SPATIAL DATA CONFIGURATION IN STATISTICAL ANALYSIS OF REGIONAL ECONOMIC AND RELATED PROBLEMS

ADVANCED STUDIES IN THEORETICAL AND APPLIED ECONOMETRICS
VOLUME 14

For a complete list of volumes in this series see final page of this volume.

Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems

by

Giuseppe Arbia

Istituto di Statistica Economica, Faculty of Statistics,
University of Rome "La Sapienza", Italy
and Fitzwilliam College, Cambridge, U.K.

KLUWER ACADEMIC PUBLISHERS
DORDRECHT / BOSTON / LONDON

Library of Congress Cataloging in Publication Data

Arbia, Giuseppe.
Spatial data configuration in statistical analysis of regional economic and related problems / Giuseppe Arbia.
p. cm. -- (Advanced studies in theoretical and applied econometrics ; v. 14)
Includes index.
ISBN 0-7923-0284-2
1. Space in economics--Statistical methods. 2. Regional economics--Statistical methods. I. Title. II. Series.
HB137.A72 1989
330'.01'5195--dc20 89-8210

ISBN-13: 978-94-010-7578-7 e-ISBN-13: 978-94-009-2395-9
DOI: 10.1007/978-94-009-2395-9

Published by Kluwer Academic Publishers,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

Kluwer Academic Publishers incorporates
the publishing programmes of
D. Reidel, Martinus Nijhoff, Dr W. Junk and MTP Press.

Sold and distributed in the U.S.A. and Canada
by Kluwer Academic Publishers,
101 Philip Drive, Norwell, MA 02061, U.S.A.

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

To Paola and Elisa

TABLE OF CONTENTS

FOREWORD BY ROBERT J BENNETT

Spatial Statistics and econometrics has long been a *Cinderella* aspect among the statistical sciences. For various reasons it seems that either its problems were too complex, or research priorities lay elsewhere. However, with the emergence of the demands of economic planning, pattern recognition, remote sensing and artificial intelligence the requirements for a sound basis of econometric theory for handling spatial data has become essential. Much of this framework has now been established following key studies by Moran, Geary, Kendall, Whittle, Granger, Ord, Besag and Ripley in Statistics, and by such researchers as Curry, Tobler, Paelinck, Haggett, Cliff, Haining and Openshaw in the applied fields of econometrics and geography. This is the lineage to which this book contributes. Yet some of the problems in spatial statistics have stubbornly remained. Particularly important amongst these has been that of the modifiable areal unit problem and the influence of spatial data configuration in general on spatial autocorrelation and associated analytical and estimation techniques.

This book represents a major contribution to the solution of these issues. The so-called *modifiable areal unit* problem arises as the result of the influence on subsequent statistical procedures of variable scales of spatial zonal data, and variable aggregation of cases within different zones. The book presents for the first time an integrated theoretical presentation of the consequences of the interaction of the effects of scale and aggregation in spatial data and their influence on the statistical properties of estimation and significance testing. The book presents the theory, evaluates its properties using simulation techniques and then works through simple hypothetical as well as empirical examples. It represents the most fundamental contribution to the modifiable areal unit problem since the work of Gehlke and Biehl in 1934 and Kendall and Yule in 1950. Hence the book makes a contribution of fundamental importance to spatial statistics and it is particularly appropriate that it should appear in this series of Advanced Studies in Theoretical and Applied Econometrics. That the scholar who has produced this work is still a young researcher with much still to offer bodes well for the development of this field.

I was pleased to have been able to support the author as his advisor on his PhD at Cambridge University, from which the book derives. It was a pleasure to assist him. Giuseppe Arbia is to be congratulated on this, his first book; and it is a pleasure to commend the work to its readers.

Professor Robert J Bennett
London School of Economics

ACKNOWLEDGEMENTS

This book derives from my PhD thesis written during my stay in Cambridge in various period of times from 1985 to 1987. I am pleased to fulfil the pleasant duty of thanking the many people who have contributed in one way or another to the preparation of this volume.

Special thanks are due to Professor R.J.Bennett of London School of Economics without whose guidance and encouragement this book would not have been completed. His constructive criticism on various drafts of the manuscript enabled me to improve it substantially. During the period in which he assisted me as a PhD advisor he has been a constant example to follow and his influence is evident throughout the book.

