Environment and Planning A, 1984, volume 16, pages 17-31

Ecological fallacies and the analysis of areal census data

S Openshaw
Department of Geography, University of Newcastle upon Tyne, Newcastle upon Tyne NE1 7RU, England
Received 7 June 1982; in revised form 23 December 1982

Abstract. In many countries census data are only reported for areal units and not at the individual level. This custom raises the spectre of ecological fallacy problems. In this paper, a 10% sample census (from the United Kingdom) and individual census data (from Italy) are used to provide an empirical demonstration of the nature and magnitude of these problems. It is concluded that ecological fallacy effects are endemic to areal census data, although their magnitude is perhaps not as large as might have been expected. The principal difficulty is that there is at present no way of predicting in advance the degree of severity likely to be associated with particular variables and particular techniques. Finally, a suggestion is made concerning how the potentially serious practical consequences can be reduced.

1 Introduction

In the United Kingdom, census data which describe the characteristics of individual persons and households are only available in an aggregate form for what are essentially arbitrary geographical areas. The areal units used to report census data (enumeration districts, census tracts, wards, local government units) have no natural or meaningful geographical identity. Census agencies have been seemingly very slow to realise, and geographers have often failed to point out, that census statistics can be biased as much by the geographical boundaries that are used to report them as by questionnaire design and the code books used to categorise the responses. These latter aspects are routinely investigated by the census agencies responsible for the collection of census data, but the former problem is ignored. Yet this problem is particularly important to many users of census data, who often, directly or indirectly, make inferences about the characteristics of individual households and the associations between them on the basis of crosstabulated frequency counts as reported for various sets of arbitrary areal units.

Census geography is therefore a subject of considerable practical relevance to many users of census data, although it has attracted little academic interest. Important geographical issues, such as the purposeful definition of meaningful areal units for reporting census data and the magnitude of aggregational biases and ecological fallacies that characterise particular sets of areal definitions, have not been thoroughly investigated or indeed subjected to much study. With the absence of relevant research it has been far too easy for census planners to use whatever areal units are most convenient to them without any concern for the underlying spatial aggregation problems.

This neglect is surprising because many of the basic problems associated with the analysis of aggregated census data have been recognised for a long time (Gehlke and Biehl, 1934; Robinson, 1950; Blalock, 1964; Hannan, 1971; Clark and Avery, 1976; Openshaw, 1977). Gehlke and Biehl (1934, page 170) have asked "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." Robinson (1950) provides empirical evidence that this extreme result

can in fact occur. Indeed, it is now known that the modifiable nature of areal units can be systematically exploited by heuristic procedures to produce a very wide range of different results, irrespective of what individual-level analysis would have produced (Openshaw, 1978a; 1978b; 1981; Openshaw and Taylor, 1979; 1981). These extreme results demonstrate the approximate range of aggregational variability inherent in areal data, but they may not be typical of the range of results likely to be produced for the restricted sets of areal units commonly used to report census data.

The significance attached to these various problems depends on the purpose behind the study. Is it to examine the characteristics of areas or is it to infer something about the characteristics of the individuals who live there? If statistical techniques are being used, the analyst should decide whether the underlying models of interest relate to the individual level or to an aggregate zonal level. If it is the former, there is no theoretical guarantee that it is possible to obtain 'good' parameter estimates for a model specified at the individual level using data from a higher level of aggregation. Very little is known about the loss of efficiency that may ensue. If the interest is in areal models, there are problems resulting from the modifiable nature of the areal units; that is, any statistical relationship may be manipulated, either intentionally or otherwise, by the choice of areal units. This also affects crossaggregation estimation. One problem is that many users of census data do not have a clear idea of what it is that they are studying and tend to mix both approaches. Another is that the aggregational properties of the various census areas are unknown and, moreover, may be variable specific and spatially invariant. Furthermore, these areal units, though neither neutral nor meaningful entities, are exogenous to all subsequent uses of the data.

These problems are potentially very serious and they directly affect the usefulness of census data. In theory they are easy problems to study since all that is required is access to spatially referenced individual data and a fast computer. However, in practice there are numerous problems of both a technical and a political nature. For instance, in the United Kingdom the 1922 Census Act prohibits the release of census data about identifiable individuals. This Act has been interpreted by civil service administrators as precluding the release of any individual data, even for anonymous individuals[1]. Ideally, these studies could be done by the Office of Population Censuses and Surveys (OPCS) without any breach of confidentiality, but the necessary resources have not yet been made available. Only after 100 years have elapsed can individual census data be studied outside of the OPCS, but the available nineteenth-century data are not particularly useful for studying problems associated with the 1981 census. Likewise, a small sample of microdata, along the lines of the US Bureau of the Census Public Use Sample, is of very limited value for aggregation research.

As a result there is often no readily available means whereby users of census data can determine whether the results, hypotheses, and conclusions obtained from the analysis of areal census data are applicable at the individual level or whether they are a characteristic of the areal units being studied. This is an important problem because individual-level inferences tend to be implicit in many applied uses of census data; for example, the identification of problem areas for planning purposes, the use of a spatial classification to identify particular client groups in marketing, and the use of areal data by sociologists to generate hypotheses at the individual level. Many of these inferences occur in descriptive studies, in which it is very easy to confuse the characteristics of areas with the characteristics of people who live there.

[1] A strange but limited exception is the longitudinal sample of individuals being studied at a gross spatial scale (Goldblatt and Fox, 1978). There are also rumours about the release of individual data from the 1971 census for Northern Ireland.

