Environment and Planning A, 1991, volume 23, pages 1025-1044

The modifiable areal unit problem in multivariate statistical analysis

A S Fotheringham, D W S Wong
The National Center for Geographic Information and Analysis, Department of Geography, Wilkeson Quad, State University of New York at Buffalo, Buffalo, NY 14261, USA
Received 11 June 1990

Abstract. In this paper the examination of the modifiable areal unit problem is extended into multivariate statistical analysis. In an investigation of the parameter estimates from a multiple linear regression model and a multiple logit regression model, conclusions are drawn about the sensitivity of such estimates to variations in scale and zoning systems. The modifiable areal unit problem is shown to be essentially unpredictable in its intensity and effects in multivariate statistical analysis and is therefore a much greater problem than in univariate or bivariate analysis. The results of this analysis are rather depressing in that they provide strong evidence of the unreliability of any multivariate analysis undertaken with data from areal units. Given that such analyses can only be expected to increase with the imminent availability of new census data both in the United Kingdom and in the USA, and the current proliferation of GIS (geographical information system) technology which permits even more access to aggregated data, this paper serves as a topical warning.

1 Introduction

Spatial analysis frequently involves the use of areal data; a common example being the analysis of socioeconomic data reported for a set of census enumeration units. One of the most stubborn problems related to the use of areal data is what is commonly referred to as the modifiable areal unit problem (MAUP): the sensitivity of analytical results to the definition of units for which data are collected.

Research within the spatial analytical framework covers a wide range, from the reporting of relatively simple univariate statistics such as the first and the second moments of a distribution and simple correlations and regressions to the calibration of more complex linear and nonlinear models. The presence of the MAUP raises our skepticism on the reliability of the results reported from an analysis of aggregated spatial data because these results are likely to vary with the level of aggregation (the scale problem) and with the configuration of the zoning system (the zoning problem). A survey of the literature reveals that the MAUP not only has an impact on traditional statistical analysis (Clark and Avery, 1976), but it has also been shown to be present in factorial ecological studies (Perle, 1977), spatial interaction modelling (Masser and Brown, 1975; 1977; Putman and Chung, 1989), input-output analysis (Blair and Miller, 1983), and location-allocation modelling (Bach, 1981; Goodchild, 1979).

In this paper, we intend to provide insights into the complexity of the MAUP in relatively advanced statistical analysis. A brief review of research related to the MAUP in univariate and bivariate analysis is presented in section 2, and a similarly brief review of the MAUP in multivariate analysis is given in section 3. Although there has been some progress in understanding the MAUP, there is still a great deal of uncertainty surrounding this topic, particularly in multivariate analysis, and aspects of this uncertainty are discussed in section 4. Although several useful empirical studies of the MAUP already exist, particularly in univariate or bivariate statistical analysis, we argue that there is a need for further empirical work in order to understand more fully the impacts of the MAUP in a multivariate setting.

Theoretical analysis of the MAUP is possible in univariate and bivariate situations, but the situation quickly becomes so complex in even relatively simple types of multivariate analysis that empirical work is likely to be the only means of understanding, at least partially, the effects of the MAUP. In sections 5, 6, and 7, we present an extensive set of results concerning the MAUP in multivariate analysis, by using census data for the Buffalo metropolitan area. The discussion is centered around the calibration of a multiple linear regression model and a multiple logit regression model.

2 The modifiable areal unit problem in univariate and bivariate analysis

Although several key studies of the MAUP exist, given its importance to and potential impact on spatial analysis the MAUP has generated surprisingly little attention in geography. Openshaw (1984) provides a comprehensive review on the early research on the MAUP beginning with the work of Gehlke and Biehl (1934), who report that the correlation coefficient for variables of absolute measurement increases when areal units are aggregated contiguously, but there is little equivalent trend for ratio or percentage variables. They also report that the trend exhibited by the correlation coefficient for absolute data also does not exist if areal units are grouped randomly without a contiguity constraint.

Blalock (1964) assesses the impacts on the correlation coefficient and slope estimate of a bivariate linear model under four different aggregation criteria: (1) random grouping, (2) grouping by independent variable, (3) grouping by dependent variable, and (4) grouping by proximity. If the first grouping method is used, no obvious systematic impact on correlation coefficients and parameter estimates is found. If the aggregation is based upon the value of the independent variable, the correlation coefficient will increase with scale, but the grouping has no systematic impact on the slope parameter. If grouping is done by dependent variable, the inflation of correlation coefficient will have the same magnitude as in the case of grouping by independent variable and the slope estimate will also increase with scale. The grouping-by-proximity procedure inflates correlation coefficients and parameter estimates slightly, but not as much as the increases generated by the grouping method applied to the independent variable.

Clark and Avery (1976) also investigated the scale effect in a bivariate regression model. However, they report that both the correlation coefficient and the slope estimate of the regression did not increase monotonically with aggregation. This phenomenon was also noted in Blalock's 1964 analysis of aggregating 150 basic spatial units (bsu) into 75, 30, 15, and 10 areal units. The directions of change in correlation coefficients and slope estimates changed at the last aggregation level. Blalock suspected that when the areal units are grouped into 15 units, most of them are not actually adjacent to each other and the groupings become more heterogeneous at this level than at the lower aggregation levels. After examining both the dependent and the independent variables, Clark and Avery (1976) point out that the variations of both variables decreased as scale increased, but at a certain level of aggregation the covariation between the two variables fell rapidly, which led to lower estimates of correlation coefficients and parameters at the highest levels of aggregation.

The reason for the general increase in correlation coefficients as the level of data aggregation increases can be understood intuitively and there is little need for empirical research. Regardless of the method of spatial data aggregation (for instance, by averaging or summing), the process involves a smoothing effect so that the variation of a variable tends to decrease as aggregation increases. As the

correlation coefficient, r_{xy}, between two variables, x and y, is determined by

$$r_{xy} = \frac{\text{cov}(x, y)}{s_x s_y}, \tag{1}$$

where $\text{cov}(x, y)$ is the covariance of x and y, and s_x and s_y represent the standard deviation of x and y, respectively; when the variances of x and y decrease, the correlation coefficient will increase if the covariance between x and y is relatively stable. The increase in the magnitude of the correlation coefficient owing to an increase in data aggregation is explained by Robinson (1950) both nonmathematically and mathematically, and Arbia (1989) provides a very insightful analysis of the problem. Variation in simple statistics such as a correlation coefficient and the slope parameter in simple linear regression are thus reasonably well understood.

One area of slight fuzziness that remains, however, concerns the role of spatial autocorrelation in determining the rate at which the variances of x and y decrease as the level of aggregation increases. To investigate the effects of spatial autocorrelation on spatial aggregation, Openshaw and Taylor (1979) generated two artificial variables yielding maximum positive and minimum spatial autocorrelation. They found that even with no spatial autocorrelation in both variables, random grouping and aggregation according to a proximity criterion will still increase the value of the correlation coefficient but that the role of spatial autocorrelation on the MAUP was unclear. Similar findings hold for the effect of spatial autocorrelation on the slope estimate, $\hat{b}$, in a simple regression model (Openshaw, 1978), which is not unexpected as

$$\hat{b} = r_{xy} \frac{s_x}{s_y}. \tag{2}$$

If the variances of x and y change at approximately the same rate as data are increasingly aggregated, variations in $\hat{b}$ and r_{xy} will be similar.

Because of the relatively simple nature of the MAUP in univariate and bivariate analyses, several attempts at providing a solution to the problem have been made. One of the earliest is the areal weighting solution proposed by Robinson (1956). This method takes into account the varying size of areal units and appears to produce relatively stable correlation coefficients and parameter estimates from simple regression at different scales. However, it was demonstrated subsequently by Thomas and Anderson (1965) that the example given by Robinson was a special case and they proposed a more general framework to weight areal data in correlation analysis.

