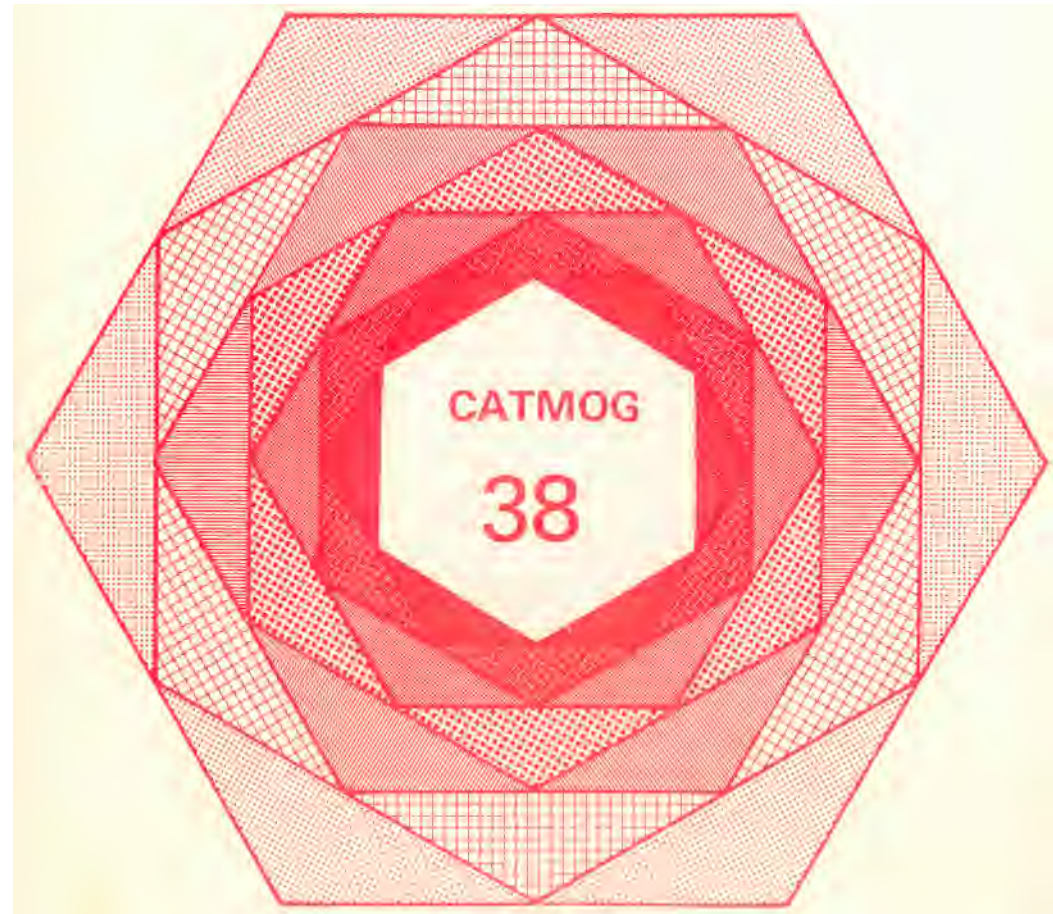
# THE MODIFIABLE AREAL UNIT PROBLEM

S. Openshaw




ISSN 0306-6142
ISBN 0 86094 134 5

© S. Openshaw

Published by Geo Books. Norwich—Printed by Headley Brothers Ltd. Kent

**CATMOG - Concepts and Techniques in Modern Geography**

CATMOG has been created to fill in a teaching need in the field of quantitative methods in undergraduate geography courses. These texts are admirable guides for teachers, yet cheap enough for student purchase as the basis of classwork. Each book is written by an author currently working with the technique or concept he describes.

1. Introduction to Markov chain analysis - L. Collins
2. Distance decay in spatial interactions - P.J. Taylor
3. Understanding canonical correlation analysis - D. Clark
4. Some theoretical and applied aspects of spatial interaction shopping models - S. Openshaw
5. An introduction to trend surface analysis - D. Unwin
6. Classification in geography - R.J. Johnston
7. An introduction to factor analysis - J.B. Goddard & A. Kirby
8. Principal components analysis - S. Daultrey
9. Causal inferences from dichotomous variables - N. Davidson
10. Introduction to the use of logit models in geography - N. Wrigley
11. Linear programming: elementary geographical applications of the transportation problem - A. Hay
12. An introduction to quadrat analysis (2nd edition) - R.W. Thomas
13. An introduction to time-geography - N.J. Thrift
14. An introduction to graph theoretical methods in geography - K.J. Tinkler
15. Linear regression in geography - R. Ferguson
16. Probability surface mapping. An introduction with examples and FORTRAN programs - N. Wrigley
17. Sampling methods for geographical research - C.J. Dixon & B. Leach
18. Questionnaires and interviews in geographical research - C.J. Dixon & B. Leach
19. Analysis of frequency distributions - V. Gardiner & G. Gardiner
20. Analysis of covariance and comparison of regression lines - J. Silk
21. An introduction to the use of simultaneous-equation regression analysis in geography - D. Todd
22. Transfer function modelling: relationship between time series variables - Pong-wai Lai
23. Stochastic processes in one dimensional series: an introduction - K.S. Richards
24. Linear programming: the Simplex method with geographical applications - James E. Killen
25. Directional statistics - G.L. Guile & J.E. Burt
26. Potential models in human geography - D.C. Rich
27. Causal modelling: the Simon-Blalock approach - D.G. Pringle
28. Statistical forecasting - R.J. Bennett
29. The British Census - J.C. Dewdney
30. The analysis of variance - J. Silk
31. Information statistics in geography - R.W. Thomas
32. Centrographic measures in geography - A. Kellerman
33. An introduction to dimensional analysis for geographers - R. Haynes
34. An introduction to Q-analysis - J. Beaumont & A. Gatrell
35. The agricultural census - United Kingdom and United States - G. Clark
36. Order-neighbour analysis - G. Aplin
37. Classification using information statistics - R.J. Johnston & R.K. Semple
38. The modifiable areal unit problem - S. Openshaw
39. Survey research in underdeveloped countries - C.J. Dixon & B.E. Leach

This series is produced by the Study Group in Quantitative methods, of the Institute of British Geographers.
For details of membership of the Study Group, write to the Institute of British Geographers, 1 Kensington Gore, London SW7 2AR, England.

The series is published by:
Geo Books, Regency House, 34 Duke Street, Norwich NR3 3AP, England, to whom all other enquiries should be addressed.

CONCEPTS AND TECHNIQUES IN MODERN GEOGRAPHY No.38

# THE MODIFIABLE AREAL UNIT PROBLEM

by

Stan Openshaw

(Newcastle University)

## CONTENTS

ACKNOWLEDGEMENTS

Particular thanks are due to Dr A.C. Gatrell, Dr K. Jones, and two anonymous referees for making many useful and extremely helpful suggestions.



## I INTRODUCTION

The usefulness of many forms of spatial study, quantitative or otherwise, depends on the nature and intrinsic meaningfulness of the objects that are under study. Geographers have a long tradition of studying data for areal units; for example, spatial objects such as zones or places or towns or regions. The problem is that ever since the demise of 'the region' as the primary object of geographical study very little concern has been expressed about the nature and definition of the spatial objects under study. As Chapman (1977) put it 'Geography has consistently and dismally failed to tackle its entitation problems, and in that more than anything else lies the root of so many of its problems' (page 7). In short insufficient thought is given to precisely what it is that is being studied.

For many purposes the zones in a zoning system constitute the objects or geographical individuals that are the basic units for the observation and measurement of spatial phenomena. It is usual in a scientific experiment that the definition of the objects of study should precede any attempts to measure their characteristics. However, this is not the case with areal data where the spatial objects only exist after data collected for one set of entities are subjected to an arbitrary aggregation to produce a set of spatial units. Consider an example about wheat and potato yields. Data for one set of entities (farms or fields) can be aggregated to produce data for a set of spatial entities (parishes or counties). In this instance spatial aggregation is necessary in order to 'create' a relevant data set. As Yule and Kendall (1950) put it '.. geographical areas chosen for the calculation of crop yields are modifiable units, and necessarily so. Since it is impossible (or at any rate agriculturally impracticable) to grow wheat and potatoes on the same piece of ground simultaneously we must, to give our investigation any meaning, consider an area containing both wheat and potatoes and this area is modifiable at choice' (page 312). What they mean is that it is necessary to use areal units that are larger than the individual field and include both wheat and potatoes so that some measure of spatial association can be computed. Obviously at the level of the individual field there is no spatial association (assuming the fields are either all wheat or all potatoes) and, therefore, the degree of spatial association depends on the nature of the areal units that are used. The definition of these geographical objects is arbitrary and (in theory) modifiable at choice; indeed, different researchers may well use different sets of units. This process of defining or creating areal units would be quite acceptable if it were performed using a fixed set of rules, or so that there was some explicit geographically meaningful basis for them. However, there are no rules for areal aggregation, no standards, and no international conventions to guide the spatial aggregation process. Quite simply, the areal units (zonal objects) used in many geographical studies are arbitrary, modifiable, and subject to the whims and fancies of whoever is doing, or did, the aggregating. It is most unfortunate that there is no standard set of spatial units.

Since any study region over which data are collected is continuous, it follows that there will be a tremendously large number of different ways by which it can be divided into non-overlapping areal units for the purpose of

spatial analysis. Viewed as a combinatorial problem, the number of different zoning systems, each of m-zones, to which data for n individuals can be aggregated becomes incredibly large even for small values of n (Keane, 1975). For example, there are approximately $10^{12}$ different aggregations of 1,000 objects into 20 groups. If the aggregation process is constrained so that the groups consist of internally contiguous objects (i.e. all the objects assigned to the same group are geographical neighbours) then this huge number is reduced, but only by a few orders of magnitude. So even with the imposition of contiguity constraints the combinatorial problem remains totally unmanageable.

Consider an example based on census data. In Tyne and Wear County there are about 1.1 million people and 300,000 households. The 1981 census uses a set of about 2,800 enumeration districts to report the results. Consider how many different sets of 2,800 zones could be used for reporting the census characteristics of 300,000 households: Moreover, there are other huge combinatorial explosions whenever a zoning system of 2,800 zones are re-aggregated to form other zoning systems with fewer zones; for example, the 258 zones used for transportation modelling and planning. There are a tremendously large number of alternative 258 zone aggregations that could be used, most (if not all) of which will yield different results.

This, then, is the crux of the modifiable areal unit problem (MAUP). There are a large number of different spatial objects that can be defined and few, if any, sets of non-modifiable units. Whereas census data are collected for essentially non-modifiable entities (people, households) they are reported for arbitrary and modifiable areal units (enumeration districts, wards, local authorities). The principal criteria used in the definition of these units are the operational requirements of the census, local political considerations, and government administration. As a result none of these census areas have any intrinsic geographical meaning. Yet it is possible, indeed very likely, that the results of any subsequent analyses depend on these definitions. If the areal units or zones are arbitrary and modifiable, then the value of any work based upon them must be in some doubt and may not possess any validity independent of the units which are being studied.

The question is, does it matter? If you change the areal basis does it have any really significant effect on the results? Do haphazard zoning systems yield haphazard results? If they do, then what can be done about it?

Consider two more examples. The definition of enterprise zones was restricted to areas with high levels of unemployment. Unemployment rates were calculated for a set of statistical reporting units known as 'travel to work areas' (TTWAs); for details see Coombes and Openshaw (1982). Unfortunately, for a few areas of the country these particular areal units provide a poor representation of labour markets and present a biased picture of levels of unemployment. For example, for some obscure reason South Tyneside (in Tyne and Wear County) was included in the same TTWA as Washington, Gateshead, Jarrow, and parts of rural Northumberland. The effect was to mix areas of very high unemployment with fairly prosperous rural areas which have no strong journey to work links; the result was to reduce the apparent level of unemployment on South Tyneside. The total June unemployment, rates for the period 1978-82 were 11.1, 10.7, 12.9, 17.2, 18.7. However, if a more geographically meaningful definition of the South Tyneside labour market is used (see Coombes et al, 1982) then the unemployment rates for South Tyneside become 13.9, 13.3, 15.9, 20.1, 20.7; more than enough to justify an enterprise



zone. The unfortunate use of the 'wrong' set of areal units has therefore deprived South Tyneside of considerable job opportunities.

A final example concerns the Parlimentary Boundary Commision which reviews the boundaries of the 520 English constituencies every 15 years. This is a pure exercise in modifying areal units. The task is to select an appropriate amalgamation of wards to create constituencies of approximately equal electoral population size, that conform as far as practicable to county boundaries and London Boroughs, and that take into account local community interests. The latest revisions (1983) were performed by manual means and have been heavily criticised because of inconsistencies in application and unequal constituency sizes. The average electorate is about 68,000 but it varies from extremes of 24,000 (Newcastle Central) to 100,000 (Buckingham). There are even larger discrepancies in neighbouring constituencies; 57,000 in Finchley but 84,000 in Wood Green. The reason is simply that only a small proportion of all alternative areal arrangements were identified (Johnston and Rossiter, 1983). For example, the 26 wards in Camden can be aggregated to form two constituencies in 878 different ways, with a maximum two percent deviation from the mean size of 68,000 (Johnston, The Times, March 15th 1983). If a political geographer had ward-level voting data, then these different constituency definitions would yield a wide range of results.

