UNRAVELLING ECOLOGICAL ANALYSIS

D. G. STEEL, M. TRANMER, AND D. HOLT

Received 10 February 2004; Accepted 25 May 2005

Ecological analysis involves analysing aggregate data for groups of individuals to make inferences about relationships at the individual level. Often the results of such analyses give badly biased estimates. This paper will consider the sources of bias in linear regression analysis using aggregate data. The role of variation of the individual level relationships between groups and the consequent within-group correlations and how these are related to auxiliary variables that characterise the differences between groups is considered. A method of adjusting ecological regression for the effects of auxiliary variables is described and evaluated using data from the 1991 Australian Census.

1. Introduction

Ecological analysis involves analysing aggregate data such as the means of a set of groups to make inferences about individual level relationships. An advantage of ecological analysis is that it uses data that are already available at relatively low cost. In an ecological analysis information from different sources may be brought together using aggregates for the same geographical areas.

Ecological analysis is a potentially 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. Ecological analysis may produce seriously biased estimates of individual level relationships, which limits its practical use.

In Section 2 we consider the targets of inference and how ecological analysis can be considered as a form of multi-level modelling. In Section 3 we consider ecological linear regression analysis within a multi-level framework and clearly identify the sources of the biases. In Section 4 we describe a method of adjusting ecological regression using individual level data for auxiliary variables. Sections 5 and 6 give an evaluation of the aggregation effects and the adjustment method. Section 7 gives a discussion.

Hindawi Publishing Corporation
Journal of Applied Mathematics and Decision Sciences
Volume 2006, Article ID 38358, Pages 1–18
DOI 10.1155/JAMDS/2006/38358

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

In ecological analysis the population is composed of groups of individual units and has a multi-level structure. Statistical analysis should be based on a statistical model that reflects this structure, that is, a multi-level model. Consider a simple two level population, the first level being the individual and a population of N individuals in which the ith individual has a vector of response variables y_i and a group indicator c_i. The population comprises M groups and the number of individuals in the gth group is N_g.

A simple two-level model for the population is

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

where ν_g is a vector of random group level effects and ϵ_i is a vector of individual level effects. The standard assumptions are

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

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

For this model

$$\begin{gathered} E[y_i] = \mu, \qquad V(y_i) = \Sigma^{(1)} + \Sigma^{(2)} = \Sigma, \\ \text{Cov}(y_i, y_j) = \Sigma^{(2)} \quad \text{for } c_i = c_j,\ i \neq j. \end{gathered} \tag{2.4}$$

Multi-level models provide a useful framework for any situation in which the process that generated the data involved groups, either through sampling or aggregation, or both. In standard multi-level modelling the targets of inference are the fixed mean parameter μ and the parameters 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. The parameters $\Sigma^{(1)}$ and $\Sigma^{(2)}$ indicate the relative importance of purely individual and purely group level effects. Estimation of the parameters usually requires a sample of groups and a sample of individuals within them and indicators that indicate to which group each individual belongs (see Goldstein [3]).

In ecological analysis the targets of inference are at the unit level but the main data available consist of group level means. The assumption is that if 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 parameters of interest 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 ecological fallacy 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. The marginal relationships may be relevant, for example if the government is planning a policy that will be applied across the whole population.

In many geographical applications there is no direct interest in individual level relationships. The focus of interest may be at a level above the individual, that is, $\Sigma^{(2)}$, but the area level analysis will include a component due to the individual level relationships in the population, that is, $\Sigma^{(1)}$.

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. Steel et al. [10] consider how multi-level models can be analysed in a number of different cases of data availability at the individual and group level.

Conventional multi-level modelling is carried out using a unit level data set which has group indicatives. We will consider the case of aggregate data consisting of group means. The group means are often based on a census of groups and individuals within them, but can also come from a sample. Assume that there exists a sample data set s of size n and that these individual data have been aggregated to provide a set, s_1, of m group means, $\overline{y}_g$, $g = 1,\ldots,m$, which are available for analysis. The number of sample individuals in each area, n_g, is also known. The overall sample mean is $\overline{y} = \Sigma_{g\in s_1} n_g \overline{y}_g / n$.

