DRAFT 4 Jan 1999

Unravelling Ecological Analysis

D. G. Steel (1), **M. Tranmer** (2) **and D. Holt** (3)

(1) School of Mathematics and Applied Statistics, University of Wollongong, NSW 2525, Australia

(2) Department of Social Statistics, University of Southampton, SO17 1BJ, UK

(3) Office for National Statistics, London, SWIV 2QQ, UK

Summary

Ecological analysis involves analysing aggregate data for groups of individuals to make inferences about relationships at the individual level. Often the results of such an analysis give badly biased estimates. While the problem has been recognised for over 60 years there has been little effort put into solving it by statisticians. Most of the developments have been done by researchers in fields who are faced with the problem of having to use aggregate data to make individual level inferences. This paper will discuss the sources of aggregation bias. The key role of variation of the individual level relationships between groups and the consequent within group correlations and how these can be explained by auxiliary variables will be considered.

1. Introduction

Ecological analysis involves analysing aggregate data such as the means or totals for a set of groups to make inferences about individual level relationships. The groups are often geographical areas such as post-code areas, but can also be institutional groups such as schools or social groups.

The main advantage of ecological analysis is that it uses data that are already available at relatively low cost. The data may come from a statistical collection such as the population census which makes a lot of data available as aggregates for geographical areas, for example Enumeration Districts (EDs). Government administrative systems provide information on a range of variables in aggregate form. For example, the Australian Taxation Office produces information on average taxable income for post-code areas. Information from different sources may be brought together by using aggregates for the same geographical areas and attempting an ecological analysis. Using Geographical Information Systems (GISs) such analyses can be easily carried out. The current interest in data mining reflects the fact that many private organizations also have large data sets. Such data can be analysed in conjunction with data from other sources using ecological analysis by using common geographical areas.

Ecological analysis is a potential valuable statistical tool, but it is subject to the ecological fallacy which arises when the results are incorrectly assumed to apply to relationships at the individual level. The potentially serious biases in the estimates of individual level relationships severely limits the practical use of ecological analysis. Also, in some cases the number of groups in the analysis will be small affecting the variances of the estimates obtained from ecological analysis. The problem of potential bias is much more important as there is usually no indication of the bias, whereas the number of groups in the analysis is known.

Most of the attempts to put ecological analysis on a firm basis have has been done by researchers, such as geographers, political scientists, sociologists, epidemiologists and economists, facing the problem in practical situations .

In section 2 we consider the issue of the targets of inference and how ecological analysis can be considered as a form of mult-level modelling. In section 3 we outline an approach based on using auxiliary variables to account for group effects. In section 4 we consider ecological linear regression analysis within a multi-level framework that clearly identifies the sources of the biases affecting ecological analysis. In section 5 we consider the important case of two by two tables. In section 6 we indicate ways forward to develop a general approach to producing reliable ecological analysis.

2. Targets of inference and ecological analysis in multi-level populations

In ecological analysis the population is composed of groups of individual units and thus have a multi-level structure. Hence any statistical analysis should be based on a statistical model that incorporates this structure, i.e. a multi-level model. For simplicity consider a two level model, the first level being the individual. Consider a population of N individuals, the ith individual has a vector of response variables y_i and an area indicator c_i . The population comprises M areas and the number of individuals in the gth area is N_g . Assume that there exists a sample data set s of size n and that these individual data have been aggregated to provide a set of m area means, $\bar{y}_g, g = 1, ..m,$ which are available for analysis. The number of sample individuals in each area n_g is also known. The overall sample mean is $\bar{y} = \dfrac{1}{n}\sum_{g \in s} n_g \bar{y}_g$.

Initially consider a multivariate situation in which there are no covariates. A simple multi-level model is

$$y_i = \mu + \nu_g + \epsilon_i \quad i \in g \tag{2.1}$$

where ν_i is a vector of random group level effects and ϵ_{ij} is a vector of individual level effects. The standard assumptions are

$$E[\nu_g] = E[\epsilon_i] = 0 \tag{2.2}$$

and

$$V(\nu_g) = \Sigma^{(2)}, \quad V(\epsilon_i) = \Sigma^{(1)}, \quad \text{and} \quad Cov(\nu_g, \epsilon_i) = 0 \tag{2.3}$$

For this model

$$E[y_i] = \mu, \quad V(y_i) = \Sigma^{(1)} + \Sigma^{(2)} = \Sigma, \quad Cov(y_i, y_j) = \Sigma^{(1)} \text{ for } c_i = c_j, i \neq j$$

The development of multi-level models initially arose from situations in which the groups have a clear potential to affect the responses of individuals. However, they provide a useful framework for any situation in which the process that generated the data involved the groups, either through sampling or aggregation, or both. We should specify a statistical model relevant to the substantive issues, and clearly define our targets of inference in terms of the parameters of this model.

There are three important cases to consider within a multi-level framework according to the type of inference required and the data available.

(1) Multi-Level Modelling. Interest is in the relationships at the unit level and the impact the groups have on these relationships. This is now the standard situation in which multi-level modelling is applied, using unit level data which include group identifiers (see Goldstein, 1986, 1995). The analysis may also include group level or contextual explanatory variables. These may be global variables, which are variables not based on characteristics of individuals in the group as well as analytical variables calculated from individual level variables.

(2) Ecological Analysis. Here interest is usually in relationships at the unit level marginal, that is unconditional, to the groups. The assumption is that if, for example, a simple random sample of individual units was available, the researcher would be happy to analyse it completely ignoring any groups in the population. The data available are at the group level.

(3) Strutural Group Level Analysis. Here interest is focussed on relationships purely at the group level. It can be important to recognize that the groups are composed of individuals and attempt to remove the individual level effects from the analysis.

While this paper is mainly concerned with (2) it is useful to clarify the commonality of and differences between the situations. In (1) and (2) the interest is in explaining the variation in individual level response variables, whereas in (3) the variation in group level response variables is of interest.

In standard multi-level modelling the targets of inference are the fixed mean parameter μ and the paramaters of the distribution of the random components $\Sigma^{(1)}$ and $\Sigma^{(2)}$. Taking the group level variance components into account enables efficient estimation of μ and calculation of appropriate estimates of standard errors. Moreover, the parameters $\Sigma^{(1)}$ and $\Sigma^{(2)}$ usually has some substantive significance by indicating the relative importance of purely individual and group level effects. Estimation of the parameters usually requires a sample of groups and a sample of individuals within them and indicators which indicate to which group each individual belongs.

In ecological analysis the main data available consist of group level means or totals which are often based on a census of groups and individuals within them, but can also come from a sample. However, the target of inference are at the unit level. Usually this is taken to mean parameters that describe the relationships between variables marginal to the groups, which in the model given by (2.1) to (2.3) would be μ and Σ. The main concern is with the ecological fallacy which in this case would occur when a covariance matrix is calculated from the group means and provides a biased estimate of Σ and functions of it such as regression and correlation coefficients. Essentially the bias arises because the group level covariance matrix has expectation that is a linear combination of $\Sigma^{(1)}$ and $\Sigma^{(2)}$ with the wrong implicit weights given to the two components (see Holt, Steel and Tranmer, 1996). To remove the bias requires estimation of $\Sigma^{(1)}$ and $\Sigma^{(2)}$.

Consideration has to be given as to whether the target of inference should be the marginal parameter Σ which characterizes the relationships between variables across the whole population ignoring the groups or the parameter $\Sigma^{(1)}$ which characterizes individual level relationships conditional on the group effects. Ecological analysis is often carried out using data for geographically defined groups which have no substantive role and which have a degree of arbitrayness in their definition. In many cases the number of groups and their boundaries could be in theory be changed. In these cicumstances having a target of inference which describes relatonships conditional on the group effects may not be appropriate. Even if the groups have a substantive role marginal relationships may be relevant, for example if the government is planning some policy that will be applied across the whole population. However, there will also be situations in which we are interested in both $\Sigma^{(1)}$ and $\Sigma^{(2)}$, or just $\Sigma^{(1)}$, and the data are available in aggregate form. In general ecological analysis is a form of multi-level modelling with aggregate data.

In many geographical applications there is no direct interest in individual level relationships. However, even if the primary interest of an analysis is at a higher level than the individual, in many situations it is important to recognise that the groups involved are composed of individuals. The focus of interest in a structural group level analysis may be at a level above the individual i.e. $\Sigma^{(2)}$, but the area level analysis will include a component due to the individual level relationships in the population i.e. $\Sigma^{(1)}$. Hence it is desirable to separate the two components.