If the reader is still unconvinced then he can attempt the following experiment. Construct an artificial set of areal units, or use a map of a few neighbouring local authorities. Assign some data to them. Compute either a few statistics for each zone (eg rates) or an overall statistic (eg correlation coefficient or mean). Now amalgamate a few zones which are contiguous, re-calculate your statistics for the aggregated data, and examine the changes. Now try to amalgamate a few more zones with the aim of either increasing or reducing the magnitude of the changes. Obviously this experiment would be easier if a microcomputer was used. However, a few hours experimentation will convince virtually anyone about the severity of the MAUP. Quite simply, different aggregations yield different results but without any systematic trends emerging that can be used for prediction or correction purposes.

What is so surising about the MAUP is that while geographers know of its existence they readily assume, in the absence of any knowledge, that it has no significant effect on their studies. An important reason for this deliberate neglect is that the validity of many applications of quantitive analyses of zonal data depends on the assumption that the MAUP does not exist and that the spatial units under study are given, meaningful, and fixed. Whilst these may be tolerable assumptions for a statistician, who may know no better, it is hardly a satisfactory basis for the application and fruther development of spatial analysis techniques in geography (Openshaw and Taylor 1981).

Although there is an almost infinite number of different ways by which a geographical region of interest can be areally divided, data are normally only presented and analysed for one particular set of units. The choice of these units is often haphazard, in that considerations such as convenience rather than geographical meaning are paramount. This uncertainty about the nature and definition of the zonal objects of spatial study is an important consequence of the MAUP. It is important because of the effects that the use of different areal units may have on the results of geographical study and

because it is endemic to all analyses of areal or zonal data. It is a major geographical problem with ramifications that need to be properly appreciated by geographers and all others interested in the analysis of spatially aggregated data. Looked at in this way, the MAUP is today one of the most important unresolved problems left in spatial analysis. There has been very little research compared with that afforded to many far less significant problems, and whilst it appears to be primarily a technical problem, it is also a major conceptual problem that is central to many aspects of geographical study.

## II AN INSOLUBLE PROBLEM OR A POTENTIALLY POWERFUL GEOGRAPHICAL TOOL?

Before examining various empirical evidence about the nature and severity of the MAUP, it is useful to consider further some of the more general aspects of the problem. This is important because the viewpoint that one adopts heavily influences thinking about how best to handle it.

It is unfortunate that so many geographers have become so completely blinkered by the concepts of conventional statistical theory and the normal science paradigm that they no longer seem to care or understand the basic geography of what they are doing. A decade or so ago this could be readily justified by the importance of introducing scientific methods and quantitative techniques into geography; the hope being expressed that the various unrealistic statistical and geographical assumptions could be relaxed later. To some extent this has happened and the significance of many purely statistical problems caused by the peculiar nature of areal data has been reduced; for example, by the development of methods capable of handling spatially autocorrelated data (Cliff and Ord, 1975). However, many of these statistical advances have been made at the expense of geographical considerations. It has perhaps been quietly overlooked that statistical techniques still cannot cope with the modifiable nature of areal data. The attempts by Griffith (1980) and others to establish a theory of spatial statistics are doomed to fail for the simple reason they begin with the assumption that the data under study are fixed!

It is humbly suggested that it is about time that quantitative geographers started to devise a body of relevant spatial analysis techniques that can cope with geographical data. The first stage in this second quantitative revolution must be the development of methods that can handle the MAUP, with the gradual replacement of many of the less relevant techniques that were originally plagiarised from a variety of disciplines in the 1960's and 1970's. If geography is to survive as a distinctive subject then it is time it stopped copying and adapting techniques imported from other disciplines and started a period of fundamentally relevant methodological innovation. There may well have to be a move away from a rather rigid and naive approach based on classical statistical theory.

The MAUP is sufficiently important that it presents a good opportunity for the development of new geographical techniques. However, before this can happen it is necessary to adopt a realistic perspective as to the importance of the MAUP to geography. To assist this process, it is useful to briefly view the modifiable areal unit problem from three different perspectives.



### (i) An insoluble problem

One very good reason for ignoring the MAUP is the belief that it is insoluble. If it really is endemic to the study of all areal data and if it really is insoluble then why not pretend it does not exist, in order to allow some analysis to be performed? This CATMOG is dedicated to those who believe in this fallacy of insolubility.

### (ii) A problem that can be assumed away

It is always possible, at least in theory, to change the nature of the MAUP by assuming it away. For example, a zonal model of trip rates made by members of a household can be re-expressed as a disaggregate model at the individual level, although there may be some problems if the zonal variables do not exist at the individual level. Other problems relate to the difficult statistical problems of identifying the nature of the underlying model implicit in aggregate level study and of estimating its unknown parameters if only aggregate data are available. It is also apparent that this approach is not always applicable; for instance, the crop yield example quoted from Yule and Kendall (1950). Even more important from a geographical point of view, it amounts to trying to write out or assume away any spatial effects and is, therefore, intrinsically non-geographical and aspatial; good statistics but very poor geography. It is of little use if the purpose of spatial study is to investigate spatial associations and it is argued that it is precisely this interest in areal phenomena that is a unique characteristic of geographical study.

### (iii) A very powerful analytical device

If the MAUP is endemic to spatial study and if it cannot simply be ignored or assumed away then methods should be developed to handle it and bring it under control in a purposeful way. Thus it can be regarded both as an opportunity for geographers to develop a new approach to spatial study based on zonal data and as a potentially very powerful geographical tool once the tremendous aggregational uncertainty or spatial freedom inherent in the MAUP can be usefully exploited for geographical purposes. Thus the MAUP is only a problem if it is viewed from a perspective that cannot handle it. It is certainly true that it is unlikely to have a precise analytical solution. However, the availability of fast super-computers opens up the possibility of seeking approximate numerical solutions. If the MAUP can be manipulated to suit particular purposes, via a kind of spatial optimisation process, then is it not possible that the resulting optimal zoning systems can be used for a number of geographical purposes and as a basis for a new approach to spatial study?

The suggestion made here is that this third perspective of the MAUP is the most appropriate one for geographers. It is, after all, a geographical problem that requires a geographical rather than a statistical solution. This CATMOG argues in favour of this view. It does so by first describing the nature and historical background of the MAUP followed by the results of a series of simple experiments involving nothing more complex than a correlation coefficient. Finally, two alternative geographical solutions are described with some discussion of areas where further developments may be expected in the future.

## III ON THE NATURE OF THE MODIFIABLE AREAL UNIT PROBLEM

### (i) Definitions

The MAUP is in reality composed of two separate but closely related problems. The first of these is the well known scale problem which is the variation in results that can often be obtained when data for one set of areal units are progressively aggregated into fewer and larger units for analysis. For example, when census enumeration districts are aggregated into wards, Districts, and Counties the results change with increasing scale. Previously, geographers have been very interested in scale problems of this sort, largely because it was thought that systematic scale effects could be easily handled.

Although scale differences are a most obvious manifestation of the MAUP there is also the problem of alternative combinations of areal units at equal or similar scales. Any variation in results due to the use of alternative units of analysis when the number of units is held constant is termed the aggregation problem (Openshaw, 1977a).

The MAUP obviously includes both these subproblems. The scale problem arises because of uncertainty about the number of zones needed for a particular study. The aggregation problem arises because of uncertainty about how the data are to be aggregated to form a given number of zones. It should be noted that for any reasonably sized data set there is considerably more spatial freedom in the choice of aggregation than there is in the choice of the number of zones.

At this stage it is worth noting that there are two different types of zonal arrangement. Most geographical studies have employed spatial aggregations based on contiguous arrangements of zones, something referred to as a zoning system. However, a zoning system is only a special case of a grouping system that incorporates a contiguity constraint. The non-contiguous case is referred to as a grouping system. The use of a contiguity constraint restricts the degree of aggregational variability but inmost practical studies it is so large anyway that it brings little real advantage other than the convenience of having zones which are formed of internally connected units.

Finally, it is noted that the MAUP is also closely involved in what is known as the ecological fallacy problem. An ecological fallacy occurs when it is inferred that results based on aggregate zonal (or grouped) data can be applied to the individuals who form the zones or groups being studied. In a geographical context the individuals can either be zones prior to a subsequent aggregation or non-modifiable entities. Obviously whether the ecological fallacy problem exists or not depends on the nature of the aggregation being used. A completely homogeneous zoning or grouping system would be free of this problem. However, most, if not all, zoning systems studied by geographers are internally heterogeneous so that the severity of any ecological fallacy problem depends largely on the nature of the aggregation being studied.

Figure 1 shows some simple examples of scale and aggregation problems. The reader is invited to investigate the effects of scale and aggregation by assigning some arbitrary values to the zones, computing a correlation



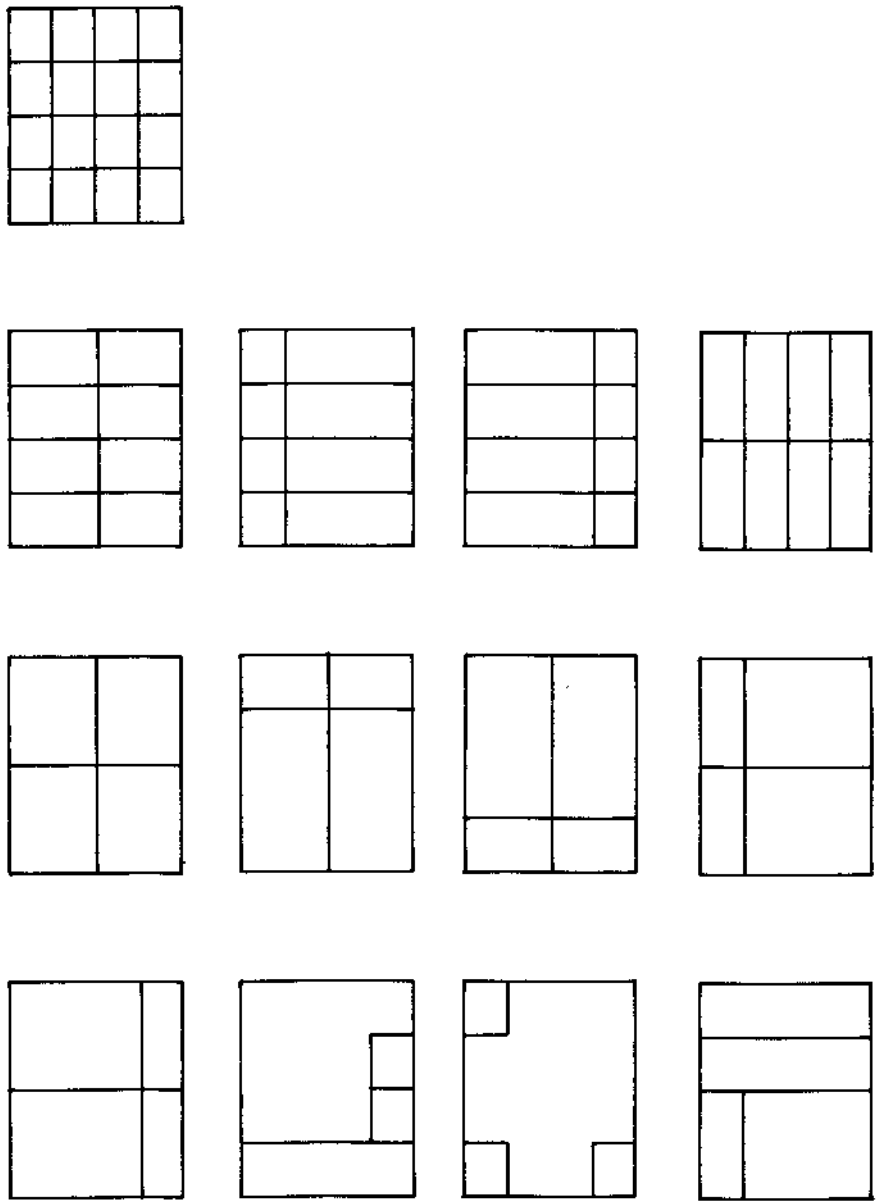


Figure 1. Alternative aggregations of 16 zones into 8 and 4 regions

coefficient (for instance) with a calculator and then repeating the process for different data aggregations. The data may be aggregated either by averaging or by adding rates, or by complete recalculation of numerators and denominators. The reader can also examine these effects.

(ii) Early correlation studies

One of the first papers to consider this type of problem is that by Gehlke and Biehl (1934). They observed that the size of the correlation coefficient increased with aggregation. The 252 census tracts in Cleveland, USA, were grouped successively into larger units of approximately the same size and subject to contiguity restrictions. The correlation between male juvenile delinquency and median monthly income was then calculated using both absolute numbers and ratios; see Table 1.

Table 1. Correlation coefficients for Cleveland, USA

| number of units | absolute numbers | rates |
|---|---|---|
| 252 | -.502 | -.516 |
| 200 | -.569 | -.504 |
| 175 | -.580 | -.480 |
| 150 | -.606 | -.475 |
| 125 | -.662 | -.563 |
| 100 | -.667 | -.524 |
| 50 | -.685 | -.579 |
| 25 | -.763 | -.621 |

The principal effect of using ratio variables is to slow down the increase of the correlation coefficient due to increasing scale by standardising for the size of areas. Gehlke and Biehl (1934) then compared these results with some random groupings of the data without contiguity restrictions; these aggregations produced correlations of -.434 for 150 zones and -.544 for 25. Random data aggregations have no systematic effect on the correlations.

