SPATIAL STATISTICS AND MODELS

Edited by

GARY L. GAILE

and

CORT J. WILLMOTT

SPATIAL STATISTICS AND MODELS

Edited by

GARY L. GAILE

Department of Geography, University of Connecticut, Storrs, Connecticut U S A

and

CORT J. WILLMOTT

Department of Geography, University of Delaware, Newark Delaware, U S A

Library of Congress Cataloging in Publication Data

Main entry under title:

Spatial statistics and models.

(Theory and decision library ; v. 40)
Includes bibliographies and index.
1. Geography—Statistical methods—Addresses, essays, lectures.
2. Spatial analysis (Statistics)—Addresses, essays, lectures. I. Gaile, Gary L. II. Willmott, Cort J., 1946– III. Series.
G70.3.S6 1984 910'.01'51 83-22984
ISBN 90-277-1618-8

Published by D. Reidel Publishing Company,
P.O. Box 17, 3300 AA Dordrecht, Holland

Sold and distributed in the U.S.A. and Canada
by Kluwer Boston Academic Publishers,
190 Old Derby Street, Hingham, MA 02043, U.S.A.

In all other countries, sold and distributed
by Kluwer Academic Publishers Group,
P.O. Box 322, 3300 AH Dordrecht, Holland

Printed in The Netherlands

GARY L. GAILE AND CORT J. WILLMOTT

INTRODUCTION

The quantitative revolution in geography has passed. The spirited debates of the past decades have in one sense, been resolved by the inclusion of quantitative techniques into the typical geographer's set of methodological tools. A new decade is upon us.

Throughout the quantitative revolution, geographers ransacked related disciplines and mathematics in order to find tools which might be applicable to problems of a spatial nature. The early success of Berry and Marble's *Spatial Analysis* and Garrison and Marble's volumes on *Quantitative Geography* is testimony to their accomplished search. New developments often depend heavily on borrowed ideas. It is only after these developments have been established that the necessary groundwork for true innovation obtains.

In the last decade, geographers significantly augmented their methodological base by developing quantitative techniques which are specifically directed towards analysis of explicitly spatial problems. It should be pointed out, however, that the explicit incorporation of space into quantitative techniques has not been the sole domain of geographers. Mathematicians, geologists, meteorologists, economists, and regional scientists have shared the geographer's interest in the spatial component of their analytical tools.

This volume is a sampling of state-of-the-art papers on topics biased towards those dealing directly or indirectly with spatial phenomena or processes. The substantive interests of the contributors are highly diverse, e.g., geology (Agterberg, Krumbein and Schuenemeyer), geomorphology (Jarvis and Mark), climatology (Balling, Rayner and Willmott), cartography (Moellering and Shepard), human geography (Gould), population and migration (Austin, Davies, Huff, Miron and Pickles), urban geography (White), economic geography (Gaile, Green, Griffith, Sheppard and Semple) and, of course, methodology (Bennett and Goodchild). Nonetheless, the authors are united in their interest in the description and explanation of spatial structure and processes through the use, development and evaluation of statistical, numerical and analytic methods. For purposes of presentation, these approaches have been segregated into predominantly process-based quantitative descriptions, i.e., "models", and predominantly general or mathematically-based

Gary L. Gaile and Cort J. Willmott (eds.), Spatial Statistics and Models, ix–x.

descriptions, i.e., "statistics". The latter category most often is derived from probability theory. Models and statistics are, by no means, mutually exclusive areas of inquiry, however.

Several main themes can be found in this volume. The vast majority of papers, for example, discuss fitting equations to data. Even though regression analyses and trend surface techniques have been around for some time, important methodological problems still have not been fully solved, e.g., estimating parameters in gravity models (Sheppard) and in linear regression (Mark) to name but a few. Simulation modeling (Rayner and White) is developing into an important spatial tool, although the methods are not as well-known or refined as traditional statistical methods. Spatial autocorrelation is a "hot topic" and it is thought to confound many spatial statistics. Miron's paper and others address this topic which until now has only been surveyed in widely scattered readings. The computer-assisted presentation of spatial data in two-space is a recurring theme — cartography in the computer age. Several of the authors herein are particularly concerned with better ways to conceptualize and operationalize this process. Classification or regionalization is a traditional and important spatial theme which, by no means, is nearly fully developed. Only recently have the methods, data and facilities (computers) become sophisticated enough to undertake these problems. Most work in this area remains to be done.