The source of the ecological bias can be identified from the model given by (2.1) to (2.3). Consider $S_{yy}^{(1)} = \Sigma_{i\in s}(y_i - \overline{y})(y_i - \overline{y})'/(n-1)$, the covariance matrix calculated from the unit level sample data and $S_{yy}^{(2)} = \Sigma_{g\in s_1} n_g(\overline{y}_g - \overline{y})(\overline{y}_g - \overline{y})'/(m-1)$, the covariance matrix calculated from the group level means, using the group sample sizes as weights. The key relationships are (see Steel and Holt, [8])

$$E\left[S_{yy}^{(2)}\right] = \Sigma^{(1)} + \overline{n}^*\Sigma^{(2)}, \qquad E\left[S_{yy}^{(1)}\right] = \Sigma^{(1)} + \left(1 - \frac{\overline{n}^0 - 1}{n-1}\right)\Sigma^{(2)}. \tag{2.5}$$

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

$$\overline{n}^* = \overline{n}\left(1 - \frac{C_n^2}{m-1}\right), \quad \overline{n}^0 = \overline{n}(1 + C_n^2) \quad \text{in which } C_n^2 = \frac{1}{m}\sum_{g\in s_1}(n_g - \overline{n})^2/\overline{n}^2 \tag{2.6}$$

is the square of the coefficient of variation of the group sample sizes. These results have some important implications. If $\Sigma^{(2)} = 0$ then the group and individual level covariance matrices have the same expectation. However, there will be a large difference in the expectations when $\overline{n}^*$ is large even if the elements of $\Sigma^{(2)}$ are much smaller than those of $\Sigma^{(1)}$, but not zero. Census Collection Districts have approximately 500 people in them, and geographical groups with much larger populations are used in ecological analysis. In these cases $S_{yy}^{(2)}$ contains very little contribution from $\Sigma^{(1)}$, but is mainly determined by $\Sigma^{(2)}$. Using $S_{yy}^{(2)}/\overline{n}^*$ to produce estimates of $\Sigma^{(2)}$ will not be badly biased if $\overline{n}^*$ is large. As an estimate of Σ the bias of $S_{yy}^{(1)}$ is $O(m^{-1})$ and it will be a reasonable estimate of the marginal individual level relationships provided m is not small. However, using $S_{yy}^{(2)}$ to estimate Σ will result in a bias of $(n^* - 1)\Sigma^{(2)}$. 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 et al. [4]). To remove the bias requires estimation of $\Sigma^{(1)}$ and $\Sigma^{(2)}$.

3. Explaining biases in ecological linear regression using a multi-level model with auxiliary variables

If individuals are allocated to groups at random 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 groups in the analysis (Steel and Holt, [9]).

In practice, individuals who live in the same area exhibit positive intra-group correlation for a variety of socio-economic characteristics. The homogeneity within groups is a key factor in the ecological fallacy. Suppose that there is a set of auxiliary 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-group homogeneity they may affect the ecological analysis greatly. The matrices $z = [z_1,\dots,z_N]'$, $c = [c_1,\dots,c_N]'$ give the values of all units in the population of size N. The ith individual has a vector of response variables y_i and a vector of explanatory variables x_i.

We will focus on the cases when there are 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.

Steel and Holt [8] considered the implication of a multi-level model with auxiliary variables for the ecological analysis of covariance matrices and correlation coefficients. They also developed a method for adjusting the analysis of aggregate data to provide less biased estimates of covariance matrices and correlation coefficients. Holt et al. [4] 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 between groups. We consider the implications of this model for ecological linear regression analysis.

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

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:

$$w_i = \mu_{w|z} + \beta'_{wz} z_i + \nu_g + \epsilon_i, \quad i \in g, \tag{3.1}$$

where

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

This model implies

$$\begin{aligned} E(w_i \mid z, c) &= \mu_{w|z} + \beta'_{wz} z_i, \\ V(w_i \mid z, c) &= \Sigma_{ww|z}^{(1)} + \Sigma_{ww|z}^{(2)} = \Sigma_{ww|z}, \\ \operatorname{Cov}(w_i, w_j \mid z, c) &= \Sigma_{ww|z}^{(2)} \quad \text{if } c_i = c_j,\ i \neq j. \end{aligned} \tag{3.3}$$