A second set of experiments was used to demonstrate that the variation in the size of the correlation coefficient was related to the size of the units involved; the smallest values being associated with the smallest units. Data from the 1910 census provided two variables (the value of farm products and the number of farmers) for 1,000 rural counties. These data were then randomly grouped by Gehlke and Biehl to yield 63 and 31 groups with the following correlations:

| n | r |
|---|---|
| 1000 | .649 |
| 63 | .859 |
| 31 | .756 |

The data were also aggregated to 40 states and 8 counties (zoning systems) with the following results:

| n | r |
|---|---|
| 40 | .725 |
| 8 | .826 |

Gehlke and Biehl conclude that 'these results raise the question whether a correlation coefficient in census tract data has any value for causal analysis.



Does it measure the inter-relation of traits in their ultimate possessors - individuals and families? A relatively high correlation might conceivably occur by census tracts when the traits so studied were completely dissociated in the individuals or families of those traits' (page 170). Finally, they asked what is probably the most important question of all concerning whether a geographical area is an entity possessing traits or merely one characteristic of a trait itself? That is to say, are areal units entities or objects that can be studied or are they merely a variable that is proxy for geographical location?

Yule and Kendall (1950) added to Gehlke and Biehl's findings, demonstrating in particular that the correlation coefficient usually tends to increase with scale. They describe how the correlations between wheat yields and potato yields for the 48 counties of England increase as spatial aggregation reduces the number of areal units and increases their size and the scale of the analysis; see Table 2.

Table 2. Correlations between wheat and potato yield (English counties)

| number of geographical areas | correlation |
|---|---|
| 48 | .2189 |
| 24 | .2963 |
| 12 | .5757 |
| 6 | .7649 |
| 3 | .9902 |

They note that we seem able to produce any value of the correlation from 0 to 1 merely by choosing an appropriate size of the unit of area for which we measure the yields. Is there any "real" correlation between wheat and potato yields or are our results illusory?' (page 311). There is little doubt that Yule and Kendall had a deep appreciation of the importance of MAUP. They recognised the difference that exists between studies based on modifiable units, such as areal units, and those based on non-modifiable units, such as the cow or the shell. Furthermore, they emphasised that in studies based on modifiable units the magnitude of a correlation will depend on the units that are used. In this vein they wrote: 'Our correlations will accordingly measure the relationship between the variates for the specified units chosen for the work. They have no absolute validity independently of these units, but are relative to them. They measure, as it were, not only the variations of the quantities under consideration, but the properties of the unit-mesh which we have imposed on the system in order to measure it' (page 312). This is a very clear early statement of the nature and importance of the MAUP.

Despite this, the correlation coefficient was still regarded as useful because the value for the 48 English counties in 1936 is a geographical and historical fact. A comparison of values for the same units over time might also be interesting. However, Yule and Kendall consider the result to be specific to the zoning system they used and that it is, therefore, not capable of scientific generalisation or for comparison with other correlations for the same variables but for different zones.

A feature of both these early studies was the observation that because the correlations are modifiable they may not provide any useful guide to individual or more spatially disaggregated levels of correlations. Robinson

(1950) provides the conclusive proof that this in fact the case. He quotes an example based on the correlation between percentage population 10 years old and over which is negro and the percentage of the same population that is illiterate; another example is based on the correlation between nativity and illiteracy. Table 3 shows the various correlations that were computed for different levels of spatial aggregation.

Table 3. Individual and ecological correlations (after Robinson, 1950)

| level of aggregation | number of units | correlations between: negro and illiteracy | nativity and illiteracy |
|---|---|---|---|
| individual | 98 million | .203 | .118 |
| state | 48 | .773 | -.526 |
| census division | 9 | .946 | -.619 |

The results are quite conclusive. There is a pronounced scale effect in that the absolute values of the correlations increase as the number of observations decrease. In addition, the aggregate values bear little resemblance to the individual values prior to spatial aggregation. Robinson concludes therefore '..there need be no correspondence between the individual correlation and the ecological correlation' (page 354). This is an important result which readily illustrates the dangers of making individual level inferences from analyses performed at an aggregate level.

A final paper that is of interest in this section is that of Blalock (1964). He describes the results of a series of experiments designed to investigate the effects of data aggregation. The correlation coefficient between differences in income for blacks and whites and percentage blacks for 150 southern USA counties was found to be 0.54. Blalock was interested in the question of what happens if the counties are grouped into larger units in various different ways. The results are shown in Table 4.

Table 4. Blalock's aggregation experiments

| number of units | random grouping | random zoning |
|---|---|---|
| 75 | .67 | .63 |
| 30 | .61 | .70 |
| 15 | .62 | .84 |
| 10 | .26 | .81 |

With random grouping systems we would expect the correlation coefficients to show no systematic scale effects. The variability in values will be due to sampling fluctuation. In this instance sampling fluctuation is in fact the aggregation component since there are a very large number of ways by which 150 objects can be randomly grouped into 75 groups or less. The apparently anomalous value of the correlation coefficient for the 10 groups is an indication of this effect; indeed, it is slightly miraculous that the other values are so uniform.

By contrast the random zoning systems will be affected by any spatial autocorrelation present in the data, so that the rising correlations with increasing scale can be regarded as the result of spatial autocorrelation whereby the zoning system retains more variance of one variable than of the



other. This is the interpretation put forward by Taylor (1977). It has also been argued that if the variables are not spatially autocorrelated then the correlation coefficient will not increase with scale. A problem with this interpretation is that Blalock ignores aggregation effects and these may easily dominate any scale effects. Some of the ideas developed by Blalock (1964) and put into a geographical context by Taylor (1977) have been tested in Openshaw and Taylor (1979). The expected systematic relationships did not emerge. The effects of the aggregational variability were simply too strong; indeed, perhaps rather alarmingly, the authors concluded that 'We have been able to find a wide range of correlations. We simply do not know why we have found them. Hence we can make no general statements about variations in correlation coefficients so that each areal unit problem must be treated individually for any specific.piece of research' (Openshaw and Taylor, 1979; p 142-143). What is meant is that the aggregational variability is not susceptible to a statistical approach since no systematic empirical regularities could be found.

(iii) More recent studies

Apart from the occasional mention, the MAUP seems to have been ignored until the problem was re-examined in the late 1970's. Openshaw (1977a) was one of the first to re-emphasise the importance of aggregation effects. An example readily shows the importance of the aggregation problem and relative insignificance of the scale problem. The data used here relate to 100 metre grid-squares for South Shields. These data could be readily aggregated to 200, 300, 400, 500, 600, 700, 800, 900 and 1 km squares. For each of these scales there are a number of alternative aggregations; for example, shifting the origin of the 100 metre lattice produces 25 different 500 metre grid-square aggregations. The resulting distribution of correlation coefficients are shown in Table 5.

Table 5. Scale and aggregation effects on the correlation between numbers of arly and Mid-Victorian houses in South Shields

| size of squares (metres) | scale effects | aggregation effects: mean correlation | aggregation effects: standard deviation |
|---|---|---|---|
| 100 | .08 | - | - |
| 200 | .21 | .31 | .11 |
| 300 | .43 | .43 | .06 |
| 400 | .28 | .47 | .11 |
| 500 | .55 | .49 | .16 |
| 600 | .45 | .52 | .16 |
| 700 | .20 | .57 | .18 |
| 800 | .56 | .58 | .18 |
| 900 | .66 | .60 | .19 |
| 1 km | .73 | .62 | .20 |

The second column shows the effects of increasing scale using only one of the possible aggregations to each scale of grid-square. The third column shows the mean correlation based on different aggregations to the same scale produced by moving the origin of the lattice. The fourth column shows the standard deviation of the correlation coefficients produced for the different aggregations to each scale. This example contradicts the claim by Evans (1981) that with grid-square data the changes in correlation coefficient are usually

consistent across a wide range of scales up to 256 km (page 55). The reason is that his 1 km squares have already smoothed the data dramatically. Finally, it is noted that the example shown in Table 5 does not consider the full extent of the aggregation problem. This would involve an examination of 10,000 alternative 100 metre squares (assuming the data being aggregated have been grid-referenced at the 1 metre level), 1,000,000 different 1 km squares, and even larger numbers of alternatives if the zones are not constrained to be square in shape.

The conclusion that can be drawn from Table 5 is that Yule and Kendall (1950) were quite correct, although they clearly underestimated the severity of the problem. Different variables can be affected by aggregation in different ways so that multivariate techniques based on correlations will tend to amplify the differences in results caused by the use of different zoning systems. As a result, the aggregation and scale variability reported for the correlation coefficient also applies to more complex multivariate methods and to many other forms of analysis. It is demonstrated later that it is not a problem that afflicts only the poor correlation coefficient.

The ecological fallacy problem has also been studied further. The principal problem here is that a detailed investigation requires access to large spatially referenced individual data sets and it is only quite recently that sufficiently powerful computers have become available_ to handle these. The ecological fallacy problem occurs because areal studies cannot distinguish between spatial associations created by the aggregation of data and real associations possessed by the individual data prior to spatial aggregation. Thus the characteristics of typical deprived urban areas need not be the same as the characteristics of the individuals who live there.

One consequence of Robinson's work was that may social scientists interpreted his warning as a rigid taboo on the use of all aggregate data; although this never extended to geography. Borgatta and Jackson (1980) pointed out that 'what happened was the assumption that, because use of aggregate data could be misleading at the individual level, every such interpretation had to be incorrect' (page 8). It is also possible that Robinson exaggerated the importance of the problem; in particular he only examined the most gross levels of aggregation. The question arises as to whether these results are typical of what might happen with finer spatial scales and more realistic zoning systems.

Recently, some further insights into this problem have come from the analysis of a random 10 per cent sample survey of all households in Sunderland and from the analysis of individual census data for part of Italy (Openshaw, 1983; Bianchi et al, 1981). A brief description of the results for Sunderland can best be examined here. These data can be studied at the individual level (8,483 households) or aggregated to polling districts (36 zones), 1 km squares (117 zones), and 500 metre squares (348 zones). A set of 54 typical indicator variables were computed. The simplest way to investigate the ecological fallacy problem is to cross-tabulate the individual and zonal correlation coefficients (Table 6).



Table 6. Cross-tabulation of individual and ecological correlations (percentage of row totals)

| individual correlations | areal correlations -1. to -.8 | -.8 to -.6 | -.6 to -.4 | -.4 to -.2 | -.2 to .0 | .0 to .2 | .2 to .4 | .4 to .6 | .6 to .8 | .8 to 1. | total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| -1. to -.8 | 100 | | | | | | | | | | 1 |
| -.8 to -.6 | 50 | 50 | | | | | | | | | 4 |
| -.6 to -.4 | 12 | 44 | 32 | 12 | | | | | | | 25 |
| -.4 to -.2 | | 9 | 36 | 34 | 15 | 4 | 1 | | | | 180 |
| -.2 to .0 | | | 4 | 32 | 39 | 18 | 5 | 1 | | | 997 |
| .0 to .2 | | | 1 | 2 | 14 | 29 | 32 | 20 | 3 | | 188 |
| .2 to .4 | | | | | | 14 | 32 | 39 | 14 | | 28 |
| .4 to .6 | | | | | | | 17 | 50 | 17 | 17 | 6 |
| .6 to .8 | | | | | | | | 50 | 50 | | 2 |
| totals | 6 | 32 | 117 | 387 | 444 | 248 | 117 | 66 | 13 | 1 | |

Sunderland 1 km squares

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| -1. to -.8 | 100 | | | | | | | | | | 1 |
| -.8 to -.6 | 75 | 0 | 25 | | | | | | | | 4 |
| -.6 to -.4 | 32 | 32 | 20 | 12 | 4 | | | | | | 25 |
| -.4 to -.2 | 7 | 27 | 31 | 16 | 14 | 4 | 0 | 1 | | | 180 |
| -.2 to .0 | | 4 | 14 | 24 | 25 | 16 | 11 | 3 | 2 | | 997 |
| .0 to .2 | | | | 3 | 9 | 19 | 26 | 28 | 14 | 1 | 188 |
| .2 to .4 | | | | 4 | 4 | 7 | 7 | 32 | 46 | | 28 |
| .4 to .6 | | | | | | | 17 | 0 | 67 | 17 | 6 |
| .6 to .8 | | | | | | | | | 50 | 50 | 2 |
| totals | 26 | 93 | 208 | 281 | 295 | 209 | 157 | 96 | 61 | 5 | |

Sunderland polling districts

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| -1. to -.8 | 100 | | | | | | | | | | 1 |
| -.8 to -.6 | 75 | 25 | | | | | | | | | 4 |
| -.6 to -.4 | 4 | 52 | 40 | 4 | | | | | | | 25 |
| -.4 to -.2 | | 3 | 37 | 47 | 12 | 1 | | | | | 180 |
| -.2 to .0 | | | 1 | 24 | 57 | 17 | 2 | | | | 997 |
| .0 to .2 | | | | 1 | 7 | 43 | 39 | 9 | 1 | | 188 |
| .2 to .4 | | | | | | | 46 | 50 | 4 | | 28 |
| .4 to .6 | | | | | | | | 67 | 33 | | 6 |
| .6 to .8 | | | | | | | | | 100 | | 2 |
| totals | 5 | 20 | 86 | 321 | 607 | 248 | 102 | 36 | 6 | | |

Sunderland 500 m squares

The size of the percentages in the diagonals gives an indication of the extent to which aggregation has either increased or decreased the magnitudes of the correlation coefficients. A comparison of the row and column totals

shows that aggregation has a flattening effect on the frequency distribution of the individual correlation coefficients. Table 6 clearly demonstrates the systematic biasing of the ecological correlations from 0 towards ±1 and that the magnitude of the bias increases with scale.