Any field of inquiry is in constant need of reanalysis and criticism. It was with this in mind that Peter Gould was invited to contribute a general commentary on the theoretical underpinnings of quantitative geography. Gould's early call for a paradigm shift in human geography provides some thoughts on the directions and challenges of the next decade.

ACKNOWLEDGEMENTS

The editors gratefully acknowledge Susan Clarke, Benito Cribari, Jeff Gaile, Dean Hanink, Clint Rowe, Junior Wells and Pat and Abby Willmott for being great people.

TABLE OF CONTENTS

PART I

SPATIAL STATISTICS

DANIEL A. GRIFFITH

THEORY OF SPATIAL STATISTICS

1. INTRODUCTION

Classical statistics is based upon sampling theory. This theory involves articulations of the concepts of statistical population, sample, sample space and probability. Meanwhile, spatial statistics is concerned with the application of sampling theory to geographic situations. It involves a translation of these four notions into a geographic context. The primary objective of this paper is to discuss these translations.

Classical statistics also deals with measurements that provide useful information about a statistical population and the convolution of these measures with sampling theory. The result is a sampling distribution from which probabilities can be derived. A second objective of this essay is to review measures that furnish useful information about statistical geographic populations.

2. DEFINING A GEOGRAPHIC POPULATION

A population is the total set of items for which a measurement is to be obtained. This set must be defined in such a way that one can clearly specify whether or not any item is a member of it. For example, all people who reside in the United States is a population that the United States Census Bureau seeks to study. A statistical population is the set of measures corresponding to a population. If a population has N items, and some attribute is measured then the statistical population will include N measures, one for each item in the parent population. For instance, the income of each person who resides in the United States is a statistical population corresponding to the foregoing parent population. Once a locational context is attached to measures the total set constitutes a statistical geographic population. Popular examples here include average income for each county in the United States or population density for each census tract in a metropolitan area.

A number of less familiar geographic populations are infrequently or often unknowingly used, too. First, consider a planar surface that is partitioned into four mutually exclusive and collectively exhaustive areal units, such as that

Gary L. Gaile and Cort J. Willmott (eds.), Spatial Statistics and Models, 3–15.

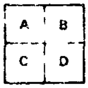

Fig. 1. One partitioning of a planar surface into four mutually exclusive and collectively exhaustive areal units.

appearing in Figure 1. The spatial distribution of values for some phenomenon over this surface constitutes a statistical geographic population. Another type of population emerges from the manner in which this planar surface was partitioned into areal units. A third type alludes to the idea of random noise versus systematic pattern in a geographic landscape.

This last category of geographic population requires additional explanation, whereas the nature of the other three populations should be self-evident. Suppose one is driving away from an urban area into a desolate region with his radio tuned into some radio station. This radio receives distinct signals in the urban area, while these signals become increasingly weaker as distance from the urban area increases. These signals represent a discernable, understandable pattern. Meanwhile, since the radio remains capable of receiving signals, it continues to pick up sound waves. But, these sound waves are from many sources, none of which is dominant, and when they mix together they are received as static, or noise. Analogously, social, economic, physical or other forces induce a given spatial distribution over a planar surface. Other factors, however, operate independent of one another, and mix with these aforementioned forces, introducing noise that yields a modified resultant distribution. Hence, the geographic population is, in reference to either this random noise component or this systematic pattern component, based upon a spatial distribution.