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

$$\Sigma_{ww} = \Sigma_{ww|z} + \beta'_{wz}\Sigma_{zz}\beta_{wz} \tag{3.4}$$

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 [8] showed

$$\begin{aligned} E[\overline{w} \mid z, c] &= \mu_w + \beta'_{wz}(\overline{z} - \mu_z), \\ E[S^{(2)}_{ww} \mid z, c] &= \Sigma_{ww} + \beta'_{wz}(S^{(2)}_{zz} - \Sigma_{zz})\beta_{wz} + (\overline{n}^* - 1)\Sigma^{(2)}_{ww|z} \\ &= \Sigma_{ww|z} + \beta'_{wz}S^{(2)}_{zz}\beta_{wz} + (\overline{n}^* - 1)\Sigma^{(2)}_{ww|z}. \end{aligned} \tag{3.5}$$

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 $O(m^{-1})$,

$$\begin{aligned} E[B^{(2)}_{yx} \mid z, c] = {} & \left[\Sigma_{xx} + \beta'_{xz}(S^{(2)}_{zz} - \Sigma_{zz})\beta_{xz} + (\overline{n}^* - 1)\Sigma^{(2)}_{xx|z}\right]^{-1} \\ & \times \left[\Sigma_{xy} + \beta'_{xz}(S^{(2)}_{zz} - \Sigma_{zz})\beta_{yz} + (\overline{n}^* - 1)\Sigma^{(2)}_{xy|z}\right]. \end{aligned} \tag{3.6}$$

Set $A = E[S^{(2)}_{xx} \mid z, c] = [\Sigma_{xx} + \beta'_{xz}(S^{(2)}_{zz} - \Sigma_{zz})\beta_{xz} + (\overline{n}^* - 1)\Sigma^{(2)}_{xx|z}]$. The model implies $\beta_{yx|z} = \Sigma^{-1}_{xx|z}\Sigma_{xy|z}$, $\beta_{zx} = \Sigma_{xx}^{-1}\Sigma_{xz}$ 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, conditional on z and c, can be shown to be (Steel, [7]):

$$A^{-1}\beta'_{xz}(S^{(2)}_{zz} - \Sigma_{zz})(\beta_{yz} - \beta_{xz}\beta_{yx}) + (\overline{n}^* - 1)A^{-1}\left[\Sigma^{(2)}_{xy|z} - \Sigma^{(2)}_{xx|z}\beta_{yx}\right] \tag{3.7}$$

$$= \left[A^{-1}\beta'_{xz}S^{(2)}_{zz} - \beta_{zx}\right]\beta_{yz|x} + (\overline{n}^* - 1)A^{-1}\left[\Sigma^{(2)}_{xy|z} - \Sigma^{(2)}_{xx|z}\beta_{yx|z}\right]. \tag{3.8}$$

The first term in the bias in (3.8) 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 (3.8) 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, that is, $\beta_{yx|z} = {\Sigma^{(2)}_{xx|z}}^{-1}\Sigma^{(2)}_{xy|z} = \beta^{(2)}_{yx|z}$. The second term in (3.8) 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 $\overline{n}^*$ which can be very large, for example when the group means are based on all individuals in the groups.

The effect of aggregation has been considered for some aggregation criterion (see Blalock, [1]). 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. In this case $A = \Sigma_{xx} + \beta'_{xz}(S_{zz}^{(2)} - \Sigma_{zz})\beta_{xz}$ and the bias is $[E[B_{zx}^{(2)} \mid z,c] - \beta_{zx}]\beta_{yz|x}$ and is entirely due to the effect of aggregation on the implied estimate of β_{zx}.