It is noted that the cross-tabulations in Table 6 do not give any indication of aggregational variability since only one aggregation at each scale was examined. That is to say, these results refer only to scale effects and it may be expected that the aggregational effects will be somewhat larger. Both are important since if these phenomena were better understood it might be possible to design improved areal definitions for reporting census data. For instance, is there a critical size for census enumeration districts which may minimise the effects of scale and aggregation on the data being aggregated? The present size is merely a reflection of the area that can be covered by a census enumerator in one day; this is hardly a meaningful variable in urban geography. It is something of a mystery why census data collecting agencies do not bother to try and resolve these very important practical questions.

These results suggest that perhaps the magnitude of the ecological fallacy problem is less than the results presented by Robinson (1950) might indicate. Certainly the changes in the magnitude of the correlation coefficient are smaller in Table 6 than in Table 3. However, this is slightly misleading since only a small percentage of all correlations in Table 6 do not have substantial and systematic biases; for the polling district data the figure is 16 per cent. Additionally, it is impossible to predict the severity of the problem without access to individual data. As a result there is no way of knowing whether a particular areal data set will yield values which are close to the individual values. A fuller discussion of empirical aspects is provided in Openshaw (1983), while Williams (1976, 1979) outlines a theoretical interpretation.

## IV THE RESULTS OF SOME AGGREGATION EXPERIMENTS

### (i) Random aggregation and the correlation coefficient

The complex nature of the MAUP suggests that further advances in our understanding of it can be most readily made by empirical experimentation. It is not denied that a theoretical approach could be rewarding; indeed, various preliminary studies have been made (Williams, 1976, 1979; Batty and Sikdar, 1982). However, the problem is proving to be exceptionally complex and it is most easily investigated by empirical means. Furthermore, the availability of high-speed computers makes it possible to design aggregation experiments of a far more comprehensive nature than would be the case if non-automated methods were being employed. Additionally, entire new numerical algorithms can be devised to explore different aspects of the aggregation problem.

The first set of experiments concerns the effects of random aggregation on the correlation coefficient. Some of the results produced by simple random aggregation experiments by Gehlke and Biehl (1934) and Blalock (1964) have already been described. The question is simply what happens if a more systematic and comprehensive series of experiments is performed. Interest is focused on purely random aggregations partly because of the historical



connections and partly because it has been suggested that the statistical distribution of a statistic due to sampling variability and its zoning distribution due to the choice of different zoning systems, are analogous. If the analogy could be proven then it would be exceptionally convenient because it would allow the standard formulae for estimating sampling errors for simple random samples to be used to provide estimates of aggregational variability, presumably under the assumption of simple random zoning. In this sampling-zoning analogy the number of zones in the zoning system would be regarded as equivalent to sample size.

For this study 1970 census data for the 99 counties in the State of Iowa, USA, are examined. Two variables are selected for analysis; the percentage vote for Republican candidates in the congressional election of 1968 and the percentage of the population over 60 years. There is nothing special about the selection of this data, it merely happened to be convenient! Openshaw and Taylor (1979) report a range of different correlations that can be produced for these variables when the 99 counties are aggregated into a number of arbitrary six zone aggregations; the values ranged from .26 for the congressional districts to .86 for a simple typology of Iowa into rural-urban types. The value of the correlation at the 99 county level is 0.34. Since the 99 counties form a complete population of Iowa counties this value can be regarded as the population correlation. The question is how well random samples and sample random zoning systems represent this population value.

Table 7 reports the means and standard deviations of the correlation coefficient for 10,000 random samples of (i) random zoning systems (randomly selected areal aggregations) with 6, 12, 18, 24, 30, 36, 42, 48, and 54 zones; and (ii) random samples (random selections of various numbers of zones) of 6, 12, 18, 24, 30, 36, 42, 48, and 54 counties. The latter provide results which approximate the values that would be obtained from standard sampling formulae. Openshaw (1977b) describes the computer algorithm used to generate the quasi-random zoning systems.

Table 7. Sampling and zoning distributions of the correlation coefficient

| number of zones | zoning distributions mean | zoning distributions standard deviation | sample size | mean | standard deviation |
|---|---|---|---|---|---|
| 6 | .36 | .218 | 6 | .31 | .429 |
| 12 | .33 | .161 | 12 | .34 | .273 |
| 18 | .33 | .139 | 18 | .34 | .209 |
| 24 | .32 | .122 | 24 | .34 | .172 |
| 30 | .33 | .110 | 30 | .34 | .144 |
| 36 | .33 | .102 | 36 | .34 | .125 |
| 42 | .33 | .092 | 42 | .34 | .109 |
| 48 | .33 | .082 | 48 | .34 | .097 |
| 54 | .33 | .073 | 54 | .34 | .086 |
| 99 | .346 | | | .346 | |

The most interesting discovery here is that scale has no systematic effect on the mean correlation coefficient. This is because the zoning systems are chosen at random so that the sample (or more precisely the zoning) estimates of the correlation coefficient approximate the population value (which for zonal data is the observed value prior to the current aggregation, ie the

99 zone value ). It should also be noted that there is considerable zoning and sampling variability about the mean values but that this reduces with increasing numbers of zones or increasing sample sizes. Finally, the standard deviations of the zoning distributions are considerably smaller than the corresponding sampling distributions but exhibit a greater degree of bias.

In these results, somewhere, are the effects of spatial autocorrelation. Most data sets exhibit positive spatial autocorrelation and the Iowa data are no exception. Spatial autocorrelation only affects the zoning distributions because aggregation takes place under contiguity restrictions. Normality or non-normality is not thought to have any important effect on these experiments.

One way of identifying the effects of spatial autocorrelation is to use data sets with different levels of spatial autocorrelation and see what effect this has on the zoning distributions. Openshaw and Taylor (1979) describe a procedure for generating artificial data for the 99 Iowa counties with the following properties: zero skewness and kurtosis to ensure normality, a correlation equal to that observed for the real Iowa data, and regression slope and intercept parameters also equal to the observed Iowa data. Three different levels of spatial autocorrelation were considered (autocorrelation is measured by Moran's I statistic for first order contiguities, see Silk (1979)): maximum negative spatial autocorrelation, MN, (the best that could be achieved were values of -.71 for the vote variable and -.57 for the old age variable), zero autocorrelation,Z, and maximum positive autocorrelation, MP, (the best that could be managed were values of .82 and .92). The same sets of 10,000 zoning systems as used for Table 7 are applied to these artificial data sets with the results shown in Table 8.

Table 8. Zoning distributions of the correlation coefficient for three different levels of spatial autocorrelation

| number of zones | MN mean | MN standard deviation | Z mean | Z standard deviation | MP mean | MP standard deviation |
|---|---|---|---|---|---|---|
| 6 | .31 | .443 | .61 | .294 | .60 | .247 |
| 12 | .30 | .370 | .47 | .263 | .52 | .176 |
| 18 | .29 | .350 | .42 | .227 | .48 | .142 |
| 24 | .31 | .309 | .40 | .192 | .44 | .121 |
| 30 | .32 | .277 | .39 | .166 | .42 | .108 |
| 36 | .32 | .242 | .38 | .146 | .40 | .098 |
| 42 | .33 | .209 | .37 | .128 | .39 | .087 |
| 48 | .33 | .183 | .36 | .112 | .38 | .080 |
| 54 | .33 | .160 | .36 | .100 | .34 | .072 |

The artificial data with negative spatial autocorrelation has the least biased results but the largest standard deviations, whereas increasing positive spatial autocorrelation produces results which are increasingly biased but with smaller standard deviations. The zero autocorrelation state confers no particular benefits.

The principal conclusion from these experiments is that the sampling-zoning analogy does not hold good. There is an additional risk involved in using standard error formulae for simple random sampling as estimates of the



aggregational variability due to the use of simple random zoning systems. An examination of a simple null hypothesis test based on the correlation coefficient shows the sort of additional risk that is involved. For a standard type I error significance level of 0.05 the value observed from the Monte Carlo experiments ranged from .10 to .22, according to the level of spatial autocorrelation and the particular variable under study.

Other problems with the sampling-zoning analogy concern the fact that most zonal data sets contain both sampling variability and aggregational variability. In addition, zonal data are unusual in that the population value for any statistic can be determined; for aggregated data this is the value of a statistic for the data prior to the current aggregation. A final problem concerns the fact that geographers have not previously shown any interest in studying purely random zoning systems; perhaps they are not thought to be meaningful entities, although it is possible also that until quite recently it was difficult to generate random zoning systems.

Another aspect of this discussion concerns the use of inferential statistical techniques with zonal data. Quite simply, it is seldom clear as to what is the nature of the hypothesis that is being tested and what, if anything, the results signify. If random zoning is not being used then in what way do zonal data constitute a sample, be it simple or complex? What is the population? A statistical answer to some of these questions is to invent a 'super population'; for example, that the Iowa data is a random sample of data for Iowa counties because it relates to one, randomly chosen, point in time. While this is easy to say, it is far less easy to identify what the significance tests mean. There is also the difficult problem of determining an appropriate set of sampling error estimation equations. The Iowa data represents a sample size of 1. Finally, it is not clear as to the geographical implications of the hypotheses that could be tested. For example, under what conditions is it possible to compare zonal estimates for one set of zones with zonal estimates for another set?

### (ii) Random aggregation and other statistics

A further consideration is whether or not the results observed for the correlation coefficient also hold good for other statistics. Perhaps the correlation coefficient is a special case. The question is therefore what scale and aggregation variability are likely to be displayed by other unstandardised statistics, such as the mean and the regression slope coefficient. Is it possible that these statistics will be less affected and more robust to aggregation effects? For example, the mean has very good large sample properties. Table 9 should dispel any fears in this direction. It illustrates some results from a regression of percentage rate for Republican candidates as a percentage of the population over 60 years of age (see page 17).

The mean statistic for the zoning distributions is only very slightly biased but still has the now characteristic small standard deviation, relative to the related sampling distributions. The regression coefficient behaves in a similar fashion to the correlation coefficient.

### (iii) Random aggregation experiments with once aggregated data

The previous experiments concerned the effects of randomly aggregating zonal data which have already been aggregated at least once previously. Most

Table 9. Zoning and sampling distributions of a mean and a regression slope statistic

| number of zones | mean old aged | | | | regression slope | | | |
|---|---|---|---|---|---|---|---|---|
| | zoning | | sampling | | zoning | | sampling | |
| | mean | std | mean | std | mean | std | mean | std |
| 6 | 14.5 | .263 | 14.5 | 1.105 | 1.55 | 1.071 | 1.13 | 1.944 |
| 12 | 14.5 | .291 | 14.5 | .747 | 1.34 | .689 | 1.23 | 1.088 |
| 18 | 14.5 | .278 | 14.5 | .591 | 1.27 | .569 | 1.23 | .800 |
| 24 | 14.5 | .267 | 14.5 | .496 | 1.25 | .484 | 1.24 | .649 |
| 30 | 14.5 | .249 | 14.5 | .422 | 1.24 | .427 | 1.24 | .536 |
| 36 | 14.5 | .230 | 14.5 | .369 | 1.23 | .389 | 1.24 | .460 |
| 42 | 14.5 | .217 | 14.5 | .323 | 1.23 | .346 | 1.24 | .396 |
| 48 | 14.5 | .202 | 14.5 | .291 | 1.23 | .307 | 1.24 | .352 |
| 54 | 14.5 | .186 | 14.5 | .255 | 1.23 | .273 | 1.25 | .312 |
| 99 | 14.5 | | 14.5 | | 1.25 | | 1.25 | |

Note: std is an abbreviation for standard deviation

data that geographers study are of this type. The question arises, therefore, as to the effects of aggregating data that have not been previously aggregated; for example, the aggregation of individual data to a zoning system. This problem is interesting partly because it is here that ecological fallacies may be created and because aggregation changes the measurement scale, usually from a nominal to a continuous form. For example, presence or absence measurements become frequencies or ratios or percentages after aggregation.

The Sunderland data are used to investigate this problem. The household data have 100 metre grid-references attached to them. For this experiment the 8,483 households can be regarded as single member zones. Notional contiguities can be generated by a Thiessen polygon program so that the individual data zones can be aggregated to form random zoning systems with 25, 50, 75, 100, 150, and 200 zones. Table 10 shows the results that were obtained for three variables which were selected to show different types of aggregational behaviour displayed by the correlation coefficient.