3. SPATIAL SAMPLING PERSPECTIVES

The first spatial sampling perspective is a classical sampling approach where replacement occurs and order is important. As such there are N^n possible samples, where n denotes the sample size. For example, if one wanted to consider all samples of size 2 for the geographic configuration appearing in Figure 1, then the areal unit combinations would be:

A,A	*B,A*	*C,A*	*D,A*
A,B	*B,B*	*C,B*	*D,B*
A,C	*B,C*	*C,C*	*D,C*
A,D	*B,D*	*C,D*	*D,D*

These sixteen pairs of areal units constitute the sample space in this case. Probabilities and the sampling distribution of measures here are described by the Central Limit Theorem and the Law of Large Numbers. This perspective is rarely used because most regions are partitioned into relatively few areal units. Urban geographers who want to say something about spatial distributions within, say, the United States urban system have employed it with census tracts as areal units, though. Their parent population is all census tracts in United States SMSAs.

The second spatial sampling perspective is similar to the classical sampling approach having no replacement and order being important. Accordingly, there are $n!/\Pi_{i=1}^{i=k}(n_i!)$ possible spatial distributions, where k denotes the number of distinct values. For instance, consider the set of values $\{0, 2, 3, 3\}$. The possible number of spatial distributions that could be constructed by convoluting this set with the geographic configuration given in Figure 1 is $4!/[(1!)\ (1!)\ (2!)] = 12$. These twelve possible arrangements, consitituting the sample space for this case, are presented in Figure 2. The sampling distribution here for any traditional statistic is a spike, as can be seen from the reporting of arithmetic means in Figure 2. In other words, since most

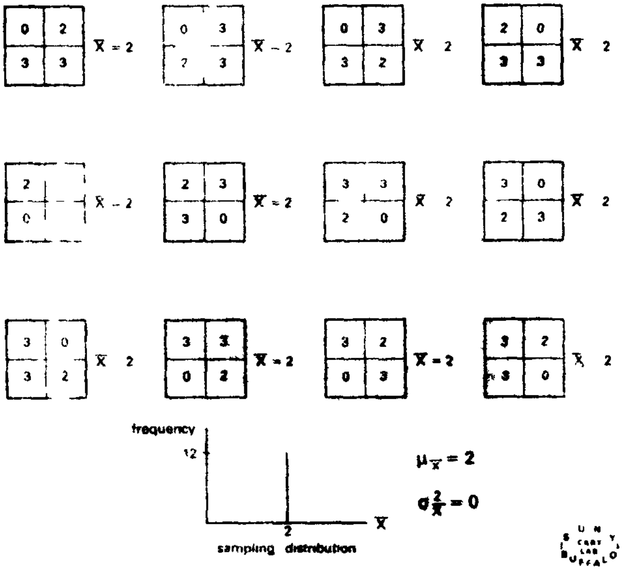

Fig. 2. All possible spatial distributions, based upon the configuration in Figure 1 and the set of values $\{0, 2, 3, 3\}$.

traditional statistics involve summation, and addition is commutative, then these statistics fail to provide any information about the spatial distribution in question. This shortcoming has led to the development of geostatistics and statistical theory based upon spatial autocorrelation. This perspective is used primarily in the latter of these two situations.

The third spatial sampling perspective refers to the manner in which a planar surface is partitioned into areal units. It is affiliated with the classical sampling approach having no replacement and order being unimportant. Consider the aforementioned set of values to refer to total population counts in areal units. Then, the total population for the region is $0 + 2 + 3 + 3 = 8$. Here, the pertinent question asks how many ways these eight people could be grouped into four areal units. The answer for this specific example is given in Table I. In general the number of possible groupings is given by

$$\sum_{i=1}^{i=k} \prod_{j=1}^{j=m} C(a_{ij}, b_{ij})$$

subject to $\Sigma_{j=1}^{j=m} b_{ij} = n$ and $a_{i1} = n$ where k is the number of possible compositions, and m is the number of combination terms for each composition.

