of spatial dependence," paper presented to the 35th Regional Science Association, Toronto, November 11-13.

Griffith, D. (1989) Pure error and lack–of–fit regression diagnostics in the presence of spatial dependence. *Sistemi Urbani*, No. 2, in press.

Haining, R., D. Griffith, and R. Bennett. (1984) A statistical approach to the problem of missing spatial data using a first–order Markov model. *The Professional Geographer*, **36**, 338-345.

Haining, R., D. Griffith, and R. Bennett. (1989) Maximum likelihood estimation with missing spatial data and with an application to remotely sensed data. *Communications in Statistics: Theory and Methods*, **18**, 1875-1894.

Martin, R. (1984) Exact maximum likelihood for incomplete data from a correlated Gaussian process. *Communications in Statistics: Theory and Methods*, **13**, 1275-1288.

Martin, R. (forthcoming). Information loss due to incomplete data from a spatial Gaussian one–parameter first–order conditional process. *Communications in Statistics: Theory and Methods*.

Monmonier, M. (1982) *Computer Assisted Cartography*. Englewood Cliffs, N. J.: Prentice Hall, pp. 123-134.

Ripley, B. (1988) *Statistical Inference for Spatial Processes*. Cambridge: Cambridge University Press.

Tobler, W. (1963) Geographic area and map projections. *Geographical Review*, **53**, 59-78.

Unwin, D., and N. Wrigley. (1987) Control point distribution in trend surface modelling revisited: an application of the concept of leverage. *Transactions of the Institute of British Geographers*, N. S. 12: 147- 160.

*Daniel A. Griffith*, Syracuse University

# A REJOINDER TO GRIFFITH'S DISCUSSION

by Graham J. G. Upton

This response is addressed to those who, like myself, decide whether a paper is worth reading on the basis of the ensuing discussion. To such readers I would comment that much of Griffith's discussion appears to be on a different paper to that which I thought I had written! In particular, I would stress that my interpolation procedures are not based on boundary lengths for precisely the reasons that Griffith indicates. Further, I do not advocate the fitting of trend surfaces—quite the reverse.

In commenting on the second section of my paper, Griffith correctly alludes to the difficulties posed by the presence of artificial "sea constituencies." However, the advent of supercomputers noted by Griffith means that this problem is conceptual, not computational. Indeed, if these artificial constituencies are excluded, then a far more compact representation will result—though a rectangular Britain would not be aesthetically pleasing. Whilst cartograms are not new, I am not persuaded that there are any existing computer–based procedures that lead to totally acceptable output.

Next, a more general observation, motivated by Griffith's discussion of CAR and SAR models. I wholeheartedly agree that the present dialogue between geographers and statisticians should be promoted. However, there can be few people as well qualified as Griffith to bridge the geography–statistics interface, and I worry that the current advances in "geometrics" may not be meaningful to the less quantitative geographer. "First neighbours" is an easy concept to grasp, whereas I cannot say the same for the "conditional autoregressive model"!

Finally (despite his misguided comments on my paper), a belated "thank you" to Dan Griffith for his generous hospitality!

## *PREAMBLE*

*Imagination is more important than knowledge.*

A. Einstein, **On Science**

*There is a particular fascination with scholarly expositions of members of the vanguard, who wander through worlds of the unknown exploring new ideas. Little is known about the intellectual realms into which they journey, their imaginations serving as guiding lights, with their writings sometimes appearing to more conservative members of their disciplines as near fantasy or science fiction. Doreian's translation of spatial autocorrelation concepts and findings into sociological contexts exemplifies this category of pioneering work. His paper derives network autocorrelation models from spatial autocorrelation models. The purpose of this paper is to apply spatial autocorrelation models to the analysis of social phenomena distributed across spatial as well as aspatial social networks. In doing so, Doreian transcribes prominent limitations of spatial autocorrelation models for network autocorrelation models. These same sentiments are expressed in Wartenberg's commentary, in which additional applications of network autocorrelation models are gleaned from evolutionary biology, ecology, and environmental epidemiology. Wartenberg's supplemental examples should help dispell the speculative nature some scholars might associate with Doreian's work.*

The Editor

# Network Autocorrelation Models: Problems and Prospects

Patrick Doreian*

*Department of Sociology, Forbes Quad 2126, University of Pittsburgh, 4200 Fifth Avenue, Pittsburgh, PA 15260, U.S.A.*

**Overview:** Network autocorrelation models draw their inspiration from, and share a common representation with, spatial autocorrelation models. The use of a weight matrix, $\mathbf{W}$, to capture network interdependencies and the statement of linear equations provide the communality. Network autocorrelation models can be used to analyze social phenomena distributed across social structures that need not be rooted in geographical space. These include the diffusion of ideas through the networks linking scientists in an invisible college, analyses of economic development for nation-states, and analyses of inter-organizational networks. Many of the problems (and responses to them) encountered in the spatial autocorrelation model literature are applicable directly to network autocorrelation models. These include discussions of boundary effects, issues of aggregation, and dynamic modeling. The problems associated with the specification and estimation of network autocorrelation models are likely to be more difficult than for the spatial case. The additional complexities stem from having to specify and model a time-dependent weight matrix $\mathbf{W}(t)$ rather than simply use $\mathbf{W}$, the necessity to model coupled processes, and the need to use qualitatively different actors linked by multiple processes.

Social scientists in general, and sociologists in particular, lay claim to the study of social phenomena. In large part, this includes analyses of social structures and social processes: social structure is generated by, and in turn constrains, the operation of social processes. If correct, it is trivial to claim that structure and the interdependence of social actors must be included in the analysis of social action. Trivial, but for the fact that it is ignored in much of contemporary social science—especially when the analysis of empirical information is included. Although this data analytic practice appears to fly in the face of empirical reality, it is straightforward to understand the reasons for it. The invention of the social survey, together with the early use of computers, permitted the creation and analysis of large data sets comprised of individual—such as people, groups or organizations—cases. Although early methods of correlations and cross tabulations have largely (but not completely) been superceded by regression, structural equation models, and log-linear models, the underlying presumption of independent data points remains the majority choice. Alas, in many empirical contexts, it does not survive close scrutiny.

* Prepared for the 1989 Symposium "Spatial Statistics: Past, Present and Future," Department of Geography, Syracuse University. The author acknowledges beneficial interaction with Daniel Griffith, including exposure to the manuscripts of those symposium lecturers preceding him, during his visit to Syracuse University; however, Griffith should not be incriminated by any errors that remain.

## 1. Spatial autocorrelation models

Social phenomena distributed across geographic space provide one arena for challenging the value and utility of models premised on the assumption of independent data points. Behavior at one geographic location need not be independent of behavior at another. Loftin and Ward (1983) provide an example using sociological data.

Their linear model was the conventional population regression function:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1.1}$$

where a vector, $\mathbf{y}$, is predicted from a set of regressors, $\mathbf{X}$, using a vector of parameters, $\boldsymbol{\beta}$.

The disturbance term, $\boldsymbol{\epsilon}$, was treated in two ways. First, it was specified as independently normally distributed (as for ordinary least squares, OLS) and second, as:

$$\boldsymbol{\epsilon} = \rho\mathbf{W}\boldsymbol{\epsilon} + \boldsymbol{\nu} \tag{1.2}$$

where $\boldsymbol{\nu} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ with $\boldsymbol{\epsilon}$ spatially autocorrelated via $\mathbf{W}$, which captured the interdependence of contiguous areal units. The conjunction of equations (1.1) and (1.2) has been called the spatial disturbances model (Doreian, 1980) as the interdependencies are considered operative on the disturbances alone. Obviously, if the parameter $\rho$ is zero or if $\mathbf{W}$ is uniformly zero, the disturbances model reduces to the OLS model. The predicted variable for Loftin and Ward was a measure of fertility. The regressors, $\mathbf{X}$, contain a set of population density measures (logarithm transforms of persons per room, rooms per unit, units per structure, and structures per acre) together with a class index and an ethnicity index. The units of analysis were 75 community areas making up Chicago. Using OLS, three of the density variables and the class index were found to be significant predictors of fertility. If the specification of (1.2) is correct, and if $\rho$ is known, equation (1.1) can be rewritten as $\mathbf{Y}^\star = \mathbf{X}^\star\boldsymbol{\beta} + \boldsymbol{\nu}$ where $\mathbf{Y}^\star = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{Y}$ and $\mathbf{X}^\star = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}$. Use of OLS, where $\mathbf{Y}^\star$ is regressed on $\mathbf{X}^\star$, provide estimates of $\boldsymbol{\beta}$ and $\sigma^2$. In general, $\rho$ is unknown. However, OLS for (1.1) will generate a spatially correlated residual, $\hat{\boldsymbol{\epsilon}}$, from which a crude estimate, $\hat{\rho}$, is obtained from regressing $\hat{\boldsymbol{\epsilon}}$ on $\mathbf{W}\hat{\boldsymbol{\epsilon}}$. This process is iterated until $\hat{\rho}$ converges. Using a procedure such as this, Loftin and Ward (1983) fitted the disturbances model, and found that both the class and ethnicity indexes were significant predictors of fertility together with, at most, a single density variable (depending on how the matrix $\mathbf{W}$ was operationalized). With an identical data set and a common model, the empirical evidence supports quite different substantive accounts depending on whether spatial autocorrelation is considered or not.

The discussion thus far treats spatial autocorrelation as a technical problem incorporated into the specification of the disturbance term. White, Burton and Dow (1981) constructed a model of the sexual division of labor in African agriculture. The core variables were female participation in agriculture, patrilocal residence, and degree of polygamy. Following the estimation of their model via OLS, it was clear that the residuals from their analysis clustered spatially. Rather than leave the analysis with an implicit unestimated $\mathbf{W}$, the authors were able to specify it from a linguistic tree constructed for versions of the Bantu language. The underlying idea was that there had been an expansion of Bantu tribes across geographic space and that the data points were not independent but were linked through similarity with regard to language. More broadly, spatial autocorrelation models have been used to deal with what has been known as Galton's Problem.