The model here allows for group effects in two ways that explain the effect of aggregation. The form of the bias in (3.7) and (3.8) 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

$$E[B_{yx}^{(2)} \mid c] = (1 - a^*)\beta_{yx}^{(1)} + a^*\beta_{yx}^{(2)}, \tag{3.9}$$

where $a^* = \overline{n}^*\delta_{xx}/\{1 + (\overline{n}^* - 1)\delta_{xx}\}$ and $\delta_{xx} = \Sigma_{xx}^{(2)}\Sigma_{xx}^{-1}$. 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 $\overline{n}^*$ is large (see Holt et al. [4])

For the case of several explanatory variables and one grouping variable

$$\begin{aligned} E[B_{zx}^{(2)} \mid z,c] = {} & \left[I + (\overline{n}^* - 1)\left(I - \frac{(Q_z - 1)\Re^2_{x|z}}{1 + \Re^2_{z|x}(Q_z - 1)} \right)\Sigma_{xx}^{-1}\Sigma_{xx|z}^{(2)} \right]^{-1} \\ & \times \beta_{zx}\frac{Q_z}{1 + \Re^2_{z|x}(Q_z - 1)}, \end{aligned} \tag{3.10}$$

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

$$E[B_{zx}^{(2)} \mid z,c] = \beta_{zx}\frac{Q_z}{1 + \Re^2_{z|x}(Q_z - 1)} \tag{3.11}$$

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, [5]) but it relies on the grouping being one dimensional.

4. An adjusted ecological regression method using auxiliary variables

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 estimates 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

$$V(\overline{w}_g \mid c) = \Sigma_{ww}^{(1)}/n_g + \Sigma_{ww}^{(2)}. \tag{4.1}$$

For example the IGLS procedure embodied in MLwiN can be used (see Goldstein, [3]). This approach relies on there being reasonable variation in the sample sizes between the groups and on the variance structure originally assumed at the individual 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 $(\overline{w}_g - \hat{\mu}_w)(\overline{w}_g - \hat{\mu}_w)'$ against $1/n_g$ where $\hat{\mu}_w$ is the current estimate of μ_w and the estimates of $\Sigma^{(1)}_{ww}$ and $\Sigma^{(2)}_{ww}$ 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 (3.1) $E[B^{(2)}_{wz} \mid z, c] = \beta_{wz}$ where $B^{(2)}_{wz} = (S^{(2)^{-1}}_{zz})S^{(2)}_{zw}$. If an estimate of the individual level population covariance matrix for z were available, possibly from another source, Steel and Holt [8] proposed the following adjusted estimator of Σ_{ww},

$$\hat{\Sigma}_{ww}(z) = S^{(2)}_{ww} + B^{(2)'}_{wz}(\hat{\Sigma}_{zz} - S^{(2)}_{zz})B^{(2)}_{wz} = S^{(2)}_{ww|z} + B^{(2)'}_{wz}\hat{\Sigma}_{zz}B^{(2)}_{wz}, \tag{4.2}$$

where $\hat{\Sigma}_{zz}$ is the estimate of Σ_{zz} calculated from individual level data. This estimator corresponds to a Pearson-type adjustment (Smith, [6]) and for Normally distributed data is the MLE when $\Sigma_{ww|z} = 0$ and $\hat{\Sigma}_{zz}$ is also the MLE. This estimator removes the aggregation bias due to z. Adjusted regression coefficients can then be calculated from $\hat{\Sigma}_{ww}(z)$, that is,

$$\hat{\beta}_{yx}(z) = \hat{\Sigma}^{-1}_{xx}(z)\hat{\Sigma}_{xy}(z). \tag{4.3}$$

The adjusted estimator replaces the components of bias in (3.7) due to $\beta'_{xz}(S^{(2)}_{zz} - \Sigma_{zz})\beta_{xz}$ and $\beta'_{xz}(S^{(2)}_{zz} - \Sigma_{zz})\beta_{yz}$ by $\beta'_{xz}(\hat{\Sigma}_{zz} - \Sigma_{zz})\beta_{xz}$ and $\beta'_{xz}(\hat{\Sigma}_{zz} - \Sigma_{zz})\beta_{yz}$ respectively. Set $\hat{A}(z) = E[\hat{\Sigma}_{xx}(z) \mid z, c] = \Sigma_{xx} + \beta'_{xz}(\hat{\Sigma}_{zz} - \Sigma_{zz})\beta_{xz} + (\overline{n}^* - 1)\Sigma^{(2)}_{xx|z}$. Then the bias of $\hat{\beta}_{yx}(z)$ is, to $O(m^{-1})$

$$[\hat{A}(z)^{-1}\beta'_{xz}\hat{\Sigma}_{zz} - \beta_{zx}]\beta_{yz|x} + (\overline{n}^* - 1)\hat{A}(z)^{-1}\left[\Sigma^{(2)}_{xy|z} - \Sigma^{(2)}_{xx|z}\beta_{yx|z}\right]. \tag{4.4}$$