In this paper, an attempt is made to illustrate some of these problems through a series of empirical experiments based on two large individual data sets. Probably the best available and most relevant UK data are a 10% random sample survey of all households in Sunderland. This was undertaken by Tyne and Wear County Council as a substitute for the 1976 census. A subset of these data was made available under various confidentiality restrictions. These data could be aggregated into polling districts, 1 km grid-squares, and 500 m grid-squares. The second data set is a 100% census survey of all households in Florence, Italy for 1971. It consists of 122342 household records which could be aggregated into 484 enumeration districts. These data were collected for the Regional Government of Tuscany at the same time as the official Italian census organised by the Italian Government's census and statistical agency, ISTAT. Whereas the data used here are not the same as the official census data, they are probably as good and the range of variables is larger.

The Italian data are far more comprehensive than the Sunderland data, although fewer small-area aggregations could be examined. My object in studying them both is to allow differences due to the peculiar nature of Italian census data to be identified. It is hoped that the various analyses performed on the Sunderland and Florence data sets will help provide some indication of the scale of any ecological fallacy problems that may be present and suggest conclusions relevant to the analysis of 1981 census data both in Italy and in the UK.

In section 2 of the paper, the Robinson (1950) correlation analyses are replicated for a wide range of variables and for smaller areal units than studied by him. In section 3, the effects of ecological fallacies on factorial studies based on individual and ecological correlation coefficients are examined. The effects on some simple regression models are considered in section 4, and in section 5 individual and spatial classifications of the same data are compared. Finally, in section 6, some conclusions drawn from the empirical studies are presented.

2 Correlation analysis

2.1 *Individual versus ecological correlations*

In a now famous paper, Robinson (1950) demonstrated that individual relationships cannot be inferred from ecological correlations based on areal data. He wrote: "there need be no correspondence between the individual correlation and the ecological correlation" (page 354). An individual correlation is one based on variables which measure the properties of indivisible objects, such as persons or households, whereas in an ecological correlation the object being studied is a group of persons living in a census enumeration district or some other areal unit and the variables are descriptive properties of areas rather than of individuals. Alker (1969) identified a number of different types of ecological fallacy that can arise from the analysis of aggregate data.

In general there is a notable absence of empirical study vis-à-vis theoretical speculation. Robinson (1950) may well have identified extreme rather than typical ecological–individual differences. His conclusions were, after all, based on the comparison of two correlation coefficients computed at the individual level and for the USA divided into eight census areas and by States, both of which are gross levels of spatial aggregation. The question arises as to what might be the typical levels of differences for a larger number of variables for those levels of aggregation most commonly used in census studies. A number of indicator variables could be computed both for the Sunderland and for the Italian data sets. A set of fifty-three variables is used to describe the 8483 households in the Sunderland data set and a set of forty variables is used for the 122342 households in the Florence data set.

The simplest demonstration of the differences between ecological and individual correlations is to crosstabulate both sets of correlations. Pearsonian correlation coefficients are used for both data sets and are comparable statistics. At the individual level the correlations are calculated from the dichotomous data; at the various areal levels these dichotomous measurements give way to rates per 10000 after aggregation. The results are shown in table 1.

The scatter of correlation coefficients gives an immediate impression of the nature and magnitude of the problem. The spread of frequencies either side of the diagonal cells provides an indication of the differences due to aggregation, although of course the results are a function of category size. An examination of the row and column frequencies shows how the distribution of the individual correlations is far more concentrated than the ecological correlations, suggesting that areal aggregation has a pronounced flattening effect. It is also noticeable that, for the range of scales examined in table 1, large differences are not very common.

The results for table 1 are repeated in table 2, but smaller grid-size categories are used. The shift away from the no-change or diagonal cells is now very pronounced. The rule is simply that areal aggregation tends to make correlations stronger, with the largest differences being recorded for those individual correlations close to zero. The effects of scale can also be observed. The Sunderland data can be aggregated to thirty-six polling districts or disaggregated into 347 grid-squares of 500 m side, compared with 117 grid-squares of 1 km side used in tables 1 and 2. Obviously, as the zones become smaller and thus more homogeneous, so the differences between the ecological and individual correlations decrease. Table 3 provides a crosstabulation of the individual and ecological correlations for polling districts and 500 m grid-squares. These results suggest that there may well be a critical scale of areal aggregation and/or a particular type of spatial aggregation that will best approximate the individual correlation values. There could be some benefits to users of census data if it were

Table 1. Crosstabulation of individual and ecological correlation coefficients (percentages of row totals).

Individual correlations: from/to	Areal correlations: from/to										Total
	−1.0/ −0.8	−0.8/ −0.6	−0.6/ −0.4	−0.4/ −0.2	−0.2/ 0.0	0.0/ 0.2	0.2/ 0.4	0.4/ 0.6	0.6/ 0.8	0.8/ 1.0	
Sunderland 1 km grid-squares (53 variables)											
−1.0/−0.8	**100**										1
−0.8/−0.6	50	**50**									4
−0.6/−0.4	12	44	**32**	12							25
−0.4/−0.2		9	36	**34**	15	4	1				180
−0.2/0.0			4	32	**39**	18	5	1			997
0.0/0.2			1	2	14	**29**	32	20	3		188
0.2/0.4						14	**32**	39	14		28
0.4/0.6							17	**50**	17	17	6
0.6/0.8								50	**50**		2
Totals	6	32	117	387	444	248	117	66	13	1	
Florence enumeration districts (40 variables)											
−1.0/−0.8	**100**										1
−0.8/−0.6		**0**									0
−0.6/−0.4			**100**								2
−0.4/−0.2	2	19	31	**24**	17	6					83
−0.2/0.0		1	7	21	**32**	23	14	2			603
0.0/0.2			1	6	10	**28**	28	22	3		78
0.2/0.4							**18**	27	55		11
0.4/0.6								**100**			1
0.6/0.8									**100**		1
0.8/1.0											0
Totals	3	21	72	154	214	167	106	33	10	0	

possible to identify as an alternative to the current use of arbitrary areal units, an optimal level of spatial resolution which minimised these differences. Although it may be difficult to identify optimal sets of areal units suitable for general use, these

Table 2. A more detailed crosstabulation of individual and ecological correlations (percentages of row totals).