Thus while these three situations have different objectives, they all require estimation of the variance components $\Sigma^{(1)}$ and $\Sigma^{(2)}$. How this can be attempted depends on the data available. The information available for analysis can consist of unit level or aggregate data, or a combination of both. Moreover, the unit level data may indicate the groups to which indivduals belong or it may not. There are five cases that can be identified in practice.

1. Unit level data with group indicators.
2. Unit level data with no group indicators.
3. Aggregate group level data.
4. Aggregate group level data and unit level data with no group indicators.
5. Aggregate group level data and unit level data with group indicators.

Conventional multi-level modelling is carried out using a unit level data set which has group indicatives. For example Maximum likelihood estimation (MLE) can be done using Iterated Generalized Least Squares (IGLS) (Goldstein, 1986).

When only aggregate data are available MLEs of the fixed parameters and the variance components can be obtained in theory. The key to separating $\Sigma^{(1)}$ and $\Sigma^{(2)}$ is the form of heteroscedasticity for the group means implies by model (1), i.e.

$$V(\bar{y}_g) = \frac{\Sigma^{(1)}}{n_g} + \Sigma^{(2)} \tag{2.4}$$

The effectiveness of this approach depends on the variation in the group sample sizes, n_g and the validity of the assumed form of the heteroscedasticity in (2.4). If the covariance matrices of either ϵ_i or ν_g are not constant but related in an unknown way to n_g then this approach will have problems.

Traditional sample survey data will often have no group indicators for reasons of confidentiality protection. Alternatively an unclustered sample design may have been used so that there is only one unit selected in any group. In this circumstance it is not possible to separate the variance components. This is not a problem if only marginal targets of inference such as Σ are of interest. However, when aggregate data are also available it can be combined with sample survey data to provide estimates of $\Sigma^{(1)}$ and $\Sigma^{(2)}$. A moments approach is based on $S_{yy}^{(1)} = \dfrac{1}{n-1}\sum\limits_{i\in s}(y_i - \bar{y})(y_i - \bar{y})'$ the covariance matrix caculated from the unit level sample data and $S_{yy}^{(2)} = \dfrac{1}{m-1}\sum\limits_{g\in s} n_g(\bar{y}_g - \bar{y})(\bar{y}_g - \bar{y})'$, the covariance matrix calculated from the group level means, using the group sample sizes as weights. The key relationships are

$$E[S_{yy}^{(2)}] = \Sigma^{(1)} + \bar{n}^*\Sigma^{(2)} \qquad \text{and} \tag{2.5}$$

$$E[S_{yy}^{(1)}] = \Sigma^{(1)} + (1 - \frac{\bar{n}^0 - 1}{n-1})\Sigma^{(2)} \tag{2.6}$$

Here $\bar{n} = n/m$ is the average number of sampled individuals per group in the sample and

$$\begin{aligned} \bar{n}^* &= \bar{n}(1 - \frac{C_n^2}{m-1}), \\ \bar{n}^0 &= \bar{n}(1 + C_n^2) \qquad \text{and} \\ C_n^2 &= \frac{1}{m}\sum_{g\in s}(n_g - \bar{n})^2/\bar{n}^2 \end{aligned}$$

is the square of the coefficient of variation of the area sample sizes. From these results unbiased estimates of $\Sigma^{(1)}$ and $\Sigma^{(2)}$ and linear functions of them such as Σ can be obtained. This approach can be used when the individual level and group level data come from different samples drawn from the same population (Tranmer and Steel, 1998).

Results (2.5) and (2.6) have some important implications. If $\Sigma^{(2)} = 0$ then the group and individual level covariance matrices have the same expectation. However, even if the elements of $\Sigma^{(2)}$ are much smaller than those of $\Sigma^{(1)}$, but not zero, then when $\bar{n}^*$ is large there will be a large difference in the expectations. For example suppose a diagonal element of $\Sigma^{(2)}$ is 0.01 time the corresponding element of $\Sigma^{(1)}$ so that group effects account for very little of the variation between individual. If $\bar{n}^* = 500$, then group level variance will have expectation approximately 6 times larger the corresponding unit level variance. The group level effects will account for $5/6ths$ of the observed group level variance. Census EDs have approximately 500 people in them , and much larger geographical groups are used in ecological analysis. In these case $S_{yy}^{(2)}$ contains very little contribution from $\Sigma^{(1)}$, but is mainly determined by $\Sigma^{(2)}$. In practice using $\frac{1}{\bar{n}^*}S_{yy}^{(2)}$ to produce estimates of genuine group level relationships will not be badly biases if $\bar{n}^*$ is large. As an estimate of Σ the bias of $S_{yy}^{(1)}$ is $O(\frac{1}{m})$ and it will be a reasonable estimate of the marginal individual level relationships provided m is not small.

Aggregate and unit level data without group indicatives can also be combined in MLn by treating each unit observation as coming from a group with sample size of one.

We will focus on the cases when there are only aggregate group level data available and when there is also a limited amount of individual level data on a few variables without any group indicators.

3. Ecological analysis using auxiliary variables

If individuals are allocated to groups at random , all of the effects of aggregation in linear statistics can be explained by the number of groups and the distribution of group sample sizes. In this circumstances there is no ecological fallacy for linear statistics, and parameters such as means, variances, regression and correlation coefficients can be unbiasedly estimated from group level data. Variances of statistics are mainly determined by the number of areas in the analysis. Confidence intervals based on aggregated data can be calculated using well known statistical methods and will be wider than those based on unit level data. Moreover, a regression or correlation analysis of m groups has the same statistical properties as an analysis based on m randomly selected individuals. Steel and Holt (1995) give rules for random aggregation.

Individuals are not allocated to groups at random. In practice, individuals who live in the same area exhibit positive intra-group correlation for a wide variety of socio-economic characteristics. There are strong socio-economic differences between groups which are confounded with the individual effects in any statistical analysis performed using aggregated data. In general we need to recognise that we are interested in a population in which the ith individual has a vector of response variables y_i , a vector of explanatory variables x_i and group indicator c_i and develop statistical models for these variables which initally at least includes the group indicators.

The variation between groups and therefore the homogeniety within groups is a key factor in the ecological fallacy. Our approach is to introduce extra variables to account for the homogeneity within areas. Suppose that there is a set of auxiliaary variables, z , that characterize the way in which individuals are clustered within the groups and, conditional on z , the observations for individuals in area g are influenced by random group level effects. The auxiliary variables in z will be called grouping variables and may only have a small effect on the individual level relationships and may not be of any direct interest. However, because of their strong within area homogeneity they may affect the ecological analysis greatly. The matrices $z = [z_1, ..., z_N]', c = [c_1, ..., c_N]'$ give the values of all units in the population. Both the explanatory and auxiliary variables may contain group level variables, although there will be identification problems if the mean of an individual level explanatory variables is used as a group level variable.

For continuous variables, the conditional probability density function (pdf) of y given x can be expressed as

$$f(y|x) = \frac{\int f(y|x,z)f(x|z)f(z)dz}{\int f(x|z)f(z)dz} \tag{3.1}$$

For models in which the parameters of the pdfs $f(y|x,z), f(x|z),$ and $f(z)$ are distinct, analysis can proceed by using individual level data to estimate the paramaters of $f(z)$ and aggregate data are used to estimate the parameters of $f(y|x,z)$ and $f(x|z)$. The models used may include random effects, although estimation will be simpler if auxiliary variables can be found which effectively remove within group homogeniety and hence the need for random effects.

4. Ecological linear regression

4.1. A multi-level for ecological regression with auxiliary variables

Steel and Holt (1996b) considered a multi-level model with auxiliary variables and considered its implications for the ecological analysis of covariance matrices and correlation coefficients obtained from them. They also developed a method for adjusting the analysis of aggregate data to provide less biased estimates of covariance matrices and correlation coefficients. Steel, Holt and Tranmer (1996) evaluated this method and were able to reduce the biases by about 70 percent by using limited amounts of individual level data for a small set of variables that help characterize the differences betweeen groups. We consider the implications of this model for ecological linear regression analysis.

The group level covariance matrices $S^{(2)}_{yy}$, $S^{(2)}_{xx}$ and $S^{(2)}_{xy}$ can be calculated using the group sample sizes as weights. These covariance matrices may be combined in $S^{(2)}_{ww}$, the covariance matrix for all of the variables, where $w = (x', y')'$. The ecological regression coefficients relating y to x are estimated by $B^{(2)}_{yx} = S^{(2)-1}_{xx} S^{(2)}_{xy}$.