Arbia (1989), in a more sophisticated treatment of the MAUP, albeit in rather restricted circumstances, argues that Robinson's areal weighting method and subsequent versions of the method are not successful in isolating the scale effect. Arbia proposes a framework which takes into account not only the size of the area, but also the interconnectedness and dependence of areal units. This is very similiar to the approach used in the study of spatial statistics and spatial autocorrelation where the value of a variable in one cell can be related not only to the values in surrounding cells but also to the configuration of the spatial system (Griffith, 1988). In Arbia's analysis, moments of variables are shown to be functions of some aspects of spatial configuration, although not necessarily at the same scale at which the moment is measured. For example, the variance of a variable at one spatial scale might be a function of the interconnectedness of areal units and the spatial autocorrelation of the variable at a more disaggregated scale. In this framework, it is also assumed that variables are derived from stationary stochastic spatial

processes, that is, the probability distribution of a random variable is constant across space. In dealing with bivariate spatial processes, Arbia extends the univariate model to incorporate the spatial dependence (cross-correlation) between the two processes. From these theoretical formulations, Arbia attempts to provide some answers to problems related to the aggregation (zoning) issue, ecological fallacy, and spatial correlograms. However, though sophisticated and useful for the univariate and bivariate cases, it is unclear how Arbia's framework can be extended to multivariate analysis. It is also doubtful whether Arbia's assumption of equal size or equal area for areal units can be relaxed and still generate the same level of insight into the MAUP.

Another 'solution' to the MAUP is based upon the concept of spatial entropy (Batty, 1974; 1976). Batty and Sikdar (1982a) focus on the relation between the information extracted from spatial data and model performance at different levels of aggregation. Usually, the performance of the model is evaluated in terms of its parameter estimates. But, in fact, the parameter estimates reflect how well the model can fit the data and the spatial variation of the data at a given scale. Batty and Sikdar (1982a) attempt to link the information of the spatial variation of the data to the model adopted by decomposing an information statistic for the data into different components associated with attributes of spatial aggregation such as density, average zonal size, dimension, and level of resolution. These components can be used to derive various types of spatial interaction model and parameter estimates. In their empirical study Batty and Sikdar (1982b) use these components, with fairly satisfactory results, to approximate parameter estimate at various aggregation levels for a simple population density model.

3 The modifiable areal unit problem in multivariate analysis

The assessments of the effects of the MAUP have not been restricted to simple bivariate models and correlation coefficient analysis. However, because of the complexity of understanding the MAUP in multivariate analysis, no theoretical work exists in this area, to the authors' knowledge, and for reasons outlined below none may be forthcoming. Blair and Miller (1983) investigate scale effects on multiregional input-output models by examining the errors introduced by the aggregation of regions. Their overall results indicate that the errors introduced by the aggregating of regions are relatively small and the worst situation is still acceptable for an input-output model. However, their empirical study was limited to a maximum of four regions, probably too few to assess fully the impacts of spatial aggregation.

In a more extended study designed to examine the factorial ecology of Detroit, Perle (1977) performed factor analysis on data at both the subcommunity and the census-tract levels. He concludes that the results are considerably different at the two levels in terms of interpreting the dimensions of the factors. The likely cause is that when the analysis shifts from the census tract level to the subcommunity level, a loss of information results, thereby reducing the dimensionality of the vector space (Perle, 1977).

Goodchild (1979) demonstrates the effects of areal aggregation on p-median and p-center classes of problems, particularly on location-allocation problems. He argues that the medians calculated at different aggregation levels are not consistent, and as a result, the output from the location-allocation model cannot provide an objective answer to the problem of optimal location. Bach (1981) also demonstrates the scale dependence of the solution of a location-allocation model.

A substantial portion of the literature on the MAUP in multivariate analyses is concerned with the application of spatial interaction models at various spatial scales.

Openshaw (1977) demonstrates the impacts of scale and zoning changes on the calibration of spatial interaction models by generating 261 twenty-two-zone interaction matrices and 87 forty-two-zone matrices from the set of 72 bsu. Considerable variation in the model goodness-of-fit and parameter estimates is reported across the various zoning systems and at the three different scales.

The Batty and Sikdar (1982a) framework described in section 2 demonstrates that a spatial entropy statistic can be decomposed into separate parts capturing the important attributes related to spatial aggregation and that these variables can be used to forecast changes in the parameter estimate of a simple population density model. This framework was later extended to the analysis of relatively complicated gravity models although the results were disappointing because of the difficulty of fitting models at high levels of aggregation, the overall accuracy of the models, and the simplicity of the approximation method used to predict the sensitivity of the parameter estimates to variations in data aggregation (Batty and Sikdar, 1982c; 1982d).

More recently, Putman and Chung (1989) report a very extensive examination of the MAUP in spatial interaction modelling by calibrating a model with six parameters. They examine several methods of aggregating data: one random aggregation (RA), and four systematic aggregation procedures consisting of equal numbers of bsu per zone (ENU), equal total area (EA), equal total population (EP), and equal numbers of low-income households per zone (HSS). From each aggregation procedure, thirty different systems of thirty zones are generated. They conclude that systematic zoning systems generally yield more consistent parameter estimates for all variables and have better goodness-of-fit measurement than the RA system. Parameter estimates for intensive or 'percentage' variables appear to be less sensitive to different zoning criteria or to zoning effect than those for extensive variables. Among the systematic aggregation procedures, EA generates systems giving the best goodness-of-fit statistic, and ENU yielded the most consistent parameter estimates, although Putman and Chung suggest that further research is needed to decide conclusively between the EA and ENU procedures.

Amrhein and Flowerdew (1989) examine the MAUP in a Poisson spatial interaction model used to analyze Canadian migration. Scale and zoning effects on parameter estimates are both examined. Their preliminary results indicate that the Poisson model may be insensitive to the scale effect up to a certain level of aggregation, but a large variation in parameter estimates is reported for different zoning systems. They also show that when the level of aggregation increases, the sensitivity of the parameter estimates to variations in zoning systems decreases. However, they are unable to determine whether their results are unique to the Canadian migration data or whether they have more widespread applicability.

4 Unresolved issues associated with the modifiable areal unit problem

Although many spatial analysts are aware of the MAUP, it is often conveniently ignored and empirical studies involving the analysis of areal data rarely mention possible scale and zoning sensitivity. This is especially true in urban analyses where census data are used extensively. For example, in a series of empirical studies on gentrification and urban revitalization by use of census-tract data (Gale, 1984; Lipton, 1977; Nelson, 1988), the impacts of the MAUP are totally ignored. Feeding census data into canned multiple regression programs is still a common practice (Sawicki, 1973) and many of these applications are used to formulate urban policy. It is still rare to find references to the MAUP in textbooks which advocate regression analysis for policy formulation (Cadwallader, 1985) and even in texts on spatial analysis and spatial statistics. With advances in GIS

(geographic information system) technology and the imminent release of important census data in the USA, Canada, and the United Kingdom, it can only be anticipated that multivariate analyses of areal unit data will increase in number; it can further be anticipated that a substantial portion of these analyses will totally ignore the MAUP.

It is thus important that further evidence of the MAUP in spatial analysis be presented and that insights into the sensitivity of analytical results to both scale and zoning variations be uncovered. This is especially important for multivariate analysis as previous research on the MAUP has been focused mainly on univariate and bivariate data. It is not clear, for example, to what extent results derived from univariate and bivariate analysis are transferable to multivariate analysis. For example, Openshaw (1978) demonstrates that the estimated slope parameter in a simple linear regression tends to increase in magnitude as data are increasingly aggregated, but it is not clear that all parameter estimates in a multiple regression will change in this way. Similarly, the impacts of spatial autocorrelation on the sensitivity of parameter estimates in multiple regression to scale and zoning variations have not been explored.