The environment plays a crucial role in a scholar's work. For this reason I wish to thank the staff of the Department of Geography in Cambridge for providing me with a stimulating atmosphere. A special mention is due to Dr. A.D. Cliff.

Thanks are also due to Dr. R.P. Haining of Sheffield University and to Dr. G.A. Young of the Statistical Laboratory in Cambridge for various comments made in many lively discussion and on various drafts of the book, to Dr.C.Kenyon of the Department of Mathematics in Cambridge for assisting me in writing some of the Fortran programs, and to the anonymous referee for enriching the work by giving numerous helpful suggestions.

Of my colleagues at the University of Rome I express my deepest gratitude to Professor A.Erba for much encouragement during many years.

A human being is a unity and the scholar is not separate from the man. During my stay at Cambridge University I was lucky enough to meet a number of very good friends that made the life in Cambridge be worthwhile. I wish to thank especially Dr. J.E. Zucchi of McGill University of Montreal and Dr. A.L. Sawaya of Sañ Paulo University for constantly reminding me that life is much more than writing a book. The proofreading and the revision of the text has been demanding. I would like to thank Mrs. Tricia Goodwin-Williams for her assistance with this. While pursuing the research reported here I was generously supported by the British Council, by the N.A.T.O. and by the Italian National Research Council (C.N.R.) through scholarships and through the project "Economia". Their financial assistance is acknowledged here. The text processing of the book took place in Fitzwilliam College in Cambridge and in the Centro Interpartimentale di Calcolo Scientifico of Rome University. Thanks are due respectively to Dr. Pearl and to Prof. Schaerf for the access to the computer facilities.

I gratefully acknowledge the permission of the publisher Pion Limited for the reproduction of Figures 2.2.4 and 5.2.1 from Cliff A.D. and Ord J.K. (1981) and Figures 2.2.5 and 6.3.3 from Openshaw S. and Taylor P.J. (1979).

Finally I acknowledge the constant help and encouragement of my beloved wife Paola and of my daughter Elisa to both of whom this book is dedicated.

Giuseppe Arbia
University of Rome

All Saints, 1988

NOTATION SHEET

The following abbreviations and symbols are used

PDF	Probability distribution function
DF	Probability density function
JPDF	Joint probability distribution function
JDF	Joint probability density function
CJPDF	Conditional joint probability distribution function
CJDF	Conditional joint probability density function
JBPDF	Joint bivariate probability distribution function
JBDF	Joint bivariate probability density function
$\sim$N	Normally Distributed
$\sim$MVN	Multivariate Normally Distributed
$\int$	Unless differently specified implies $\int_{-\infty}^{\infty}$
Π	Unless differently specified implies $\Pi_{i=1}^{n}$
Σ	Unless differently specified implies $\sum_{i=1}^{n}$
$\Sigma_{(2)}$; $\Sigma_{(3)}$; $\Sigma_{(4)}$	$\Sigma_j \Sigma_{k\ j\neq k}$; $\Sigma_j \Sigma_k \Sigma_{l\ j\neq k\neq l}$; $\Sigma_j \Sigma_k \Sigma_l \Sigma_{p\ j\neq k\neq l\neq p}$
$O(n^{-1})$	Terms of order n^{-1}
G(i)	A group of sites indexed by i
G	Is a grouping matrix
g_{ij}	Is the typical element of **G** such that $g_{ij}=1$ if $j \in G(i)$ and zero otherwise
r_i	Is the cardinality of G(i)
N(j)	The set of neighbors of site j
W	Is the connectivity matrix of a spatial system
w_{jk}	Is the typical element of **W** such that $w_{jk}=1$ if $k \in N(j)$ and zero otherwise
v_j	Are the number of neighbors of site j (or nodality) $v_j = \Sigma_k w_{jk}$

A	Is the total number of joints ("connectivity") of a spatial system $A = \sum_{(2)} w_{ij} = \sum v_j$
v^*	Is the average connectivity $v^* = A\, n^{-1}$
A_i	Is the connectivity within group i $A_i = \sum_{k\,\varepsilon\, G(i)} \sum_{j\,\varepsilon G(i)} w_{jk}$
t_{il}	Is the connectivity between group i and group l
i,j,k,l,p	Integer used as summation indices
n	Integer indicating the number of sites at an *individual-process* level
m	Integer indicating the number of sites at a *group-process* level.
X,Y	Are vectors or matrices written boldface
X	Is a random variable at an *individual-process* level
X^*	Is a random variable at a *group-process* level

As to general mathematical statistical concepts, the terminology is chosen in accordance with Kendall and Stuart (1976). In particular:

-Upper case (e.g. X,Y) indicate stochastic processes, while lower case (e.g. x,y) indicate the realizations of the same processes.