Maximum likelihood can be used to estimate the disturbances model. Using the notation $\mathbf{A} = \mathbf{I} - \rho\mathbf{W}$, the log–likelihood function can be written as:

$$\ln(\mathbf{L}) = \text{constant} - (N/2)\ln(\sigma^2) - \frac{1}{2\sigma^2}[\mathbf{y}'\mathbf{A}'\mathbf{A}\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{A}'\mathbf{A}\mathbf{X}\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{A}'\mathbf{X}\boldsymbol{\beta}] + \ln|\mathbf{A}| \quad (1.3)$$

From this it is straightforward to establish the following estimation equations:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{A}'\mathbf{A}\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}'\mathbf{A}\mathbf{y} \quad (1.4)$$

and

$$\hat{\sigma}^2 = [\mathbf{y}'\mathbf{A}'\mathbf{A}\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{A}'\mathbf{A}\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{A}'\mathbf{A}\mathbf{X}\boldsymbol{\beta}]/\mathbf{N} \quad (1.5)$$

Substitution of (1.4) and (1.5) into (1.3) yields the concentrated log–likelihood function from which a value of $\rho$ is obtained: $\hat{\rho}$ minimizes

$$\ln(\mathbf{y}'\mathbf{A}'\mathbf{P}\mathbf{A}\mathbf{y}) - (2/N)\ln|\mathbf{A}|$$

where $\mathbf{P} = \mathbf{I} - (\mathbf{A}\mathbf{X})\{(\mathbf{A}\mathbf{X})'(\mathbf{A}\mathbf{X})\}^{-1}(\mathbf{A}\mathbf{X})'$ and $\ln|\mathbf{A}| = \Sigma\ln(1 - \rho\lambda_i)$ with $\{\lambda_i\}$ being the eigenvalues of $\mathbf{W}$. With $\hat{\rho}$ as the estimate of $\rho$, its value can be substituted into (1.4) to get $\hat{\boldsymbol{\beta}}$, and into (1.5) to get $\hat{\sigma}^2$. Approximate standard errors for $\hat{\rho}$, $\hat{\boldsymbol{\beta}}$, and $\hat{\sigma}^2$ are obtained from the variance–covariance matrix obtained from the second partial derivatives of the log–likelihood function. Details of the procedure can be found in Ord (1975), Doreian (1980), or Upton and Fingleton (1985). An alternative to the spatial disturbances model is the spatial effects model where the spatial interdependence is incorporated directly into the statement of the model. A motivating example is found in Mitchell (1969) in a study of the spatial distribution of rebel control—or government control—for the HUK rebellion in the Philippines. It is clear that control of one area has immediate consequence for those areas contiguous with it or easily reached from it. Equation (1.1) is replaced by:

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (1.6)$$

where $\boldsymbol{\epsilon} \sim \mathbf{N}/(\mathbf{0}, \sigma^2\mathbf{I})$ and the log–likelihood function is

$$\ln(y) = \text{constant} - (N/2)\ln\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y}'\mathbf{A}'\mathbf{A}\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{A}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}) + \ln|\mathbf{A}| \quad (1.7)$$

It is straightforward to establish

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z} \quad (1.8)$$

and

$$\hat{\sigma}^2 = (1/N)\mathbf{z}'\mathbf{M}\mathbf{z} \quad (1.9)$$

where $\mathbf{z} = \mathbf{A}\mathbf{y} = (\mathbf{I} - \rho\mathbf{W})\mathbf{y}$ and $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. The spatial effects parameter, $\hat{\rho}$, minimizes

$$\ln(\mathbf{y}'\mathbf{M}\mathbf{y} - 2\rho\mathbf{y}'\mathbf{M}\mathbf{W}\mathbf{y} + \rho^2\mathbf{y}'\mathbf{W}'\mathbf{W}\mathbf{y}) - (2/N)\Sigma\ln(1 - \rho\lambda_i)$$

An approximate variance–covariance matrix, as before, is obtained by use of the second partial derivatives of the log–likelihood function. Details are found in Ord (1975) and Doreian (1981).

## 2. Network autocorrelation models

Network autocorrelation models are, in essence, an extension of spatial autocorrelation models to phenomena where the interdependence among the structural units is generated directly through the operation of some social process (instead of some geographically distributed process). This definition is not intended as a slight upon spatial autocorrelation models. Indeed, the spatially distributed examples of Loftin and Ward, White, *et al.*, and Mitchell all provide a powerful motivation for considering and incorporating more general interdependencies between social actors. The connecting link between network and spatial autocorrelation models is found in the representation of the matrix **W**. The ways in which **W** is constructed in terms of contiguity, accessibility, or common boundaries can be seen as variations of a sociometric scheme where interdependence need not rest directly on geographical, or even physical, characteristics.

### 2.1. Example 1: scientific values

Science produces empirically validated knowledge. This knowledge, together with guesses, conjectures, and other ideas is distributed across disciplines and specialties. While scientists are, in the main, geographically dispersed they do work within "invisible colleges" which have an internal and stratified structure. The stratification is determined in large part by the publication of scientists in reputable journals. Although journals are among the central institutions of science, there is nothing that guarantees a journal's reputation. Minimally, it depends on the level of interest maintained in it within a scientific community. Further, interest in a journal rests on the extent to which it is seen as publishing significant work. Such a chain of reasoning verges on the circular as concepts like 'reputable,' 'interest' and 'significant' rest on communal standards within a scientific community. Burt and Doreian (1982) argue that these characteristics are maintained by one or more social psychological processes whereby scientists socialize each other. Moreover, these two researchers show that these processes are mediated by the internal structure of a scientific community. If this argument is correct, then a research strategy whereby scientists are sampled from an invisible college and solicited for their views concerning the important journals of their field without taking into account the social relations among those scientists is problematic. Values, in addition to knowledge, are transmitted over a social structure.

### 2.2. Example 2: dependency theory

Following World War II there was considerable interest in patterns of economic development among Third World countries. Within the sociological literature, economic development models have been seen as inadequate because they "are based on the implicit assumption that countries represent separate systems of economic production" (Rubinson, 1976). Dependency theory models have been constructed as a way of overcoming this limitation, and most variants of this approach assume further that all countries are part of a single system of production that contains multiple political units within it. Dependent variables such as rate of economic growth, level of economic development, and societal inequality have been linked to variables measuring First World penetration of Third and Fourth World countries. Various mechanisms and models have been specified and, when estimated, appear to support arguments that the receipt of developmental aid and the receipt of foreign capital are inimical to the interests of most Third and Fourth World nations in the world system of nations. Of

course, this claim has been challenged. It is rather odd that the proponents of these theories, at least in the version of American quantitative sociology, fall back on regression models to sift the empirical information. By all of the arguments of the dependency theorists, the world is an interdependent system and, one would presume, ought to be modeled as such.

As examples accumulate whereby linear models can lead to mistaken inference if spatial autocorrelation is omitted, then in a more general network context there is the serious risk that classical regression models estimated with data depicting nation states (in an interdependent system) are vulnerable to the same kinds of mistaken inference. While the theory is inherently structural, the procedures for estimating model parameters are not. As most social phenomena occur in structural contexts, the problem may be more general. The received wisdom among social network analysts is that social structure makes a difference and must be included in the analysis of most social phenomena. But this claim may be little more than received dogma and it behooves network analysts to spell out the way in which network autocorrelation models could be constructed and estimated.

### 2.3. Defining W for network autocorrelation models

In the context of spatial autocorrelation, Upton and Fingleton (1985) remark "As ever, the choice of **W** is essentially arbitrary ..." The remark is as daunting as it is frank. One of the more common forms of defining the weight matrix, **W**, for spatial autocorrelation models is to start with a matrix, **C**, that represents whether or not areas are contiguous. This binary **C** is often made row stochastic to form the matrix **W** (which then has 1 as the largest eigenvalue). Formally, this is no different to using the conventional sociometric representation of the structure of a group and turning it into a row stochastic matrix. Initial explorations of network autocorrelation models have tended to do this.

In the spatial autocorrelation literature, distances between the centroids of the areal units, together with the specification of a distance–decay model have been used. Similarly, for strongly connected graphs, it would be possible to use graph theoretic distances (of geodesics) in the specification of network autocorrelation models.

These suggested examples, motivated by successes found in spatial autocorrelation models, are imitations that stay very close to the spatial case. While network autocorrelation models imitating spatial models have had some success, it is clear that they need to draw their inspiration from social network ideas.

In geographical examples, contiguity and accessibility are frequently mutually redundant and change slowly. This is not true for most social networks. In social networks, reachability within a graph may provide a genuinely new basis for measuring interdependence. This may be especially true for valued graphs where the matrix elements represent the strength of a link between two actors. Reachability at level $n$ (Doreian, 1974) considers all paths between pairs of actors with a view to finding the path with the largest minimal element n in the path. In essence, a threshold filter is put over the sociomatrix to restrict attention to only those links above a certain level. Actors reachable at one level are not be reachable at another higher level unless there is a path between them whose links are above the threshold value. However, this is only a tiny step from the spatial foundations.

## 2.4. Equivalence

Of the many ideas that emerged within network analysis during the 1970s and 1980s, the notion of equivalence has captured most of the attention. The sociometric origins of network analysis are seen in the study of small groups. As the computing technology available to social analysts enabled the study of larger systems, it became clear that very large networks verge on the incomprehensible.

This fueled the desire for simpler representations. Of more interest was the idea that networks among individuals (be they people, groups, organizations, or states) could be seen as empirical instantiations of simpler and more fundamental structures. Thus, if it were possible to lay out stringent criteria, large networks could be distilled for their structural essence. The first concept of equivalence, in intuitive terms, was one were two actors are equivalent if they are connected in exactly the same fashion to the rest of the network. Structurally, two such actors are indistinguishable and can be merged to a common position. Formally, the specification of structural equivalence is:

In a graph $< \mathbf{P}, \mathbf{R} >$ made up of a set of actors, $\mathbf{P}$, and a social relation, $\mathbf{R}$, an equivalence, $\mathbf{E}$, is a structural equivalence if and only if for all distinct actors $\mathbf{a}, \mathbf{b}, \mathbf{c} \epsilon \mathbf{P}$, $\mathbf{aEb}$ implies

(i) **aRb** if and only if **bRa**;

(ii) **aRc** if and only if **bRc**;

(iii) **cRa** if and only if **cRb**; and,

(iv) **aRa** implies **aRb** (White and Reitz, 1983).

In principle, any social network can be reduced to a set of structurally non–equivalent positions that are each occupied by structurally equivalent actors. However, this intuition is of little value in practice as there are very few *exact* structural equivalences in social networks. The pragmatic response to this dilemma has been to develop methods that measure the extent to which each pair of nodes is equivalent, and then to mobilize some clustering algorithm. Each structural position (location) is now occupied not by actors that are exactly equivalent, but by actors that are sufficiently close to being equivalent. The use of a measure of equivalence and a clustering algorithm permits an analyst to establish a partition of the actors in a network.