Previous studies of the MAUP in multivariate situations have not been particularly fruitful. Frameworks proposed by Batty and Sammons (1978) and later by Batty and Sikdar (1982a) fail to handle the complexity posed by multivariate analysis where the interactions among variables become quite convoluted. For example, consider the estimator of β_2 in a simple-looking multiple regression model of the form

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 \,, \tag{3}$$

which is

$$\hat{\beta}_2 = \frac{\text{var}(x_1)\text{cov}(x_2, y) - \text{cov}(x_1, y)\text{cov}(x_1, x_2)}{\text{var}(x_1)\text{var}(x_2) - [\text{cov}(x_1, x_2)]^2} \,, \tag{4}$$

where var (.) is the variance of the variable inside the parentheses, and cov (..) is the covariance of the two variables inside the parentheses. In order to predict the sensitivity of $\hat{\beta}_2$ to change in scale or zone definition, one would have to predict the relative changes in var(x_2), var(x_2), cov(x_1, x_2), cov(x_1, y), and cov(x_2, y)! Although empirical research on the sensitivity of parameter estimates in multivariate analysis might guide important theoretical research in this area (such as Arbia's contribution to univariate and bivariate analyses) it would appear to be extremely difficult, if not impossible, to anticipate analytically the effect of scale change or zoning-system change on parameter estimates in multivariate analysis. Empirical studies therefore offer us perhaps the only possibility to discover insights into the nature of parameter sensitivity in multivariate analysis. In what follows we examine the effects of the MAUP in the calibration of multiple linear and logit regression models.

5 Analytical structure

To investigate scale and zoning effects in multivariate analysis, two models were calibrated with a data set aggregated to a variety of spatial scales and zoning systems. The first is a linear regression model relating the mean family income, in dollars, of an areal unit ($\bar{I}^{\text{F}}$) to four independent variables: the percentage of the population in a unit who are homeowners (P^{own}); the percentage of the population who are blue-collar workers (P^{blue}); the percentage of the population who are black (P^{black}); and the percentage who are aged over 65 years (P^{eld}). That is,

$$\bar{I}^{\text{F}} = \beta_0 + \beta_1 P^{\text{own}} + \beta_2 P^{\text{blue}} + \beta_3 P^{\text{black}} + \beta_4 P^{\text{eld}} \,. \tag{5}$$

A priori, it was hypothesized that $\bar{I}^{\text{F}}$ would be positively related to P^{own} and negatively related to P^{blue}, P^{black}, and P^{eld}.

The second is a logit regression model relating the logged logit transformation of the proportion of owner-occupied housing within an areal unit to the proportion who are blue-collar workers, the mean family income in tens of thousands of dollars, and the median house value in hundreds of thousands of dollars (H^{mv}). That is,

$$\ln \frac{P^{own}}{1-P^{own}} = \beta_0 + \beta_1 P^{black} + \beta_2 P^{blue} + \beta_3 \bar{I}^{F} + \beta_4 H^{mv} , \quad (6)$$

where a priori it was hypothesized that the proportion of owner-occupied housing within an areal unit would be positively related to $\bar{I}^{F}$ and H^{mv} and negatively related to P^{black} and P^{blue}. The logit formulation is employed here because the dependent variable, the proportion of owner-occupied housing, ranges between zero and one and therefore invalidates the use of linear regression (Wrigley, 1985).

The two equations reflect the interests of one of the authors with the relationship between housing tenure and income but were chosen primarily for two reasons. The first is that they both represent highly plausible relationships that could easily form the basis of an investigation by an unwary researcher interested in either the spatial distribution of mean family income or the proportion of owner-occupied housing. The second is that both equations could form the basis of policy decisions made about income or homeownership determinants. One of our goals is to demonstrate the unreliability of the estimates of the parameters in these two models when they are calibrated with aggregated data; this unreliability takes on an added dimension when the results of the calibration could quite plausibly be used to determine public policy.

The base-level data used to calibrate equations (5) and (6) were obtained for 871 block groups in the Buffalo Metropolitan Area from the 1980 US census (USBC, 1981a; 1981b; 1981c). Data at the census-tract level were also obtained from the same census (USBC, 1983). To give a sense of the detail of this study, the spatial distribution of the 871 block groups is shown in figure 1. All subsequent data sets are aggregates of these units.

The average population sizes of a block group and a census tract are 1045 and 4175, respectively. Both these sets of reporting units are frequently employed in spatial analysis, with analysis performed at the level of the census tract probably

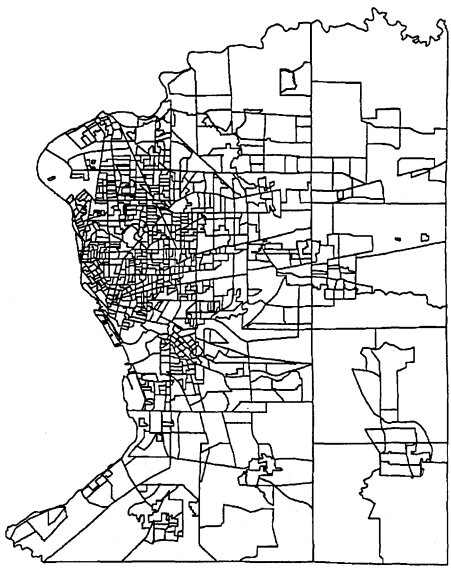

Figure 1. Spatial distribution of the 871 block groups for the Buffalo Metropolitan Area.

being the more common of the two. To give an idea of the MAUP in multivariate analysis, equations (5) and (6) were calibrated with data from the 871 block groups and the 218 census tracts. The results are shown in table 1.

In order to appreciate more fully the rather detailed examination of the MAUP that is presented in the following two sections, it is useful at this stage to interpret briefly the results in table 1. In table 1(a) the parameter estimates for the multiple regression equation are all associated with variables that are proportions and can therefore be interpreted as follows: for every increase in variable x_i of 0.1, the predicted mean family income changes by $\$\beta_1/10$. For instance, at the block-group level, an increase of 0.1 in the proportion of owner-occupied housing would increase the predicted mean family income by \$1312. Similar changes in P^{blue}, P^{black}, and P^{eld} would reduce the predicted mean family income by \$892, \$503, and \$308, respectively. Although the signs of the parameters are consistent across the two sets of spatial units, to anticipate our results somewhat, there are some differences which cause concern. For instance, whereas at the block-group level an increase of 0.1 in the proportion of owner-occupiers produces a decrease in the predicted $\bar{I}^{\text{F}}$ of \$1312, at the census-tract level the same increase produces a decrease in the predicted $\bar{I}^{\text{F}}$ of \$2266. Although the percentage of the population who are black is highly significant in explaining the distribution of $\bar{I}^{\text{F}}$ at the block-group level, it is barely significant at the census-tract level. Clearly, there would be difficulties in relying on these results to formulate policies. It is interesting to note that the four independent variables explain over 81% of the variance of the dependent variable at the census-tract level but only 37% at the block-group level.

The dependent variable in equation (6) is the logarithm of an odds ratio: the proportion of owner-occupied housing divided by the proportion of all other housing. A positive parameter estimate associated with an independent variable thus indicates that increasing values of that variable are associated with increasing proportions of owner-occupied housing. Thus, at the block-group level, owner-occupied housing appears to be inversely related to the proportion of black population and positively

Table 1. A comparison of block groups and census tracts as units of analysis: (a) multiple linear regression model (dependent variable: $\bar{I}^{\text{F}}$), and (b) multiple logit regression model (dependent variable, P^{own}).

Variable[a]	Block groups	Census tracts
(a)		
N	871	218
R^2	0.37	0.82
P^{own}	13120 (15.3)	22664 (24.1)
P^{blue}	−8919 (−6.2)	−11800 (−7.5)
P^{black}	−5032 (−6.1)	−1848 (−2.1)
P^{eld}	−3075 (−1.9)	−5943 (−2.0)
(b)		
N	871	218
P^{black}	−0.299 (−2.08)	0.005 (0.02)
P^{blue}	0.384 (1.37)	1.075 (1.49)
$\bar{I}^{\text{F}}$	0.417 (4.47)	0.953 (4.15)
H^{MV}	0.533 (1.34)	−0.371 (−0.49)

Note: values in parentheses are the t-values of the parameter estimates; no reliable goodness-of-fit statistic was available for the logit regression model.

[a] P^{own}, P^{blue}, P^{black}, P^{eld} are the percentages of the population who are homeowners, blue-collar workers, black, and aged over 65 years, respectively; $\bar{I}^{\text{F}}$ and H^{MV} are the mean family income and the median house value (in dollars) respectively.

related to mean family income. The other two parameter estimates are not significantly different from zero. At the census-tract level, the only significant variable is $\bar{I}^{F}$, the parameter estimate of which doubled in magnitude.