Table 10. Zoning distributions for once aggregated data for Sunderland

| number of zones | variable 1 | | variable 2 | | variable 3 | |
|---|---|---|---|---|---|---|
| | mean | std | mean | std | mean | std |
| 25 | .79 | .045 | -.93 | .015 | -.94 | .014 |
| 50 | .82 | .034 | -.92 | .015 | -.92 | .017 |
| 75 | .83 | .026 | -.92 | .015 | -.91 | .016 |
| 100 | .84 | .026 | -.92 | .015 | -.90 | .020 |
| 150 | .83 | .022 | -.91 | .016 | -.88 | .013 |
| 200 | .82 | .022 | -.91 | .015 | -.87 | .018 |
| individual correlation | .42 | | -.81 | | -.57 | |

Note: std is an abbreviation for standard deviation



These results are superficially similar to those reported for the re-aggregation of already aggregated data. One difference is the smaller relative sizes of the standard deviations of the zoning distributions in Table 10. This could well reflect the use of small sample sizes; computer times for a sample of 100 different individual data aggregations amounted to 2 hours of CPU time on an IBM 370/168. The use of notional contiguities and a sample data set may also have contributed to reducing the expected range of aggregation effects. Most of the results for the other 50 variables which were examined tended to have zoning distribution means of the correlation coefficient which are similar to the 1 km zonal values. Nevertheless, the results again show that zonal correlations need not correspond to the individual level correlations and that a 'good' zoning system for one variable can be quite 'poor' for another, at least in terms of the differences between ecological and individual correlation coefficients. It is still confidently expected that the aggregational variability in the range of possible results due to the choice of the first zoning system will exceed that of any subsequent re-aggregations of the data, although the current experiment did not show it. Even if this assumption can be disproven, it is highly likely that the choice of the first zoning system has a crucial effect on the severity of any subsequent ecological fallacies and that, as far as practicable, the design of this zoning system should be optimised to minimise these effects. It may be that the possible benefits are slight or are offset by the computer costs that are involved, but until we try we shall never know.

### (iv) Identifying the limits of the MAUP

So far attention has been restricted to investigating the variability in results due to purely random spatial aggregations. The question now arises as to what are the worst case or, real limits of aggregation effects if we are perverse enough to look and know bow to find them. The existence of electoral boundary gerrymandering has been known about in political geography for over 170 years, ever since the famous 1810 gerrymander (Taylor and Johnston, 1979; pages 371-374). However, it is only recently that its general implications for spatial analysis have beep recognised (Openshaw, 1977a, 1977c, 1978b). By searching for the approximate limits of the range of aggregation effects it is possible to demonstrate the magnitude and severity of the MAUP.

Openshaw (1977a) uses a heuristic procedure, of a type similar to iterative relocation algorithms in cluster analysis, to optimise any general function by manipulating the zoning systems. This method provides an approximate solution to what is termed the Automatic Zoning Problem; the algorithm is called the Automatic Zoning Procedure (AZP). The basic algorithm is best described in general terms as consisting of a series of steps.

Step 1. Decide how many regions are required in the final aggregation.

Step 2. Generate a random zoning system with this number of regions.

Step 3. Randomly select one of these regions and proceed around its boundary measuring the effects on the objective function of moving zones from the bordering regions into it.

Step 4. 'Once an improvement is recorded for the objective function which is being optimised, then check whether the move is possible; that is, it must not destroy the internal contiguity of the region from which a zone is being moved; either reject or accept the move.

Step 5. Once all the members of a region have been examined return to step 3 to process another region; if all regions have been examined then go to step 6.

Step 6. If one or more moves have been made then return to step 3 otherwise stop.

In this algorithm the initial data are assumed to relate to a set of zones and these zones are to be aggregated into a smaller number of large zones which for purposes of clarity are termed regions. For example, the 99 Iowa Counties form a set of 99 zones which can be aggregated into 6 regions. The aggregation is performed in such a way so as to approximately optimise an objective function whilst ensuring that all the zones assigned to the same region are internally connected or contiguous. The objective function can be any general function and it need not be continuous. For example, the aim may be to maximise or minimise a correlation coefficient between two variables in order to identify the approximate limits of variability due to the MAUP. The AZP algorithm is a heuristic procedure which experience has shown can readily solve many types of optimal zoning problems although there is no guarantee that it will always find the global optimum; indeed with this type of problem there can be no certainty that there is a unique global optimum to be found. For most problems it probably gets fairly close to a 'good' local optimum; large problems are easier to solve than small ones. No doubt the heuristics could be further improved; for example, by the incorporation of a multiple simultaneous move heuristic; but at present this is not the most important problem. More important was the discovery of how to incorporate a constraint handling procedure (Openshaw, 1978b), because together with fast computers this made possible the application of the AZP algorithm to a wide range of region building problems.

Returning to the correlation coefficient, this can be used as the objective function in the AZP and attempts made to seek zoning systems that either maximise it or minimise it. This can be regarded as an exercise in applied gerrymandering or, if you prefer, spatial engineering of zoning systems. The dramatic results are shown in Table 11. Even for the 99 Iowa zones, a small data set by current standards, a very wide range of results can be obtained. The amount of aggregational variability, or spatial freedom, will be even greater with larger data sets and is probably some exponential function of the aggregation factors involved. Nevertheless, for a 6 region aggregation of the 99 Iowa counties the range of possible correlations is between -.99 and +.99. It is also possible that many of the intermediate results can be obtained; for example, a zoning system with a correlation of 0.5 or -0.334. Different amounts of spatial autocorrelation have no noticeable effects.

Table 11. Some approximate limits of the correlation coefficient due to different aggregations of the Iowa data

| number of zones | Iowa data | | MN data | | Z data | | MP data | |
|---|---|---|---|---|---|---|---|---|
| | min r | max r | min r | max r | min r | max r | min r | max r |
| 6 | -.99 | .99 | -.99 | .99 | -.99 | .99 | -.99 | .99 |
| 12 | -.99 | .99 | -.97 | .99 | -.99 | .99 | -.98 | .99 |
| 18 | -.97 | .99 | -.97 | .99 | -.97 | .99 | -.92 | .99 |
| 24 | -.92 | .99. | -.98 | .99 | -.90 | .99 | -.89 | .98 |
| 30 | -.73 | .98 | -.93 | .98 | -.86 | .98 | -.78 | .95 |
| 36 | -.71 | .96 | -.93 | ..98 | -.80 | .98 | -.61 | .93 |
| 42 | -.55 | .95 | -.92 | .97 | -.79 | .96 | -.52 | .93 |
| 48 | -.50 | .90 | -.87 | .96 | -.66 | .95 | -.39 | .89 |
| 54 | -.42 | .82 | -.85 | .95 | -.52 | .91 | -.32 | .88 |

Notes: based on best of five different random zoning systems used as starting aggregations. MN, Z, MP are the three artificial Iowa data sets (see Table 8)



Yule and Kendall (1950), in a prophetic statement, warn against the development of zonal manipulation procedures of the kind used here. They write 'the student should not now go to the other extreme and claim that, since a large range of values of correlation coefficients may be obtained according to the choice of a modifiable unit, a particular value has no significance' (page 312). Perhaps they did not realise that such a wide range of aggregation effects were present or did not know how to find them in a systematic fashion. Instead what they mean is that significance of the correlation coefficient depends on the meaningfulness of the areal units on which it is based. Perhaps they thought, rather naively, that counties are a sensible spatial unit for the study of crop yield relationships whereas arbitrary aggregations of the counties to maximise a correlation coefficient would not be. It is a shame that Yule and Kendall's work on the modifiable areal unit problem did not continue past this point. Perhaps it could not, because the problem rapidly becomes one of trying to assess the degree of meaningfulness associated with different geographical definitions for a particular purpose. In general terms it is an impossible problem; for example, how would we go about determining whether counties are an appropriate unit by which to study crop yield relationships or indeed anything?

Some critics of the optimal zoning results have suggested that it only works when applied to correlation coefficients and that in any case the optimal zoning systems will be of the most peculiar shapes and sizes. This latter point is examined later. The first is simply incorrect. The performance and parameter estimates of a variety of linear and nonlinear models have also been shown to vary between wide limits (Openshaw, 1977c, 1978a, 1978b).. Some models, for instance interaction models, are highly sensitive since the pattern of trips that these models try to represent depends on the zoning systems used. A simple example based on the linear regression model should help emphasise the importance of the MAUP. The AZP can be used to produce zoning systems which generate data to either maximise or minimise best statistical estimates of the slope coefficient in a regression model based on the Iowa data (Openshaw, 1978a). In this experiment every time a change is made to the zoning system by the AZP the parameters are re-estimated. Two different parameter estimation procedures are used; one based on ordinary least squares the other on a robust line fitting procedure in the style of Tukey (1977) and described in McNeil (1977); the purpose is to avoid making normal linear regression model assumptions. The results are shown in Table 12 and two of the 12 region zoning systems are shown in Figure 2.

Table 12. Approximate limits of regression slope coefficients due to different aggregations of the Iowa data

| number of zones | Ordinary least squares estimation of slope | | robust line fitting estimation of slope | |
|---|---|---|---|---|
| | minimise | maximise | minimise | maximise |
| 6 | -121 | 27 | -84 | 22 |
| 12 | -24 | 12 | -34 | 42 |
| 18 | -12 | 12 | -14 | 16 |
| 24 | -8 | 10 | -11 | 14 |
| 30 | -5 | 7 | -12 | 12 |
| 36 | -4 | 6 | -8 | 10 |
| 42 | -3 | 5 | -5 | 8 |
| 48 | -2 | 4 | -4 | 6 |
| 54 | -1 | 4 | -2 | 6 |

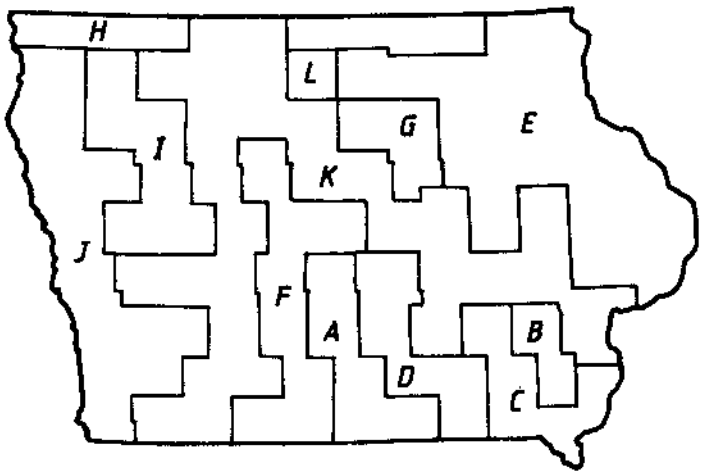



Figure 2a. Zoning system that minimises the regression slope coefficient (-24, r = -.25)

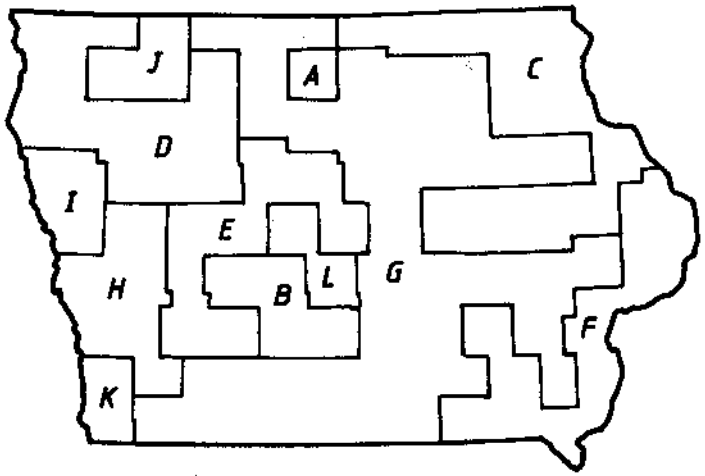



Figure 2b. Zoning system that maximises the regression slope coefficient (12, r = .87)



The propensity that many geographers have shown for attributing substantive interpretations to the slope coefficients in regression models should be greatly diminished by these results. For example, the value of the slope coefficient in distance decay models clearly reflects the zoning system as well as behaviour patterns. It is likely that more complex models, including entropy maximising spatial interaction models, will also suffer from similar effects as that displayed by these linear regression models. Currently, there is no evidence to the contrary.

If the slope coefficients can be made to vary then the performance of the models can also be made to vary by changing the zoning systems being studied. This has already been demonstrated by maximising or minimising correlation coefficients. The same effects can be observed for a different goodness of fit statistic. Table 13 shows maximum and minimum levels of model performance as measured by the mean absolute error for the Iowa regression models.

Table 13. Best and worst fit Iowa regression models

| number of zones | mean absolute error worst fit | best fit |
|---|---|---|
| 6 | 14.8 | .02 |
| 12 | 15.3 | .8 |
| 18 | 15.0 | .7 |
| 24 | 14.3 | 1.6 |
| 30 | 12.4 | 1.9 |
| 36 | 12.2 | 2.2 |
| 42 | 11.5 | 2.5 |
| 48 | 10.7 | 3.2 |
| 54 | 10.3 | 3.6 |

Figure 3 shows the geometry of two 12 zone systems that maximise and minimise the mean absolute error. In these experiments the objective function used in the AZP is the mean absolute error goodness of fit statistic and the model parameters are re-estimated using a robust line fitting procedure every time the zoning system changes. The range in results reported here is due solely to the nature of the zoning systems that are used.

It is now thought likely that no spatial model or method of analysis can escape the effects of the MAUP. It is also by no means certain that some methods or models will be better than others in their sensitivity to the MAUP. It would seem that the range of results tends to be data specific and that it may be impossible or unwise to try and make any general conclusions other than the observation that the MAUP is endemic to all spatially aggregated data and will affect all methods of analysis based upon such data. Its importance depends on the data and the aggregation factors involved.