Various authors have pointed out arithmetic relationships between ecological regression coefficients and the individual level regression coefficients within each area and the marginal regression coefficients. The marginal individual level covariance matrices $S_{yy}^{(1)}$, $S_{xx}^{(1)}$ and $S_{xy}^{(1)}$ may be combined in $S_{ww}^{(1)}$. For each group define $S_{wwg}^{(1)} = \frac{1}{n_g - 1} \sum_{i \in s,g} (w_i - \bar{w}_g)(w_i - \bar{w}_g)'$ which has components $S_{yyg}^{(1)}$, $S_{xxg}^{(1)}$ and $S_{xyg}^{(1)}$. The marginal individual level regression coefficients would be estimated by $B_{yx}^{(1)} = S_{xx}^{(1)-1} S_{xy}^{(1)}$ and the sample regression coefficients in the gth group are $B_{yxg}^{(1)} = S_{xxg}^{(1)}{}^{-1} S_{xyg}^{(1)}$. Define $B_{yxW}^{(1)} = S_{xxW}^{(1)^{-1}} S_{xyW}^{(1)}$ as the average within groups regression coefficents, then

$$B_{yx}^{(2)} = \frac{n-1}{m-1} S_{xx}^{(2)-1} S_{xx}^{(1)} B_{yx}^{(1)} - \frac{n-m}{m-1} S_{xx}^{(2)-1} S_{xxW}^{(1)} B_{yxW}^{(1)}$$

Hence $B_{yx}^{(2)} = B_{yx}^{(1)}$ if, and only if, $B_{yxW}^{(1)} = B_{yx}^{(1)}$. In general there is no reason to expect $B_{yx}^{(2)}$ to be close to $B_{yx}^{(1)}$. The difference will be partly determined by n/m and how the explanatory variables are clustered within the groups.

The model given in (2.1) to (2.3) is expanded to include x and z by assuming the following model conditional on z and the groups used:

$$w_i = \mu_{w|z} + \beta'_{wz} z_i + \nu_g + \epsilon_i, \quad i \in g \tag{4.1}$$

$$\text{where } V(\nu_g \mid z, c) = \Sigma_{ww|z}^{(2)} \text{ and } V(\epsilon_i \mid z, c) = \Sigma_{ww|z}^{(1)}. \tag{3.2}$$

This model implies

$$E(w_i \mid z, c) = \mu_{w|z} + \beta'_{wz} z_i, \tag{4.3}$$

$$V(w_i \mid z, c) = \Sigma_{ww|z}^{(1)} + \Sigma_{ww|z}^{(2)} = \Sigma_{ww|z} \text{ and} \tag{4.4}$$

$$Cov(w_i, w_j \mid z, c) = \Sigma_{ww|z}^{(2)} \text{ if } \ c_i = c_j, \ \ i \neq j$$

The matrix $\Sigma_{ww|z}^{(2)}$ has components $\Sigma_{xx|z}^{(2)}, \Sigma_{xy|z}^{(2)}$ and $\Sigma_{yy|z}^{(2)}$ and $\beta'_{wz} = (\beta_{xz}, \beta_{yz})'$. Assuming $V(z_i) = \Sigma_{zz}$ then the marginal covariance matrix is

$$\Sigma_{ww} = \Sigma_{ww|z} + \beta'_{wz} \Sigma_{zz} \beta_{wz} \tag{4.5}$$

which has components Σ_{xx}, Σ_{xy} and Σ_{yy} . The target of inference is $\beta_{yx} = \Sigma_{xx}^{-1} \Sigma_{xy}$.

Under this model, Steel and Holt (1996b) showed

$$E[\bar{w} \mid z, c] = \mu_w + \beta'_{wz}(\bar{z} - \mu_z) \quad \text{and}$$

$$E[S^{(2)}_{ww} \mid z, c] = \Sigma_{ww} + \beta'_{wz}(S^{(2)}_{zz} - \Sigma_{zz})\beta_{wz} + (\bar{n}^* - 1)\Sigma^{(2)}_{ww|z}$$
$$= \Sigma_{ww|z} + \beta'_{wz}S^{(2)}_{zz}\beta_{wz} + (\bar{n}^* - 1)\Sigma^{(2)}_{ww|z}$$

Providing that the variance of $S^{(2)}_{ww}$ is $O(m^{-1})$ the expectation of the ecological regression coefficients can be obtained by replacing $S^{(2)}_{yy}$ and $S^{(2)}_{xy}$ by their expectations, to give, to $O(m^{-1})$,

$$E[B^{(2)}_{yx} \mid z, c] = [\Sigma_{xx} + \beta'_{xz}(S^{(2)}_{zz} - \Sigma_{zz})\beta_{xz} + (\bar{n}^* - 1)\Sigma^{(2)}_{xx|z}]^{-1}$$
$$\times [\Sigma_{xy} + \beta'_{xz}(S^{(2)}_{zz} - \Sigma_{zz})\beta_{yz} + (\bar{n}^* - 1)\Sigma^{(2)}_{xy|z}] \quad (4.6)$$

Set $A = E[S^{(2)}_{xx} \mid z, c] = [\Sigma_{xx} + \beta'_{xz}(S^{(2)}_{zz} - \Sigma_{zz})\beta_{xz} + (\bar{n}^* - 1)\Sigma^{(2)}_{xx|z}]$. The model implies $\beta_{yx|z} = \Sigma^{-1}_{xx|z}\Sigma_{xy|z}$, $\beta_{zx} = \Sigma^{-1}_{xx}\Sigma_{xz}$ and if we define $\beta_{yz|x} = \beta_{yz} - \beta_{xz}\beta_{yx|z}$, then $\beta_{yx} = \beta_{yx|z} + \beta_{zx}\beta_{yz|x}$. The resulting bias can be shown to be.

$$[A^{-1}\beta'_{xz}(S^{(2)}_{zz} - \Sigma_{zz})(\beta_{yz} - \beta_{xz}\beta_{yx})] + (\bar{n}^* - 1)A^{-1}[\Sigma^{(2)}_{xy|z} - \Sigma^{(2)}_{xx|z}\beta_{yx}] \quad (4.7)$$

$$= [A^{-1}\beta'_{xz}S^{(2)}_{zz} - \beta_{zx}]\beta_{yz|x} + (\bar{n}^* - 1)A^{-1}[\Sigma^{(2)}_{xy|z} - \Sigma^{(2)}_{xx|z}\beta_{yx|z}] \quad (4.8)$$

The first term in the bias in (2) will disappear if either $\beta_{xz} = 0$ or $\beta_{yz|x} = 0$, that is if the explanatory and grouping variables are unrelated or if the response variables have no relationship with the grouping variables once the explanatory variables included in the model are taken into account. Since $E[B^{(2)}_{zx} \mid z, c] = A^{-1}\beta'_{xz}S^{(2)}_{zz}$ the first term in (4.8) is due to the bias of $B^{(2)}_{zx}$ in estimating β_{zx}. The second term in the bias in (3.8) will disappear if $\Sigma^{(2)}_{xy|z} = \Sigma^{(2)}_{xx|z}\beta_{yx|z}$, that is if conditional on the grouping variables, the covariance between the values of the response and explanatory variables for different individuals in the same group is solely due to the covariance of the explanatory variables within the same group and the relationship between y and x for the same individual. This condition is equivalent to the population regression coefficients relating y to x, conditional on the grouping variables, being the same at the individual and group level i.e. $\beta_{yx|z} = \Sigma^{(2)}_{xx|z}{}^{-1}\Sigma^{(2)}_{xy|z} = \beta^{(2)}_{yx|z}$. This term will also disappear if $\Sigma^{(2)}_{xy|z} = 0$ and $\Sigma^{(2)}_{xx|z} = 0$, when there are no random effects conditional on z. The second term in the bias involves $\bar{n}^*$ which can be very large, for example when the group means are based on a census of all indivduals in the groups.

In the sociological literature the effect of aggregation has often been condsidered by considering some aggregation criterion. In this model this idea can be represented by all the grouping effect operating through the auxiliary variables and there being no group level effects conditional on them. In this case $A = \Sigma_{xx} + \beta'_{xz}(S^{(2)}_{zz} - \Sigma_{zz})\beta_{xz}$ and the bias is

$$[A^{-1}\beta'_{xz}S^{(2)}_{zz} - \beta_{zx}]\beta_{yz|x} = [E[B^{(2)}_{zx} \mid z, c] - \beta_{zx}]\beta_{yz|x}$$

and is entirely due to the effect of aggregation on the implied estimate of β_{zx}.

We can conceptually shift the causes of the aggregation effects between the random effects and auxiliary variables. As more auxiliary variables are included the components of $\Sigma^{(2)}_{ww|z}$ change. How much of the aggregation effects are due to auxiliary variables and how much to the random effects is essentially a thoeretical issue, except when we have individual level information on the auxiliary variables, as shown in section 4.2.