Individual correlations: from/to	Areal correlations: from/to															Total
	−0.7/ −0.6	−0.6/ −0.5	−0.5/ −0.4	−0.4/ −0.3	−0.3/ −0.2	−0.2/ −0.1	−0.1/ 0.0	0.0/ 0.1	0.1/ 0.2	0.2/ 0.3	0.3/ 0.4	0.4/ 0.5	0.5/ 0.6	0.6/ 0.7	0.7/ 0.8	
Sunderland 1 km grid-squares (53 variables)																
−0.4/−0.3	29	29	6	**29**												17
−0.3/−0.2	6	18	18	29	**8**	6	14									49
−0.2/−0.1	2	2	10	22	25	**17**	10	5	1	5	1	1				131
−0.1/0.0			1	6	17	27	**25**	11	7	3	1	1	1			575
0.0/0.1					5	11	20	**23**	14	16	6	3	1	1		422
0.1/0.2					1	1	6	11	**18**	15	15	14	11	3	3	125
0.2/0.3					2	0	2	6	8	**13**	14	22	22	10	0	63
0.3/0.4										18	**12**	29	24	6	12	17
Totals	12	20	33	84	158	229	260	184	129	119	61	56	44	22	9	
Florence enumeration districts (40 variables)																
−0.4/−0.3			50	**50**												2
−0.3/−0.1	11	28	6	28	**11**	6	6									18
−0.4/−0.1	6	8	9	22	17	**9**	14	6	5	3						65
−0.1/0.0		2	8	8	16	17	**19**	14	8	4	4	0	1			255
0.0/0.1				1	4	10	13	**17**	16	15	12	8	3			348
0.1/0.2			2	0	2	5	7	7	**21**	13	13	12	8	8	2	61
0.2/0.3										**6**	18	24	18	24	6	17
0.3/0.4											**13**	13	0	13	38	8
Totals	6	15	29	43	67	87	108	106	94	73	65	41	19	14	7	

Table 3. Crosstabulation of individual and ecological correlations using polling districts and 500 m grid-squares for Sunderland (percentages of row totals).

Individual correlations: from/to	Areal correlations: from/to										Total
	−1.0/ −0.8	−0.8/ −0.6	−0.6/ −0.4	−0.4/ −0.2	−0.2/ 0.0	0.0/ 0.2	0.2/ 0.4	0.4/ 0.6	0.6/ 0.8	0.8/ 1.0	
Polling districts											
−1.0/−0.8	**100**										1
−0.8/−0.6	75	**0**	25								4
−0.6/−0.4	32	32	**20**	12	4						25
−0.4/−0.2	7	27	31	**16**	14	4	0	1			180
−0.2/0.0		4	14	24	**25**	16	11	3	2		997
0.0/0.2				3	9	**19**	26	28	14	1	188
0.2/0.4				4	4	7	**7**	32	46		28
0.4/0.6							17	**0**	67	17	6
0.6/0.8									**50**	50	2
0.8/1.0										**0**	0
Totals	26	93	208	281	295	209	157	96	61	5	
500 m grid-squares											
−1.0/−0.8	**100**										1
−0.8/−0.6	75	**25**									4
−0.6/−0.4	4	52	**40**	4							25
−0.4/−0.2		3	37	**47**	12	1					180
−0.2/0.0			1	24	**57**	17	2				997
0.0/0.2				1	7	**43**	39	9	1		188
0.2/0.4							**46**	50	4		28
0.4/0.6								**67**	33		6
0.6/0.8									**100**		2
0.8/1.0										**0**	0
Totals	5	20	86	321	607	248	102	36	6	0	

difficulties cannot justify the continued use of areal systems which are basically haphazard, in that nothing explicit is optimised.

2.2 *Can ecological correlations be corrected for aggregation effects?*
Goodman (1959) describes a correction procedure that he thought could be applied to ecological correlations between variables exhibiting a strong linear relationship. A related but simpler technique is to regress the ecological and individual correlations to obtain a statistical model that may be used to predict individual correlations from the values of ecological correlations. When applied to the Sunderland data this resulted in an intercept of zero, and slope coefficients of 0.25 for polling districts, 0.37 for 1 km grid-squares, and 0.52 for 500 m grid-squares. The associated r^2 values were 0.48, 0.54, and 0.75, respectively. The same technique applied to the Florence data also yielded an intercept of zero, but with a slope coefficient of 0.25; the r^2 value was 0.47. The size of the slope coefficients gives an indication of the inflating effects of areal aggregation. In general terms the effect of using these models to correct the ecological correlations is to increase the diagonal-cell percentages. The best results were obtained for the 1 km grid-square data; the other data sets tended to be overcorrected, so that the distribution of the estimated individual correlations was too peaked. A number of other regression models were investigated which sought relationships between various moments of the data distributions and the change in correlation brought about by aggregation, but no better models were found.