The results in table 1 give a very preliminary idea of the MAUP in multivariate analysis. To examine this problem in more detail, two sets of analyses are undertaken: the first is focused on the scale dependency of parameter estimates; the second is focused on the zoning sensitivity, keeping the scale constant. Both sets of analyses involve the aggregation of the 871 block groups into other sets of areal units. In all cases, the 871 units are aggregated according to a random procedure subject to a contiguity constraint whereby a set of seed units equal to the number of units required by the aggregation procedure is selected and contiguous units are added to these seeds by a random process until all the units have been assigned. The data for each cluster are then derived from the constituent units of the cluster. In the case of the variables that are proportions, this is achieved by aggregating the actual numbers and dividing by the aggregated population or the aggregated number of dwelling units. In the case of the income and housing cost variables, the initial value reported for the block groups is a median and each aggregated value is formed by calculating the mean of the medians of the constituent block groups. Each new set of units thus represents a plausible set of reporting units for the base-level data reported for the 871 block groups. We now describe the results on the scale and zoning problems.

6 The scale problem

To examine the sensitivity of the parameter estimates from equations (5) and (6) to changes in the scale of the data used to calibrate the models, the 871 basic reporting units were aggregated in the manner described above to 800, 400, 200, 100, 50, and 25 areal units. At each level of aggregation, 20 different zoning systems were created to remove the bias of reporting results for only one set of areal units at each aggregation level. Thus, for each level of aggregation, except the 871-zone system, the models in equations (5) and (6) were calibrated separately for 20 different configurations of the same data. This results in 121 different estimates for each parameter. The results are shown in figures 2 and 3.

If the MAUP did not exist, there would be relatively little variation in each of the sets of 121 parameter estimates and certainly no systematic variation. It is quite evident from figures 2 and 3 that modification of the areal units from which data are collected does create a severe problem for parameter estimation and reliability. In figure 2 the variations of the 4 parameter estimates from the multiple linear regression model in equation (5) are depicted, and several trends are worth noting. For 2 of the 4 sets of parameters (the P^{eld} and P^{blue} estimates) the estimates vary *systematically* as the number of areal units decreases, becoming more negative in both cases. In the case of the P^{eld} estimates, for example, the value when the model is calibrated on 871 areal units is -3075, which is not significantly different from zero at the 95% confidence limit, whereas the mean estimate when the data are aggregated to only 25 units is -26540 and all 20 estimates at this level are significantly different from zero. In the former case, an increase of 0.1 in the proportion of elderly in a unit would create a decrease in the predicted mean family income of only \$308, whereas in the latter case a similar increase would produce a decrease of \$2654. For the other two sets of parameters, the P^{own} and P^{black} estimates, there is relatively little systematic variation, although the estimates for P^{black} appear to become less negative as aggregation increases. We explore these trends in more detail below.

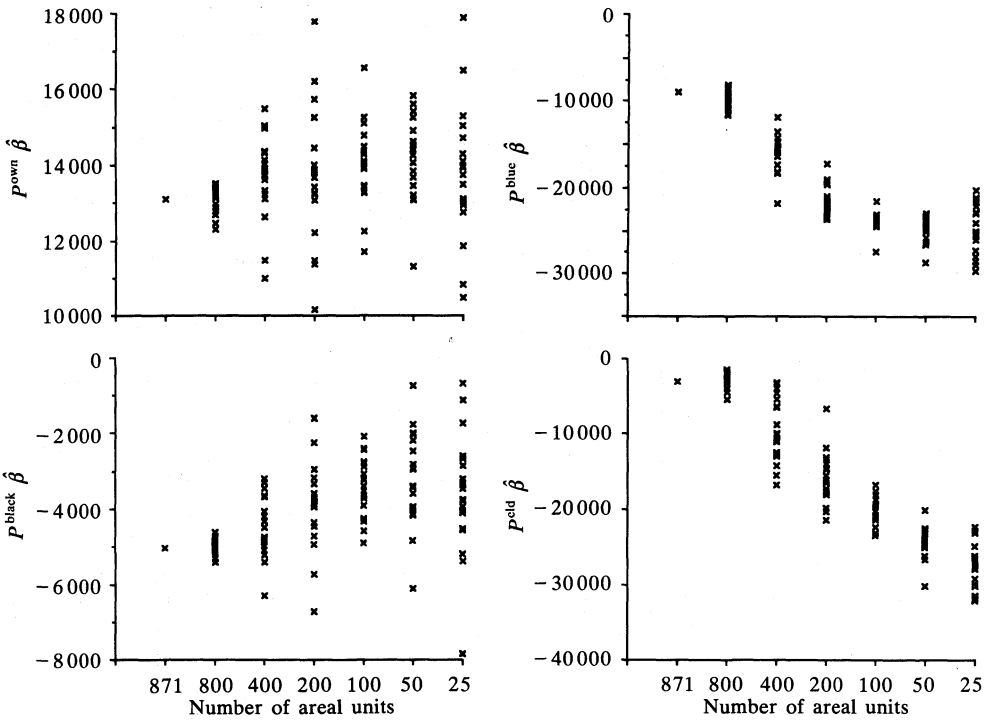

Figure 2. Variations in parameter estimates with scale changes from a multiple regression model. Note: see table 1, footnote a.

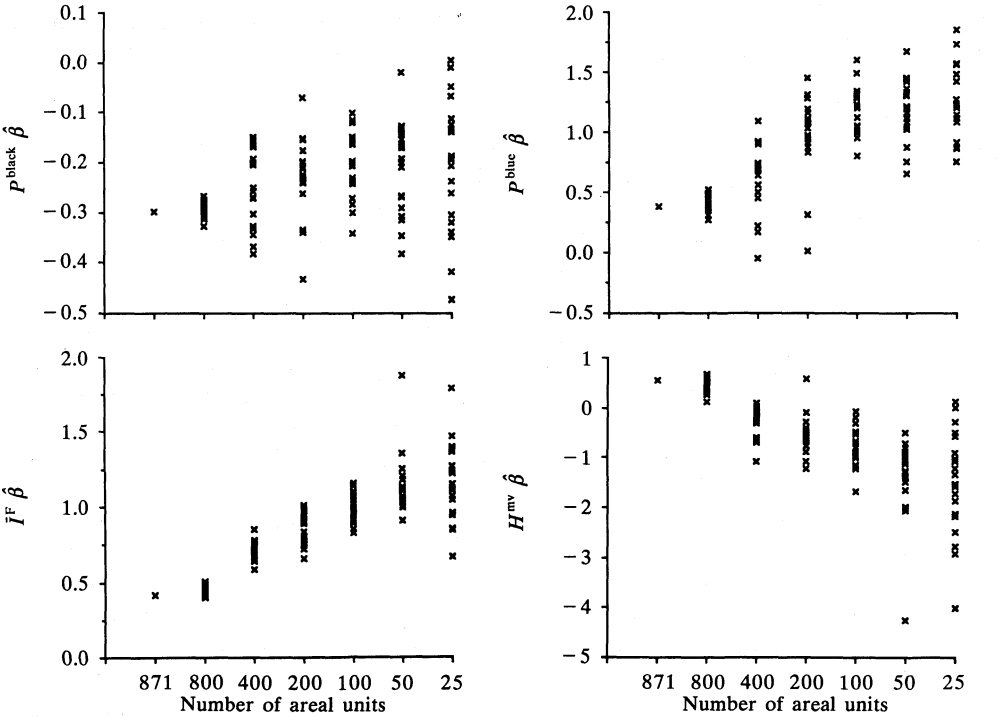

Figure 3. Variations in parameter estimates with scale changes from a multiple logit regression model. Note: see table 1, footnote a.

Although it might appear from figure 2 that the parameter estimates associated with the variables P^{own} and P^{black} exhibit a greater degree of sensitivity to zoning-system variations, this is primarily because these parameter estimates exhibit slightly greater stability as the areal units are aggregated and this affects the vertical scale of the graphs. The range of the parameter estimate for P^{own} for 25 units is 17910 to 10430 whereas the range of the estimate for P^{blue} at that level is -20130 to -30040. The effects of zoning-system variation are examined in more detail in the subsequent section.