The model specified in (4.1) and (4.2) relates to the joint distribution of y and x and gives a general and flexible framework for considering aggregation effects. The implications of using a joint modelling approach can be illustrated by the case where there is one explanatory variable, no grouping variables and Normal distributions for all random terms. Let $\delta_{xx} = \Sigma^{-1}_{xx}\Sigma^{(2)}_{xx}$ denote the population intra-group correlation for x which reflects the within group homogeniety of x, then $\beta_{yx} = (1-\delta_{xx})\beta^{(1)}_{yx} + \delta_{xx}\beta^{(2)}_{yx}$ and

$$\begin{aligned} E[y_i|x_i] &= \mu_y + [(1-\delta_{xx})\beta^{(1)}_{yx} + \delta_{xx}\beta^{(2)}_{yx}](x_i - \mu_x) \\ &= \mu_y + \beta_{yx}(x_i - \mu_x) \end{aligned} \tag{4.9}$$

which is the standard unit level regression model. However, when we condition on the values of the explanatory variables in the group we get the other units in the group influencing the ith individual. For the gth group which had n_g units in the data set, let $a^g = \dfrac{n_g\delta_{xx}}{1+(n_g-1)\delta_{xx}}$, and $\underset{\sim}{x}_g$ be the vector of values of the explanatory variable in the group, then

$$E[y_i|\underset{\sim}{x}_g] = \mu_y + \beta^{(1)}_{yx}(x_i - \mu_x) + a^g[\beta^{(2)}_{yx} - \beta^{(1)}_{yx}](\bar{x}_g - \mu_x) \tag{4.10}$$

Hence

$$E[\bar{y}_g|\underset{\sim}{x}_g] = \mu_y + [(1-a^g)\beta^{(1)}_{yx} + a^g\beta^{(2)}_{yx}](\bar{x}_g - \mu_x) \tag{4.3}$$

The result (4.10) shows that the joint model formulation has implications for standard multilevel modelling using individual level data. It suggests that the group level means of the explanatory variables should be included on the analysis. This is sometimes done as a way of reflecting the effect of substantively interesting compositional effects (see Goldstein, 1995 p31). The results here suggest that it should be done as a matter of course to allow for the possibility of group level random effects affecting both the dependent and explanatory variables.

These results show that the second term in the bias of (4.8) arises because the unexplained random effects produce an effect related to the mean of the explanatory variables and when aggregation occurs this becomes confounded with the individual level effect. Goldstein (1995, p65) notes that in the presence of a contextual or compositional effect aggregate analysis cannot separate the individual and contextual effect.

The model here allows for group effects in two ways and is more general than those proposed elsewhere to explain the effect of aggregation. The form of the bias in (4.7) and (4.8) suggests that it will not be possible to reach any general conclusions about the size or likely direction of the biases. However, the general formulas for the bias can be applied to some special cases.

In this case of one explanatory variable and no grouping variables

$$E[B_{yx}^{(2)}|c] = (1-a^*)\beta_{yx}^{(1)} + a^*\beta_{yx}^{(2)}$$

where $a^* = \dfrac{\bar{n}^*\delta_{xx}}{1+(\bar{n}^*-1)\delta_{xx}}$ Comparing this result with (4.9) we see that the effect of aggregation is to shift the weight given to the population regression parameters towards the group level. Even a small value of δ_{xx} can lead to a considerable shift if $\bar{n}^*$ is large. For example suppose $\delta_{xx} = 0.05$, then the contribution of $\beta_{yx}^{(2)}$ to β_{yx} is 5 percent, but if $\bar{n}^* = 101$ the contribution to $E[B_{yx}^{(2)}|c]$ is 83 percent. Holt, Steel and Tranmer (1996) conducted an analysis of a range of variables from the 1986 Australian population census. Two levels of groups were considered with values of $\bar{n}^*$ of 580 and 9600 respectively. They found that for most variables the value of δ_{xx} was less than 0.05 but a^* generally exceeded 0.8 and 0.95 for the two levels of groups respectively.

For the case of several explanatory variable and one grouping variable

$$E[B_{zx}^{(2)}|z,c] = [I+(\bar{n}^*-1)(I-\frac{(Q_z-1)\Re_{x|z}^2}{1+\Re_{z|x}^2(Q_z-1)})\Sigma_{xx}^{-1}\Sigma_{xx|z}^{(2)}]^{-1}\beta_{zx}\frac{Q_z}{1+\Re_{z|x}^2(Q_z-1)}$$

where $\Re_{z|x}^2 = \Sigma_{zz}^{-1}\Sigma_{zx}\Sigma_{xx}^{-1}\Sigma_{xz}$ and $\Re_{x|z}^2 = \Sigma_{xx}^{-1}\Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{zx}$ are the population multiple correlation coefficent between z and x , and x and z respectively and $Q_z = S_z^{(2)}/S_z^{(1)}$. If $\Sigma_{xx|z}^{(2)} = 0$ then

$$E[B_{zx}^{(2)}|z,c] = \beta_{zx}\frac{Q_z}{1+\Re_{z|x}^2(Q_z-1)}$$

and the factor will exceed 1 provided Q_z exceeds 1. There is an amplification effect on the contribution of $\beta_{zx}\beta_{yz|x}$. This has been noted before (e.g Smith, 1977) but it relies on the grouping essentially being one dimensional.

4.2 An adjusted ecological regression method

The discussion above has identified the causes of the ecological fallacy as the grouping effects associated with the auxiliary variables and the remaining group level variance components. We now consider methods to produce esimates of β_{yx} from aggregate data. One approach is based on the variance structure for the group means implied by the model when there are no auxiliary variables

$$V(\bar{w}_g) = \Sigma_{ww}^{(1)}/n_g + \Sigma_{ww}^{(2)}$$

. For example the IGLS procedure emobodied in MLn can be used (see Goldstein, 1995). This approach relies on their being reasonable variation in the sample sizes between the groups and on the variance structure originally assumed at the inidividual level leading to variances which have a component that is constant and one which is proportional to $1/n_g$. At each step of the iterative process the method regresses $(\bar{w}_g - \hat{\mu}_w)(\bar{w}_g - \hat{\mu}_w)'$ against $\frac{1}{n_g}$ where $\hat{\mu}_w$ is the current estimates of μ_w and the estimates of $\Sigma_{ww}^{(1)}$ and $\Sigma_{ww}^{(2)}$ are the resulting regression coefficients.

Another approach is to assume that a set of z variables can be identified that explain much of the aggregation effect on the variables of interest. If individual level data on these variables are available, the aggregation bias due to these z variables may be estimated. Under (4.1) $E[B_{wz}^{(2)} \mid z, c] = \beta_{wz}$ where $B_{wz}^{(2)} = S_{zz}^{(2)-1} S_{zw}^{(2)}$. If an estimate of the individual level population covariance matrix for z were available, possibly from another source, Steel and Holt (1996) proposed the following adjusted estimator of Σ_{ww},

$$\begin{aligned}\widehat{\Sigma}_{ww}(z) &= S_{ww}^{(2)} + B_{wz}^{(2)\prime}(\widehat{\Sigma}_{zz} - S_{zz}^{(2)})B_{wz}^{(2)} \\ &= S_{ww|z}^{(2)} + B_{wz}^{(2)\prime}\widehat{\Sigma}_{zz}B_{wz}^{(2)}\end{aligned}$$

where $\widehat{\Sigma}_{zz}$ is the estimate of Σ_{zz} calculated from individual level data. This estimator corresponds to a Pearson type adjustment (Smith, 1989) and for Normally distributed data is the MLE. This estimator removes the aggregation bias due to z. Adjusted regression coefficients can then be calculated from $\widehat{\Sigma}_{ww}(z)$, i.e.

$$\widehat{\beta}_{yx}(z) = \widehat{\Sigma}_{xx}^{-1}(z)\widehat{\Sigma}_{xy}(z)$$

The adjusted estimator replaces the components of bias in (4.7) due to $\beta'_{xz}(S^{(2)}_{zz} - \Sigma_{zz})\beta_{xz}$ and $\beta'_{xz}(S^{(2)}_{zz} - \Sigma_{zz})\beta_{yz}$ by $\beta'_{xz}(\widehat{\Sigma}_{zz} - \Sigma_{zz})\beta_{xz}$ and $\beta'_{xz}(\widehat{\Sigma}_{zz} - \Sigma_{zz})\beta_{yz}$ respectively. Set $\widehat{A}(z) = E[\hat{\Sigma}_{xx}(z)|z, c] = [\Sigma_{xx} + \beta'_{xz}(\widehat{\Sigma}_{zz} - \Sigma_{zz})\beta_{xz} + (\bar{n}^* - 1)\Sigma^{(2)}_{xx|z}]$, then the bias of $\widehat{\beta}_{yx}(z)$ is, to $O(\frac{1}{m})$

$$[\widehat{A}(z)^{-1}\beta'_{xz}\widehat{\Sigma}_{zz} - \beta_{zx}]\beta_{yz|x} + (\bar{n}^* - 1)\widehat{A}(z)^{-1}[\Sigma^{(2)}_{xy|z} - \Sigma^{(2)}_{xx|z}\beta_{yx|z}] \qquad (4.12)$$