A problem with these regression model approaches is that the predicted correlations are not constrained to lie between -1 and $+1$, the bounds of a correlation coefficient. A number of nonlinear logit, arctangent, and S-shaped functions were investigated which asymptoted to the desired limits, but again no better models could be devised. The basic problem is that there is a degree of random scatter that cannot be accounted for. It is apparent that there are a number of different factors which act as mechanisms for the aggregation effect. For example, individual variables with small total frequencies can exhibit the largest range of aggregation effects when aggregated and correlated with variables which have large total frequencies, but this phenomenon is nonlinear and complex.

It seems then that there will be no simple solution to this problem. As Langbein and Lichtman (1978, page 61) put it: "Investigators will find no philosopher's stone for transmuting information about groups into conclusions about individuals", although for certain purposes special techniques may help a little (Duncan and Davis, 1953; Johnston, 1976). Instead it would seem more sensible to advocate the use of spatial engineering techniques to create sets of areal units with specific properties, be they of a statistical or numerical nature or expressed in terms of qualitative geographical sensibility criteria.

3 Factorial studies

3.1 *Individual versus ecological factorial studies*
A very common analytical technique applied to census data is that of social-area analysis, usually in the form of factor analysis (Clark and Gleave, 1973). Ecological fallacies are perhaps most visible with these methods, since the interpretation of the factors is often in terms of characteristics assumed to exist at the individual level. It is interesting, therefore, to investigate whether the differences between ecological and individual correlations shown in tables 1 to 3 produce different factor interpretations and whether multivariate rather than bivariate analysis is better able to identify the magnitude of the problem. If the inflating effect of spatial aggregation on the correlation coefficient is largely systematic, then perhaps there may be little or no major differences in interpretation. Alternatively, if the effects are variable specific and highly complex, different results could well emerge.

The sizes of the eigenvalues extracted from the correlation matrices for the various individual and areal data sets provide a measure of the inflating effects of spatial aggregation. For the individual Sunderland data, eighteen eigenvalues exceeded unity and accounted for 62.5% of the variance in the original correlation matrix. After aggregation to 500 m grid-squares, these values changed to fifteen eigenvalues and 73.4%, for the 1 km grid-squares to fourteen eigenvalues and 82.5%, and for the polling districts to eight eigenvalues and 85.0%. For the individual Italian data, fourteen eigenvalues exceeded unity and accounted for 57.8% of the variance. After aggregation to enumeration districts, eight eigenvalues accounted for 73.7% of the variance. It is apparent that the smaller number of factors for the areal data sets is the result of

Table 4. Congruence coefficients between individual factors and the most similar areal factor.

Individual factor number	Factor numbers ranked by size of eigenvalue			
	Florence enumeration districts[a]	Sunderland		
		500 m grid-squares[a]	1 km grid-squares[a]	polling districts[a]
1	1 (0.69)	1 (0.90)	2 (0.77)	2 (0.56)
2	6 (0.82)	4 (0.75)	14 (0.67)	1 (0.75)
3	1 (0.45)	3 (0.84)	4 (0.72)	2 (0.53)
4	4 (0.80)	5 (0.96)	6 (0.87)	3 (0.80)
5	1 (0.72)	2 (0.86)	3 (0.81)	1 (0.55)
6	2 (0.34)	9 (0.93)	10 (0.79)	3 (0.29)
7	2 (0.35)	7 (0.61)	1 (0.58)	1 (0.58)
8	5 (0.79)	8 (0.68)	5 (0.47)	5 (0.56)
9	1 (0.38)	4 (0.70)	7 (0.68)	1 (0.68)
10	5 (0.47)	6 (0.74)	5 (0.62)	4 (0.74)
11	2 (0.79)	10 (0.85)	2 (0.67)	2 (0.32)
12	6 (0.35)	7 (0.57)	1 (0.54)	7 (0.25)
13	8 (0.77)	15 (0.53)	2 (0.42)	6 (0.65)
14	3 (0.66)	15 (0.56)	5 (0.35)	5 (0.31)
15	- -	13 (0.26)	8 (0.56)	8 (0.31)
16	- -	5 (0.49)	6 (0.36)	3 (0.51)
17	- -	8 (0.49)	9 (0.82)	5 (0.51)
18	- -	13 (0.56)	12 (0.31)	5 (0.21)

[a] Figures in brackets are congruence coefficients.

Table 5. Size distribution of congruence coefficients between the individual and areal factors.

Coefficient size	Congruence coefficients			
	Florence enumeration districts[a]	Sunderland		
		500 m grid-squares[a]	1 km grid-squares[a]	polling districts[a]
Over 0.9	0 (0)	3 (3)	0 (0)	0 (0)
0.8-0.9	2 (2)	3 (3)	3 (3)	1 (1)
0.7-0.8	4 (4)	3 (3)	3 (3)	2 (2)
0.6-0.7	2 (2)	5 (2)	10 (4)	2 (2)
0.5-0.6	0 (0)	8 (4)	3 (3)	7 (7)
Under 0.5	104 (6)	248 (3)	233 (5)	132 (6)

[a] Figures in brackets are best values for each individual factor.

variables, which were either not associated or were not strongly associated at the individual level, becoming areally associated with increasing scales of aggregation. It is of interest, therefore, to investigate what happens to the factor loadings.

For present purposes it is sufficient to use factor loadings obtained by applying a VARIMAX rotation with iterative estimates of communalities. The number of factors is determined by applying the eigenvalue-of-one rule of thumb. The simplest way of comparing the patterns of individual and areal factor loadings is by computing congruence coefficients (Harman, 1966) for all pairwise combinations of factors. Congruence coefficients can be interpreted as analogous to correlation coefficients. High values (over 0.7) are indicative of a high degree of similarity between factors; lower values (0.5–0.7) a poor fit; and less than 0.5, no fit [after Johnston (1973)].