-Greek letters indicate the stochastic process moments (e.g. μ,σ,γ,ρ), while Latin letters are used to indicate the corresponding sample moments (m,s,c,r)

For the moments of any order the general form is as follows:

${}_{XY}M_{ij}{}^{e}$

where M is a bivariate moment, X (and Y) are the variate(s) to which the moment refers, i(and j) are the observation(s) to which it refers (j is omitted in the case of univariate moments; when the process is stationary both i and j are omitted) and e is the exponent.

Everybody knows that geography is about maps

(Unwin, 1981; p.1)

The fundamental notion in statistical theory is that of group or aggregate

(Kendall & Stuart, 1976; p.1)

1. Introduction: spatial effects and the role of configuration of data

1.1. OBJECTIVES AND APPROACHES

Spatial data in regional economic analysis have two distinctive features. First of all they cannot be thought as randomly generated from the classical urn models; rather they are dependent in that "the value of, say, prosperity in one region gives statistical information about the likely value of adjacent areas "(Unwin and Hepple, 1974). This is usually referred to as the *spatial autocorrelation* problem (Cliff and Ord , 1981). Secondly they are constituted by aggregation of the characteristics of individuals within portions of space. The population in a country is the sum of the individuals living in that country, the total income of a region is the sum of the income of the population in that region: the per-capita consumption of an area is the mean of the individual consumption of that area, and so on. However the borders of the zones in which a study area is divided are not just *divinely given*; rather there is a very large number of different ways in which the individuals can be aggregated to form areal data. Analysis is often made more complex by the common situation faced in geographical investigation that the variable of interest is recorded on a system of irregular collecting areas. Thus the study area is divided into territorial units which are, in general, different in size and shape and which connect to one another in an arbitrary and irregular way. The problem of the arbitrariness and irregularity of spatial units will be referred to as the *modifiable areal unit problem* following Openshaw and Taylor (1979).
The interaction between these two intrinsic features of spatial data, between spatial autocorrelation and modifiable areal units, creates a number of problems if we seek to use the observed data to estimate statistical relationships. These problems have been recognised extensively in the literature and there is now a large number of empirical investigations of the effects of different spatial data configurations. Some of these will be reviewed in Chapter 2. However no theoretical explanation has yet been provided which allows prediction and control for the effects of spatial data configuration during geographical investigations.

The main aim of this book is to propose a statistical framework within which the spatial autocorrelation problem is taken into account, to explain, and therefore to control for, the effects of the modifiable areal unit problem. In addition a number of related problems will also be investigated. Furthermore the whole work is based on the conviction that a statistical approach to spatial series in regional economics has to be, at the same time, truly geographical incorporating *in some way* the information contained in the map on which the data are laid. This is what we call the *configuration of spatial data*. We use this term to describe the variety of relevant information which fully specify the geography of a particular situation: the links of neighbourhood between regions, their shape, their size, and their relative and absolute location in the study area.
The emphasis throughout the book is on theory which allows us to explain and suggests the ways to controlling for some spatial data effects. However, where possible, we will also propose solutions to reduce or to eliminate these effects.