Suppose that $\widehat{\Sigma}_{zz}$ is an estimate based on a individual level sample invovling m_0 first stage units then for many sample designs $\widehat{\Sigma}_{zz} = \Sigma_{zz} + O(\frac{1}{m_0})$, and so to $O(\frac{1}{m_0})$ the bias of $\widehat{\beta}_{yx}(z)$ is

$$\widehat{A}(z)^{-1}(\bar{n}^* - 1)\Sigma^{(2)}_{xx|z}[\beta^{(2)}_{yx|z} - \beta_{yx}] \qquad (4.13)$$

$$= [\widehat{A}(z)^{-1}\Sigma_{xx} - I]\beta_{zx}\beta_{yz|x} + \widehat{A}(z)^{-1}(\bar{n}^* - 1)\Sigma^{(2)}_{xx|z}[\beta^{(2)}_{yx|z} - \beta_{yx|z}] \qquad (4.14)$$

It is not necessary for the individual level data to contain group identifiers, only that it permitted estimation of Σ_{zz} . If $\Sigma_{xx|z} = 0$ then the bias of $\widehat{\beta}_{yx}(z)$ is $O(\frac{1}{m_0})$

The adjusted estimator can be rewritten as

$$\widehat{\beta}_{yx}(z) = B^{(2)}_{yx|z} + \widehat{\beta}_{zx}(z)B^{(2)}_{yz|x} \qquad (6.4)$$

where $\widehat{\beta}_{zx}(z) = \widehat{\Sigma}^{-1}_{xx}(z)B^{(2)\prime}_{xz}\widehat{\Sigma}_{zz}$. Corresponding decompositions apply at the group and individual levels:

$$B^{(2)}_{yx}(z) = B^{(2)}_{yx|z} + B^{(2)}_{zx}B^{(2)}_{yz|x}$$

$$B^{(1)}_{yx}(z) = B^{(1)}_{yx|z} + B^{(1)}_{zx}B^{(1)}_{yz|x}$$

The adjustment is trying to correct for the bias in the estimation of β_{zx} by replacing $B^{(2)}_{zx}$ by $\widehat{\beta}_{zx}(z)$.

The bias due to the conditional variance components $\Sigma^{(2)}_{ww|z}$ remains. The two approaches can be combined. Multilevel modelling with aggregate data can be used to produce estimates of $\Sigma^{(2)}_{ww|z}$, $\Sigma^{(1)}_{ww|z}$ and maximum likelihood estimates of β_{yz} and β_{xz} . These can be combined to produce an estimate of β_{yx} which accounts for the conditional variance components. That is we calculate

$$\widehat{\Sigma}_{ww} = \widehat{\Sigma}^{(1)}_{ww|z} + \widehat{\Sigma}^{(2)}_{ww|z} + \widehat{\beta}_{wz}\widehat{\Sigma}_{zz}\widehat{\beta}_{wz}$$

and then use the relevant components of $\widehat{\Sigma}_{ww}$ i.e $\widehat{B}_{yx} = \widehat{\Sigma}^{-1}_{xx}\widehat{\Sigma}_{xy}$. However, this approach still relies on the use of purely aggregate data to estimate variance components.

4.3 Evaluation of aggregation effects and adjusted ecological regression methods

The data

An empirical investigation into the effects of aggregation on multiple regression analysis was carried out using data from the Australian 1991 Population Census for the city of Adelaide. Group level data were available in the form of totals for the 1711 census Collection Districts (CDs), which contain an average of about 250 dwellings. Our analysis was confined to people aged 15 or more and there was an average of about 450 such people per CD. To enable an evaluation to be carried out we also used data from the households sample file (HSF) which is a one percent sample of households, and the people within them, released from the population census.

The evaluation concentrated on the dependent variable of personal income. This variable is collected in 14 ranges but was treated for this evaluation as a continuous variable by giving each person the mid point of the range. The following variables were considered as possible explanatory variables: marital status, sex, degree, employed-manual occupation, employed-managerial or professional occupation, employed-other, unemployed, born Australia, born UK and four age categories. The adjustment variables considered were: age 45 to 59, age 60+, owner occupied, renting from government, housing type.

Aggregation effects on variances and bivariate statistics

The aggregation effect on the variance of each variable, which is the ratio of the group level to unit level variance, i.e $Q_{aa} = S_{aa}^{(2)}/S_{aa}^{(1)}$ are given in Table 1, along with the associated estimate of the intra-CD correlation $\hat{\delta}_{aa}$. All the variables have experienced some aggregation effect, ranging from 2.65 for sex to 171.1 for renting from government. A small amount of within group correlation can lead to very large aggregation effects on statistics because of the large number of individuals within the areas. The variables considered as potential auxiliary variables generally have the larger aggregation effects. This is one reason for selecting these particular variables. It is usually possible to calculate Q_{aa} for a range of variables since a reasonable idea of the individual level variance can often be obtained from other published data. For a dichotomous variable all that is required is an estimate of the population proportion.

Tables 2a, 2b and 2c summarize the effect of the aggregation on the analysis of bivariate covariances, correlations and regression coefficients between income and each of the explanatory and auxiliary variables. The ED level correlations are generally of the same sign but larger than the corresponding individual level correlation. In most cases the change is sufficient to affect the substantive interpretation. The affect of aggregation on the regression coefficents are similar. Examination of plots of $r_{ab}^{(2)}$ against $r_{ab}^{(1)}$ for all the variables in this analysis showed the usual amplification of group level correlations and resulting S shape noted by Steel, Holt and Tranmer (1996), although there a number of cases in which the correlations have different signs at the two levels.

Mutivariate aggregation effects and CGV analysis

The aggregation effect on each of the variables or each pair of variables does not completely characterize the grouping and aggregation in a multivariate situation. Steel and Holt (1996b) introduced the idea of Canonical Grouping Variables (CGVs) as a way of identifying the important variables associated with the grouping of a population. Suppose unit and CD level covariance matrices $S^{(1)}$ and $S^{(2)}$ have been calculated for a set of variables. The CGVs for CDs are obtained from the eigenvectors $d_1^{(2)}, ..., d_p^{(2)}$ of $S^{(1)-1}S^{(2)}$ with associated eigenvalues $\theta_1^{(2)}, ..., \theta_p^{(2)}$. Define the matrix $D^{(2)} = [d_1^{(2)}, ..., d_p^{(2)}]$ then the CGVs are defined by $U = D'Y$ and have some properties that are useful in interpreting the grouping structure of CDs. The CGVs have covariance matrix $diag(\theta_l^{(2)})$ at the CD level and I_p at the individual level. Subject to the constraints of being mutually uncorrelated at the indvidual and CD level the CGVs have successively the maximum aggregation effect and therefore maximum intra-CD correlation.

The matrix $S^{(1)-1}S^{(2)}$ is an extension of the univariate aggregation effect Q_{aa} and the eigenvalues give the aggregation effect of each of the mutually orthogonal grouping dimensions in the set of variables being considered. Useful summary measures of the aggregation effects in multivariate data are given by $\bar{\theta} = \frac{1}{p}\sum_l \theta_l^{(2)}$ = $\frac{1}{p} trace[S^{(1)-1}S^{(2)}]$ and $\bar{Q} = \frac{1}{p}\sum_a Q_{aa}$. The quantity $\sum_1^q \theta_l^{(2)} - 1$ is the amount of aggregation effect that can be associated with the first q CGVs and is an upper limit to the aggregation effect that any q adjustment variables can remove.

Considering all the variables together, i.e. (y, x, z), gave $\bar{\theta} = 27.7$ and $\bar{Q} = 33.7$. The first four CGVs accounted for 83 percent of the total aggregation effect. Examing the coefficients for the first four CGVs, shows that the first corresponds to renting from government, the second is owner occupied and housing type, the third is aged 60+ and the fourth is a combination of income, degree and managerial or professional occupation.