In table 4, the best-match areal factors for each individual factor are reported. These results indicate that there is a small number of good fits and a large proportion of poor or no fits. The size distribution of the full set of congruence coefficients is given in table 5. If an arbitrary threshold of 0.7 and over is used to indicate a 'close fit', then the best match is for the Sunderland 500 m grid-square data with nine factors; next are the Florence enumeration districts and the Sunderland 1 km grid-square data with six factors, followed by the Sunderland polling districts with three factors.

A few individual factors have a weak association with two or more areal factors. For example, factor 1 for the Florence enumeration districts has some similarity with individual factors 1 and 5, and factor 1 for the polling districts has some similarities with individual factors 2, 5, 7, and 9. This is to be expected, as spatial aggregation brings together unrelated sets of variables. It may be that different factor rotations would be more successful at identifying these effects. Generally, however, it seems that the principal effect of spatial aggregation of census data is to create new factors by bringing together variables that were not strongly associated at the individual level. A more insidious effect is to change the strength of the relationships between most variables and this influences the nature and the significance attached to most factors. In practice not all these effects and differences can be detected; such is the level of subjectivity in the art of factor labelling.

4 Regression models

4.1 *A superficial comparison of some simple individual and ecological models*

The question now arises as to how well regression models can cope with these aggregation problems and more especially whether models which have been built as descriptions of areal associations have any relevance at the individual level, and vice versa. A major problem here concerns the construction of the same model form to handle different levels of measurement; the individual data are mainly dichotomous, but the areal data have continuous measurement scales. This is a problem because regression models that can handle categorical data cannot handle continuous data without recoding the data, and the results may well depend on the recodings that are used. The possibility that differences in model structure may affect the results needs to be kept in mind throughout this section.

For the purposes of this paper, a small number of the variables used in section 2 are selected as dependent variables which a subset of the remaining variables could be used to predict. For the areal data, a standard stepwise regression procedure was used to identify the 'best' regression models with exactly six variables. These predictor variables are then used to build an equivalent regression model with the individual-level data. The method used for this latter task is known as multiple classification analysis (MCA) (see Andrews et al, 1973). MCA offers a form of regression

analysis for variables with categorical measurement scales. Finally, to complete the crossaggregation comparisons, a method known as SEARCH [alias the automatic interaction detector (AID)] was used to identify the best possible individual-level models. SEARCH offers an automated method for building models of relationships from variables with categorical measurement scales (see Sonquist and Morgan, 1964; Sonquist et al, 1974). It can be regarded as being a form of stepwise regression that operates on individual-level data, although, of course, it has much more to offer. Here it is used merely to identify the first six different variables that emerge as being important. The analysis is restricted to only six variables, simply to avoid giving the individual-level models an unfair advantage; they can readily handle far larger numbers of predictor variables than stepwise regression because they do not suffer from multicollinearity problems. These best six variables from SEARCH are then used by MCA to provide a regression model in a form that can be compared with the areal data results. Both MCA and SEARCH are available in the OSIRIS IV statistical package.

4.2 *Comparison of models' goodness-of-fit*

Initially attention is restricted to considering the relative performances of the various models. How well do the variables selected by stepwise regression as being important for the aggregated data perform at the individual level? This latter result can be

Table 6. Comparison of model performances.

Dependent variable	Model[a]	Adjusted r^2 values[b]				
		individual data	500 m grid-squares	1 km grid-squares	polling districts	enumeration districts
Sunderland data						
Presence or absence of colour TV	SR	-	0.57	0.69	0.74	
	MCA	-	0.17	0.13	0.15	
	SEARCH	0.23 (0.20)	-	-	-	
	MR	-	0.47	0.33	0.55	
Moved in during last five years	SR	-	0.30	0.46	0.86	
	MCA	-	0.15	0.06	0.14	
	SEARCH	0.20 (0.15)	-	-	-	
	MR	-	0.23	0.14	0.82	
Rooms per person	SR	-	0.43	0.56	0.80	
	MCA	-	0.21	0.18	0.20	
	SEARCH	0.24 (0.21)	-	-	-	
	MR	-	0.32	0.30	0.75	
Florence data						
Agricultural workers	SR	-				0.58
	MCA	-				0.03
	SEARCH	0.18 (0.03)				-
	MR	-				0.41
Retired persons	SR	-				0.78
	MCA	-				0.40
	SEARCH	0.46 (0.42)				-
	MR	-				0.76
Overcrowded households	SR	-				0.82
	MCA	-				0.23
	SEARCH	0.39 (0.32)				-
	MR	-				0.80

[a] SR represents stepwise regression; MR represents multiple regression.
[b] Figures in brackets are r^2 values for SEARCH runs stopped after six variables had been selected.

compared with the best result achieved by SEARCH–MCA using only the individual-level data. The question then is how well would a multiple regression model perform if these same variables were used with the areal data. These latter results can then be compared with the initial stepwise regression runs. Thus three different comparisons can be made in an attempt to find out whether different variables are important at different levels of spatial aggregation.

In table 6, the results of these comparisons for a selection of dependent variables are reported. The stepwise regression results for Sunderland show, once again, the expected scale effects, in that the value of the r^2 coefficient increases with the size of zone (Blalock, 1964). Likewise, the performances of all the individual-level models are far poorer than those for the areal models. There is nothing really surprising about this, since it is a feature of all disaggregate models that disaggregation increases the level of random variation that cannot be accounted for by the models.