### (v) Spatial calibration of a statistical model

One interesting, albeit mischievous, development is the use of the AZP to provide a uniquely geographical approach to estimating the unknown parameters in statistical models. The conventional approach to estimating the slope and intercept parameters in a linear regression model is as follows:

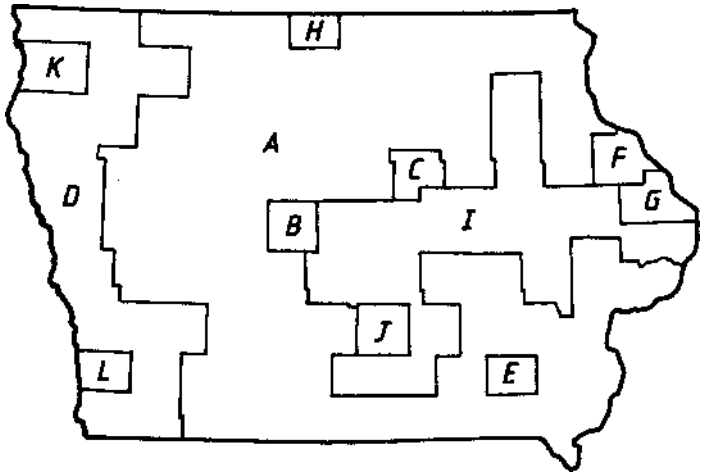



Figure 3 a. Zoning system that produces the worst possible fit (r = -.058, mean absolute deviation = 15.97, regression intercept is 60.325 and slope coefficient is -.287)

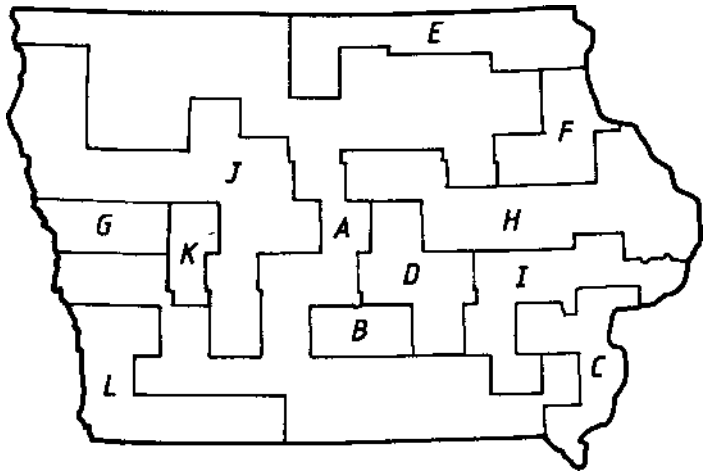



Figure 3 b. Zoning system that produces best possible fit (r = -.997, mean absolute deviation = .322, regression intercept is -9.054 and slope coefficient is 4.713)



(1) carefully specify a model; (2) select a 'good' parameter estimation procedure to yield unbiased estimates of the parameters; and (3) apply the model to a zonal data set. The choice of the latter whilst not toally haphazard (ie any data set) is usually based on a convenient data set (ie virtually any data set) for an arbitrary set of zones. This procedure is poor in its geography because of the heresy committed when the data are chosen. The results could well have a haphazard look about them since no attempt has been made to control the scale and aggregational variability inherent in the initial choice of a convenient data set and, in any case, no means are available for taking these aspects into account.

The analogous purely geographical alternative is to: (1) haphazardly pick some convenient values for the undetermined parameters; (2) fit the model by manipulating the data to fit by optimising the zoning system. The end result will be similar to the statistical approach except that the initially arbitrary parameter values may now have the properties of good estimators. This geographical approach contradicts the normal science paradigm as it is currently practised but perhaps this is a necessary violation if we are to escape from the bogus assumption of fixed zonal data. A statistician would regard this geographical approach as unscientific gerrymandering, an exercise in playing with numbers. But could any geographer possibly recommend the former statistical approach given its inability to control for the aggregational variability in zonal data? Both approaches are possible and ideally some means should be found to combine them.

Consider an example which demonstrates the potential power of the purely geographical approach. Suppose for the Iowa regression model it is decided to hold the parameters fixed at some completely arbitrary values; perhaps geographical theory or prior knowledge could be used to suggest sensible values. The objective is to fit the model by manipulating the zoning system.

Table 14. Spatial calibration of a linear regression model by seeking optimal 6,12, 18, 24, and 30 zone aggregations of the Iowa data

| target parameters | | zoning systems which fit a model to these parameters | zoning systems which produce data that yield the target parameters |
|---|---|---|---|
| intercept | slope | | |
| 41.46 | 2.00 | 12 18 | none |
| 41.46 | 1.75 | 12 18 24 30 | 18 24 30 |
| 41.46 | 1.50 | 6 12 18 24 30 | 6 12 18 24 30 |
| 41.46 | 1.25 | 6 12 18 24 30 | 6 12 18 24 30 |
| 41.46 | 1.00 | 6 12 18 | 6 12 18 24 30 |
| 41.46 | 0.75 | 6 12 18 24 | 18 |
| 41.46 | 0.50 | none | none |
| 60.0 | 1.25 | none | none |
| 50.0 | 1.25 | 12 | none |
| 40.0 | 1.25 | 6 12 18 24 30 | 6 12 18 24 30 |
| 30.0 | 1.25 | 12 18 | 12 18 24 30 |
| 20.0 | 1.25 | 12 | none |
| 10.0 | 1.25 | none | none |

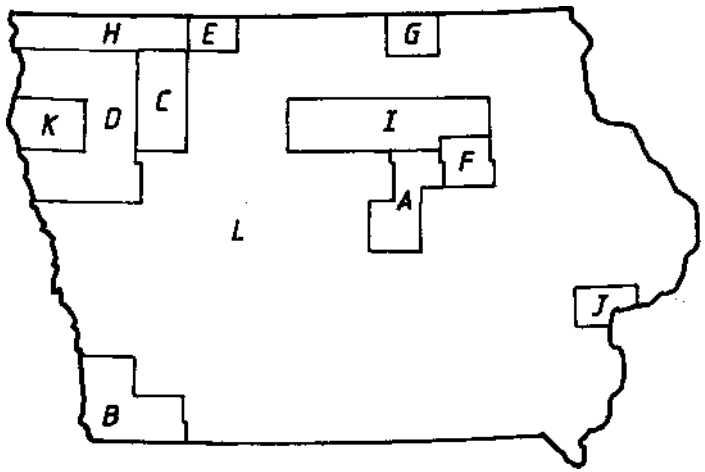



Figure 4a. Zoning system that fits a model with arbitrary intercept and slope of 41.4 and 2 (actual 42.4 and 1.90)

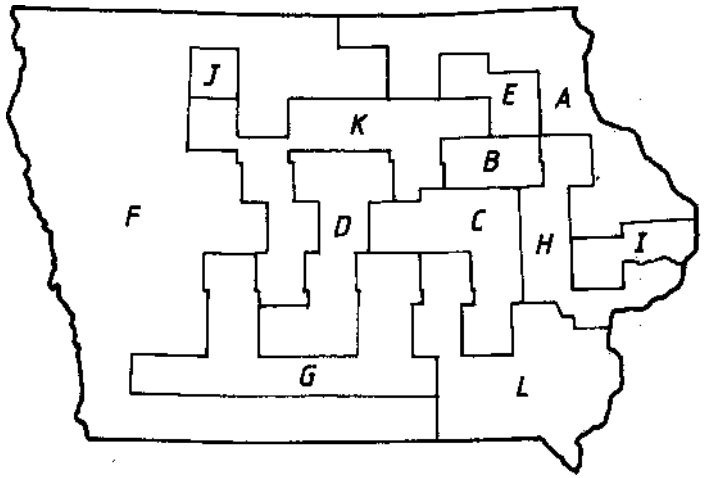



Figure 4b. Zoning system that fits a model with arbitrary intercept and slope of 41.4 and 1.0 (actual 42.0 and .98)



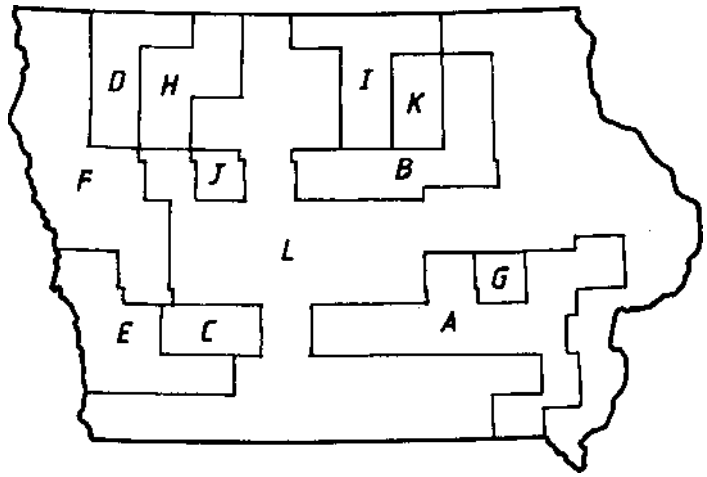



Figure 5a. Zoning system that fits model with arbitrary intercept and slope of 50 and 1.25 (actual 48.4 and 1.26)

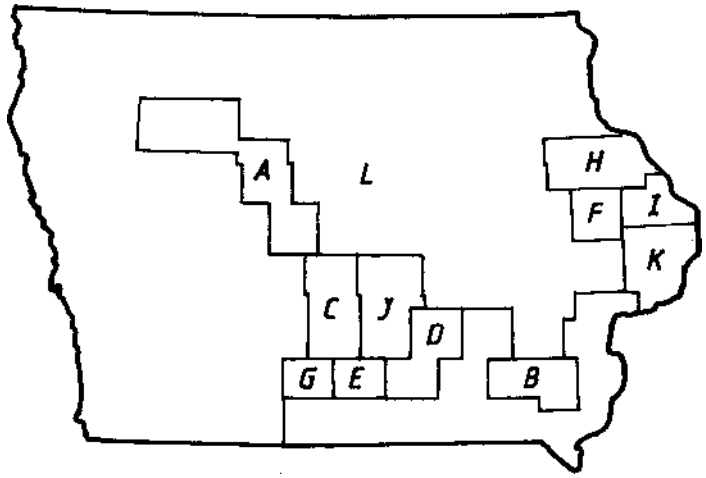



Figure 5b. Zoning system that fits model with arbitrary intercept and slope of 30 and 1.25 (actual 30.3 and 1.31)

Suppose that two sets of runs are performed; the first holds the intercept at the 99 zone level and systematically varies the slope coefficient; the second holds the slope coefficient at the 99 zone level and systematically varies the intercept. An alternative approach to fitting these models is to minimise the difference between the target parameters and the values estimated for a particular zoning system. Both sets of results are shown in Table 14 (page 27) with some of the zones being reproduced in Figures 4 and 5 (pp. 28, 29).

The decisions as to whether an acceptable level of fit is achieved are arbitrary. Nevertheless, it is suggested that quite reasonable levels of fit have been achieved. It is particularly noticeable that the 99 zone intercept and slope parameters (41.46 and 1.25) can be matched at all five levels of aggregation and that these zoning systems have zero aggregation effects. A robust data fitting procedure was used for Table 14. Similar results can be obtained for ordinary least squares regression, indeed rather more zoning systems would be judged to fit the target parameters.

One use of spatial calibration is to test specific geographical hypotheses about the nature of the results that may be expected; this is elaborated upon later. The argument here is that these empirical results demonstrate that the statistical and geographical aspects of spatial analysis need to be integrated. Zone design is in many ways a geographical complement to the statistical process of parameter estimation and with zonal data they cannot be separated if meaningful geographical results are to be obtained. This viewpoint is controversial since it implies that a large number of geographical studies are: (1) inherently non-geographical, (2) based on haphazard zoning systems with little direct control over aggregation effects; and (3) otherwise seriously flawed. The logic of this argument leads inextricably to a very different paradigm for spatial study than that currently used; this is examined later.

(vi) .. but do the optimal zoning systems look nice?

A final consideration concerns the nature of the optimal zoning systems shown in Figures 2 to 5. It can be argued, with some justification, that for reasons not yet investigated or understood, the aggregational properties of the 'real' zoning systems that geographers use are not as bad as the perverse optimal zoning systems that the AZP can identify. Perhaps the use of zones that look 'nice' or are based on regularly shaped units or convenient administrative definitions may avoid the extremes of the MAUP that have been identified in the various aggregation experiments. At the limits this is certainly true but the real problem is that the aggregational properties of nearly all ad hoc zoning systems are simply unknown. Additionally, it is difficult to establish any spatial benchmark against which the performance of alternative zoning systems can be measured. Geometric criteria, shape and size are not particularly relevant because it is the characteristics of the data and not the zones themselves that is important. The only absolute benchmark is the same data at a pre-aggregation or individual level and the characteristics of the latter are seldom known or available for analysis.

In principle it really does not matter what shape zones have since it is the relationship between zonal boundaries and the micro-level patterns which they detect and report that is the subject of spatial analysis. If the assumption of an isotropic plain were applicable then obviously a



geometrically regular set of zones at a carefully selected scale would be most relevant. However, given the very uneven, lumpy, and discontinuous nature of real world patterns it is not at all obvious as to why zoning systems should possess geometric regularity, and if they do what advantages this brings over the sorts of shapes described in Figures 2 to 5.