1.2. AN OVERVIEW OF THEORETICAL PROBLEMS

A first manifestation of the modifiable areal unit problem is the **scale problem** or the problem of the level of resolution. Consider, for example, Figure 1.1.a which shows the map of the British counties and Figure 1.1.b which displays the map of British regions . If we are examining a single variable and we want to study, say, the spatial inequality of its distribution through its variance, our conclusions will depend on the scale we choose. For example the distribution of income can be close to a situation of equity at a regional level, but very unequal at a county level. Furthermore if we consider associations between two or more variables it is well known (Yule and Kendall, 1950) that the correlation coefficient changes with different scales of areal units. This applies either to direct correlation analysis, or to indirect uses with multivariate techniques based on correlation, such as factor analysis. As a matter of fact it is not unusual in the literature to find data aggregated to a level so that high correlation is demonstrated.
A second, closely related aspect of the modifiable areal unit problem is the **aggregation problem**. Consider, for example, Figures 1.2.a and 1.2.b. In both cases we are dealing with the map of Italy divided into 32 territorial units. However the borders between the units are different in the two cases. Again the value of statistical measures, like the variance and the correlation, changes greatly with changing zone boundaries.
A third analytical problem arises when aggregated data are the only source available while the object of the study are individual characteristics and relationships. This is often the case when dealing with Census data tracts or with electoral results. In Figures 1.3.a and 1.3.b are shown the percentage of votes of the Communist Party in the 1987 Italian political elections and, respectively, the percentage of population over 75 years in the 32 polling districts. Looking at Figures 1.3.a and 1.3.b one may conclude that the probability of voting Communist Party is higher for elderly people. However this conclusion is fallacious. In fact it has been shown in several studies (Robinson, 1950) that the correlation measured using areal data is never a substitute for individual correlation. This is usually referred to in the literature as the **ecological fallacy** problem.

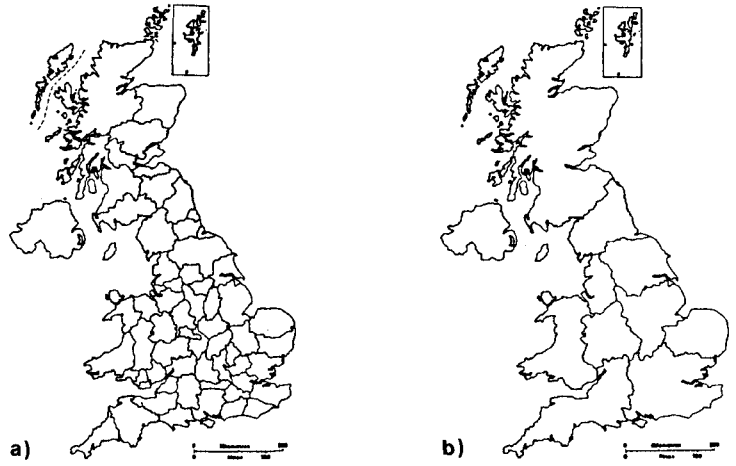

Figure 1.1: Map of Great Britain at two different scale levels:(a) Counties, (b)Regions.

Figure 1.2: Two alternative aggregations of the Italian provincie in 32 larger areas

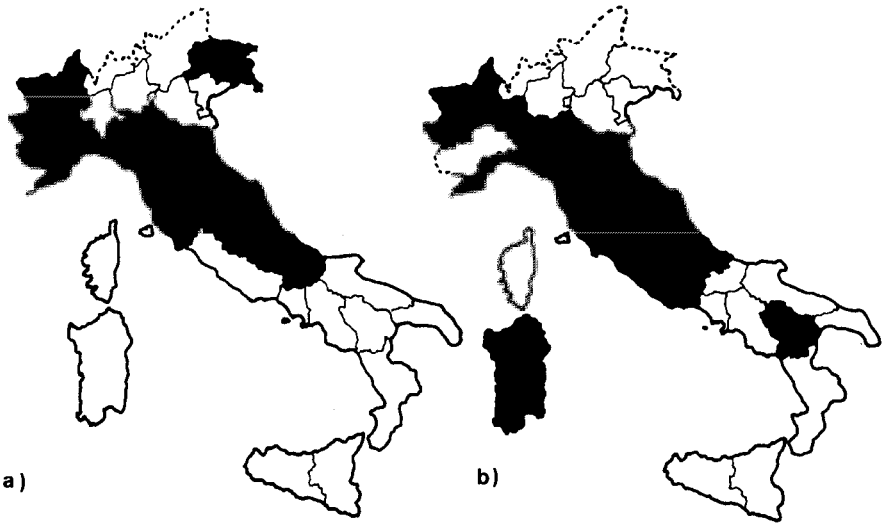

Figure 1.3 Percentage of votes of the Communist Party in the 1987 Italian political elections (a) and percentage of population over 75 years (b) in 1981 Italian Census in 32 polling districts. The polling districts with values above the average are shaded.