TABLE I
The number of ways eight items can be allocated to four areal units

Composition	Combination		Spatial distributions
8, 0, 0, 0	C(8, 8)	= 1	4
7, 1, 0, 0	C(8, 7) C(1, 1)	= 8	12
6, 2, 0, 0	C(8, 6) C(2, 2)	= 28	12
6, 1, 1, 0	C(8, 6) C(2, 1) C(1, 1)	= 56	24
5, 3, 0, 0	C(8, 5) C(3, 3)	= 56	12
5, 2, 1, 0	C(8, 5) C(3, 2) C(1, 1)	= 168	24
5, 1, 1, 1	C(8, 5) C(3, 1) C(2, 1) C(1, 1)	= 336	24
4, 4, 0, 0	C(8, 4) C(4, 4)	= 70	12
4, 3, 1, 0	C(8, 4) C(4, 3) C(1, 1)	= 280	24
4, 2, 2, 0	C(8, 4) C(4, 2) C(2, 2)	= 420	24
4, 2, 1, 1	C(8, 4) C(4, 2) C(2, 1) C(1, 1)	= 840	24
3 3, 2, 0	C(8, 3) C(5, 3) C(2, 2)	= 560	24
3, 3, 1, 1	C(8, 3) C(5, 3) C(2, 1) C(1, 1)	= 1120	24
3, 2, 2, 1	C(8, 3) C(5, 2) C(3, 2) C(1, 1)	= 1680	24
2, 2, 2, 2	C(8, 2) C(6, 2) C(4, 2) C(2, 2)	= 2520	24
Total		8143	***

Constructing the corresponding sampling distribution is a very tedious task, especially without the aid of a computer.

The fourth spatial sampling perspective to be outlined here stems from the classical stratified sampling approach. A sample space is constructed for each of p areal units where replacement occurs and order is important. For any areal unit i, then, there are $N_i^{n_i}$ possible samples of size n_i. The total number of spatial distributions in this case is $\Pi_{i=1}^{i=p} N_i^{n_i}$. One simple example for this perspective is presented in Figure 3. Because $N_i = 2$ and $n_i = 2\ \forall i$, there are 256 possible values. Since some values repeat, though, only 81 distinct spatial distributions exist. The probability of observing these distributions is not the same; in other words, they are not equally likely distributions. The resulting sampling distribution is not a spike, and is associated with the Central Limit Theorem for multivariate analysis. It is employed extensively in research based upon information collected on the "long form" the United States Bureau of the Census distributes for selected attributes, and then compiles and publishes at the census tract or block level.

The last major spatial sampling perspective shares a close affinity with stochastic processes theory. It views any value x_i in a geographic distribution as being the weighted average of neighboring areal unit values, plus a random or noise component. Moreover,

$$x_i = \rho \sum_{j=1}^{j=p} w_{ij} x_j + u_i \tag{1}$$

where ρ = the degree of similarity for juxtaposed areal unit values; $\Sigma_{j=1}^{j=p} w_{ij} = 1$, $w_{ij} \geq 0\ \forall_{i,j}$; x_j = the value for areal unit j; p = the number of areal units; and, u_i = a noise component associated with areal unit i. The parameter ρ in equation (1) is related to the notion of spatial autocorrelation, and determines the prominance of pattern in a spatial distribution. Using matrix notation equation (1) becomes $\mathbf{X} = \rho\mathbf{WX} + \mathbf{U}$. Hence, $\mathbf{U}$ is given by

$$\mathbf{X} = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{U} \tag{2}$$

where $\mathbf{I}$ is the identity matrix. Because ρ is unknown in most cases, spatial autocorrelation indices may be used to obtain $\hat{\rho}$, and then $\hat{\mathbf{U}}$. Measures calculated for all possible vectors $\hat{\mathbf{U}}$ compose the sampling distribution. This perspective is exemplified in Figure 3, which is the sample space of noise term spatial distributions corresponding to the sample space presented in Figure 2. The sampling distribution for $\bar{u}$ is not a spike, as can be seen from the reporting of these arithmetic means in Figure 4. In other words, although