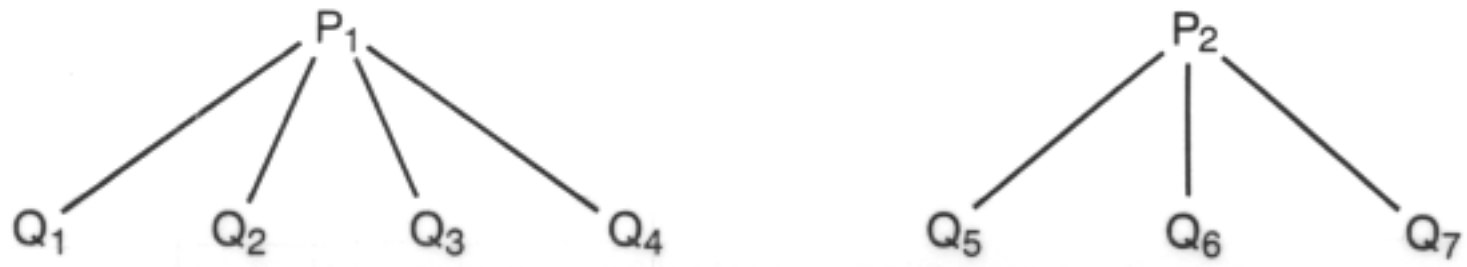
For this idea to be mobilized in a network autocorrelation model, the underlying intuition is that equivalent actors are subject to equivalent processes that affect them by virtue of their occupancy of the same position in the network. Consider Figure 1, where the two nodes $\mathbf{P_1}$ and $\mathbf{P_2}$ send ties to non–overlapping sets of other actors.

This generic picture can be illustrated by the following examples. First, $\mathbf{P_1}$ and $\mathbf{P_2}$ are distinct parents linked to their respective children. A second situation could see $\mathbf{P_1}$ and $\mathbf{P_2}$ as former colonial powers linked to sets of their former colonies. For some variable of interest, it could be that $\mathbf{P_1}$ and $\mathbf{Q_1}$ are similar as a result of the dyadic link between them. Both could believe in "the empire" where the elite of $\mathbf{Q_1}$ have migrated from $\mathbf{P_1}$. An alternative view of a structural process would be one where $\mathbf{Q_1}$ through $\mathbf{Q_4}$ are similar, although there may be no direct link between them. Clearly, $\mathbf{Q_1}$ through $\mathbf{Q_4}$ are structurally equivalent and would be subject to the same process, for example, due to an unfavorable trading relations with the colonial power. Under the first model, $\mathbf{P_1}$ and $\mathbf{Q_1}$ through

$\mathbf{Q_4}$, would all be similar, while under the second (structural equivalence) representation $\mathbf{Q_1}$ through $\mathbf{Q_4}$ would be similar by virtue of being structurally equivalent but, for some selected variable of interest, could be quite distinct from $\mathbf{P_1}$. Empirically, if the variable of interest was measured and all five actors were close, then there would be support for the cohesion argument. Alternatively, if $\mathbf{P_1}$ was quite different from $\mathbf{Q_1}$ through $\mathbf{Q_4}$, which in turn are similar to each other, then the suggestion would be that a structural equivalence mechanism rather than a cohesion mechanism was at work.

**Figure 1.**

Illustration of structural and regular equivalences.




A generalization of structural equivalence is regular equivalence where objects are regularly equivalent if they are equivalently connected to equivalent others. More formally, this can be expressed as

In a graph $< \mathbf{P}, \mathbf{R} >$ (defined above) an equivalence is a regular equivalence if and only if for all actors $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d} \epsilon \mathbf{P}$, $\mathbf{aEb}$ implies

(i) aRc implies there exists d $\epsilon$ P such that bRd and dEc; and

(ii) cRa implies there exist d $\epsilon$ P such that dRb and dEc (White and Reitz, 1983).

Using Figure 1, $\mathbf{Q_5}$ through $\mathbf{Q_7}$ are structurally equivalent by virtue of being connected to $\mathbf{P_2}$. However, $\mathbf{Q_5}$ through $\mathbf{Q_7}$ are not structurally equivalent to $\mathbf{Q_1}$ through $\mathbf{Q_4}$, since $\mathbf{P_1}$ is distinct from $\mathbf{P_2}$. But, it is clear that $\mathbf{Q_1}$ through $\mathbf{Q_4}$ are connected to $\mathbf{P_1}$ in the same way that $\mathbf{Q_5}$ through $\mathbf{Q_7}$ are connected to $\mathbf{P_2}$. Conversely, $\mathbf{P_1}$ is connected to $\mathbf{Q_1}$ through $\mathbf{Q_4}$ in the same fashion that $\mathbf{P_2}$ is connected to $\mathbf{Q_5}$ through $\mathbf{Q_7}$. In short, $\mathbf{P_1}$ and $\mathbf{P_2}$ are regularly equivalent while $\mathbf{Q_1}$ through $\mathbf{Q_7}$ are regularly equivalent. Although $\mathbf{Q_5}$ through $\mathbf{Q_7}$ are connected to a different colonial power, it could be argued that they are subject to the same process as it applies to all colonies regardless of the identity of, colonial powers. Similarly, parents occupy a role while, in relation to them, children occupy a complementary role. If a process is mediated by regular equivalence, then one would expect that $\mathbf{Q_1}$ through $\mathbf{Q_7}$ would be similar with regard to some variable while $\mathbf{P_1}$ and $\mathbf{P_2}$ would be similar to each other. It is worth noting that structural equivalence is a special case of regular equivalence in the sense that structurally equivalent actors are also regularly equivalent, but not vice versa.

Thus far, this discussion has sketched out a cohesion mechanism, a structural equivalence mechanism, and a regular equivalence mechanism. The implicit assumption is that if we know which mechanism is at work, then we can construct an appropriate weight matrix $\mathbf{W}$ that captures the interdependency among the actors. If $e_{\mathbf{ij}}$ denotes the extent of equivalence of $\mathbf{i}$ and $\mathbf{j}$ in some social structure, this may suffice for the interdependence measure. Alternatively, the set of $\{e_{\mathbf{ij}}\}$ can be normalized in some fashion, for example:

$$w_{ij} = \frac{\max\{e_{ij}\} - e_{ij}}{\Sigma[\max\{e_{ij}\} - e_{ij}]}$$

Of course, there may be other ways in which the weight matrix can be constructed. At face value, the wry comment of Upton and Fingleton concerning the arbitrariness of $\mathbf{W}$ is pertinent. However, if the cohesion, or structural equivalence, or regular equivalence mechanisms can be specified in *advance*, it is possible to construct the appropriate $\mathbf{W}$ on *substantive* grounds. In principle, the way is then clear to mobilize all of the statistical machinery found within the rubric of spatial autocorrelation to formulate, estimate and test network autocorrelation models. [1]

Following Anselin (1988, pp. 34-5), a family of network autocorrelation models can be specified: [2]

$$\mathbf{y} = \rho_1 \mathbf{W}_1 \mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad (2.1a)$$

$$\boldsymbol{\epsilon} = \rho_2 \mathbf{W}_2 \boldsymbol{\epsilon} + \boldsymbol{\nu} \qquad (2.1b)$$

where $\boldsymbol{\nu} \sim \mathbf{N}(\mathbf{0}, \sigma^2\mathbf{I})$. There are three special versions of the generic model specified in (2.1). When $\rho_1 = \rho_2 = \mathbf{0}$, equation (2.1a) reduces to the usual OLS population regression function while (2.1b) becomes the conventional specification of a normally distributed error term. For $\rho_1 = \mathbf{0}$ and $\rho_2 \neq \mathbf{0}$, equation (2.1) is the network disturbances model. Finally, when $\rho_1 \neq \mathbf{0}$ and $\rho_2 = \mathbf{0}$ we have the network effects model.

## 3. Issues stemming from network autocorrelation models

### 3.1. Multiple processes

If a cohesion model can be unequivocally specified it can be estimated and interpreted. Similarly, if an equivalence model can be specified, then it too can be estimated and interpreted. However, some network analysts posit a sharp distinction between cohesion models and equivalence models. If rendering a decision as to whether a cohesion process or an equivalence process is at work is necessary, it is a major disadvantage to use a model with a single regime of network effects. For example, Burt and Doreian (1982) estimated separately a cohesion model and a structural equivalence model in a study of the distribution of evaluations of major journals by scientists in a specific field. The relative performance of the two models were considered through an analysis of the residuals remaining when the separate analyses had been conducted. As the two mechanisms take the form of rival hypotheses it seems preferable to examine them competitively and directly. Rather than fit the two models separately and examine their residuals, it is preferable to have a model where both processes are explicitly included. Similarly, for a debate between structural equivalence mechanisms and regular equivalence mechanism, it would be desirable to build a model with

both present. Theoretically speaking, this task can be carried out in the following way (Doreian, 1989a) depending upon the practical issues involved in estimating a model with two regimes of network effects.

A model with two regimes of effects[3] autocorrelation can be written as

$$\mathbf{y} = \rho_1 \mathbf{W_1 y} + \rho_2 \mathbf{W_2 y} + \mathbf{X\beta} + \boldsymbol{\epsilon} \tag{3.1}$$

with $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I})$. With $\mathbf{A} = \mathbf{I} - \rho_1 \mathbf{W}_1 - \rho_2 \mathbf{W}_2$ and $|\mathbf{A}|$ as the Jacobian of the transformation from $\boldsymbol{\epsilon}$ to $\mathbf{y}$, the log–likelihood function can be written as

$$\ln(\mathbf{L}) = \text{constant} - (N/2)\ln\sigma^2 - \frac{1}{2\sigma^2}[\mathbf{z'z} - \mathbf{2\beta' X' z} + \boldsymbol{\beta}' \mathbf{X' X \beta}] + \ln|\mathbf{A}| \tag{3.2}$$