The results in figure 3, which are obtained from the calibration of the logit regression model in equation (6), show similar susceptibility to data aggregation. The parameter estimates associated with the variables P^{black}, P^{blue}, and $\bar{I}^{F}$ systematically increase as the data are aggregated; the estimates associated with the variable H^{mv} systematically decreases. Given the similarity of results from the two models, the subsequent analyses will concentrate on the more familiar linear regression model.

In order to shed more light on the sensitivity of the parameter estimates to variations in data aggregation, it would be useful to determine if the differences discussed above and depicted in figures 2 and 3 are statistically significant. For instance, the apparent greater sensitivity to variations in aggregation level of the parameter estimates associated with the variables P^{blue} and P^{eld} in figure 2 may not translate into statistically significant variations because of differences in the standard errors of the estimates. It could also be possible, although unlikely, that the large differences reported in figure 2 are not statistically significant; in which case the MAUP would not exist.

Information on the significance of the variations in figure 2 is provided in figure 4.[1] The parameter estimates are arranged somewhat differently in that the zoning-system variation, depicted vertically in figure 2, is now depicted horizontally within each level of aggregation so that a confidence interval around each estimate can be added. This is done by placing symbols (a square above and a cross below) 1.5 standard errors either side of the parameter estimate. The reason for selecting 1.5 standard errors is because the focus of interpretation is on whether estimates are significantly different from each other. If the lower 1.5 standard error symbol of one parameter estimate is above the upper 1.5 standard error symbol of another, the two are very likely to be significantly different at the 95% confidence level according to the standard difference in means test:

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\mathrm{SE}|\hat{\beta}_1 - \hat{\beta}_2|}, \tag{7}$$

where SE is the standard error. Thus, at a glance, it is relatively easy to see that the parameter estimates for P^{own} and P^{black} are much less sensitive to changes in data aggregation than are the estimates for P^{blue} and P^{eld}. In the case of the former, there are few places in figures 4(a) and 4(c) where the crosses lie above the squares and hence the MAUP is virtually absent. In the case of the latter, there are many places in figures 4(b) and 4(d) where the parameter estimates are significantly different. Compare, for instance, the parameter estimates at the 800-unit level with those from the 100-unit level; for both variables the parameter estimates are significantly larger at the more disaggregated level. For these variables the MAUP is evidently quite severe.

[1] A note of caution is advisable here. If the unobserved error terms in the model are spatially autocorrelated, the ordinary least squares estimators of the variances of the parameter estimates will be biased, although the actual parameter estimates remain unbiased (Miron, 1984).

All four variables exhibit the trend that as the data become increasingly aggregated, the standard errors of the parameter estimates increase, which is not unexpected as the standard error depends in part on the number of observations used in the calibration. The sensitivity of the parameter estimates to variations in zoning system also seems to increase as the level of aggregation increases, which is also not unexpected given that the MAUP is an aggregation problem. However, in general, there appears to be an encouraging stability of estimates within any level of aggregation, although this topic is pursued further in the subsequent section.

Despite the above, there are relatively few trends common to the behaviour of the four sets of parameter estimates under varying levels of data aggregation. What, for instance, causes the estimates for the variables P^{blue} and P^{eld} in figures 4(b) and 4(d) to exhibit such different behavior from that of the estimates for the variables P^{own} and P^{black} in figures 4(a) and 4(c)? Is there any link with the spatial autocorrelation of the variables? In order to examine any possible relationship between variations in the level of spatial autocorrelation exhibited by a variable and the variations exhibited by its parameter estimates, Moran's I coefficient was calculated for all five variables in equation (5) for every data set. These values are depicted in figure 5 for the dependent variable, mean family income, and in figure 6 for the four independent variables.

For four of the five variables, P^{black} being the exception, the graphs of the Moran coefficient have an approximately normal shape, with spatial autocorrelation being highest for intermediate levels of aggregation. In the case of the P^{black} variable, spatial autocorrelation is highest at the lowest levels of aggregation and decreases monotonically as the data become increasingly aggregated. This suggests that the P^{black} variable is the most concentrated at a local scale whereas the other variables

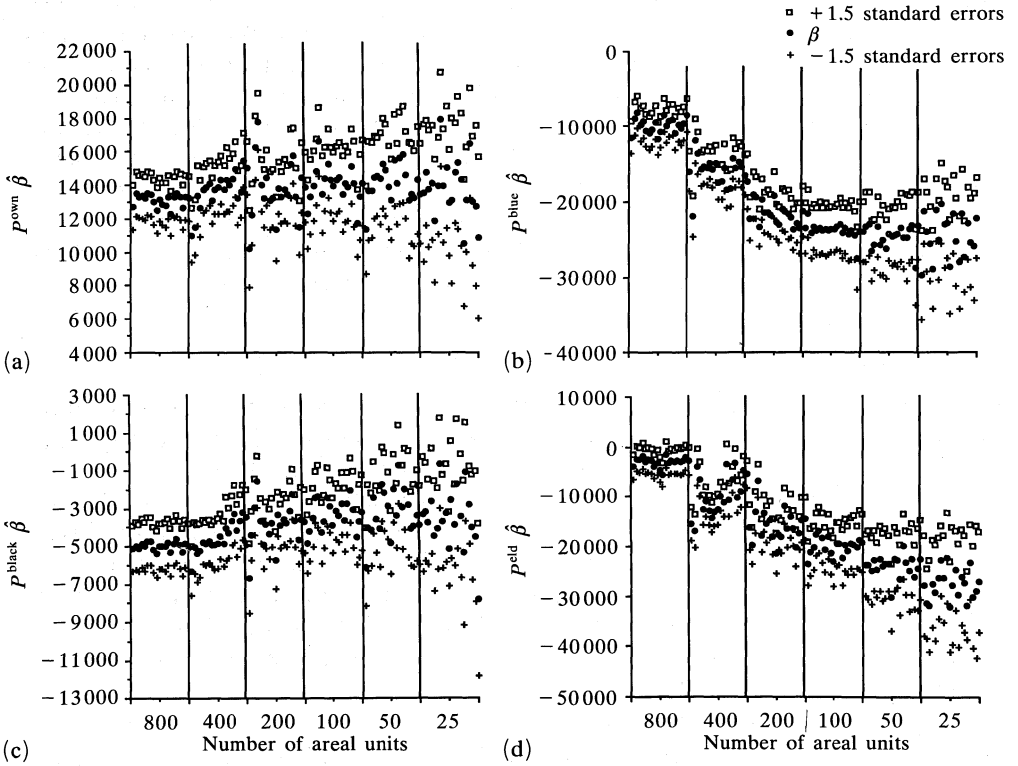

Figure 4. Confidence intervals around the parameter estimates for the variables, (a) P^{own}, (b) P^{blue}, (c) P^{black}, and (d) P^{eld}. Note: see table 1, footnote a.

exhibit local differences which diminish as the data are aggregated. It is clear that there is no such thing as an optimal zoning system for all variables in terms of minimizing spatial autocorrelation.

By comparing figures 5 and 6 with figure 2, it does not appear that there is any connection between the level of spatial autocorrelation of any of the variables and the severity of the MAUP, either as a scale problem or as a zoning problem. For instance, the behavior of the parameter estimates for the variables P^{own} and P^{black} in figure 2 correspond closely, yet the trends in the spatial autocorrelation of the two variables in figure 6 is quite different. Similarly, the fact that the level of spatial autocorrelation is generally higher for P^{blue} than for P^{eld} does not appear to affect the sensitivity of the parameter estimates to scale or zoning-system changes in any substantial way.

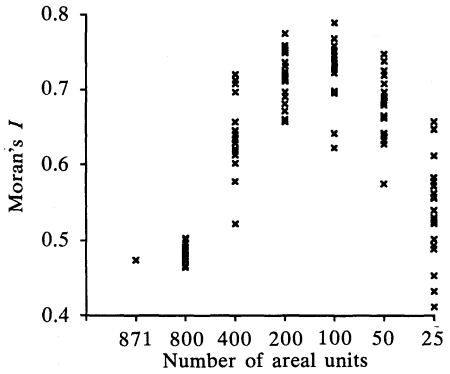

Figure 5. Moran's I for the mean family income, $\bar{I}^F$.