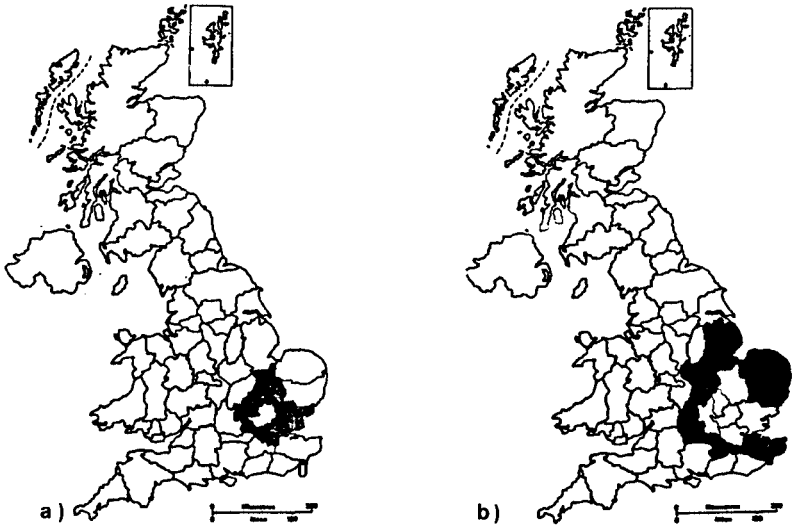

Figure 1.4: First order neighbours (a) and second order neighbours (b) of a reference area.

While there are several other problems relating to the analysis of areal data, the problem of estimating a **spatial correlogram** merits special attention. The concept of the correlogram has been borrowed in the spatial literature from the time series analysis. Figure 1.4.a shows the first-order neighbours of a reference area, while Figure 1.4.b displays the second-order neighbours of the same area. Higher-order neighbours can be defined in a similar fashion. While it is clear that the dependence is strongest between immediate neighbouring areas a certain degree of dependence may be present among higher-order neighbours. This has been shown to be an alternative way of looking at the scale problem (Cliff and Ord, 1981, p.123).
However, unlike the case of a time series where each observation depends only on past observations, here dependence extends in all directions. This fact poses new problems for the estimation of the correlogram.
We will show in the next chapters that the problem of estimating a spatial correlogram of a spatial series shares the same characteristic as the other problems mentioned in this section; therefore it can be treated within the same methodological framework.
In all the problems which we have briefly introduced here, a paramount role is played by the configuration of the spatial data i.e. by the fact that each unit on which a variable is recorded possesses its own unique size and shape and connects to other units in an irregular way. The investigation of the effects of data configuration within the context of spatial autocorrelation and modifiable areal units leads us to the need for an appropriate methodology. We introduce this below.

1.3. A SKETCH OF THE METHODOLOGY

Throughout the book we will assume that a good representation of reality can be obtained by assuming that a spatial series is a finite realization of a stochastic process in two dimensions. To make any progress with such an approach we are forced to accept as a starting point the hypothesis of normality and stationarity of the process itself. (Some theoretical results, however, can be applied to other kind of processes). There are a number of practical situations in which it is not unreasonable to assume that the process has a stationary multivariate normal distribution. When this is not the case we will assume that it is always possible to subdivide the areal units into sufficiently small units (ultimately the individual economic agents) where the assumption of stationarity becomes more plausible. The advantage of this approach is that we are able to study the effects on the moments of the generating process of any statistical manipulation of the original data.
For example, dealing with the scale problem, provided we know the distribution of the generating process at a county level and provided we know the way in which data are aggregated up to the regional level, we can specify the distribution of the generating process at a regional level. In doing so we are also able to exploit all the relevant information about the configuration of the spatial data.
In each chapter attention is given to explaining classical results found in the literature and to confirming these theoretical results by analysis of sets of real economic data. Furthermore, in some cases, artificial computer-generated data are also examined. For example the theoretical results obtained in Chapter 7 on the *ecological fallacy* problem, are tested by exploiting a microsimulation approach where the individual behaviour is simulated and the results aggregated up to a geographical level. In all the simulation studies new computer programs had to be written. The listing of these programs, written in FORTRAN-77 on the IBM-3081 system of Cambridge University Computer Centre, are contained in the appendices to the relevant chapters.