Notationally, use of MLE leads to the same estimation equation for $\boldsymbol{\beta}$ as before [see equation (1.8)]:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'z} \tag{3.3}$$

similarly for $\sigma^2$;

$$\hat{\sigma}^2 = (\mathbf{z'z} - \mathbf{2\beta' X' z} + \boldsymbol{\beta}' \mathbf{X' X \beta})/\mathbf{N} \tag{3.4}$$

with $\mathbf{z} = \mathbf{Ay}$. The iterative estimation for the $\rho_i$ is more complex, as is the asymptotic variance–covariance matrix for obtaining approximate standard errors for the estimated parameters (with $\mathbf{B}_i = \mathbf{W}_i \mathbf{A}^{-1}$ for $i = 1, 2$):

$$V(\omega, \rho_1, \rho_2, \boldsymbol{\beta}) = \omega^2 \begin{pmatrix} N/2 & \omega\, tr(\mathbf{B}_1) & \omega\, tr(\mathbf{B}_2) & \mathbf{0} \\ \omega\, tr(\mathbf{B}_1) & \mathcal{B}_{11} & \mathcal{B}_{12} & \omega X' \mathbf{B}_1 \mathbf{X\beta} \\ \omega\, tr(\mathbf{B}_2) & \mathcal{B}_{21} & \mathcal{B}_{22} & \omega \mathbf{X' B}_2 \mathbf{X\beta} \\ \mathbf{0}' & \omega \boldsymbol{\beta}' \mathbf{X' B}_1' \mathbf{X} & \omega \boldsymbol{\beta}' \mathbf{X' B}_2' \mathbf{X} & \omega \mathbf{X'X} \end{pmatrix}^{-1} \tag{3.5}$$

where

$$\begin{aligned}
\mathcal{B}_{11} &= \omega^2[tr(\mathbf{B}_1'\mathbf{B}_1) + tr(\mathbf{B}_1)^2] + \omega \boldsymbol{\beta}' \mathbf{X}' \mathbf{B}_1' \mathbf{B}_1 \mathbf{X\beta} \\
\mathcal{B}_{12} &= \omega^2[tr(\mathbf{B}_1'\mathbf{B}_2) + tr(\mathbf{B}_1\mathbf{B}_2)] + \omega \boldsymbol{\beta}' \mathbf{X}' \mathbf{B}_2' \mathbf{B}_1 \mathbf{X\beta} \\
\mathcal{B}_{21} &= \omega^2[tr(\mathbf{B}_2'\mathbf{B}_1) + tr(\mathbf{B}_2\mathbf{B}_1)] + \omega \boldsymbol{\beta}' \mathbf{X}' \mathbf{B}_1' \mathbf{B}_2 \mathbf{X\beta} \\
\mathcal{B}_{22} &= \omega^2[tr(\mathbf{B}_2'\mathbf{B}_2) + tr(\mathbf{B}_2)^2] + \omega \boldsymbol{\beta}' \mathbf{X}' \mathbf{B}_2' \mathbf{B}_2 \mathbf{X\beta}.
\end{aligned}$$

### 3.2. Distinct types of actors

The two regime model of network effects is plausible for a community of scientists in an invisible college. Scientific leadership, and its corresponding material and psychic rewards, accumulate and evolve over a scientific career. Even if a scientific elite can be distinguished from the bulk of the members of an invisible college, the model retains plausibility. But for the motivating example of nation states bound into a global system, plausibility is stretched even for the two regime model. If colonies can be distinguished from colonial powers, or if core states, semi–peripheral states, and peripheral states are subject to distinct (but coupled) processes, the two regime model as stated loses its plausibility. The crux of the problem is that the qualitative distinctions on the actors may need to be incorporated into the model.

One rather simple, but effective, approach to this problem takes the form of incorporating the distinctions into the matrix $\mathbf{X}$. Snyder and Kick (1979) observe that world system theorists and dependency theorists have proposed no adequate operational criteria for classifying countries into the world system positions. Various *ad hoc* definitions have been proposed, and it is not surprising that the empirical status of some nations, for example Spain, is completely ambiguous as to which position contains them. Snyder and Kick's solution to this problem rests on the conceptualization of structural equivalence. The world system conceptualization (Wallerstein, 1974) is fundamentally structural and it seems reasonable to try and distill world system positions from structural data.

Snyder and Kick's (1979) proposal is straight forward: use structural data to generate structural positions. They used four tie types—trade relations, treaties, exchange of diplomats and military interventions—for the structural data. Each relation generates a nation–by–nation matrix. These matrices were stacked and analyzed jointly to obtain a partition in terms of structural equivalence by use of CONCOR (Breiger, Boorman, and Arabie, 1975).[4] The authors identified ten non–equivalent positions, across which 118 nations were distributed. Using world system terms, one position was clearly the core, another three positions can be viewed as belonging to the semi–periphery, with the remaining six blocks as parts of the periphery.

In 1979, network autocorrelation models were not an option.[5] Working within a regression framework, Snyder and Kick included the structural data by means of a set of dummy variables (omitting one to avoid an exact linear dependence among the regressors). Regression models, or rather, the parameter estimates and inferential decisions, are frequently challenged. Snyder and Kick's work was no exception—but the basis for the critics' objections did not include issues of network autocorrelation. Considerations of regression diagnostics and curvilinear relations suggested that Snyder and Kick's analysis (and theoretical arguments) were not supported. However, when Nolan (1983) reanalyzed the Snyder and Kick data using only three positions–core, semi–periphery, periphery—the initial findings were supported. Clearly, there are limitations to the number of dummy variables that can be included in a regression to capture structural positions when the data points are interdependent.

While using dummy variables to represent qualitative differences between nation states is direct and practical, it is not clear that the underlying mechanisms of the world system are adequately modeled. For models of income inequality or economic growth of nation states the incorporation of dummy variables amounts to little more than the fitting of mean values for the nations of different sectors (together with slope shifts if necessary). The central idea of the world system theorists is that nations of the various positions are locked into reciprocal mechanisms advantageous to one group and disadvantageous for another. It seems more appropriate to generate directly a model that focuses on the mechanisms themselves.

One simple way to incorporate two distinct types of actor is to partition $\mathbf{y}$, $\mathbf{X}$, $\mathbf{W}$, and $\boldsymbol{\epsilon}$ so that an effects model is written as:

$$\begin{pmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{pmatrix} = \begin{pmatrix}\rho_1\mathbf{I}_1 & \mathbf{0}\\ \mathbf{0} & \rho_2\mathbf{I}_2\end{pmatrix}\begin{pmatrix}\mathbf{W}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{W}_2\end{pmatrix}\begin{pmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{pmatrix} + \begin{pmatrix}\mathbf{X}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{X}_2\end{pmatrix}\begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{pmatrix} + \begin{pmatrix}\boldsymbol{\epsilon}_1\\ \boldsymbol{\epsilon}_2\end{pmatrix} \tag{3.6}$$

in obvious notation. Ignoring the network autocorrelation term, this is exactly the two population model of Zellner's (1962) seemingly unrelated regression model, treated at length

in Theil (1971). The block diagonal form of $\mathbf{W}$ means that the eigenvalues of $\mathbf{W_1}$ and $\mathbf{W_2}$ can be determined separately and used to give the eigenvalues of $\mathbf{W}$. Although, the two types of actors have distinct equations:

$$\begin{aligned} \mathbf{y}_1 &= \rho_1 \mathbf{W}_1 \mathbf{y}_1 + \mathbf{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1 \\ \mathbf{y}_2 &= \rho_2 \mathbf{W}_2 \mathbf{y}_2 + \mathbf{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{\epsilon}_2 \end{aligned} \tag{3.7}$$

the maximum likelihood method can be used directly with the known eigenvalues of $\mathbf{W}$. In the interpretation of a partitioned $\mathbf{W}$, the two types of actors are kept distinct. For nations of Type $\mathbf{i}$, only the values of $\mathbf{y}$ for other Type $\mathbf{i}$ nations are used in the prediction equations. Using the nations example, with $\mathbf{y}$ a measure of economic growth, the core and non-core nations can have distinct $\mathbf{W_i}$ and $\rho_\mathbf{i}$. Within the two classes of actors the weights, and, by implication, the underlying processes, can be quite different and can be differentially important (depending upon the values of $\rho_\mathbf{i}$). At face value, a $\mathbf{W}$ constructed via regular equivalence could take the partitioned form given the core nations would be maximally like each other and maximally unlike non-core nations. Similarly non-core nations will be maximally like each other and unlike core nations with regard to (regular) position in the network.

However, dependency arguments go beyond saying that there are distinct processes for core and non-core nations. The claim is that First World nations benefit at the expense of Third and Fourth world nations. By repatriation of capital and extraction of profits, First World corporations and nations benefit while Third and Fourth world nations lose not only their resources, but also control over resources, and hence suffer from a distorted development. If this argument is correct (or indeed if the counter argument that all nations benefit is correct), then the model stated in equation (3.6) becomes inadequate. The weight matrix still can be partitioned but it would take a more complicated form, *i.e.*,

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{21} & \mathbf{W}_{22} \end{pmatrix}$$

The role of $\mathbf{W_{11}}$ is that of $\mathbf{W_1}$ while $\mathbf{W_{22}}$ plays the same role as $\mathbf{W_2}$. However, the matrices of greater interest will be $\mathbf{W_{12}}$ and $\mathbf{W_{21}}$ as they represent the way in which the classes of nations have an impact on each other. In tandem with the partitioned form of $\mathbf{W}$, these are four network effects parameters. The within position parameters are $\rho_{11}$ and $\rho_{22}$, while $\rho_{12}$ and $\rho_{21}$ are the between position parameters. Letting

$$\mathbf{R} = \begin{pmatrix} \rho_{11} & \rho_{12} \\ \rho_{21} & \rho_{22} \end{pmatrix}$$

the model can be stated as:

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} = \mathbf{R} \otimes \mathbf{W} \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} + \begin{pmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \end{pmatrix} \tag{3.8}$$

where $\otimes$ is the Kronecker product. If only First world nations benefit then $\rho_{21}$ would be positive and $\rho_{12}$ would be zero or negative (or at least smaller than $\rho_{21}$). If all nations benefit, then both $\rho_{12}$ and $\rho_{21}$ would be positive. Patterns inside the $\mathbf{W_{ij}}$ also could be of interest. The major problem with the model of (3.8) is that it may be intractable. The

MLE methods thus far, rest on the simple representation of $\ln|\mathbf{A}|$ as $\Sigma(1-\rho\lambda_i)$ where the $\{\lambda_i\}$ are the eigenvalues of $\mathbf{W}$.

It would appear that this simple decomposition is prohibited by the partitioned form of $\mathbf{W}$ and 4 network autocorrelation parameters unless either $\rho_{21}$ or $\rho_{12}$ is zero. If, from an estimation view, this model is intractable, then the straightforward use of dummy variables, obtained from the structure of the network, as proposed by Snyder and Kick (1979) has great appeal.

### 3.3. Boundary effects

As defined earlier, a social network is a set of actors, $\mathbf{P}$, over which one (or more) social relation(s) are defined. Although the existence of $\mathbf{P}$ is taken as given, the empirical problem of locating its boundaries remains a persistent and vexing problem. For virtually all $\mathbf{P}$, we know there are network ties that cross the boundary between $\mathbf{P}$ and all other actors. Consider Figure 2 where the upper panel shows a network from which a sub–network (second panel) has been extracted for a network analysis. If an influence process is at work, and if it is activated through the network ties, then actors g, h, i, j, and k are beyond the boundary of $\mathbf{P}$ (made up of a, b, c, d, and e). The actors on the boundary of $\mathbf{P}$ are affected by actors in the (selected) network as well as some actors outside $\mathbf{P}$. As an exact analogue of the spatial case, this network example confronts the same boundary problems as in spatial systems. At face value, then, network autocorrelation models can benefit from the experience generated through the use of spatial autocorrelation models. Unfortunately, this understates the problems found in network autocorrelation models.