Likewise it is not apparent why neutral or locationally arbitrary areal units, for example grid-squares, should be of any interest in geography. Since we have the means to design zoning systems which are optimal for a given purpose, should we not be seeking to use these zoning systems as a means of investigating further the relationships under study. An analogy with a television aerial seems most appropriate. You could use an aerial designed for a radio and perhaps receive a poor picture. You could build your own to the most beautiful geometric design and get no picture at all. You could design an aerial to produce the best possible picture without worrying too much about aesthetics. The zone design problem is broadly analogous to an aerial. The zoning systems acts as a detector of spatial patterns and the patterns that are detected and their distinctiveness depend on its design. Surely no geographer can be content to use zoning systems produced by others or seek to use nice looking zones purely on aesthetic grounds without any regard for their performance as pattern detectors.

## V POSSIBLE SOLUTIONS

(i) No philosopher's stone

It is not thought likely that a general solution can be found that will allow existing methods to be used as if the MAUP did not exist. The problem is far too complex, it is difficult to investigate by analytical means, and its inherent geographical nature makes it unlikely that a statistical solution will emerge or if it does that it will suffice.

The simplest solution to the MAUP is to pretend it does not exist and hope that the results being produced for ad hoc zoning systems will still be meaningful or least interpretable. This view is implicit, by the lack of any explicit statements to the contrary, in much geographical work. For example, the performance of a mathematical model depends partly on its specification and partly on the zoning system that is used. There is often an elaborate body of theory to help with the model specification problem but little or no guidance is available to aid the choice of zoning system. Likewise many quantitative geography texts describe the existence of the MAUP but offer little or no advice as how best to use the techniques that are described to study data for modifiable units.

It is also fortuitous that ad hoc zoning systems often produce plausible results despite the neglect afforded to the careful definition of areal entities. However, it should be noted that the general absence of comparative studies may have helped disguise the extent to which zone-dependent regularities are being uncovered. The principal example sometimes quoted to demonstrate that the choice of zoning system is of little consequence is that of factorial ecologies where it seems that the major structural relationships between sets of social variables are relatively free from zoning effects. Whilst zonal invariance may be useful for some purposes, it is also slightly

worrying that so many social area analyses should be so similar despite cultural and other important differences. Perhaps a combination of closed number set problems and correlated denominators have combined to determine the results. It may also be that more sensitive methods and carefully engineered zoning systems would detect very different spatial patterns.

The problem is not that geographers have failed to realise that the MAUP exists, only that they do not know what to do about it. Perhaps mistakenly, they have opted to concentrate on the more tractable statistical problems presented by the analysis of spatial data whilst neglecting the more geographical ones. The pioneering work on spatial autocorrelation by Cliff and Ord (1973) and on space-time processes by Bennett (1979) are good examples. They provide elegant solutions to complex statistical problems concerned with the spatial, and temporal, dependency of zonal data but in so doing they deny the existence of the MAUP. For example, the expected moments of Cliff and Ord's spatial autocorrelation statistic can be computed under two different sets of assumptions, both of which assume that zonal data are fixed. Yet spatial autocorrelation is a characteristic of zonal data which is dependent on the choice of a particular zoning system. It can be varied by manipulating the zoning system.

A final consideration is that when geographers express concern about zoning systems it is mainly a reflection of problems of data comparability. It is suggested that the current naive approach depends on two major assumptions which are both incorrect. First, that the results will be substantially the same even if different areal units are used. A corollary of this argument would be that meaningful results can be obtained for virtually any set of arbitrary areal units; this view is widely held. The aggregation experiments reported in this section disproves this assumption. Second, that geographers have little or no control over the zoning systems for which data are available so that it is not practical to consider zoning systems as anything other than fixed. This is an over-simplification because it is always possible to seek to re-aggregate zonal data in order to find a 'better' set of areal units and thus recover from the effects of the initial aggregation. Why not exploit the modifiable nature of areal units rather than passively accept whatever zonal manipulations others perform on their behalf? Sadly, it seems that many geographers are happier if they do not know about the effects of the zonal manipulations that they or others perform.

(ii) Non-geographical solutions

The most convenient solution is to accept the normal science view that zoning systems should be independent of the phenomena they are used to report. This would allow the selection of areal units to be independent of the subsequent analysis, and would partly justify the status quo. However, this is at best an inherently non-geographical approach. The areal units being studied should be meaningful in some way which is relevant to the purpose of the study; therefore, it is argued that zoning systems cannot logically be independent of the phenomena they represent. In this context independence implies irrelevance. Nevertheless, a number of arbitrary zone design criteria have been suggested; for instance, approximate equality of population and zone shape compaction (Sammons, 1976; 1979); multiple design criteria (Masser and Brown, 1978); and information statistics (Batty, 1978; Batty and Sammons, 1978). However, it is not apparent why geographers should only be interested in areal units of a regular shape and size or in what way an information



statistic is an appropriate measure of the performance of a zoning system. More to the point, how do you decide which criteria to use? How do you know if its use is successful?

An example demonstrates the arbitrariness of these and other general purpose zone design criteria. Table 15 shows the effects on the Iowa correlation coefficient of the following:

(i) the equal area, population, and compaction criteria of Sammons (1976);

(ii) the spatial entropy statistic of Batty and Sammons (1978);

(iii) the minimum within-zone heterogeneity criteria of Cliff et al (1975);

(iv) the maximum independent variable variance criteria (Cramer, 1964; Hannan (1971);

(v) the maximum relative variation of the independent variable Blalock, 1964);

(vi) the minimum standard error of the regression slope coefficient (Williams, 1976).

All these criteria were formulated as objective functions for the AZP and solutions obtained.

Table 15. Effects of different zone design criteria on the Iowa correlation coefficient

| | number of zones | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| design criteria | 6 | 12 | 18 | 24 | 30 | 36 | 42 | 48 | 54 |
| equal area | .40 | .34 | .31 | .35 | .39 | .48 | .24 | .33 | .32 |
| equal population | .88 | .72 | .63 | .56 | .59 | .47 | .50 | .40 | .55 |
| equal density | -.03 | .71 | .52 | .52 | .53 | .53 | .53 | .46 | .56 |
| compact zones | .30 | .12 | .25 | .30 | .46 | .03 | .42 | .26 | .21 |
| spatial entropy | .90 | .21 | .26 | .28 | .54 | .26 | .33 | .43 | .46 |
| zonal homogeneity | .49 | .26 | .42 | .45 | .37 | .28 | .31 | .31 | .33 |
| independent variable variation | .64 | .50 | .42 | .39 | .54 | .44 | .42 | .38 | .33 |
| relative variation | .68 | .65 | .40 | .54 | .47 | .47 | .35 | .27 | .42 |
| standard error of slope | .99 | .99 | .97 | .97 | .95 | .93 | .90 | .85 | .81 |

Table 15 demonstrates that different criteria merely produce different results. At best some of the criteria reduce the systematic effects of scale but the levels of correlation largely reflect the nature of the criteria. Since the choice of criteria are arbitrary, then so too are the results. Worse still, the criteria are independent of any particular purpose so that the results are largely meaningless.

(iii) A traditional geographical solution

Looked at in another way, the MAUP is fairly trivial. All that is needed is for geographers to agree upon what constitutes the objects of geographical enquiry. The MAUP exists because of uncertainty as to what are the spatial entities which are being studied. Remove that uncertainty and the problem disappears. Unfortunately, this task of identifying meaningful geographical entities is a difficult one for many geographers to face because of the traditional regional geography connotations. Additionally, different definitions will be needed for different purposes.

The best examples of this approach have been the use of functional region definitions of urban areas for studying census data (Spence et al, 1982; Coombes et al, 1982). The justification here is that local authority definitions are best provided by functional region definitions. This solution clearly works well only when there is sufficient geographical knowledge to define with a high degree of precision the sorts of areal units that are most sensible for a particular purpose. There are many areas in geography where it cannot be applied. Furthermore, this approach only removes the aggregational uncertainty, the effects of the MAUP still survive and condition the results.

(iv) Towards a new methodology for spatial study

Once it is accepted that the results of studying zonal data depend on the particular zoning system that is being used, then it is no longer possible to continue using a normal science paradigm. The data are not fixed so that the results depend, at least in part, on the areal units that are being studied; units which are essentially arbitrary and modifiable. The selection of areal units, or zoning systems, cannot therefore be separate from, or independent of, the purpose and process of a particular spatial analysis; indeed it must be an integral part of it. This view conflicts with the current use of scientific methods and statistical techniques in geography, and for these reasons many geographers would refuse to consider it to be a viable proposition.

Let us continue with the heresy a little longer. The problem is to invent a new paradigm for spatial study which can explicitly handle the geography of the MAUP. The most obvious approach is to reverse the normal science paradigm. Instead of meekly accepting whatever result the choice of a haphazard zoning systems happens to produce, it is necessary to start by specifying precisely what outcome is expected. This can take the form of a hypothesis. If the desired result can be attained by solving the associated automatic zoning problem, that what are the limits on both the range of results and the range of zoning systems that produce similar outcomes? What if anything does the geography of these optimal zoning systems tell us about the hypothesis being studied? If the desired result cannot be attained without violating either statistical assumptions or geographical factors, then the associated hypothesis must be rejected.

The following methodology is suggested as being appropriate for a geographical solution to the MAUP.

STEP 1. Define the purpose of the study in an explicit fashion. This can be done by speculating as to what outcome is expected given prior knowledge or what outcome is desired. For example, does a model that fits data in Nevada also work in Iowa? This desired result would be expressed as a hypothesis and set up as an objective function for the AZP. For example, to find out if a model fits the Iowa data, minimise the model errors using the AZP. If the question is whether a particular set of parameters can provide an acceptable level of performance then again solve the associated AZP.

STEP 2. Try to obtain the desired result by identifying zoning systems which approximately optimise the appropriate objective function using the AZP. For example, minimise the differences between a set of target factor loadings and values produced for a particular zoning systems; the purpose here might be to investigate whether a set of social area analysis results for one area also apply to another.



STEP 3. Decide what the results mean in a statistical sense, if this is appropriate, as well as in terms of the geography of the optimal zoning systems. Has the target result(s) been achieved with a tolerable degree of error? If not, then the associated hypothesis must either be rejected or changed, so return to STEP 1. If the results are acceptable, then have any important statistical assumptions been violated? If constraints are needed then go to STEP 4. What does the zoning system tell us about the geography of the study area? The zoning system makes visible the interaction between the data being aggregated and the hypothesis being studied and a study of the nature of the zones may be very useful. How did the AZP optimise the objective function? Is there a trivial spatial solution? If the number of zones are changed what effect does this have? Finally, are the optimal zoning systems satisfactory from a geographical point of view?

STEP 4. It may be necessary to introduce constraints to impose restrictions on either the nature of the zones or on the properties of the data they generate. These constraints are in addition to the usual contiguity restrictions necessary to ensure that the zones are internally connected. The AZP can handle either equality or inequality constraints. These additional constraints represent a potentially important interface between the geography and statistics of spatial study; for example, constraints to ensure zero spatially autocorrelated data and a maximum zone size. The feasibility of the additional constraints is partly related to the aggregation factors involved; when large numbers of zones are being aggregated then a large number of complex constraints can often be satisfied.

STEP 5. Now solve the constrained automatic zoning problem. If a satisfactory result is found then return to STEP 3 for interpretation. If not, then examine the consequences of failing to satisfy some or all of the constraints. The extent to which various constraints can or cannot be satisfied may also provide useful information about the nature of the problem under study. In most cases an iterative process of experimentation is probably needed.

Clearly the statistical power of this new approach is considerably less than that promised by conventional methods. Hypothesis testing is used here as a device for introducing an explicit purpose into the process of spatial study: This is necessary so that whatever zonal entities are identified they should be both purpose related and geographically meaningful. The new paradigm is likely to be most useful when comparative studies are being performed.

Consider the previous correlation analysis for Iowa. Simply reporting the level of correlation is not very useful because the result is zone-dependent. Similarly, it is no use testing the null hypothesis that the correlation coefficient is significantly different from zero. Consider the related linear regression model. If no prior information other than a model specification is available, then a wide range of different results can be obtained depending on the choice of zoning system. However, as we move through the new paradigm, the introduction of various statistical and geographical constraints reduces the range of alternatives, although there may still be a number of different results that require interpretation and explanation.