What is particularly interesting here is that the performances of the MCA models using variables identified by the stepwise regression procedure with the areal data are often comparable with the performances of the best SEARCH models[(2)]. Likewise, the first six variables identified by the SEARCH procedure generally perform well in multiple regressions with the areal data. There are some exceptions, but the results in table 6 do seem to suggest that the same variables can be used at both scales of analysis and that the loss of efficiency, as measured in terms of r^2 values, is small when committing both ecological and individualistic fallacies. These conclusions apply both to the Sunderland and to the Florence data sets.

4.3 *Comparison of the predictor variables selected at different levels of aggregation*
Further support for this view comes from a comparison of the variables selected as being the best predictors both at the individual and in the areal scales (see table 7). There is a remarkable degree of correspondence, bearing in mind the very different nature of the stepwise regression and the SEARCH procedures used to define the best predictor variables. This, of course, does not mean that parameter estimates made at the areal level can be applied at the individual level or that different areal aggregations will provide similar results; both these topics require further investigation. Nevertheless, these tentative findings are interesting in that it would appear that the worst-case ecological fallacy situation may not occur. Alternatively, it may be that,

Table 7. Common predictor variables (out of six) identified by the SEARCH procedure and by the stepwise regression.

Dependent variable	Areal units			
	500 m grid-squares	1 km grid-squares	polling districts	enumeration districts
Sunderland data				
Presence or absence of colour TV	3	3	3	-
Moved in during last five years	5	4	3	-
Persons per room	4	4	4	-
Florence data				
Agricultural workers	-	-	-	3
Retired persons	-	-	-	4
Overcrowded households	-	-	-	2

(2) The SEARCH models also take into account interaction effects which have no equivalent in stepwise regression procedures. These effects are included in the r^2 values of the SEARCH model but not in the related MCA results.

in the small number of examples considered here, the important variables dominate both scales of aggregation. Further empirical investigation with a wider range of data sets is clearly required before any firm conclusions can be reached.

5 Classification methods

5.1 *Individual versus areal data classifications*

A final technique is that of cluster analysis. This method has been, and is being, widely used to provide census data classifications for policy purposes; for example, the identification of deprivation areas as a starting point for area-based policies aimed at ameliorating the effects of multiple deprivation. The stereotyped image of unemployment, poor housing, a lack of basic amenities, and low socioeconomic status is one result of these studies. It is not denied that areas with these and related deprivation characteristics exist, but doubts are expressed whether area-based profiles adequately describe the characteristics of the people who live there. Is a person deprived merely because he lives in the same enumeration district as a deprived person? Does multiple deprivation exist at the individual or the areal level? The problem here is that aggregate census data cannot distinguish between deprived areas and deprived people (Openshaw and Cullingford, 1979). The more general question, therefore, is how good is a classification of areas as a description of the people who live there?

One way of answering this question is to compare two classifications of the same data, one at the individual level and the other at the areal level. This task is made difficult by the computational problems involved in the classification of large individual data sets; for example, there are 122342 households in Florence to be classified. Openshaw (1980; 1982a) described classification algorithms and a suite of computer programs which can classify data sets of virtually any size, provided they have dichotomous measurement scales. The areal data sets are more easily classified using standard clustering techniques; for example, that developed by Openshaw (1982b) to classify 1981 census data.

For this paper, only the Florence data are classified. It was thought better to employ a population data set to avoid possible sampling problems associated with the Sunderland data. Details of both the areal and individual classifications are described in Bianchi et al (1980; 1983). Here attention is restricted to the results of a crosstabulation of the households in Florence by cluster both in the individual and in the enumeration district classifications. In table 8, the results are presented for

Table 8. Crosstabulation of households by clusters in the individual and the enumeration district classifications.

Enumeration district cluster-codes	Percentage of households in individual-data cluster-codes:							
	1	2	3	4	5	6	7	8
1	5	8	8	9	29	7	15	19
2	4	19	17	17	9	21	5	9
3	12	17	24	11	7	10	12	6
4	5	11	13	14	17	13	11	16
5	10	13	16	11	11	12	14	13
6	19	9	8	12	14	12	7	20
7	3	9	10	15	18	16	11	16
8	15	11	14	6	10	10	18	17
9	14	10	10	9	13	11	11	21
10	10	11	9	17	14	16	7	16
11	6	17	24	15	8	13	10	6

a comparison of an eleven-cluster enumeration district classification and an eight-cluster individual-data classification.

These crosstabulations show that there is very little correspondence between the two classifications, yet both have been found to be meaningful in terms of a number of criteria (see Bianchi et al, 1983). The individual classification successfully identifies the major groupings that were expected to occur, given current knowledge about Italian social structure, and the enumeration district classification provides a generally accepted description of the spatial structure of Florence. That such large differences exist between the two classifications is not too surprising, since the objects being classified (households and arbitrary areal units) are different. In other words, unless the areas are completely homogeneous, the results will be different. Nevertheless, the scale of the observed differences is large, bearing in mind that enumeration districts are the finest areal units that are usually available for census analysis. It is apparent, therefore, that classification methods present the greatest opportunities for constructing ecological fallacies by inference.

Table 9. Percentage of households in enumeration districts belonging to a particular areal cluster by membership of each individual-data cluster.

Enumeration district number	Percentage of households in individual-data cluster-codes:							
	1	2	3	4	5	6	7	8
1	2	15	21	12	12	19	8	7
2	7	6	12	8	16	13	18	18
3	7	15	0	15	34	19	7	0
4	4	9	11	20	8	17	11	15
5	4	8	10	14	15	19	9	16
6	3	6	8	8	30	15	11	16
7	0	13	3	18	18	24	8	12
8	1	10	7	15	20	12	10	22
9	4	10	13	22	15	13	8	11
10	2	7	8	21	15	16	10	18
11	2	9	7	10	23	12	6	27
12	0	6	13	18	14	14	17	13

Table 10. Performance of SEARCH models trying to predict cluster-codes from the individual data.