For a cohesion based model, f is affected little by the boundary location as its entire ego–network is contained within the sampled network. Similarly for e. Of course, actors outside $\mathbf{P}$ can have an impact on f, but only through actors in the studied network. However, for positional models (for example, both structural and regular equivalence), mis–specification of the boundary is extremely consequential as the position of an actor is made up of all ties (present and absent) across the *whole* network. If the graph of Figure 2 (a) is accurate, then the positions of all actors in panel (b) are changed by the omission of network members. In particular, the construction of $\mathbf{W}$ will change dramatically and the estimated network autocorrelation model is likely to be misleading.

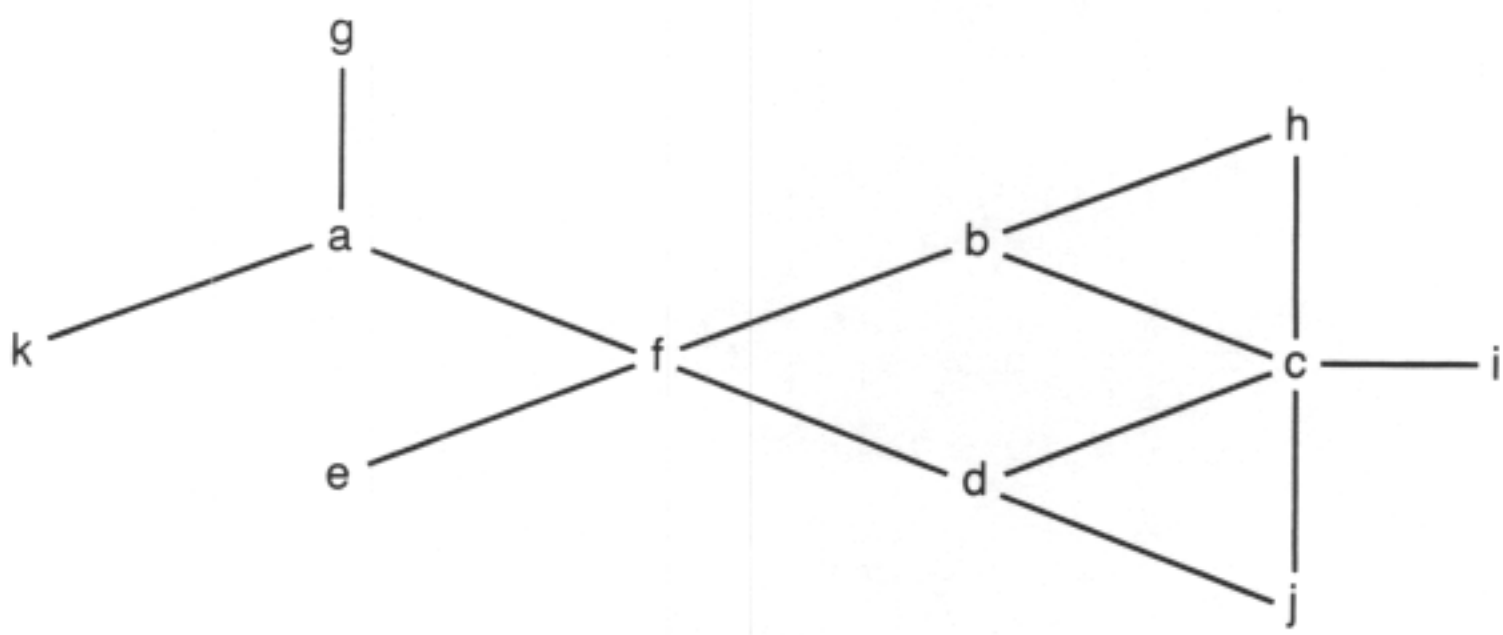
A special case of the boundary problem is the omission of an actor from $\mathbf{P}$. Taking Figure 2(b) as the true group, it is possible that data are not collected from an actor known to be in $\mathbf{P}$—say, by oversight, respondent refusal, *etc.* Imagine that data are not present for f in Figure 2(b). The structure that remains [Figure 2(c)] is radically changed. So much so, that any network analysis of the data is pointless. In contrast, omitting e is far less consequential for a subsequent network autocorrelation model. Anselin (1988) draws on work of Griffith (1983, 1985) to point to a way in which the consequences of boundary effects can be studied.[6]

Consider a network made up of two parts, the nodes of a particular group, denoted $G$, and nodes not in $G$, but in the wider network, denoted $H$. For a network autocorrelation model, as specified in (1.6), recognition of included and excluded actors leads to

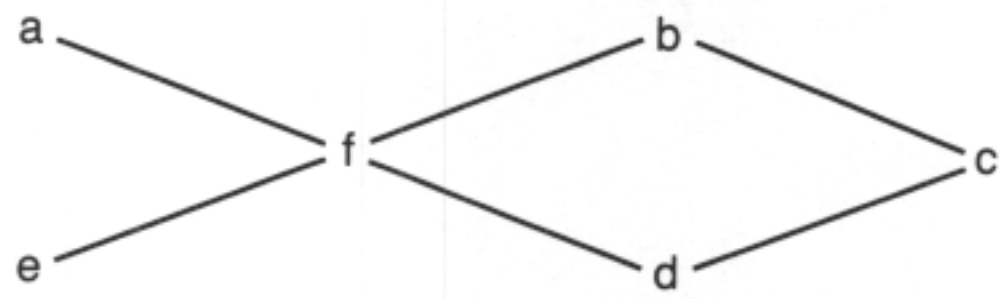
$$\begin{pmatrix}\mathbf{y}_G\\ \mathbf{y}_H\end{pmatrix}=\rho\begin{pmatrix}\mathbf{W}_{GG}\ \mathbf{W}_{GH}\\ \mathbf{W}_{HG}\ \mathbf{W}_{HH}\end{pmatrix}\begin{pmatrix}\mathbf{y}_G\\ \mathbf{y}_H\end{pmatrix}+\begin{pmatrix}\mathbf{X}_G\\ \mathbf{X}_H\end{pmatrix}\boldsymbol{\beta}+\begin{pmatrix}\boldsymbol{\epsilon}_G\\ \boldsymbol{\epsilon}_H\end{pmatrix} \tag{3.9}$$

**Figure 2.**

Examples of boundary problems in networks.




(a) An intact social network




(b) A subgroup embedded in the social network


(c) The subgroup with a missing node

with the obvious partitions of $\mathbf{W}$ and $\mathbf{y}$. The actual model estimated, with no recognition of $H$, would be

$$\mathbf{y} = \rho \mathbf{W}_{GG} \mathbf{Y}_G + \mathbf{X}_G \boldsymbol{\beta} + \boldsymbol{\epsilon}_G \tag{3.10}$$

while from (3.8) the corresponding equation is (Anselin, 1988, p. 175):

$$\mathbf{y}_G = \rho \mathbf{W}_{GG} \mathbf{Y}_G + \rho \mathbf{W}_{GH} \mathbf{y}_H + \mathbf{X}_G \boldsymbol{\beta} + \boldsymbol{\epsilon}_G \tag{3.11}$$

Re–writing, this becomes:

$$\mathbf{y} = \rho \mathbf{W}_{GG} \mathbf{Y}_G + \mathbf{X}_G \boldsymbol{\beta} + (\rho \mathbf{W}_{GH} \mathbf{y}_H + \boldsymbol{\epsilon}_G) \tag{3.12}$$

There is then an unknown network dependence term $(\rho \mathbf{W}_{GH} \mathbf{Y}_H + \boldsymbol{\epsilon}_G)$ with a distinct network autocorrelation structure. This error term is unlikely to have zero mean, nor will it be spherical (Anselin, 1988, p. 176). Further complicating matters is the fact that $\mathbf{Y}_G$ and $\mathbf{Y}_H$ are interdependent (equation 3.8) so that the error term in (3.11) is no longer independent of $\mathbf{Y}_G$.

Clearly, one line of attack is use (3.8)–(3.12) as the basis for simulation studies. Another is substantive and empirical. Consider the example of a network of social service agencies dealing with, in one way or another, the population of children and youth. This network is distributed across many sectors including mental health, health, criminal justice, social welfare, education, and employment sectors (Doreian, Woodard and Musa, 1989). Boundaries within and between these sectors are fuzzy, but cores and boundaries can be specified. A k–core is a set of nodes in a connected graph, such that considering only the nodes in the k–core, each node has in–degree and out–degree of at least k. As k defines a threshold, boundaries can be established experimentally, within and between sectors, with a view to examining the consequences of excluding sets of agencies.

This operationalization of boundaries rests on a particular data structure, obtained by a snow–ball sampling scheme. Starting with the central set of agencies (acknowledged by all as in the core), directors and staff are asked to list the other agencies they need to interact with in order to service their clientele. When another organization is cited enough times, it is added to the agency list, and data are obtained from it until no more organizations are added. With a low threshold, the network is expanded well into the peripheral agencies and beyond any reasonable boundary to the network. Data are then available on agencies outside $\mathbf{P}$, and boundary effects can be examined in the relevant context of the network.

### 3.4. Aggregation issues

Given a network, equivalence ideas are used to provide two complimentary reductions:

(i) to join nodes together in a single position, and generate a set of non–equivalent positions (blocks); and,

(ii) simultaneously, collapse relations between nodes to define relations between the constructed blocks.

The initial formulation of structural equivalence (Lorrain and White, 1971) was given in terms of category theory where the product of morphisms was crucial. If one morphism represents "mother of" and a second morphism represents "brother of," then their compound (product) will be a morphism (the product is closed) that ought to correspond exactly to

"maternal uncle of." In its initial formalization, the construction of positions via collapsing of objects and morphisms jointly became impractical for all but small networks. The available algorithms for getting a structural equivalence partition are all rather crude attempts to create a practical partition which retain as much of the initial conception as possible. However, the partitioning of nodes and links is done in sequence—a partition of the nodes then collapsing ties. In fact, both the nodes and the ties are aggregated in terms of some equivalence conception. Consider the example in Figure 3 in terms of regular equivalence. The graph of 9 nodes can be reduced to 4 blocks and the ties between blocks are constructed from the ties between nodes, in each block, to nodes in the other blocks. The rows and columns of sociomatrix C, in Figure 3 have been permuted so that block members are together.

The image matrix can be constructed in a variety of ways depending on the criteria chosen. For example, one criterion could be that the presence of any link between blocks suffices to define a tie between blocks, or that the density of ties between blocks exceeds some threshold value (usually the overall density for the tie in the network). Under either criterion, the image matrix is the 4–by–4 matrix in Figure 3. Apart from providing a simpler and more easily understood network, the hope is that image matrices can form the building blocks for a structural theory of relations.