For example, suppose we wish to see whether a correlation of 0.8 between old age and Republican voters reported from some other study area also

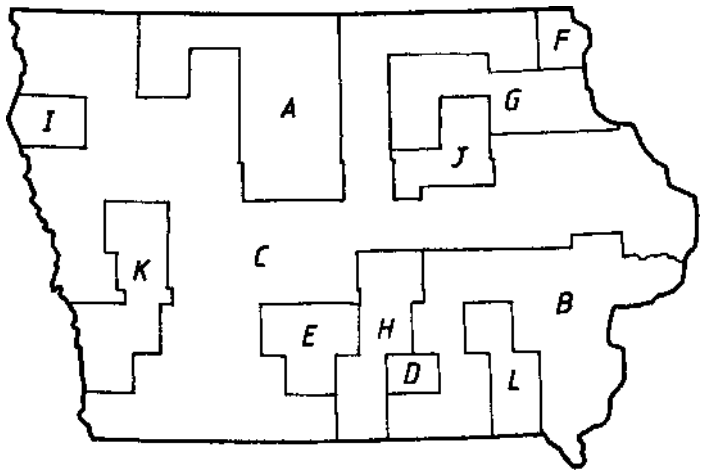



Figure 6a. Zoning system that minimises the correlation coefficient subject to constraints (r = -.928, intercept = 5.0, slope = -3.12, spatial autocorrelation of residuals = 0.0, spatial autocorrelation of independent variable = 0.0, homoscedasticity = 0.0)

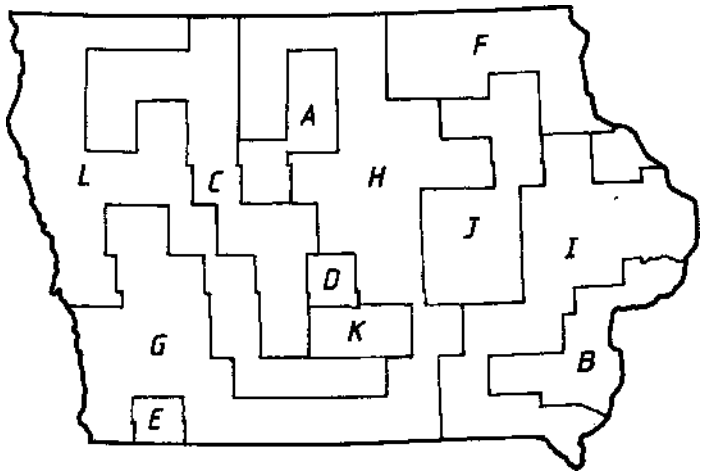



Figure 6b. Zoning system that maximises the correlation coefficient subject to constraints (r = .993, intercept = -8.598, slope = 4.654, spatial autocorrelation of residuals = 0.0, spatial autocorrelation of independent variable = 0.0, homoscedasticity = 0.0)



occurs in Iowa. Suppose also that you want the associated linear regression model to satisfy various statistical assumptions; specifically, that the residuals have zero spatial autocorrelation, that the spatial autocorrelation of the predictor variable is zero, that the mean residual is zero, and that the rank correlation between the absolute residuals and the independent variable is also zero (a residual homoscedasticity constraint). The zoning systems shown in Figure 6 satisfy these assumptions and the range of correlation is still large, between -.928 and +.993. All we can conclude from this is that there is so much aggregational variability in these data that the results are not meaningful despite the undoubted high degree of statistical significance. Clearly additional constraints are needed and there should be some basis for these restrictions. For example, if zonal population sizes are restricted to about plus or minus 15 per cent of the average, then the range of correlations is reduced to between .28 and .94. If area is used instead of population then the range is slightly wider; -.06 to .81. The problem here is that there is no real basis for these size constraints. The best strategy would be to combine the Iowa data with another data set ( viz for which the correlation of 0.8 was obtained). The AZP would be used to identify optimal zoning systems for both data sets simultaneously (the contiguity constraints would keep their zones apart). The resulting map patterns for a global correlation of 0.8 could then be examined.

The idea then is to use the optimal zoning approach to test hypotheses by manipulating the aggregation process. Instead of asking whether a result obtained in study area A is different from a result for study area B, it is necessary to consider the range of results that can be produced for both A and B. Instead of trying to fit a model to an arbitrary zoning system, it is necessary to consider which zoning systems provide the best results and to consider what properties they, or the aggregated data they produce, should have. The map patterns produced by optimal zoning systems for particular purposes may themselves contribute to the spatial analysis process.

## VI CONCLUSIONS

It has been argued that the MAUP is a fundamental geographical problem that is endemic to all studies of spatially aggregated data. It is a geographical fact of life that the results of spatial study will always depend on the areal units that are being studied. This being so it is time that geographers started to develop methods of analysis capable of accommodating and even exploiting this situation. One possible statistical approach could be based on the role of sampling in statistical inference with the aim of developing further the obvious similarities between the operations of sampling and zoning. However, the results reported earlier indicate that the analogy is a poor one and that there are geographical obstacles to developing it further; for example, we may have to use random zoning systems. Instead it has been argued that the MAUP is fundamentally a geographical phenomenon that is most unlikely to be solved by geographers who are blinkered by both a statistical perspective and fervent adherence to a paradigm that denies the very existence of the problem.

This CATMOG has described the first shaky steps in the development of a new methodology for spatial study which is explicitly based on and around the purposeful and deliberate engineering of zoning systems. This is viewed

as having the potential to open up an entirely new approach to the study of spatial data as well as offering a general methodological framework into which any existing model or technique can be incorporated. It is argued that this constitutes the beginning of a new era which will be characterised by the development of more relevant and more appropriate core of geographical analysis techniques. It would seem that the adoption of an exceptionalist position is a basic prerequisite for this development to take place.

Critics will argue that the 'cure' in the form of AZP appears to be no better than the disease and will inevitably result in difficulties in making generalisations outside of a particular zoning system for a particular data set. The answer to this latter problem is straightforward. All that need be done is to incorporate several different data sets in the same AZP problem formulation. They will remain separate entities by virtue of having no contiguity links but they will be linked through the definition of global constraints (other than contiguities) and through a common objective function. The answer to the first point is self-evident. The widespread and serious impact of the MAUP on spatial study has been convincingly demonstrated so it is no longer possible to simply ignore it. Thus it would seem that methods which cannot cope with the MAUP should not be used. Currently there are no convincing alternative methods for handling spatially grouped data in a statistically sound framework. So why not investigate more radical non-statistical frameworks and what could possibly be better for a geographer than a purely geographical approach?

The consequences of seriously accepting this challenge may well be fundamental changes in the manner by which geographers analyse spatial data. There has to be an admission of an approach to spatial study that is tantamount to operating the normal science paradigm in reverse. This is clearly non-scientific according to any contemporary liberal definition. It would seem that while few geographers would question the utility of using the AZP to identify ranges of possible results due to the MAUP, few have so far shown any enthusiasm for going any further let alone consider the unimaginable horrors of scientific heresy. However, it is likely that the former will inexorably lead to the latter. There have been paradigm shifts before in science so why not a new one designed specially for geographers? It should be appreciated that the AZP and its associated methodology offers as yet the only practical working solution to the MAUP. There can be no real doubts about its geographical nature but perhaps it is too geographical for many modern geographers.

It is suggested therefore that the prospect is gradually dawning that the MAUP is not so much an insoluble problem but rather a powerful analytical tool ideally suited for probing the structure of areal data sets. The growing speed of computers opens up the tremendous potential offered by heuristic solution procedures, such as the AZP, to identify the most appropriate zoning systems for any particular purpose without having to solve currently intractable theoretical and analytical problems. That is to say, we do not as yet fully understand the problem and we are certainly no way near to being able to develop a calculus to handle it, but the problem can be solved or turned around using what are essentially Monte Carlo optimisation methods. Currently much can be done with small data sets and fairly complex models or with larger data sets and simple models. Very soon it will be possible to routinely apply the same methods to any spatial data set and any model or function, no matter how complex. When this happens often enough then a new geographical revolution will surely have occurred.



## BIBLIOGRAPHY


Batty, M. (1978, Speculations on an information theoretic approach to spatial representation. in: *Spatial representation and spatial interaction,* (eds) I. Masser and P.J.B. Brown, (Martinus Nijhoff; Leiden), pp 115-147.

Batty, M. and Sammons, R. (1978), On searching for the most informative spatial pattern. *Environment and Planning A,* 10, pp 747-749.

Batty, M. and Sikdar, P.K. (1982), Spatial aggregation in gravity models. 1. An information-theoretic framework. *Environment and Planning A,* 14, pp 377-405.

Bianchi, G., Openshaw, S., Scattoni, P., Sforzi F. and Wymer, C. (1981), Analisi dell'area sociale: comparazione delle classificazioni condotte su dati medi per sezioni di censimento e su data individuali. (Paper presented at 2nd Italian Regional Science Conference, Napoli, October 19th-21st.

Bennett,R.J. (1979), *Spatial time series: analysis, forecasting and control,* (Pion: London).

Blalock, H. (1964), *Causal inferences in nonexperimental research,* (University of North Carolina Press: Chapel Hill).

Borgatta, E.F. and Jackson, D.J. (1980), *Aggregate data: analysis and interpretation,* (Sage Publications: Beverly Hills).

Chapman, G.P. (1977), *Human and environmental systems: a geographer's appraisal,* (Academic Press; New York).

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

Cliff, A.D. and Ord, J.K. (1975), Model building and the analysis of spatial pattern in human geography, *Journal of the Royal Statistical Society Ser B,* 37, pp 297-348.

Cliff, A.D., Haggett, P., Ord, J.K., Bassett, K. and Davies, R. (1975), *Elements of spatial structure: a quantitative approach,* (Cambridge University Press: London).

Coombes, M.G., Dixon, J.S., Goddard, J.B., Openshaw, S. and Taylor, P.J. (1982), Functional regions for the population census of Great Britain, in: *Geography and the Urban Environment,* (eds) D.T. Herbert and R.J. Johnston, (Wiley: London), 5, pp 63-112.

Coombes, M.G. and Openshaw, S. (1982), The use and definition of travel to work areas in Great Britain: some comments, *Regional Studies,* 16, pp 141-149.

Cramer, J.S. (1964), Efficient grouping, regression, and correlation in Engel curve analysis. *Journal of the American Statistical Association,* 59, pp 233-250.

Evans, I.S. (1981), Census data handling. in: *Quantitative Geography: a British View,* (eds) N. Wrigley and R.J. Bennett, (Routledge and Kegan Paul: London), pp 46-59.

Gehlke, C.E. and Biehl, H. (1934), Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association, Supplement,* 29, pp 169-170.

Griffith, D.A. (1980), Towards a theory of spatial statistics, *Geographical Analysis,* 12, pp 325-339.

Hannan, M.T. (1971), *Aggregation and disaggregation in sociology,* (Lexington Books: Lexington Mass.).

Johnston, R.D. and Rossiter, D.J. (1982), Constituency building, political representation and electoral bias in urban England. in: *Geography and the Urban Environment,* (eds) D.T. Herbert and R.J. Johnston, (Wiley: London), 5, pp 113-156.

Keans, M. (1975), The size of the region-building problem. *Environment and Planning A,* 7, pp 575-577.

Masser, I. and Brown, P.J.B. (1978), *Spatial representation and spatial interaction,* (Martinus Nijhoff: Leiden).

McNeil, D.R. (1977), *Interactive data analysis,* (Wiley: London).

Openshaw, S. (1977a), A geographical solution to scale and aggregation problems in region-building, partitioning, and spatial modelling. *Transactions of the Institute of British Geographers, New series,* 2, pp 459-472.

Openshaw, S. (1977b), Algorithm 3: a procedure to generate pseudo-random aggregations of N zones into M zones, where M is less than N'. *Environment and Planning A,* 9, pp 1423-1428.

Openshaw, S. (1977c), Optimal zoning systems for spatial interaction models. *Environment and Planning A,* 9, pp 169-184.

Openshaw, S. (1978a), An empirical study of some zone design criteria. *Environment and Planning A,* 10, pp 781-794.

Openshaw, S. (1978b), An optimal zoning approach to the study of spatially aggregated data. in: *Spatial representation and spatial interaction,* (eds) I. Masser and P.J.B. Brown, (Martinus Nijhoff: Leiden).

Openshaw, S. (1981), Le problem de l'aggregation spatiale en geographie. *L'espace Geographique,* 1, pp 15-24.

Openshaw, S. (1983), Ecological fallacies and the analysis of areal census data. *Environment and Planning A,* (forthcoming)

Openshaw, S. and Taylor, P.J. (1979), A million or so correlation coefficients: three experiments on the modifiable areal unit problem. in: *Statistical methods in the spatial sciences,* (ed) N. Wrigley, (Pion: London), pp 127-144.

Openshaw, S. and Taylor, P.J. (1981), The modifiable areal unit problem. in: *Quantitative geography: a British View,* (eds) N. Wrigley and R.J. Bennett, (Routledge and Kegan Paul: London), pp 60-70.

Robinson, A.H. (1950), Ecological correlation and the behaviour of individuals. *American Sociological Review,* 15, pp 351-357.

Sammons, R. (1976), *zoning systems for spatial models,* (Redding Geographical Paper 52, Department of Geography: Reading University).

Sammons, R. (1979), Zone definition in spatial modelling. in: *Resources and Planning,* (eds) B. Goodall and A. Kirby, (Pergamon: Oxford), pp 77-100•



Silk, J. (1979), *Statistical concepts in geography,* (Allen and Unwin: London).

Spence, N., Gillespie, A., Goddard, J.B., Kennett, S. Pinch, S. and Williams A. (1982), *British cities: an analysis of urban change,* (Pergamon: Oxford)

Taylor, P.J. (1977), *Quantitative methods in geography,* (Houghton Mifflin: Boston).

Taylor, P.J. and Johnston, R.J. (1979), *Geography of elections,* (Penguin: Harmondsworth).

Tukey, J.W. (1977), *Exploratory data analysis,* (Addison-Wesley: Reading, Mass.).

Williams, I.N. (1976), Optimistic theory validation from spatially grouped regression: theoretical aspects. *Transactions of the Martin Centre,* 1, pp 113-145.

Williams, I.N. (1979), Some implications of the use of spatially grouped data. in: *Towards the dynamic analysis of spatial systems* (eds) R.L. Martin, R.J. Bennett, and N.J. Thrift, (Pion: London), pp 53-64.

Yule, G.U. and Kendall, M.G. (1950), *An introduction to the theory of statistics,* (Griffin: London).