1.4. AN OUTLINE OF THE BOOK

The plan of the book is as follows. Before entering into the theoretical discussion, we first introduce in more detail the set of problems of concern. In Chapter 2 we review the results found in the literature on the effects of the spatial configuration of data on statistical analysis. In Chapter 3 we seek to clarify the concept of *spatial configuration of data* in order to introduce the essential concepts and notation used in the rest of the book.
Chapter 4 introduces the methodological basis for the theoretical contribution of the book. Here we introduce some basic notions of spatial stochastic processes and we derive the fundamental theoretical results used in later chapters.
Chapter 5 is concerned with the analysis of the problems arising in statistical analysis of **univariate** spatial series. In particular in sections 5.2 and 5.3 we analyse the scale problem in situations involving a single variable, while in section 5.4 the aggregation problem is attacked. The effects of the configuration of spatial data in **bivariate** statistical analysis of spatial series are discussed in Chapters 6 and 7. In Chapter 6 we consider again the scale and aggregation problems in correlation analysis of two spatial series, whereas in Chapter 7 we consider the *ecological fallacy* problem. Chapter 8 is devoted to discussing the problems of estimating the spatial correlogram in irregular lattices. Finally the appendices to the book contain some of the data used for our empirical studies (Appendices A.1 to A.4) and a review of the methods available for simulating two-dimensional random surfaces (Appendix A.5).

1.5. OMITTED TOPICS

It is clear that the aim of this book is not to give an exhaustive account of all the problems that occur when a statistical analysis is performed with data which are distributed over space. Consequently a number of topics are omitted from the discussion.
First of all, while the most common situation in regional economics is that of variables recorded in *irregular collecting areas*, there are situations in which the actual location of geographical entities is of interest. This is the case of microanalytical studies in the spatial economy such as the detection of housing patterns or the study of the location of industrial plants or of public facilities (Wilson and Bennett, 1985). The study of *point patterns* of this kind are not considered here, although some indication is given of the way in which the methodological framework developed in this book could be extended to deal with these situations.
Secondly, data are always considered for a single cross-section in time; consequently no attention is given to problems related to *spatial-time series* (Bennett, 1979).
Thirdly, the effects of spatial configuration of data in the analysis of more than one series is restricted to bivariate regression analysis while no consideration is given to *multivariate techniques* like factor analysis or principal component analysis (Streitberg, 1978).
Finally, other interesting topics like the *missing data* problem (Haining, Griffith and Bennett, 1984) or *spatial sampling* (Cochran, 1963; Ripley, 1981) are also omitted because of lack of space.

2. Theoretical problems motivation

2.1. INTRODUCTION

In this chapter we wish to discuss to a deeper extent the theoretical problems reviewed in Chapter 1. In particular in Section 2.2 the modifiable areal unit problem will be analysed in its two manifestations of *scale* and *aggregation*. Section 2.3 contains a review of the literature on the e*cological fallacy*. Finally the problem of estimating a *spatial correlogram* will be discussed in Section 2.4.

2.2. THE MODIFIABLE AREAL UNIT PROBLEM

2.2.1. The nature of the modifiable areal unit problem: modifiable versus unmodifiable units

Yule and Kendall (1950) lucidly introduce a fundamental distinction between two different kinds of data to which a statistical analysis may be applied.
It sometimes happens that we have to deal with statistical units which cannot be further decomposed into smaller units. Consider, for example, the case in which we measure the income or the level of consumption of a single economic agent. The ultimate unit of analysis is the individual whose income and consumption is a "unique non-modifiable numerical measurement" (Yule and Kendall, 1950). It is not conceivable to divide the individual economic agent into smaller units. The same thing is true if we consider, for example, the age of an individual, or the level of production of a single firm or the price of a commodity. This kind of data are related to units which are intrinsically *non-modifiable* .
The same is not true when dealing with spatial units which are, in contrast, *modifiable*. "Since it is impossible, or at any rate impracticable, to grow wheat and potatoes on the same piece of ground simultaneously we must, to give our investigation a meaning, consider an area containing both wheat and potatoes; and this area is modifiable at choice"(op.cit.; p.312). Similar examples may be found in economics where, for example, a regional unemployment rate, the percentage of commuters, the total male population, the votes cast for a political candidate, can only be referred to a modifiable geographical unit. The relevance of this distinction is that the value of any statistical measure "will, in general, depend on the unit chosen if that unit is modifiable" (Yule and Kendall, 1950).
Openshaw and Taylor (1979) distinguished two aspects of the modifiable areal unit problem in a geographical context. The first aspect relates to the different results one may get in statistical analysis with the same set of data grouped at different scale levels (e.g. counties or regions) and it is therefore referred to as the *scale problem*; the second considers the variability of results not due to variations in the size of the areas,

but rather to the shape of them. This is therefore referred to as the *aggregation problem*. We will deal with the two aspects separately in the next two sections.