The results of a CGV analysis of (y, x) gave $\bar{\theta} = 13.6$ and $\bar{Q} = 14.5$. The first five CGVs accounted for 82 percent of the total aggregation effect. Examining the coefficients for the first five CGVs, shows that the first corresponds mainly to aged 60+ contrasted with married, the second is a combination of income, degree and managerial or professional occupation, the third is aged 60+ and born UK, the fourth is married contrasted with born UK and the fifth is aged 45-59.

Aggregation effects on multiple regression

Standard multiple regression models were estimated using the HSF data and the CD data, weighted by CD population size. The results are summarized in Table 3. The R^2 of the CD level equation, 0.880, is much larger than that of the individual level equation, 0.496. However, the CD level R^2 is indicating how much of the variation in CD mean income is being explained. Generally the regression coefficients estimated at the two levels are of the same sign the exceptions being married, which is non-significant at the individual level, and the coefficient for aged 20-29. The values can be very different at the two levels, with the CD level coefficients being larger than the corresponding individual level coefficients in some cases and smaller in others. The differences are often considerable, for example the coefficient for degree increases from 8471 to 21700. The average absolute difference was 4533.

The difference between the two estimated models can also be examined by comparing their fit at the individual level. The fitted value based on the individual level model is $\hat{y}_i^{(1)} = B_{yx}^{(1)\prime} x_i$ and that based on the CD level model is $\hat{y}_i^{(2)} = B_{yx}^{(2)\prime} x_i$. The usual estimate of the residual variance is $\frac{1}{n-p} \sum_{i \in s} (y_i - \hat{y}_i^{(1)})^2$, which was 10351^2 and this can be compared with $\frac{1}{n-p} \sum_{i \in s} (y_i - \hat{y}_i^{(2)})^2$, which was 12113^2. Using the CD level equation to predict individual level income gave an R^2 of 0.310 compared with 0.496 for the individual level regression equation.

Other variables could be added to the model but the R^2 obtained was considered acceptable and this sort of model is indicative of what researchers might use in practice. The R^2 obtained at the individual level is consistent with those found in other studies of income (e.g Davies, et al, 1997). As with all regression models there are likely to be variables with some explanatory power omitted from the model, however, this reflects the world of practical data analysis. We were concerned with looking at the effect of aggregation and the effectiveness of methods for adjusting for aggregation effects when a reasonable but not necessarily perfect statistical model is being used. The log transformation was also tried for the income variable but did not result in an appreciably better fit.

The estimates and associated estimated standard errors obtained at the two levels are different and hence the assessment of their statistical significance. Using a ten percent significance level the coefficients for married, aged45-59 and aged60+ were non-significant in the individual level equation. In the CD level equation the coefficients for unemployed, manual occupation, aged15-19 and aged 45-50 were non-significant. The estimated standard errors of coefficients at the CD level were between 1.19 and 3.65 times larger those estimated at the individual level. The changes in the estimated residual mean squared error and the degrees of freedom imply an increase of 3.23. For all the coefficients except female the increase is less than 3.23, which is due to the effect of aggregation on S_{xx}.

Adjusted CD level regression

The CGV analysis suggests which variables have strong grouping effects. In considering potential adjustment variables we also need to consider those variables for which it is reasonable to expect individual level data might be available. Because the adjustment relies on obtaining a good estimate of the unit level covariance matrix of the adjustment variables we need to keep the number of variables small. By choosing variables that characterize much of the difference between CDs we hope to have variables that will perform effectively in a range of situations. Based on these considerations the evaluation concentrated on the following auxiliary variables:owner occupied, renting from government, housing type, aged45-59 an aged 60+.

To assess how well these variables perform in removing aggregation effects $\widehat{\Sigma}_{ww}(z)$ was calculated. The resulting adjusted aggregation effects $\widehat{Q}_{aa}(z) = \widehat{\Sigma}_{aa}(z)/S_{aa}^{(1)}$ are given in column five of Table 1. The ratio $\widehat{Q}_{aa}(z)/Q_{aa}$ is given in the last column of Table 1 and indicates that these adjustment variables remove between 9 and 75 percent of the aggregation effect. For the income variable the reduction is 32 percent and the average reduction is 52 percent. These values tell us the effect of the adjustment for each variable sepraratelly. A CGV analyses based on $S_{ww}^{(1)-1}\widehat{\Sigma}_{ww}(z)$ gives an overall assessment of the amount of the aggregation effect of the dependent and explanatory variables that is removed by these adjustment variables. Because they are also used as adjustment variables the explanatory variables aged45-59 and aged 60+ were not included in this CGV analysis. The reduction in $\bar{\theta}$ was 51 percent. Examination of the coefficients resulting from the CGV analysis showed that the first CGV remaining after adjustment was mainly associated with income, degree and a managerial or professional occupation. The second CGV was mainly associated with being born in the UK. The first two CGVs accounted for most of the remainin aggregation effects. An analysis of the CGVs based on $S_{xx}^{(1)-1}\widehat{\Sigma}_{xx}(z)$ gave similar results, with income disappearing from the first CGV. These results suggest that the adjustment variables considered account for about half of the aggregation. Comparing the results of the CGV analysis of (y,x) before and after adjustment suggests that the auxiliary variables used have accounted for the first grouping dimension but not the second. Much of the remaining aggregation effects are associated with income and indicators of relatively high socio-economic status such as having a degree or managerial or professional occupation. For these variables the reduction in the aggregation effects of 33 and 23 percent respectively. This is consistent with the first CGV in (y,x) accounting for 55 percent of the aggregation effects.

Tables 2a, 2b and 2c show the adjusted covariances and associated regression and correlation coefficients calculated from $\widehat{\Sigma}_{ww}(z)$. There is a reduction in aggregation effects of covariances, although many remain large. For some variables the improvement is marginal. In some cases there is an over-adjustment, shown by aggregation effects less than one. There is over-adjustment for the correlation of income with the auxiliary variables and no improvement for the correlation involving the explanatory variables. Similar results apply for the regressions coefficients. The relationships between the explanantory and auxiliary variables is a factor in the ecological bias. Graphical analysis of the relationships between all the variables in the analysis showed that for the covariances there is dramatic improvements for those involving the auxiliary variables and evidence of improvements for the covariances between the explanatory variables, but with considerable differences remaining. A similar picture emerges for correlations, with those invovling the auxiliary variables being improved considerably, but those between income and the explanatory variables and between the explanatory variables experiencing some small improvement but not enough to give reasonable estimates of the individual level coefficients. A similar picture is evident for simple linear regression coefficients.

The estimates of the regression equation obtained from $\widehat{\Sigma}_{ww}(z)$ are given in Table 4. In general the adjusted CD regression coefficients are no closer than those for the original CD level regression equation. The resulting adjustment of R^2 which is still considerably higher than than those in the individual level equation indicating that the adjustment is not working well. The measure of fit at the individual level gives an R^2 of 0.284 compared with 0.310 for the unadjusted equation, so according to this measure the adjustment has had a small detrimental effect. The average absolute difference between the CD and individual level coefficients has also increased slightly to 4771.

While the adjustment has eliminated about half the aggregation effects in the dependent and explanatory variables it has not resulted in reducing the difference between the CD level and individual level regression equations. To clarify the source of these biases we decomposed the components of the bias of the unadjusted ED regression coefficients into elements that can be ascribed to the adjustment variables and those that cannot. Equation (6.4) shows that the adjustment procedure will be effective if $B^{(2)}_{yx|z} = B^{(1)}_{yx|z}$, $B^{(2)}_{yz|x} = B^{(1)}_{yz|x}$ and $\widehat{\beta}_{zx}(z) = B^{(1)}_{zx}$. Table 5 gives the values of $B^{(1)}_{yx|z}$, $B^{(2)}_{yx|z}$, $B^{(1)}_{yz|x}$ and $B^{(2)}_{yz|x}$ for the explanatory variables and auxiliary variables considered. The coefficients in $B^{(1)}_{yx|z}$ and $B^{(2)}_{yx|z}$ are generally very different and the average absolute difference is 4919. Inclusion of the auxiliary variables in the regression has had no appreciable effect on the aggregation effect on the regression coefficients and the R^2 is still considerably larger at the CD level than the individual level. The coefficients in $B^{(1)}_{yz|x}$ and $B^{(2)}_{yz|x}$ are also very different and the average absolute difference is 1676.