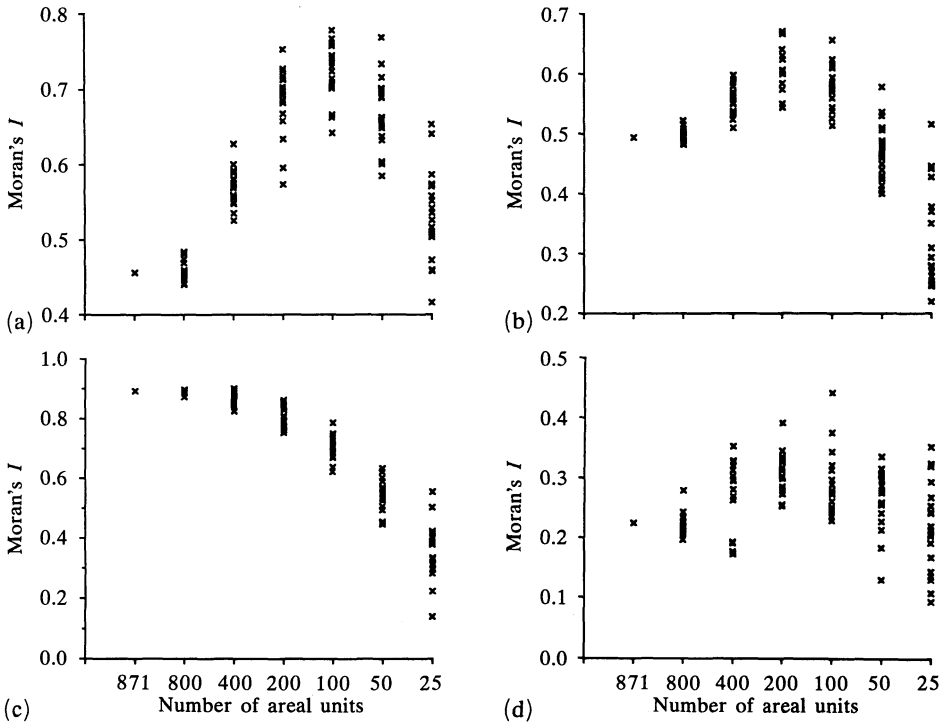

Figure 6. Moran's I for (a) P^{own}, (b) P^{blue}, (c) P^{black}, and (d) P^{eld}. Note: see table 1, footnote a.

As a footnote to the scale problem, we report in figure 7 the coefficient of multiple determination for each calibration of equation (5). The trend echoes that found in bivariate regression with the goodness of fit of the model apparently increasing rapidly as the data are aggregated. What perhaps is surprising is the rapidity of the increase in the goodness-of-fit statistic, which averages approximately 0.4 at the 800-zone level to approximately 0.85 at the 100-zone level. It is interesting to note the fairly low variance of the statistics at a particular level of data aggregation: it seems impossible to get a 'good fit' at the 800-zone level and equally impossible to get a 'bad fit' at the everything above the 200-zone level! Clearly, the coefficient of determination is as reliable as the parameter estimates reported in an analysis of data drawn from areal units.

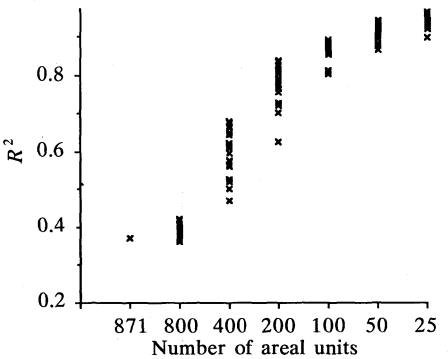

Figure 7. Variations in R^2 from a multiple linear regression model with scale changes.

7 The zoning problem

Although the focus of section 6 is on the scale problem, some initial evidence on the zoning problem in multivariate analysis is also presented. At each level of aggregation, the models in equations (5) and (6) were calibrated with data from 20 different zoning systems. These initial results suggest that generally the parameter estimates do not appear to be as sensitive to the zoning system design as they are to the level of aggregation. However, there are exceptions to this general observation and the results are hardly definitive being based on only 20 different zoning systems at each level. In order to examine the zoning problem more comprehensively, a detailed analysis was undertaken for one level of aggregation, that corresponding to the 218-zone system of census tracts. Data for the 871 block groups were aggregated into 218 zones in 150 different ways by using the random aggregation procedure with a contiguity constraint described above (section 5). The dispersion of the four sets of parameter estimates is described in figure 8 where the value of the parameter estimate is graphed against its t value where t is the parameter estimate divided by its standard error. With 213 degrees of freedom, absolute values of t above 1.96 are significantly different from zero at the 95% confidence limit.

For the variable P^{own}, all 150 of the parameter estimates from the 218-zone systems are significantly positive with t values ranging between 5 and 15, fairly conclusive evidence that the percentage of owner-occupied housing is a good indicator of mean family income. The range of the parameter estimates also does not seem unduly extreme: an increase of 0.1 in P^{own} produces an increase in predicted mean family income ranging from \$1000 to \$2000. The zoning problem appears to be more severe for the variables P^{blue} and P^{black}: for both variables the estimates exhibit a large variation and although most are generally significantly negative, there are several estimates which are not significantly different from zero.

In the case of P^{blue}, for example, it is possible to conclude with one set of zones that the percentage of blue-collar workers does not affect mean family income, whereas, with another, one could report that an increase in P^{blue} of 0.1 will reduce the predicted $\bar{I}^F$ by over \$20 000! The parameter estimates for the variable P^{eld} perhaps exhibit the most extreme behavior. Although most of the estimates are not significantly different from zero, several are significantly negative whereas two are significantly positive. With one zoning scheme it is therefore possible to conclude that compared with the average, the elderly have significantly *lower* incomes, ceteris paribus, whereas, with another, it can equally legitimately be concluded that the elderly have significantly *higher* incomes, ceteris paribus.

To compare the parameter estimates of the 150 internally generated zoning systems with those from the census tracts generated by the US Bureau of the Census, the 151 estimates for each variable are depicted vertically in figure 9 along with those from the block-group level analysis.[2]

The aggregation of the block-group data into census tracts produces parameter estimates for the variables P^{blue} and P^{eld} which are reasonably representative of the

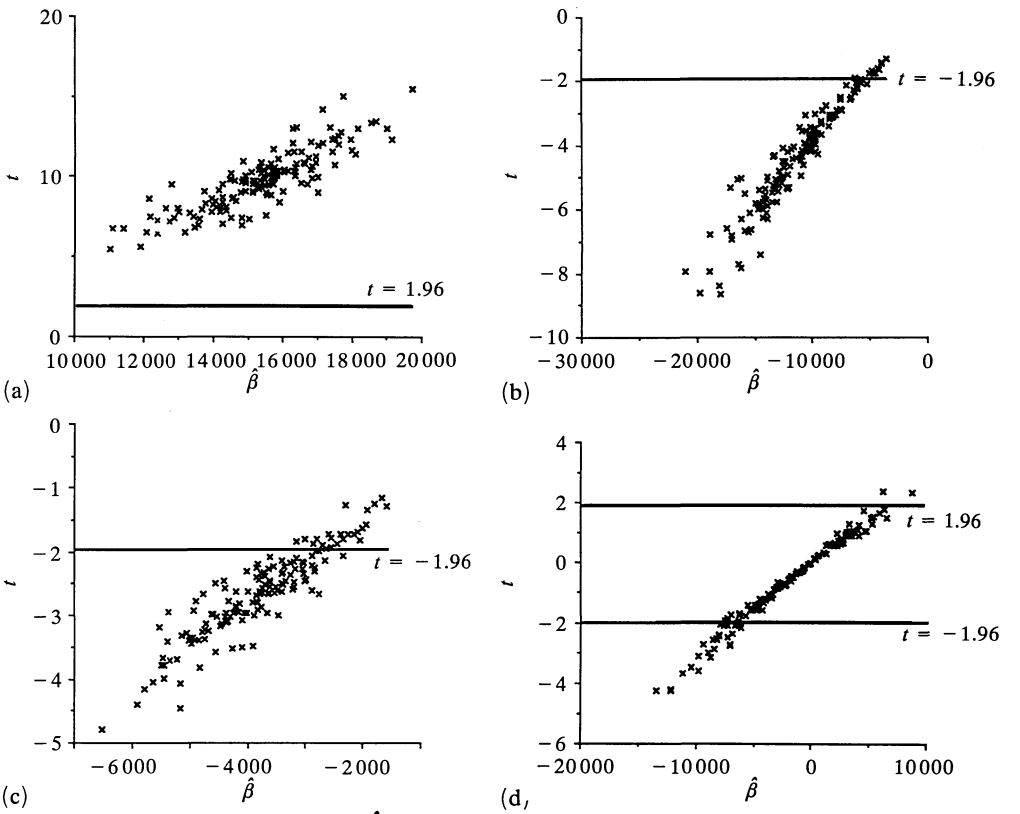

Figure 8. Scatterplots of t against $\hat{\beta}$ for 150 representations of the block-group data aggregated to 218 zones for (a) P^{own}, (b) P^{blue}, (c) P^{black}, and (d) P^{eld}. Note: see table 1, footnote a.