In terms of network autocorrelation, each block is fundamental and the nodes in that block provide indicators of it.[7] Rather than formulate an autocorrelation model in terms of individual nodes, such a model can be formulated in terms of blocks—if they are fundamental. This, however, leads directly to problems of aggregation as nodes are aggregated into positions. If an actor is placed incorrectly in a block, then the aggregation will have the same spill–over problem described by Anselin (1988, p. 12), generating network dependence. Serious as this is, there is yet another aggregation problem stemming from the aggregation of ties between blocks. The risk here is very serious as it leads to inaccurate [8] construction of $\mathbf{W}$.

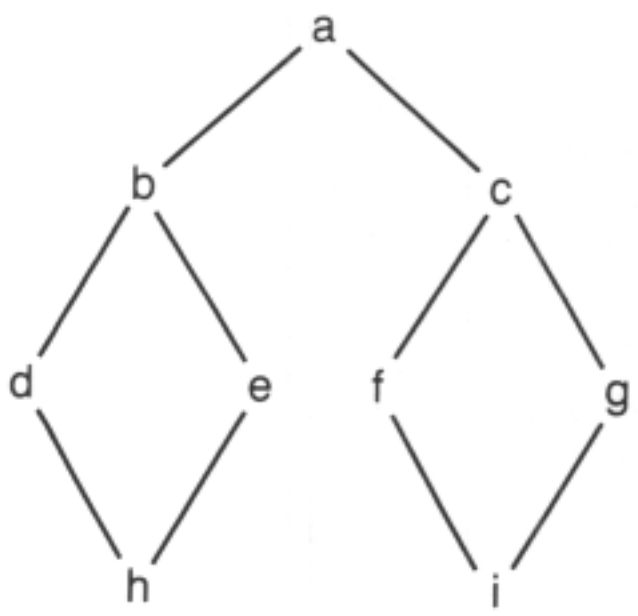
### 3.5. Mixtures of processes

The discussion of two regimes of network effects was couched in terms of rival structural mechanisms–cohesion versus structure equivalence, or structural equivalence versus regular equivalence–where the linear model provided an inferential framework. Two rival structural accounts, were competitively examined. Of course, one outcome could well be that $\rho_1$ and $\rho_2$ are both non–zero and that both mechanisms, via $\mathbf{W}_1$ and $\mathbf{W}_2$, are operative in generating the $\mathbf{y}$ as it is distributed over the network. However, it may be necessary to go beyond this formulation to one where the mechanisms are explicitly coupled.

Consider Figure 4, where the nodes represent political actors and the lines represent strong political ties in a hypothetical graph. It is a graph whose structure renders the issue of deciding *between* structural versus regular equivalence fruitless (cf. Doreian, 1988). When structural equivalence is considered, the partition yields two political alliances (Figure 4, upper panel). In addition, when regular equivalence is considered, the partition returned is a complementary one and provides additional insight into the structure of the group. Actors f and g have an integrative role between the alliances; actors a, c, d, k, and i all provide further integration within these alliances; and actors b, e, h, j, and l play no structural role beyond being peripheral and buried in a larger grouping. Given the structural equivalence partition, the regular partition also is coherent. The actors f and g are boundary spanners of

**Figure 3.**

Homomorphic reduction of a network to an image.




(a) A nine node network

| | | K | L | | M | | | | N | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | a | b | c | d | e | f | g | h | i |
| K | a | | 1 | | | | | | | |
| L | b | 1 | | | 1 | 1 | | | | |
| | c | 1 | | | | | 1 | 1 | | |
| M | d | | 1 | | | | | | 1 | |
| | e | | 1 | | | | | | 1 | |
| | f | | | 1 | | | | | | 1 |
| | g | | | 1 | | | | | | 1 |
| N | h | | | | 1 | 1 | | | | |
| | i | | | | | | 1 | 1 | | |

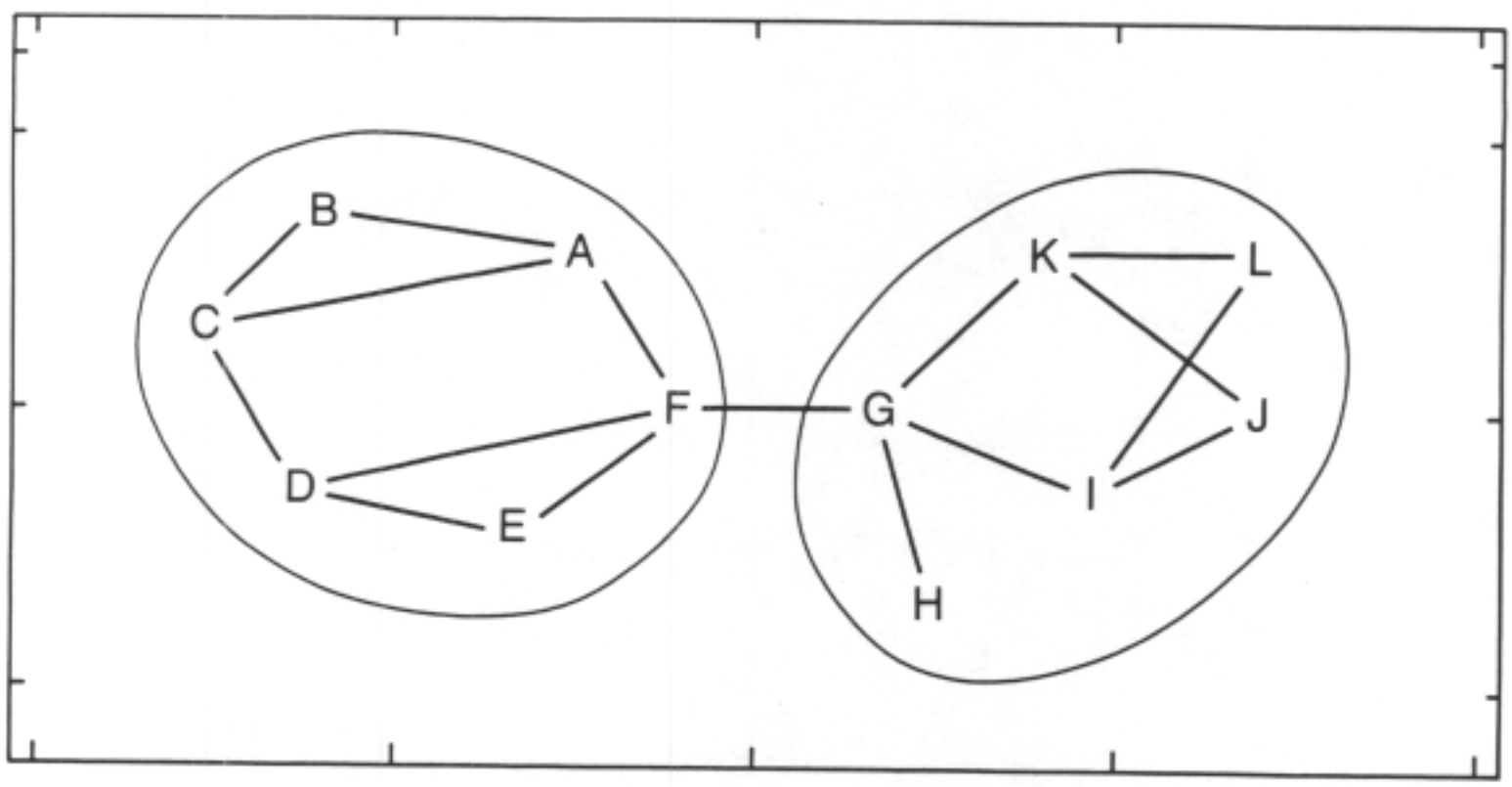
(b) Permuted sociomatrix

| | K | L | M | N |
|---|---|---|---|---|
| K | 0 | 1 | 0 | 0 |
| L | 1 | 0 | 1 | 0 |
| M | 0 | 1 | 0 | 1 |
| N | 0 | 0 | 1 | 0 |

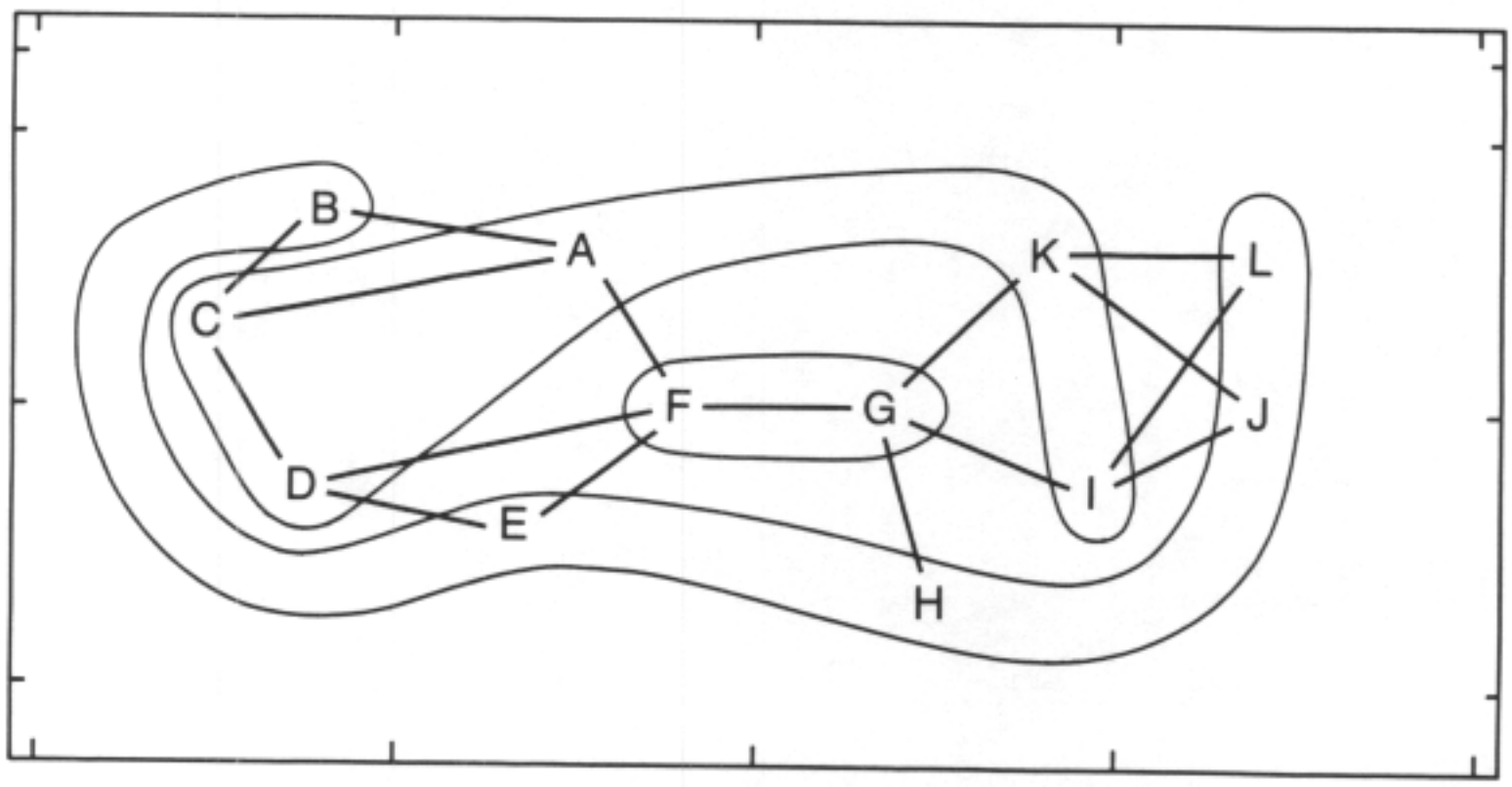

(c) Image matrix under regular equivalence and image graph

**Figure 4.**

Two complementary partitions of graph nodes.