Classification and cluster	Sum of squares explained	Classification and cluster	Performance[a]
Individual-data classification	86.7	*Enumeration district classification*	0.9
Cluster 1	69.6	Cluster 1	0.0*
Cluster 2	74.8	Cluster 2	0.8
Cluster 3	80.2	Cluster 3	0.9
Cluster 4	77.0	Cluster 4	1.1
Cluster 5	70.5	Cluster 5	0.0*
Cluster 6	90.2	Cluster 6	0.0*
Cluster 7	74.9	Cluster 7	2.5
Cluster 8	90.5	Cluster 8	6.2
		Cluster 9	1.1
		Cluster 10	1.8
		Cluster 11	4.3
		Enumeration district classification with thirty clusters	1.9

* No split of the data explained 0.8 or more of the total sum-of-squares of the dependent variable.

A further illustration of the differences is provided by examining a highly distinctive cluster of enumeration districts which describes the characteristics of agricultural areas. The households in the enumeration districts assigned to this cluster are cross-tabulated by their membership of the individual clusters in table 9. Clusters 5 and 8 in the individual classification have the highest spatial concentration of agricultural workers, but there is still a widespread membership of other individual clusters.

A final demonstration of the differences is provided by the results of a series of attempts, using SEARCH, to find a relationship between the cluster-codes and the individual data. In these models the dependent variable has a categorical measurement scale in the range 1 to 8 for the individual-data classification cluster-code and in the range 1 to 11 for the enumeration district cluster-code. All households assigned to the same enumeration district cluster share the same cluster-code. The aim is to predict membership of each cluster-code by using SEARCH to identify relationships between these cluster-codes and the individual-data variables. As shown in table 10, these SEARCH models are able to account for a large part of the sum-of-squares of the cluster-codes of the individual classification but very little of the cluster-codes of the enumeration district classification. The use of thirty instead of eleven clusters results in only a minute improvement in predictive performance. These findings emphasise, once again, the dangers of using area-based classifications as a description of individual households.

5.2 *An explanation and a suggestion for a new census statistic*

These cluster analysis results are most useful because they are more easily understood than either correlation, factor analytical, or regression methods. What appears to be happening is that the area classification gives most emphasis to average areal characteristics. The households that contribute most to these areal profiles usually belong to two or more individual clusters. Aggregation combines and mixes a number of different frequency distributions of attributes from the individual data. On occasions the households that contribute most to the areal profile may constitute a minority; for example, if there is a spatial concentration of households with very distinctive characteristics. The extent to which these processes occur depends on the relationship between census-area boundaries and the spatial distribution of households belonging to individual clusters.

In addition to the averaging effects of spatial aggregation creating unnatural variable associations, the noise- and data-reducing properties of classification methods (in common with other statistical techniques) provide a selective amplification or filtering of the averaged-out data. Cluster analysis will emphasise areas with distinctive characteristics, even if the percentage occurrences of a variable at the individual level is low. Likewise, it will give little or no emphasis to variables which have similar levels in most areas. For example, the agricultural-area cluster is spatially highly distinctive, even though it provides a poor representation of the individual households who live in such areas. Similarly, some individual clusters never appear in the area classification because the spatial distribution of their constituent households is not geographically concentrated. For example, there are two old-person clusters in the individual-level classification, but none in the area classification. This is not surprising when the range of percentage old-persons is from 30% to 49%, but in the two individual clusters, 89% and 90%. The primary distinction between the two classifications, then, is that the area classification emphasises geographical concentrations whereas the individual one is completely aspatial.

If both levels of classification are available, they can be used in a complementary manner. However, there is a strong case to be made for using the individual classification to report for each enumeration district the numbers of households in

each individual cluster. These derived statistics would offer an easy substitute for microcensus data since they summarise a large number of individual attributes. More importantly, they would give an indication of the spread of socioeconomic and demographic characteristics within an enumeration district. This would allow any aggregational basis in the areal data to be identified and the dangers of ecological fallacies would then be reduced. For example, it would be possible to count the numbers of 'deprived' households living in a 'deprived' area by reference to the number of households in certain clusters in the individual classification. Furthermore, because the individual-level classifications are aspatial, and thus independent of census geography, they would offer a new dimension to census analysis. These new statistics could be mapped, and it is perhaps not unlikely that they could result in a better understanding of the characteristics of urban and rural areas.

6 Conclusions

In this paper, the results are described of a series of empirical studies designed to demonstrate the magnitude and nature of ecological fallacy problems associated with the analysis of aggregate census data. The methods examined here are fairly simple-minded and are typical of the types of analyses often performed on census data. The results confirm that the ecological fallacy problem is important, but its severity depends on the methods of analysis employed, on the mode of interpretation afforded to the results, and even on the choice of variables. The problem is that at present there is no way of being able to predict or determine whether a particular areal data set is going to yield results which are close to the individual values. For example, a correlation coefficient which is at the extremes of tables 1 or 2 would produce very misleading results, and, because only aggregate data are available from the census in many countries, there is no way of knowing how large this effect might be.

One solution would be for a new set of derived census variables to be added to areal data sets. These new variables would give an indication of the distribution of household types within an area according to an individual-level cluster analysis. Obviously this would not solve all problems, but it would be a useful addition to existing census variables.