The adjustment procedure is built around replacing $B^{(2)}_{zx}$ by $\widehat{\beta}_{zx}(z)$. Table 6 gives the values of $B^{(1)}_{zx}$ $B^{(2)}_{zx}$ and $\widehat{\beta}_{zx}(z)$. This shows some beneficial effect and the adjusted R^2 values are closer to the individual level, although they are lower, suggesting some over-adjustment. The adjustment procedure replaces $B^{(2)}_{zx} B^{(2)}_{yz|x}$ by $\widehat{\beta}_{zx}(z) B^{(2)}_{yz|x}$ and these values and the corresponding individual level values are given in Table 7. This shows that with the exception of the intercept the adjusted CD values are considerably closer to the individual level values than the CD level values. The adjustment has had some beneficial effect in the estimation of $\beta_{zx}\beta_{yz|x}$ and the bias of the adjusted estimators is mainly due to difference between the estimates of $\beta_{yx|z}$. To some extent the adjustment has adjustedthe component of bias it is designed to reduce. The remaining biases mean that the overall effect are largely unaffected. It appears that conditioning on the auxiliary variables has not sufficiently reduced the biases due to the random effects.

Attempts were made to estimate the remaining variance components from purely aggregate data using MLn but this proved unsuccessful. Plots of the squares of the residuals against the inverse of the population sizes of groups showed that there was not always a increasing trend that would be needed to obtain sensible estimates.

These results are disappointing and suggest that to reduce the bias in the adjusted ecological regression it would be necessary to identify other potential auxiliary variables so that considerably more than half of the aggregation effects are accounted for. A diagnostic that indicates that there might be difficulties in obtaining a good adjustment is the aggregation effects remaining after the adjustment which are given in Table 1. While many have been reduced considerably, several are still quite large.

The multilevel model which incorporates grouping variables and random effects provides a general framework through which the causes of ecological biases can be explained. Using a limited number of auxiliary variables it is possible to explain about half the aggregation effects in income and a number of explanatory variables. Using individual level data on these adjustment variables the aggregation effects due to these variables can be removed. However, the resulting adjusted regression coefficients are no less biased. This suggests that for this adjustment approach to work well it is necessary to find adjustment variables that account for a very large proportion of the aggregation effects.

The CGV analysis shows that after allowing for the auxiliary variables considered there were residual grouping effects that seemed to be associated with indicators of higher socio-economic status. This suggest that we should attempt to find further auxiliary variables that account for these grouping effects and for which it would be reasonable to expect that the required individual level data to be available. In practice there are always going to be some residual group level effects and so we need methods that can satisfactorily account for them.

5. Ecological analysis in two by two tables

In social research the variables of interest are usually categorical at the individual level. Often the group level means are proportions corresponding to dichotomous variables at the individual level. The case of two such variables leads to estimating the conditional probabilities in two by two tables. Various methods of ecological inference in this situation were evaluated by Cleave, Brown and Payne (1995). These included ecological linear regression as originally suggested by Goodman (1953,1959) and a method based on an Aggregated Compound Multinonial (ACM) model. The ACM model assumed that the frequencies in each group have independent multinomial distributions and used a Dirichlet compounding distribution. Their evaluation favoured the ACM based method but they recognized that it is not easy to implement.

King (1997) proposed a new, computer intensive, method of ecological analysis for categorical data. This method models the conditional probabilities of Y given X as random effects with a joint truncated Normal distribution and also exploits the constraints implied by the group means for each group. We develop adjusted ecological analysis procedures using auxiliary variables to account for the variation of the relationship between groups and using individual level data on these variables. These methods are compared using data from the 1991 UK and Australian population census.

5.1 Adjusted ecological analysis for dichotomous variables

Let N_{abg} be the number of individuals for which $y_i = a$ and $x_i = b$. We use $" + "$ to indicate summation over a subscript. Define $P_{ab|g} = N_{abg}/N_{++g}$ as the finite population proportions for group g. The conditional proportions are $P_{a|bg} = N_{abg}/N_{+bg}$ for group g and $P_{a|b} = N_{ab+}/N_{+b+}$ for the overall population.

In our application the group level data consist of the means $\bar{y}_g$ and $\bar{x}_g$ based on the population in the gth groupfor $g = 1, ..M$ and so we will concentrate on this case, but all the methods considered can be used with sample based means. The quantities that we wish to estimates are $P_{1|0}$ and $P_{1|1}$. In some cases there may also be interest in $P_{1|0g}$ and $P_{1|0g}$ for $g = 1, ..M$

The standard approach in this situation is ecological linear regression (Goodman 1953, 1959) which is built around the relationship

$$\bar{y}_g = P_{1|0g}(1 - \bar{x}_g) + P_{1|1g}\bar{x}_g \tag{5.1}$$

If $P_{1|0g}$ and $P_{1|1g}$ are random variables with $E[P_{1|0g}|\bar{x}_g] = \pi_{1|0}$ and $E[P_{1|1g}|\bar{x}_g] = \pi_{1|1}$ then a linear regression of $\bar{Y}_g$ on $\bar{X}_g$ gives unbiased estimates of $\pi_{1|0}$ and $\pi_{1|1}$. This approach can produce estimates outside [0,1] and simple direct use of this model is not usually recommended. This approach can be extended to allow for the group level conditional probabilities to depend on covariates. In both cases the covariates are used at the group level to explain some of the variation of $P_{1|0g}$ and $P_{1|0g}$ across areas.

King's ecological inference method is also built around equation (1) and he noted some important relationships that arise from it :

$$P_{1|0g} = \frac{\bar{y}_g}{1 - \bar{x}_g} - \frac{\bar{x}_g}{1 - \bar{x}_g} P_{1|1g} \tag{5.2}$$

which is called a tomography line and

$$max(0, \frac{\bar{y}_g - \bar{x}_g}{1 - \bar{x}_g}) \leq P_{1|0g} \leq min(1, \frac{\bar{y}_g}{1 - \bar{x}_g}) \tag{9}$$

$$max(0, \frac{\bar{y}_g - (1-\bar{x}_g)}{\bar{x}_g}) \leq P_{1|1g} \leq min(1, \frac{\bar{x}_g}{\bar{x}_g}) \qquad (10)$$

These relationships have been noted before for the two-way tables at the population level and form the basis of the method of bounds (see Duncan and Davis, 1953). King (1997) suggested applying these relationships at the group level to assist in producing estimates. The group level conditional proportions $P_{1|0g}$ and $P_{1|1g}$ are assumed to have a joint Normal distribution which is truncated so that they each lie in the [0,1] interval. The method incorporates the fact that $\bar{y}_g$ and $\bar{x}_g$ imply bounds for $P_{1|1g}$ and $P_{1|0g}$. This approach produces estimates of $P_{1|1g}$ and $P_{1|0g}$ which can then be combined to produce estimates of $P_{1|1}$ and $P_{1|0}$. This method can incorporate group level covariates to help model the variation between groups.

Assume that there is a single auxiliary variable z and that y, x and z are all categorical. The target of inference is the conditional probability distribution of y given x, $P(y|x)$. Our approach is based on

$$P(y|x) = \frac{\sum_z P(y|x,z)P(x|z)P(z)}{\sum_z P(x|z)P(z)} \qquad (5.5)$$

Estimation of $P(y|x)$ is then based on equation (5.5) with the individual level data being used to estimate $P(z)$ and using linear regression based on aggregate data to estimate $P(y|x,z)$ and $P(x|z)$. Without any auxiliary variables this produces the standard Goodman ecological regression. Our hope is that with inclusion of appropriate auxiliary variables much of the group effects can be removed and the ecological linear regression will provide reasonable estimates.

For a single categorical auxiliary variable the information needed to calculate the adjusted marginal probabilites using (5.5) is only the proportions in each category. This could be calculated from the weighted group level data or could come from some other source, such as a survey. If two or more categorical auxiliary variables are used then the marginal cross tabulations of these variables are required, unless further assumptions can be made concerning the relationship between the auxiliary variables. No individual level data about the variables of direct interest are used.

An adjusted correlation approach can also be developed for two dichotomous variables and is based on the fact that the correlation combined with the marginal totals determines the proportions in the cross classification i.e

$$P_{11+} = R_{yx}\sqrt{P_{1++}(1-P_{1++})P_{+1+}(1-P_{+1+})} + P_{1++}P_{+1+} \qquad (5.6)$$

where R_{yx} is the correlation coefficient based on the table of proportions P_{ab+}. The adjusted matrix $\widehat{\Sigma}_{ww}(z)$ can used to produce adjusted estimates of the correlation $R_{yx}(z)$ which can then be substituted into equation (5.6). While this method is based on linear relationships it has the advantage that it enables use of information about several auxiliary variables even if only the two way cross tabulations are available.

The potential benefit of using group level information about covariates is discussed by Cleave, Brown and Payne (1995) and King (1997). In both cases the covariates are used at the group level to explain some of the variation in the random coefficients characterizing the relationship between the variables across the groups. Our approach is motivated by the idea that a large part of the variation in the relationships across areas is also due to compositional effects and can be removed by inclusion of the auxiliary variables.