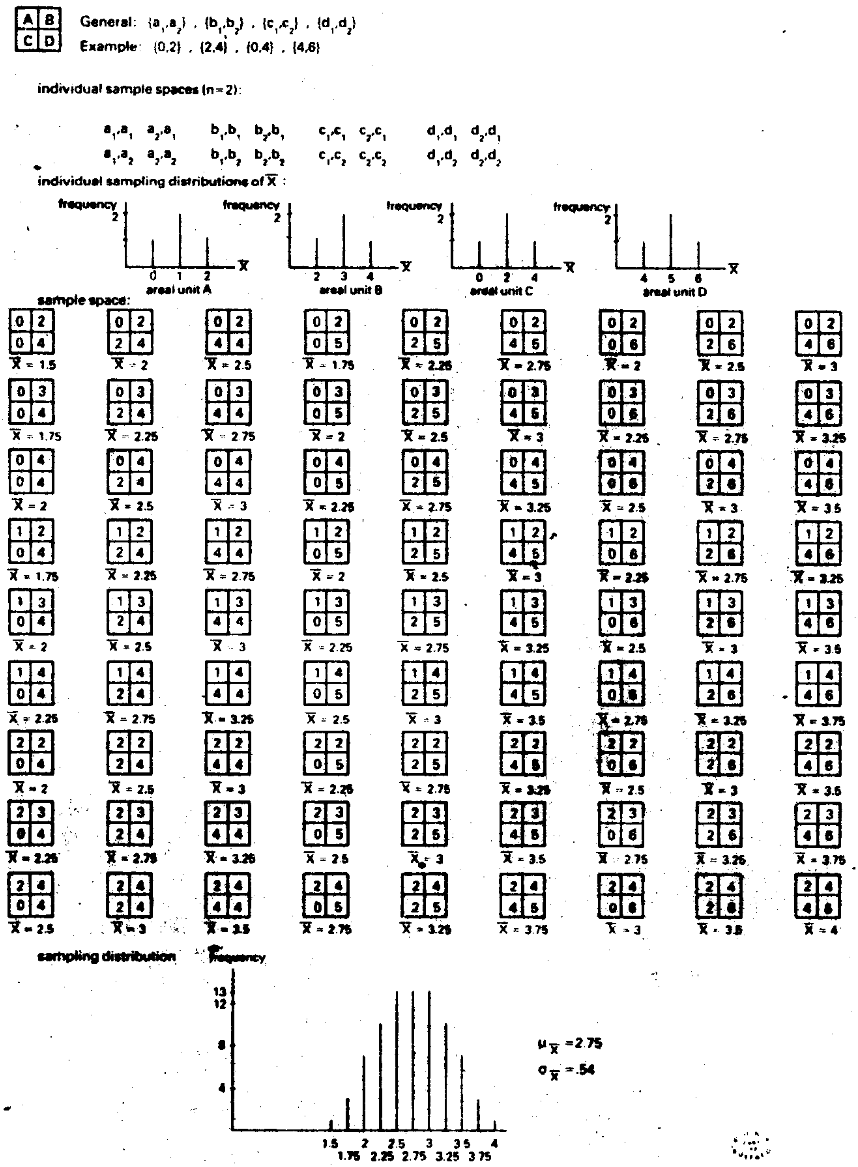

Fig. 3. **All possible spatial distributions for a stratified random sample example.**