Finally, two further areas of research are suggested. The first is to examine more closely the effects of data aggregation on the accuracy of parameter estimates. This implies that models can be built at the individual level with parameters that can be estimated from areal data. If possible, this would provide the basis for a statistical solution to the areal aggregation problem. The second area of research involves the further testing of a wider range of methods both on individual and on areal data sets. The purpose would be to provide further empirical evidence about aggregation effects so as to persuade census agencies to take a greater interest in these problems, particularly the need for the most careful design of spatial frameworks for the presentation of aggregate census data.

Acknowledgements. Colin Wymer of the University of Newcastle upon Tyne kindly prepared the Florence data sets. Giuliano Bianchi of the Istituto Regionale per la Programmazione Economica della Toscana made possible the analysis of the Italian data. Two anonymous referees and Ron Johnston of the University of Sheffield provided numerous comments on an earlier version of the paper.

References

Alker H R, 1969, "A typology of ecological fallacies" in *Quantitative Ecological Analysis in the Social Sciences* Eds M Doggan, S Rokkan (MIT Press, Cambridge, MA) pp 69-86

Andrews F M, Morgan J N, Sonquist J A, Klem L, 1973 *Multiple Classification Analysis* Institute for Social Research, The University of Michigan, Ann Arbor, Michigan 48109, IL

Bianchi G, Openshaw S, Scattoni P, Sforzi F, 1980, "Probleme di zonizzazione: l'identificazione di aree sociali a scale urbana" in *Nuovi Contributi allo Studio dello Sviluppo Economico della Toscana* (Istituto Regionale per la Programmazione Economica della Toscana, Florence) chapter 9

Bianchi G, Openshaw S, Scattoni P, Sforzi F, Wymer C, 1983 *Analisi dell'Area Sociale: Comparazione delle Classificazioni Condotte su Dati Medi Sezioni do Censimento e su Dati Individuali* Proceedings of the Italian Regional Science Conference, Naples, Italy, October 1981 (forthcoming)

Blalock H M, 1964 *Causal Inferences in Nonexperimental Research* (University of North Carolina Press, Chapel Hill, NC)

Clark B D, Gleave M G (Eds), 1973 *Social Patterns in Cities* Institute of British Geographers, Special Publication 5 (Alden Press, Oxford)

Clark W A V, Avery K L, 1976, "The effects of data aggregation in statistical analysis" *Geographical Analysis* **8** 428–438

Duncan O D, Davis B, 1953, "An alternative to ecological correlation" *American Sociological Review* **18** 665–666

Gehlke C E, Biehl K, 1934, "Certain effects of grouping upon the size of the correlation coefficient in census tract material" *Journal of the American Statistical Association* **29** 169–170

Goldblatt P, Fox J, 1978, "Household mortality from the OPCS longitudinal study" *Population Trends* **14** 20–28

Goodman L A, 1959, "Some alternatives to ecological correlation" *American Journal of Sociology* **64** 610–625

Hannan M T, 1971 *Aggregation and Disaggregation* (D C Heath, Lexington, MA)

Harman H H, 1966 *Modern Factor Analysis* (University of Chicago Press, Chicago, IL)

Johnston R J, 1973, "Residential differentiation in major New Zealand urban areas: a comparative factorial ecology" in *Social Patterns in Cities* Eds B D Clark, M B Gleave, Institute of British Geographers, Special Publication 5 (Alden Press, Oxford) pp 143–167

Johnston R J, 1976, "Areal studies, ecological studies, and social patterns in cities" *Transactions of the Institute of British Geographers, New Series* **1** 118–122

Langbein L I, Lichtman A J, 1978 *Ecological Inference* (Sage, Beverly Hills, CA)

Openshaw S, 1977, "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** 359–472

Openshaw S, 1978a, "An empirical study of some zone-design criteria" *Environment and Planning A* **10** 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, P J B Brown (Martinus Nijhoff, The Hague) pp 93–113

Openshaw S, 1980, "Monothetic divisive algorithms for classifying large data sets" in *Proceedings in Computational Statistics COMPSTAT 1980* Eds M M Barritt, D Wishart (Physica, Vienna) pp 419–425

Openshaw S, 1981, "Le problem de l'aggregation spatiale en geographie" *L'Espace Geographie* **1** 15–24

Openshaw S, 1982a, "A suite of FORTRAN programs for the classification of individual census data (CID)" mimeograph, Department of Geography, University of Newcastle upon Tyne, Newcastle upon Tyne, England

Openshaw S, 1982b, "A suite of portable FORTRAN IV programs (CCP) for classifying 1981 census data for areas: an introduction and user guide" project report, Centre for Urban and Regional Development Studies, University of Newcastle upon Tyne, Newcastle upon Tyne, England

Openshaw S, Cullingford D, 1979, "Deprived places or deprived people: a study of aggregation effects inherent in area based policies" DP-28, Centre for Urban and Regional Development Studies, University of Newcastle upon Tyne, Newcastle upon Tyne, England

Openshaw S, Taylor P J, 1979, "A million or so correlation coefficients: three experiments on the modifiable areal unit problem" in *Statistical Applications in the Spatial Sciences* Ed. N Wrigley (Pion, London) pp 127–144

Openshaw S, Taylor P J, 1981, "The modifiable areal unit problem" in *Quantitative Geography* Eds N Wrigley, R J Bennett (Routledge and Kegan Paul, Henley-on-Thames, Oxon) pp 60–70

Robinson A H, 1950, "Ecological correlation and the behaviour of individuals" *American Sociological Review* **15** 351–357

Sonquist J A, Baker E L, Morgan J N, 1974 *Searching for Structure* (Institute for Social Research, University of Michigan)

Sonquist J A, Morgan J N, 1964 *The Detection of Interaction Effects* (Institute for Social Research, University of Michigan)