5.2 Empirical evaluation

We used individual and group level data from the 1991 UK and Australian population censuses to evaluate the methods. For the UK group level data were obatained for EDs from the Small Area Statistics data base and individual level data wrere available from a 2 percent Sample of Anonymized Records (SAR) for Local Authority Districts (LADs). The evaluation used data for the LAD of Manchester which contains 897 EDs and 7613 individuals in the SAR, of which 5802 are aged 16 or more. The Australian 1991 population census data for Adelaide described in section 4.3 was also used.

In this evaluation we considered the following estimators of the conditional probabilities $P_{1|0}$ and $P_{1|1}$.

(a) SAR relative frequencies

(b) Adjusted ecological linear regression.

(c) Adjusted correlation approach.

(d) King's ecological inference method.

The evaluation considered the following tables from the 1991 UK population census employed by marital status, employed by LLTI, and employed by marital status. The auxiliary variables considered were age 45-59, age 60+, living in owner occupied dwelling, renting from local authority. These auxiliary variables were chosen beacause of their success in removing aggregation effects in correlations in the evaluation by Steel, Holt and Tranmer (1996)

For the evaluation using the 1991 Australian population census the dependent variable was income divided into two categories with approximately 50 percent of the population in each. The following tables from the 1991 Australian population census were considered: income by whether or not a person had a degree, income by sex, income by whether born in Australia. The auxiliary variables considered were age 40-59, age60+, living in owner occupied dwelling, renting from government, housing type.

Table 8 to 13 gives the estimates of the conditional probabilities $P_{1|0}$ and $P_{1|1}$ using the methods listed above. The methods were tried with and without auxiliary variables. The 'true' values of the probabilities as estimated from the SAR are also given. The adjusted ecological regression and adjusted correlation methods were implemented using the SAS package and King's method was implemented using EzI software, developed by King and colleagues. There are only limited results of King's method with auxiliary variables as we experienced difficulties with the estimation procedure converging in many cases invovling one or more auxilary variables. However, the spirit of King's method is to account for much of the variation across groups through the random effects used in the model.

In Table 8, when adjustment variables are not included, the ecological linear estimates are considerably different from the SAR values. The inclusion of owner occupied housing as an adjustment variable leads to ecological linear estimates that are much closer to their true values. Used as a single adjustment variable 'aged 60 and over' (age60+) does not improve the estimates. The estimates obtained by the adjusted correlation method do not perform as well as the adjusted ecological linear method. In this example, King's method without covariates gives results similar to those for the ecological linear method without covariates. However, when the owner occupied auxiliary variable is used, the estimated probabilities are close to the SAR values and are closer than those obtained using the adjusted methods.

For the estimates of the proportion unemployed by marital status given in Table 9, the unadjusted ecological linear method provide poor estimates, with an out of range estimate for $P_{1|1}$. Including the variable owner occupied as an adjustment variable in the ecological linear approach leads to some improvement. The adjusted correlation method works reasonable well for combinations of adjustment variables that include housing tenure and age. King's method without covariates works better in this case than in the previous example. This may be because differences in the the proportions unemployed and married across the CDs lead to more informative constraints and both the conditional probabilities are close to zero. However, the estimates from King's method still have some bias. When owner occupied is used as an auxiliary variable, the estimates obtained from Kings method are very close to their true values.

In Table 10 King's method works very well for estimating $P_{1|0}$ without any auxiliary variables. The linear regression method is imporved by use of owner occupied as an auxiliary variable, but is still larger than the SAR value. For $P_{1|1}$ King's estimate is not very close, but the other methods give negative estimates which are imporved by the use of auxiliary variable but still remain negative. This demonstrates the problem that the ecological linear method has when the probabilty is close to zero or one because there is nothing in the method to force estimates in this range. The low value of King's estimate here suggests that the use of the bounds (5.3) and (5.4) is largely responsible for it being closer than the other estimates.

In Table 11 the linear method works slightly better than King's when no auxilairy variables are used. When owner occupied is used as an auxiliary variable the ecological linear regression method gives estimates close to the SAR values and better than the those from King's method.

King's method and the linear regression method perform similarly for the estimation of $P_{1|0}$ in Table 12. As $P_{1|1}$ is close to one the linear regression method tends to give estimates greater than one. King's method gives an estimate less than one, as it has to be, but is higher than the SAR value, again suggesting that the bounds have been the main factor in generating the estimate.

In Table 13 King's and the linear regression method give very similar estimates, which for $P_{1|1}$ are close to the SAR value but for $P_{1|0}$ still have some bias. Inclusion of the owner occupied variable improves the King estimate of $P_{1|1}$ but not the adjusted linear regression estimate.

In these examples King' method and the adjusted ecological linear regression methods give similar results except when the probabilities being estimated are getting near zero or one. The problem with the adjusted eclogical linear method is that nothing has been done to constrain the estimates appropriately and so out of range estimates can be produced. This is an advantage of King's method. These results highlight the importance of using auxiliary variables in the ecological estimation procedures to obtain reasonable estimates for both King's and the other methods. The choice of auxiliary variables used, either singly or in combination, is important and methods of identifying effective adjustment variables need to be developed. In this example, the housing characteristics are the most powerful adjustment variables. If appropriate auxiliary variables are used, quite simple methods can perform almost as well as more sophisticated ones. We expect that the adjusted ecological methods can be extended to also include random effects within the sort of multilevel framework developed, for example by Goldstein (1995). Modifying these methods to fulfil the constraints given by (5.7) and (5.4) will probably improve their performance

6 Future developments

The basic problems affecting ecological analysis are due the variation of relationships between groups which may be related to the explanatory variables and homogeneity of variables within groups. To unravel ecological analysis we first need realistic models at the individual level that reflects these features. Two main avenues for doing this are to include other variables that partly explain the between group variation and within group homogeneity and structures for the random components that include group level effects. Methods that use only one of these avenues are unlikely to be successful. Our results for linear regression suggest that including a small number of auxiliary variables can explain a lot of the within group homogeneity but that a significant amount will always remain Hence, methods for account for the remaining homogeneity due to group level effect need to be developed. The result for two by two tables also show the value of using auxiliary variables. Hence, future efforts should focus on methods for identifying such variables.

Acknowledgements

This work was supported by the Economic and Science Research Council (Grant number R 000236135).

References

Cleave, N., Brown, P.J. and C. D. Payne (1995). Evaluation of Methods for Ecological Inference. *Journal of the Royal Statistical Society, A,* **158**, pp 55 - 72

Davies, H., Joshi, H. and Clarke, L. (1997). Is it cash that the deprived are short of? *J. Royal Statistical Society*, Series A, 160, 107-126.

Goldstein, H. (1986), Multi level mixed model analysis using iterative genralized least squares. Biometrika, 73(1), 43-56

Duncan, O.D., and Davis, B (1953). An alternative to Ecological Correlation. Amer., Sociological Review, 18, 665-666.

Goldstein, H. (1995), Multi level Models in Educational and Sociological Research, Griffin, London. Goodman, L.A. (1953). Ecological regression and behaviour of individuals. *American Sociological Review*, **18**, 663-666.

Goodman, L.A. (1959). Some alternatives to Ecological regression. *American Journal of*

Sociological, **64**, 610-624.

King, G. (1997). *A Solution to the Ecological Inference Problem :Reconstructing Individual Behavior from Aggregate Data*. Princeton Univ Press

Holt, D, Steel, D., and Tranmer, M. (1996) ”Area Homogeneity and the Modifiable Areal Unit Problem” . Geographical Systems., Vol 3, pp181 -200.

Holt, D, Steel, D., Tranmer, M. and Wrigley, N (1996) ”Aggregation and Ecological Effects in Geographically Based Data”. Geographical Analysis,. Vol 28, no3, pp244- 262

Smith, K.(1977). Another look at the clustering perspective on aggregation problems, *Sociological Methods and Research*, 5, 289-310.

Smith, T.M.F. (1989). Chapter 6.5 The Pearson adjustment for multivariate normal model in Skinner, C., Holt, D. and Smith, T.M.F. *Analysi of Complex Surveys, John Wiley Sons, London*.

Steel, D. and Holt, D. (1996a), ”Rules For Random Aggregation”. Environment and Planning , A. Vol 28, no6, pp 957-978.

Steel, D. and Holt, D. (1996b), ”Analysing and Adjusting Aggregation Effects: The Ecological Fallacy Revisited”. The International Statistical Review , Vol64, no1, pp 39 -60.

Steel, D., Holt, D. and Tranmer, M. (1996), ”Making Unit -Level Inference from Aggregate Data”. Survey Methodology. Vol 22, no1, pp 3-15.