[2] The parameter estimates reported here for census tracts are marginally different from those reported in table 1. The values in table 1 are obtained by use of data reported at the census-tract level; the values in figure 9 are obtained by use of data aggregated from the block-group data files. The difference results from the definition of the $\bar{I}^F$ variable, which when reported for the census tracts is the median family income but when reported from the aggregated block-group data is the mean of the block-group medians.

150 values obtained when internally generated zones are used. The census tract estimates for the variables P^{own} and P^{black}, however, are clearly highly unusual both in terms of the other 150 representations of data at this level and in terms of the more disaggregated block-group representation. The results suggest that the aggregation of block groups into census tracts produces a rather biased 218-zone coverage for P^{own} and P^{black} compared with the 150 aggregations generated in this study.

It could be argued that the census tracts are designed with greater internal homogeneity, and hence greater external heterogeneity, than are the systems generated in this study and that the criteria for homogeneity are more likely to include variables such as the proportions of homeowners and blacks rather than blue-collar workers and elderly (USBC, 1980). However, a comparison of the variances of the four variables for the 150 internally generated systems and the set of census tracts, shown in figure 10, does not support this argument. At the census-tract level, the two variables with highly unusual parameter estimates, P^{own} and P^{black}, do not exhibit extreme variances and whereas one is greater than the average of the internally generated zones, the other is less. Of the two variables where the census parameter estimates are fairly centrally located in the distributions shown in figure 9, P^{blue} and P^{eld}, the between-census-tract variance is centrally located on figure 10 for the former, whereas the latter has an extremely low variance. It would thus appear that the internally generated zones in this study do not display any unusual heterogeneity compared with census tracts and in the case of one variable, P^{eld}, the census tracts exhibit much greater heterogeneity.

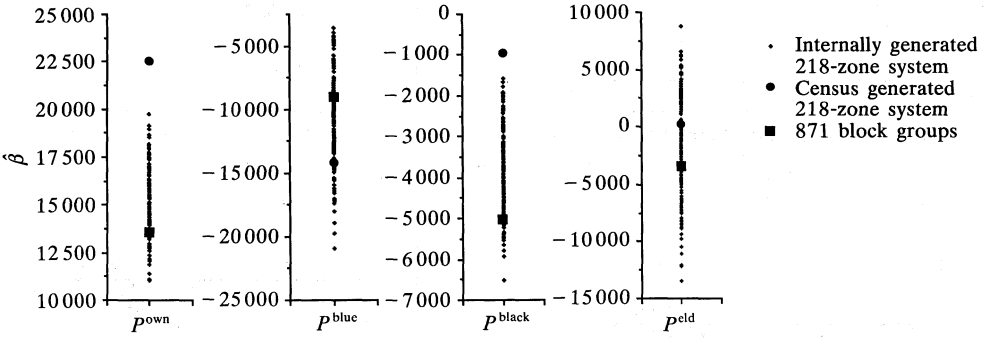

Figure 9. Parameter estimates for 152 representations of the block-group data. Note: see table 1, footnote a.

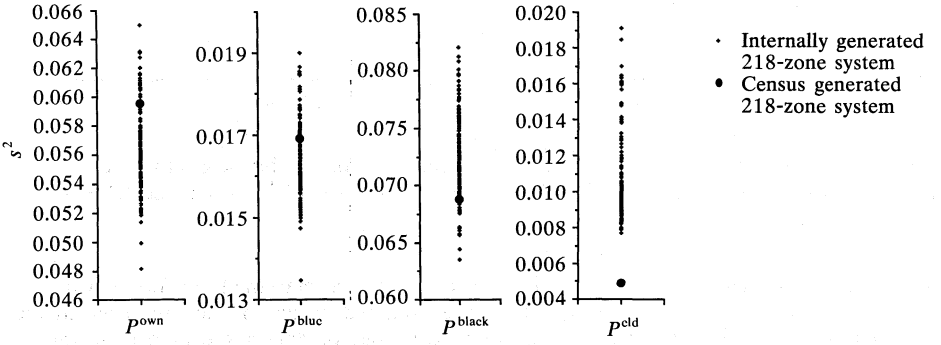

Figure 10. Variances, s^2, for the 151 representations of the block-group data. Note: see table 1, footnote a.

The results in figure 10 thus demonstrate two points: the unusual census-tract results in figure 9 do not appear to be a function of internal homogeneity; and the zone-generation system employed here does not produce highly unusual sets of zones.

8 Summary

The modifiable areal unit problem, the sensitivity of model calibration to variations in scale and zoning systems, is shown to produce highly unreliable results in the multivariate analysis of data drawn from areal units. Given that a large number of geographical studies make use of such data, perhaps the most common being the analysis of aggregated census data, this is rather worrying. In the foreseeable future, given the lack of individual-level data, the imminent publication of census figures in the USA, Canada, and the United Kingdom, and the increasing capabilities of geographical information systems to undertake multivariate analyses of areal data, the MAUP should become of even greater interest to spatial analysts. It therefore seems especially timely to alert potential users to the MAUP in multivariate analysis.

In this paper, the sensitivity of the calibration results of two multivariate models, a multiple linear regression model, and a multiple logit regression model to variations in scale and zoning systems has been explored. A data set at the block-group level for the Buffalo Metropolitan Area was used to calibrate the models. The results are rather depressing. It has been shown, for example, that in an analysis of the spatial distribution of mean family income, an 800-zone data system yields the result that an increase of 0.1 in the proportion of elderly would create a decrease in the predicted mean family income of only \$308. When these data are aggregated to only 25 zones, the results suggest that the same increase would produce a decrease in predicted mean family income of \$2654. An analysis of goodness of fit in the same model shows that there is an alarming increase in the coefficient of multiple determination as data are increasingly aggregated. The results suggest that it is possible to find any desired level of accuracy simply by aggregating the data sufficiently.

Similar variations in results are demonstrated when different zoning systems are used at the same level of aggregation. A total of 150 different zoning systems were created, each containing 218 zones, by different aggregations of the original 871 block groups. The same multiple linear regression model was calibrated for each data set and the results compared with those obtained from use of census-tract data (another 218-zone system aggregated from the block-group data). From the same data set it was possible to report from one zoning system that the proportion of blue-collar workers does not affect mean family income, whereas from another it was possible to report that an increase of 0.1 in the proportion of blue-collar workers would reduce the predicted mean family income by over \$20000. In the case of the influence of the proportion of elderly on mean family income, it was possible to report from one zoning system that there was a significant positive relationship, from another that there was no significant relationship, and from yet another than there was a significant negative relationship!

When one considers that models very similar to those calibrated here are used to formulate policies concerning income distribution and housing, the results take on added significance. It is clearly possible to find almost any desired result by aggregating the data in different ways. The reliance on the results from one particular zoning system, be they census tracts, block groups, or some other similar aggregation of individual-level data, must be highly suspect.