(a) Partition based on structural equivalence




(b) Partition based on regular equivalence

two systems. If the political dynamics involve taking a position with regard to some issue, it is clear that an account based on the cleavage between coalitions is important. But to mediate the conflict, the boundary spanners play a critical role and serve as conduits into their own coalitions. One would expect that boundary spanners are more moderate in their political views than those having no integrative role. There may be a cohesion mechanism inside the alliances and regular equivalence mechanism between them.

## 4. Dynamic models

Most analyses of networks are cross–sectional and avoid many issues raised by the inclusion of time. As noted by Barnes and Harary (1983), this is a serious omission. An empirical situation is likely to exhibit change in three possible ways:

(i) changes in the values of variables characterizing the nodes;

(ii) changes in the network ties; or,

(iii) changes in the nodal attributes together with changes in the network.

Only the first appears to be straightforward. Differential equations (or difference equations) can be used to model changes in the nodal attributes from within the perspective of structural control (Doreian and Hummon, 1976). In essence, equation (4.1) is derived as a model of an equilibriating mechanism:

$$\frac{dy(t)}{dt} = \gamma[\mathbf{y}^{\star}(t) - \mathbf{y}(t)] \tag{4.1}$$

where $\gamma$ is sensitivity parameter. Equation (4.1) can be integrated and the solution system is used as a set of estimation equations. The connection to network autocorrelation (Doreian, 1989a) comes from specifying

$$\mathbf{y}^{\star}(t) = \rho\mathbf{W}\mathbf{y}(t) + \mathbf{X}(t)\boldsymbol{\beta} + \boldsymbol{\epsilon}(t) \tag{4.2}$$

the network effects model (with one or two regimes) is used to model the control values and the integrated process equation gives the estimation equation. There is literature on space–time models (see Upton and Fingleton, 1985; Anselin, 1988). Modeling the changes of nodal properties is difficult, but the problems inherent in modeling the changes of $\mathbf{W}_{ij}$ seem much harder. The problem is one of modelling $\mathbf{W}(t)$ and is both technical and substantive.

To model change in the network ties it may be best to use specific structural theories as a source. For example, structural balance (Cartwright and Harary, 1963) can be mobilized to study change from the premise that social actors prefer balance to imbalance. A sketch of doing this in a dynamics perspective is provided by Hunter (1979).

A second example can be taken from the literature on interorganizational networks where there are hypotheses concerning the formulation (and continuance and dissolution) of inter–organizational ties. Thus, occupational diversity, internal flexibility, professionalization, and large budget size all are seen as conducive to the formation of inter–agency ties. Common definitions of problem areas and agreement on issues of concern both lead to higher quality ties between agencies. Also, the greater the level of turbulence in the environment, the fewer the co–operative ties. Many hypotheses can be compiled to provide a theoretical basis for studying change in the composition of networks.

In terms of network autocorrelation models, network dynamics involve consideration of $\mathbf{y}(t)$, $\mathbf{X}(t)$ and $\mathbf{W}(t)$. Relative to the spatial case, it may be that the volatility of social networks and their inherently changing character will make it more difficult to build and estimate network autocorrelation models.

## 5. Conclusion

Network autocorrelation models draw their inspiration from the success of spatial autocorrelation modeling efforts as the connection between them is direct (via $\mathbf{W}$). The common focus for the two efforts is the recognition that data points are interdependent. The problem of modeling interdependent systems in geographic space is isomorphic to the problem of using network tools in 'social space'. It follows that each group can learn from the other. Thus far, geography, in the vanguard of autocorrelation modeling efforts, has forged the foundations for modeling interdependent systems. However, as more social scientists recognize the need to incorporate social structure into regression and other analyses, more people will be working on the common problem. In principle, when solutions are found outside geography they would be helpful for geographers.

Some of the problems inherent in network autocorrelation models seem much more severe than in the spatial case. Social networks change much more quickly than spatial configurations. The really tough problem is the specification of the process by which $\mathbf{W}(t)$ changes. If this problem is solved, it will benefit all social science—including geography. Perhaps, at some stage, we can envision coupling the two concerns of interdependence modeling and deal with the social processes in geographic space.

## 6. References

Anselin, L. (1988) *Spatial Econometrics: Methods and Models*, Boston: Kluwer Academic Publishers.

Barnes, J. and F. Harary (1983) Graph theory and network analysis, *Social Networks*, **5**, 235-244.

Brandsma, A. and R. Ketellapper (1979) A biparametric approach to spatial autocorrelation, *Environment and Planning A*, **11**, 51-58.

Breiger, R., S. Boorman, P. Arabie (1975) An algorithm for clustering relational data with applications to social network analysis and comparison to multidimensional scaling, *Journal of Mathematical Sociology*, **12**, 328-383.

Burt, Ronald S. (1989) STRUCTURE, Research Program in Structural Analysis, Center for the Social Sciences, Columbia University.

Burt, Ronald S. (1976) Positions in networks, *Social Forces*, **55**, 93-122.

Burt, R. and P. Doreian (1982) Testing a structural model of action: conformity with respect to journal norms in elite methodology sociology, *Quality and Quantity*, **16**, 109-150.

Cartwright, D. and F. Harary (1956) Structural balance: a generalization of Heider's theory, *Psychological Review*, **63**, 277-293.

Cliff, A. and K. Ord (1981) *Spatial Processes, Models and Applications*, London: Pion.

Doreian, P. (1989a) Estimating models with two regimes of network effects, in M. Kochen (ed.) *The Small World*, Chapter 14: Ablex.

Doreian, P. (1989b) Mapping networks through time, in J. Weesie and H. Flap (eds.) *Networks Through Time*, Utrecht: University of Utrecht, forthcoming.

Doreian, P. (1988) Equivalence in a social network, *Journal of Mathematical Sociology*, **13**, 243-282.

Doreian, P. (1981) Estimating linear models with spatially distributed data, *Sociological Methodology*, 359-388.

Doreian, P. (1980) Linear models with spatially distributed data, spatial disturbances or spatial effects, *Sociological Methods Research*, **9**, 29-60.

Doreian, P. (1974) On the connectivity of valued graphs, *Journal of Mathematical Sociology*, **3**, 245-258.

Doreian, P., D. Musa, and K. Woodard (1989) Detecting cores and boundaries in social networks, Working Paper, University Center for Social and Urban Research, University of Pittsburgh.

Griffith, D. (1985) An evaluation of correction techniques for boundary effects in spatial statistics: contemporary methods, *Geographical Analysis*, **17**, 81-8.

Griffith, D. (1983) The boundary value problem in spatial statistical analysis, *Journal of Regional Science*, **23**, 377-87.

Holland, P. and S. Leinhardt (eds.) *Perspectives on Social Network Research*, New York: Academic Press.

Hunter, J. (1979) Toward a general framework for dynamic theories of sentiment and small groups derived from theories of attitude change, in P. Holland and S. Leinhardt (eds.), *Perspectives on Social Network Research*, New York: Academic Press, pp. 223-238.

Loftin, C. and S. Ward (1983) A spatial autocorrelation model of the effects of population density, *American Sociological Review*, **48**, 121-128.

Lorrain, F. and H. White (1971) Structural equivalence of individuals in social networks, *Journal of Mathematical Sociology*, **1**, 49-80.

Mitchell, E. (1969) Some econometrics of the Huk rebellion, *American Political Science Review*, **63**, 1159-1171.

Nolan, P. (1983) Status in the world system, economic and equality, and economic growth, *American Journal of Sociology*, **89**, 410-419.

Ord, K. (1975) Estimation methods for models of spatial interaction, *Journal of the American Statistical Association*, **70**, 120-126.

Rubinson, R. (1976) The world–economy and the distribution of income within states: a cross national study, *American Sociological Review*, **41**, 638-659.

Snyder, D. and E. Kick (1979) Structural position in the world system on economic growth, 1955-1970: a multiple network analysis of transnational interactions, *American Journal of Sociology*, **84**, 1096-1126.

Theil, H. (1971) *Principles of Econometrics*, New York: Wiley.

Upton, G. and F. Fingleton (1985) *Spatial Data Analysis by Example*, vol. 1., New York: Wiley.

Wallerstein, I.W. (1974) *The Modern World–System*, New York: Academic Press.

White, D.R. and K.P. Reitz (1983) Graph and semi-group homomorphisms on networks of relations *Social Networks*, **5**, 193-234.

White, D.R. and M.L. Burton and M.M. Dow (1981) Sexual division of labor in African agriculture: A network autocorrelation analysis, *American Anthropologist* **83**, (4): 824-849.

Zellner, A. (1962) An efficient method for estimates of seemingly unrelated regressions, *Journal of the American Statistical Association*, **57**, 348-368.