2.2.2. The scale problem

The scale problem in geographical studies arise for different reasons. First of all, human society is organized in territorial units usually arranged into nested hierarchies like town, counties, regions and states; it is therefore common to have to cope with data aggregated in terms of such spatial units. In these cases the problem arises of choosing the level of resolution that describes best the phenomenon under study (Moellering and Tobler, 1972). A second situation arises from the procedure in geographical studies of superimposing on a map a regular grid of contiguous quadrats in order to investigate the pattern of a particular phenomenon (Greig-Smith, 1952). This procedure is most commonly used in plant ecology, but there are examples also in human geography (see for example, Cliff and Ord,1981; p.133).
In both the regular and the irregular case the scale problem can be summarized by saying that "generalizations made at one level do not necessarily hold at another level, and that conclusions we derived at one scale may be invalid at another" (Haggett, 1965).

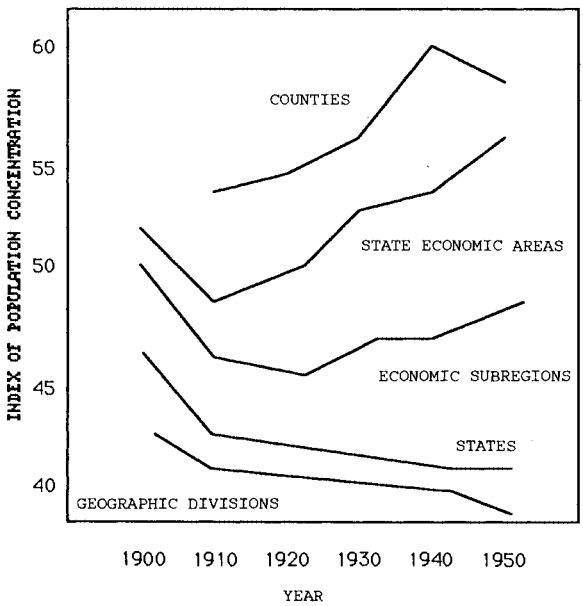

Figure 2.2.1: Index of population concentration in U.S. from 1900 to 1950. Source Duncan et al. (1961).

To give an idea of how serious the problem of scale is, let us start with an example. Figure 2.2.1 shows the index of population concentration, for different areal subdivisions of the United States through time. The figure (taken from Duncan et al. 1961, p.86) shows that the population concentration has increased over the 50 years considered, if we measure it at a *county* level or at a *state economic area* level. In contrast the index remains approximately constant at the level of *economic subregion*. Finally the population concentration decreases if we measure it at a *state* level or at the level of *geographic divisions*. This example shows that because of the modifiable areal unit problem we are not even able to give an answer in a conclusive way to the simple question: "what happened to the population concentration in the U.S. in the period considered?"

The problem has been recognised for a long time and researchers have concentrated on its effects on statistical measures, particularly on the correlation coefficient and on the variance. In a short paper Gehlke and Biehl (1934) studied the effects on the correlation coefficient between male juvenile delinquency and median monthly income in 252 census tracts in Cleveland. The correlation coefficient was first computed for the 252 census tracts and then for the same data grouped successively into fewer and larger units with the contiguity constraint. The results are shown in Table 2.1 for the raw data and for ratios. It is useful to remark, however, that due to the loss of degrees of freedom, the estimates of the correlation coefficient become less reliable as sample size decreases.

The main conclusion is that, in both cases, the correlation coefficient increases monotonically in absolute value. When census tracts are grouped at random rather than by contiguity, the two authors obtained the results displayed in Table 2.2 .

In this case no systematic scale effect occurs. The result is interesting and suggests that the dependance between areas can be at the origin of the problem: dependence plays a role when we group contiguous areas; in contrast, it is eliminated when we group areas at random.

number of units	Correlation coefficient	
	Raw data	Ratios
252	-0.502	-0.516
200	-0.569	-0.504
175	-0.580	-0.480
150	-0.606	-0.475
125	-0.662	-0.563
100	-0.667	-0.524
50	-0.685	-0.579
25	-0.763	-0.621

Table 2.1 : Correlation coefficients at different scale levels where data are grouped with the contiguity constraint, for juvenile delinquency and monthly income in Cleveland , USA. Source Gehlke and Biehl (1934).