Although the extent of the MAUP in multivariate analysis could perhaps have been anticipated given other studies in univariate and bivariate analysis, it is

demonstrated here that the effects of the MAUP in multivariate analysis, unlike those in univariate and bivariate analysis, are essentially unpredictable. Even in the simplest multiple regression containing only two independent variables, the interaction between changes in variances and covariances cannot be anticipated. In this study, for example, in which we utilize a linear model with four independent variables, it could not be anticipated that some parameter estimates would increase in absolute magnitude as the level of data aggregation increased, whereas others would decrease. An examination of spatial autocorrelation did not suggest any link between the level of spatial autocorrelation of a variable and its sensitivity to the MAUP.

Given that the MAUP makes any single multivariate analysis of aggregated data highly suspect, are there any solutions? Clearly one way of assessing the importance of the MAUP and to document its effects on calibration results is to report results at different levels of aggregation and with different zoning systems at the same scale. Such reporting could be made much easier with the increased use of GIS technology. It is easy to envisage, for instance, a technology that will allow the user to rezone data, to aggregate it in a specified number of ways, and to report a summary of calibration results for each of the different zoning systems. The focus of reporting such results could then change from a single scale and for one particular zoning scheme towards an analysis of how the results change with scale and zoning scheme. Are some results more stable than others and why?

Another solution would be to avoid the use of aggregated data where possible as the MAUP is a product of aggregation. However, this is clearly not a viable solution for many types of analysis, particularly those that use census data and where confidentiality becomes a problem for the use of highly disaggregated data. There are also some types of data for which basic units, such as individuals, do not exist. For instance, suppose data on population density is being analyzed for an urban area. Given that density is obtained by dividing population within an areal unit by the area of that unit, no limit to the disaggregation of the data can be reached. Similar problems also arise in the analysis of many physical processes such as those affecting soil properties or water quality. What, for instance, is a basic unit of soil in which to measure cation-exchange capacity or porosity?

A third possible solution is to create 'optimal' zoning systems (Moellering and Tobler, 1972; Openshaw, 1978; 1984). However, what constitutes 'optimal' in terms of multivariate anlaysis is likely to be rather subjective. A zoning system that is optimal for one variable, in terms of maximizing interzonal variation and minimizing intrazonal variation, for example, might not be optimal for another. It would be necessary to derive some overall optimizing criterion and then Openshaw's suggestion of jointly estimating both the zoning system and the model parameters would be worth exploring (Openshaw, 1978). Further discussion of some of these topics is provided in Fotheringham (1989).

On the one hand, this paper thus paints a rather depressing picture for the future of the multivariate analysis of aggregated spatial data. Calibration results from one set of areal units are highly suspect and should not be relied upon to draw any substantive conclusions about the underlying relationships being examined. On the other hand, the MAUP does provide spatial analysts with a fascinating and intriguing challenge, the solution to which would form the basis for an important body of spatial theory.

Acknowledgements. The authors would like to express their thanks for the financial support of the National Center for Geographic Information and Analysis (NCGIA) under NSF grant SES-8810917. This work was prompted by the first author's attendance at the Specialist Meetings of the Research Initiatives on The Accuracy of Spatial Databases and on Multiple Representations, sponsored by the NCGIA.

References

Amrhein C G, Flowerdew R, 1989, "The effect of data aggregation on a Poisson regression model of Canadian migration", in *Accuracy of Spatial Databases* Eds M F Goodchild, S Gopal (Taylor and Francis, London) pp 229-238

Arbia G, 1989 *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems* (Kluwer, Dordrecht)

Bach L, 1981, "The problem of aggregation and distance for analyses of accessibility and access opportunity in location-allocation models" *Environment and Planning A* **13** 955-978

Batty M, 1974, "Spatial entropy" *Geographical Analysis* **6** 1-31

Batty M, 1976, "Entropy in spatial aggregation" *Geographical Analysis* **8** 1-21

Batty M, Sammons R, 1978, "On searching for the most informative spatial pattern" *Environment and Planning A* **10** 747-779

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

Batty M, Sikdar P K, 1982b, "Spatial aggregation in gravity models: 2. One-dimensional population density models" *Environment and Planning A* **14** 525-553

Batty M, Sikdar P K, 1982c, "Spatial aggregation in gravity models: 3. Two-dimensional trip distribution and location modles" *Environment and Planning A* **14** 629-658

Batty M, Sikdar P K, 1982d, "Spatial aggregation in gravity models: 4. Generalisations and large-scale applications" *Environment and Planning A* **14** 795-822

Blair P, Miller R E, 1983, "Spatial aggregation in multiregional input-output models" *Environment and Planning A* **15** 187-206

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

Cadwallader M T, 1985 *Analytical Urban Geography* (Prentice-Hall, Englewood Cliffs, NJ)

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

Fotheringham A S, 1989, "Scale-independent spatial anlaysis", in *Accuracy of Spatial Databases* Eds M F Goodchild, S Gopal (Taylor and Francis, London) pp 221-228

Gale D E, 1984 *Neighborhood Revitalization and the Postindustrial City: A Multinational Perspective* (Lexington Books, Lexington, MA)

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 Supplement* **29** 169-170

Goodchild M F, 1979, "The aggregation problem in location-allocation" *Geographical Analysis* **11** 240-255

Griffith D A, 1988 *Advanced Spatial Statistics* (Kluwer, Dordrecht)

Lipton S G, 1977, "Evidence of central-city revival" *Journal of the American Institute of Planners* **43** 136-147

Masser I, Brown P J B, 1975, "Hierarchical aggregation procedures for interaction data" *Environment and Planning A* **7** 509-523

Masser I, Brown P J B, 1977, "Spatial representation and spatial interaction" *Papers of the Regional Science Association* **38** 71-92

Miron J, 1984, "Spatial autocorrelation in regression analysis: a beginner's guide", in *Spatial Statistics and Models* Eds G L Gaile, C J Willmott (Kluwer, Dordrecht) pp 201-222

Moellering H, Tobler W, 1972, "Geographical variances" *Geographical Analysis* **4** 34-50

Nelson K P, 1988 *Gentrification and Distressed Cities: An Assessment of Trends in Intrametropolitan Migration* (University of Wisconsin Press, Madison, WI)

Openshaw S, 1977, "Optimal zoning systems for spatial interaction models" *Environment and Planning A* **9** 169-184

Openshaw S, 1978, "An empirical study of some zone-design criteria" *Environment and Planning A* **10** 781-794

Openshaw S, 1984 *Concepts and Techniques in Modern Geography, Number 38. The Modifiable Areal Unit Problem* (Geo Books, Norwich)

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

Perle E D, 1977, "Scale changes and impacts on factorial ecology structures" *Environment and Planning A* **9** 549-558

Putman S H, Chung S-H, 1989, "Effects of spatial systems design on spatial interaction models. 1: The spatial definition problem" *Environment and Planning A* **21** 27-46

Robinson A H, 1956, "The necessity of weighting values in correlation analysis of areal data" *Annals of the Association of American Geographers* **46** 233-236

Robinson W S, 1950, "Ecological correlations and the behavior of individuals" *American Sociological Review* **15** 351-357

Sawicki D S, 1973, "Studies of aggregated areal data—problems of statistical inference" *Land Economics* **49** 109-114

Thomas E N, Anderson D L, 1965, "Additional comments on weighting values in correlation analysis of areal data" *Annals of the Association of American Geographers* **55** 492-505

USBC, 1980 *Census '80: Continuing the Factfinder Tradition* US Bureau of the Census, by C P Kaplan, T L Van Valey and Associates, Washington, DC (US Government Printing Office, Washington, DC)

USBC, 1981a *Census of Population and Housing, 1980: Summary Tape File 1 Technical Documentation* US Bureau of the Census (US Government Printing Office, Washington, DC)

USBC, 1981b *Census of Population and Housing, 1980: Summary Taype File 1A (New York)* (US Bureau of the Census, Washington, DC)

USBC, 1981c *Census of Population and Housing, 1980: Summary Tape File 3A (New York)* (US Bureau of the Census, Washington, DC)

USBC, 1983 *1980 Census of Population and Housing, Census Tracts for Buffalo, NY SMSA* US Bureau of the Census (US Government Printing Office, Washington, DC)

Wrigley N, 1985 *Categorical Data Analysis for Geographers and Environmental Scientists* (Longman, Harlow, Essex)