## NOTES

1. More precisely, a specific model is treated in this fashion. Given autocorrelated data, it is very difficult to distinguish an effects model from a disturbances model *in the data*. The choice between the models should be made prior to the data analysis (Doreian, 1980).
2. Anselin's specification is broader than (2.1) with $\boldsymbol{\nu} \sim N(0, \Omega)$ where $\Omega$ is diagonal and heteroskedastic.
3. Brandsma and Ketellapper (1979) contains the parallel treatment for two regimes of disturbance autocorrelation.
4. Attaining the goal of delineating world system positions on the sole basis of the ties among nations, is a major methodological advance. However, it does not remove the inherent ambiguity of the "verbal" classifications: the boundaries between positions retain fuzziness. CONCOR is a splitting clustering procedure where the analyst can stop splitting clusters at any point. The issue of fuzzy boundaries becomes more complicated when other structural equivalence algorithms are used. Another popular algorithm is STRUCTURE (Burt, 1989) where the distance between positions is measured as the Euclidean distance between vectors made up of a (sending) row and a (receiving) column. These distances are clustered to delineate positions. The two methods can, and frequently do, provide distinct partitions. It is an open problem as to what network properties either algorithm is responsive (Doreian, 1988).
5. Even now, a model with 4 regimes of network effects is very impractical, if not impossible.
6. Anselin (1988, pp. 174-5) shows that when spatial units are omitted from spatially autocorrelated models, the impact of the excluded areas is not confined to the boundary areas. For a cohesion model, however, this is less consequential.
7. Burt (1976) takes this one step further by defining types of positions–primary, broken, sycophant, *etc.*–and measuring the extent to which each node (not block) occupies a type of position. These variables are used in LISREL to build models of network phenomena. Forgotten in such analyses is the notion of network autocorrelation which may compromise the whole use of LISREL. However, something very important in Burt's approach is the use of confirmatory factor analysis to check that the nodes put into a common block/position are indicators of that position.
8. This is not necessarily a mis–specification of $\mathbf{W}$, as the analyst may correctly specify $\mathbf{W}$ in terms of, say, structural equivalence. When $\mathbf{W}$ is constructed, however, the mis–assignment of nodes corrupts it.

# DISCUSSION

## "Network autocorrelation models: problems and prospects"

## by Patrick Doreian

Network autocorrelation models are a fusion of two methodologies from associated fields. Network models are sociological tools designed to categorize social interaction along prescribed pathways among social forces. As models of social process, they have given rise to much insightful analysis and interpretation. Spatial autocorrelation models are constructs designed to describe statistical interdependence among geographic neighbors. As geographical constructs, they have been extremely useful in describing and explaining spatial structural dependence (*e. g.*, Cliff and Ord, 1981). Patrick Doreian discusses the fusion of these two concepts into network autocorrelation models, tools that can be used to study interactions along social networks, accommodating the interdependence of network nodes. Modeling these interdependencies, he argues, will improve the accuracy and reliability of analytic network models. In essence, it will make the models more accurate and realistic.

Doreian's approach is innovative in that it goes beyond a simple application of geographic methodology to sociological problems. He adapts the method to the specifics of his problem, using the construct of geographic structure (or nodal links) to constrain the sociological models of interaction. This fusion should lead to models that are more representative of the true processes under study. Rather than being constrained by the geographic model, Doreian has modified the concept of network models to better fit current views of social interaction.

In biology and medicine, similar growth through cross–disciplinary fertilization with quantitative geography is possible. Not only can the concept and models of nonindependence be used, but the explicit use of geographic information adds new insight. I now will draw examples from evolutionary theory, ecology and epidemiology to illustrate this point.

Much of the theory of evolutionary biology is based upon the assessment of the genetic structure of populations. This structure is determined by the countervailing forces of natural selection, reproductive recombination (genetic drift) and mutation that are mediated by demographic and environmental influences. In essence, one looks to see which organisms are most similar and which are not, and then one tries to explain these differences as occurring through chance variation (recombination and mutation) or some selective force. Chance mutation does not lend itself well to geographic modeling (at the population scale) as it is thought to occur randomly through the genome (for a contrary view, one should see recent work by Cairns, Overbaugh, and Miller, 1988). Both recombination and selection, however, can have strong geographic components, and the study of the resulting geographic patterns has led to much insightful evolutionary analysis (*e. g.*, Endler, 1977). More recently, evolutionary biologists have begun to model the spatial pattern of the environment, and it is in this realm that I anticipate the most important advances will be gained by using geographic models in conjunction with evolutionary processes.

A second illustration stems from ecology, where succession is one of the principal paradigms of ecological thought. Plant communities vary over space and time in response to changing environmental conditions, and succession is the pathway over which these changes occur. For a long time succession was thought to be a unidirectional, temporal progression

from well dispersed, rapid growth, short lived, ephemeral species (*e. g.*, weeds) to poorly dispersed, slow growth, long lived species (*e. g.*, oak, pine and redwood trees). Then ecologists noticed that succession was not unidirectional, but rather, depending upon local conditions and disturbance rates, could head in a variety of directions. More recently, ecologists have hypothesized that succession can occur spatially as well, and that it is simply the response of plants to a changing environment. And yet, few models have been built that adequately capture plant succession. Those that do succeed incorporate some component of geographic structure.

In trophic ecology, food webs have been used as schematic depictions for community structure. Dating back to the early days of ecology (*e. g.*, Lotka, 1925; Elton, 1927; Lindemann, 1942; Odum, 1969), food web diagrams have been used to show species–species interactions, predator–prey relationships, energy flow along trophic pathways, and other aspects of community structure and function. As qualitative tools, interactions or flows between nodes (*e. g.*, species) are shown as connected lines, while lack of direct interaction or flow is denoted by the absence of a path. Sometimes quantitative estimates of flows along links are provided to show the strength of flow, where these estimates represent broad–scale averages over time. Recently, quantitative interest in the size, structure and complexity of food webs has arisen (*e. g.*, Cohen, 1978; Cohen, Bariand and Newman, in press; DeAngelis, Post and Sugihara, 1983; Pimm, 1982). By comparing length and size (number of nodes) and structure of the foods across habitats, communities and biomes, researchers have drawn inferences about theoretical ecology regarding species interactions within these groupings.

To date, most studies of food webs have concentrated on the binary connection matrices describing species interactions that are similar to Doreian's social networks. While field research has documented the existence of these links, few quantitative evaluations have been undertaken. One intriguing approach for investigating the functioning of food webs would be to statistically model environmental and food web dynamics. For example, for a terrestrial system, one could monitor population densities, and light, temperature, moisture and nutrient levels over time. Using the food web model for this system (with 1s on the diagonal), one could fit Doreian's network autocorrelation model [his equation (2.1)] to the data. This would fit parameters to the food web links that would be useful for descriptive purposes, as well as allowing for perturbation analyses to be undertaken. At the current time only binary networks (*i. e.*, qualitative models) and purely theoretical, quantitative models have been explored. By adding nodal interdependence to derive quantitative, data–based models, it is likely that ecologists can achieve a more fine–scaled resolution to species relationships.

The final example comes from environmental epidemiology, which is a rapidly growing field of investigation. As the development of synthetic chemicals expanded after World War II, and the public's awareness of the ubiquity and potential hazard of these substances has grown, health scientists are assuming an increasing role in their characterization and study. Epidemiologists who study patterns of disease have been confronted with a new paradigm of disease causation, and slowly are adapting their methodology to address these issues.

The standard model of infectious disease causation is that an infectious agent (or vector) is a source of exposure and risk. Once inside the host organism, the infection is a biological entity that grows. The infectious agent (or vector) has a period and strength of infectivity and those coming into contact with a carrier of the agent may develop the disease. Their

probability of illness is mediated by the length of exposure, the route of exposure, the strength of the infectious agent, the individual's own susceptibility, and other unknown risk factors. Generally, this model is simplified. The strength of the agent is a function of the disease that is chosen for study, and hence held constant within a study. Each route of exposure is considered important or not, and accordingly included or not included in the analysis. But, its importance is not scaled. And the unknown risk factors are considered to be distributed randomly through the population and of no predictive importance for the group under study. Therefore, disease incidence and prevalence models only are based upon the probability of contact via 'risky pathways' and the length of exposure. To make projections about disease spread and the probability of an epidemic under these models, one must develop only a history of contact among the individuals or population under consideration, and the agent. However, the strength, distribution and magnitude of the agent may be affected by exogenous factors, such as weather or food source. These factors often are omitted for infectious disease models because they are unknown, or their range of variations is thought to be sufficiently small as to not affect the model markedly. Further, the ability to characterize the variation of these factors is limited. Additionally, infectious diseases tend to be acute and have short latencies or induction periods (AIDS being a notable exception). This temporal compression facilitates their study.

One observation is that the study of infectious diseases would benefit from models analogous to Doreian's network models. Contact models are binary connection matrices that fail to accommodate other parameters of infectivity. By assessing infectivity, one could derive estimates for many of these parameters that would increase our knowledge of the disease process. While some such models exits, most ignore this approach (and few model the interaction).

In environmental epidemiology, the nature of exposure and disease is more complicated. As with infectious disease, probability of incidence also is based upon length of exposure, route of exposure, strength of agent, susceptibility, and other risk factors. However, in this case the strength of the agent is not constant, but varies by the type of agent, its exposure pathway and its concentration. Environmental epidemiologists often model exposure as a function of proximity to a source of pollution and the frequency with which one encounters the source, as well as the strength (or concentration) of the source. Most sources of exposure are geographically coherent in space. For example, they may be plumes downwind or downstream from a source, or parts of a community drawing water from the same source. We can study the disease process by comparing spatial patterns of the exposure agent to those of disease. Our recent work in disease cluster investigation has found that models consistent with environmental exposures may give different results than models consistent with infectious exposures (Wartenberg and Greenberg in press).

In summary, the utilization of geographic information in biology and medicine may lead to enhanced analysis of problematic situations. Most models fail to accommodate spatial (or dependency) structure, and thus obscure a certain level of resolution. By taking advantage of this information and using models of the sort Doreian proposes, investigators in these fields stand to gain substantially.

## References

Cairns, J., Overbaugh, J., Miller, S. (1988) The origin of mutants. *Nature*, **335**, 142-145.

Cliff, A., and J. Ord. (1981) *Spatial Processes: Models and Applications*. London: Pion.

Cohen, J. (1978) *Food Webs and Niche Space*. Princeton, NJ: Princeton University Press.

Cohen, J., F. Bariand, and C. Newman. *Community Food Webs: Data and Theory*. New York: Springer–Verlag (in press).

DeAngelis, D., W. Post, and G. Sugihara (eds.). (1983) *Current Trends in Food Web Theory* (ORNL–5983). Oak Ridge, TN: Oak Ridge National Laboratory.

Elton, C. (1927) *Animal Ecology*. New York: MacMillan.

Endler, J. (1977) *Geographic Variation, Speciation, and Clines*. Princeton, NJ: Princeton University Press.

Lindemann, R. (1942) *The trophic–dynamic aspect of ecology*. Ecology, **23**, 399-418.

Lotka, A. (1925) *Elements of Physical Biology*. Baltimore: Williams and Wilkins.

Odum, E. (1969) The strategy of ecosystem development. *Science*, **164**, 262-270.

Pimm, S. (1982) *Food Webs*. London: Chapman and Hall.

Wartenberg, D., and M. Greenberg. (1989) Methodological problems in investigating disease clusters. Unpublished manuscript submitted to *Archives of Environmental Health*.

*Daniel Wartenberg*, Robert Wood Johnson Medical School

## Index