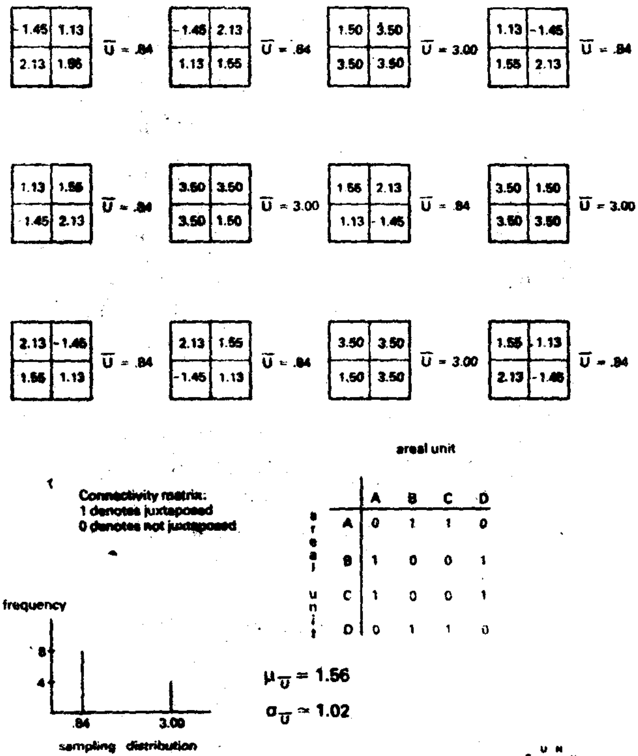

Fig. 4. The noise component for those spatial distributions appearing in Figure 2.

$\mu_{\bar{x}} \equiv \bar{x}$, the accompanying $\mu_{\bar{u}} \not\equiv \bar{u}$. This perspective has been utilized to compare statistics such as $\bar{x}$ and $\bar{u}$ in order to determine how important the spatial pattern component is.

4. SPATIAL AUTOCORRELATION

As the last sampling perspective indicates, classical statistics fail to provide any information about geographic distributions. Figure 2 exemplifies this problem. Regardless of which of the twelve possible spatial distributions is considered $\bar{x} = 2$, $s^2 = 2$, and so forth. Attempts to take the underlying configuration into account have led to the development of geostatistics, planar statistics, and directional statistics. Each of these three types will be subsequently discussed.

Another problem that became recognized is associated with the nature of information rendered by an application of classical statistics to geographic situations One of the fundamental assumptions of classical statistics is that the elements of some population take on numerical values in an independent fashion This assumption frequently is violated in spatial statistical situations. Moreover, the value of some phenomenon in a given areal unit tends to be related to those values of this phenomenon taken on by juxtaposed areal units. The last sampling perspective is the one basically employed here. That area of spatial statistics that has emerged is known as spatial autocorrelation. Literally speaking, the spatial term of this phrase refers to a geographical dependence structure for observations. The term correlation refers to a relationship between entities, and the prefix auto- refers to the fact that a single variable is being related to itself. The body of spatial statistics that is evolving through this conception characterizes a statistical distribution not only by its mean μ and its variance σ^2, but also by its geographic configuration **W** and its spatial autocorrelation ρ. This topic of spatial statistics is treated first because it has profound implications for geostatistics, planar statistics, and directional statistics. A more detailed discussion of it appears elsewhere in this volume (Miron, 1984).

The spatial autocorrelation viewpoint does not ignore randomness. Rather, it maintains that a geographic distribution is composed of both pattern and random error Consequently, x_i may be decomposed in accordance with equation (1), where ρ is the spatial autocorrelation parameter. Equation (2) uses the spatial linear operator $(\mathbf{I} - \rho\mathbf{W})^{-1}$, which is like an input/output multiplier in that it accounts for all direct and indirect effects flowing over a surface, transforming a random distribution into a partially patterned one. This transformation affects the type of information yielded by a statistical analysis.

Using the calculus of expectations, let $E(x_i) = \mu_x$ and $E(u_i) = \mu_u$. Now from equation (1)

$$E(x_i) = E(\rho \sum_{j=1}^{j=p} w_{ij}x_j + u_i)$$

$$\mu_x = \rho \sum_{j=1}^{j=p} w_{ij}\, E(x_j) + E(u_i) = \rho\mu_x + \mu_u$$

$$\mu_x - \rho\mu_x = \mu_u$$

$$\therefore \mu_x = (1 - \rho)^{-1}\, \mu_u \tag{3}$$