Suppose that $\hat{\Sigma}_{zz}$ is an estimate based on a individual level sample involving m_0 first stage units. Then for many sample designs $\hat{\Sigma}_{zz} = \Sigma_{zz} + O(m_0^{-1})$, and so to $O(1/m_0)$ the bias of $\hat{\beta}_{yx}(z)$ is

$$\begin{aligned}&\hat{A}(z)^{-1}(\overline{n}^* - 1)\Sigma^{(2)}_{xx|z}\left[\beta^{(2)}_{yx|z} - \beta_{yx}\right] \\ &\quad = [\hat{A}(z)^{-1}\Sigma_{xx} - I]\beta_{zx}\beta_{yz|x} + \hat{A}(z)^{-1}(\overline{n}^* - 1)\Sigma^{(2)}_{xx|z}\left[\beta^{(2)}_{yx|z} - \beta_{yx|z}\right].\end{aligned} \tag{4.5}$$

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 $\hat{\beta}_{yx}(z)$ is $O(m_0^{-1})$.

The adjusted estimator can be rewritten as

$$\hat{\beta}_{yx}(z) = B^{(2)}_{yx|z} + \hat{\beta}_{zx}(z)B^{(2)}_{yz|x}, \tag{4.6}$$

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

$$B^{(2)}_{yx}(z) = B^{(2)}_{yx|z} + B^{(2)}_{zx}B^{(2)}_{yz|x} \qquad B^{(1)}_{yx}(z) = B^{(1)}_{yx|z} + B^{(1)}_{zx}B^{(1)}_{yz|x}. \tag{4.7}$$

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

The bias due to the conditional variance components $\Sigma_{ww|z}^{(2)}$ remains. The two approaches can be combined. Multilevel modelling with aggregate data can be used to produce estimates of $\Sigma_{ww|z}^{(2)}$, $\Sigma_{ww|z}^{(1)}$ 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, calculate

$$\widehat{\Sigma}_{ww} = \widehat{\Sigma}_{ww|z}^{(1)} + \widehat{\Sigma}_{ww|z}^{(2)} + \widehat{\beta}_{wz}\widehat{\Sigma}_{zz}\widehat{\beta}_{wz} \tag{4.8}$$

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

5. Evaluation of aggregation effects in ecological regression

5.1. 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). The 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 census households sample file (HSF) which is a one percent sample of households, and the people within them.

The evaluation concentrated on the dependent variable of personal income. This variable is collected in 14 ranges but was treated 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, possessing a degree, employed-manual occupation, employed-managerial or professional occupation, employed-other, unemployed, born in Australia, born in UK and four age categories. The auxiliary variables considered were: age 45 to 59, age 60+, owner occupied, renting from government, housing type.

5.2. 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, that is, $Q_a = S_{aa}^{(2)}/S_{aa}^{(1)}$ are given in Table 5.1, along with the associated estimate of the intra-CD correlation $\widehat{\delta}_{aa}$. All the variables 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 variances 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_a 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 5.2, 5.3, and 5.4 summarize the effect of aggregation on the analysis of bivariate covariances, correlations and regression coefficients between income and each of the explanatory and auxiliary variables. The CD level correlations are generally of the same sign

Table 5.1. Summary of aggregation effects on variances.

Variable	Mean	Before adjustment Aggregation effect	Before adjustment Intra-class correlation	After adjustment Aggregation effect	After adjustment Intra-class correlation	Ratio of age effects
Income	17186.0	23.4	0.050	15.9	0.033	0.68
Renting Govt	0.10	171.1	0.380	1.0	0.000	—
Housing type	0.90	109.5	0.243	1.0	0.000	—
Owner occupied	0.69	88.8	0.196	1.0	0.000	—
Age 60+	0.22	35.0	0.076	1.0	0.000	—
Manager prof	0.16	21.1	0.045	14.2	0.029	0.67
Married	0.55	20.0	0.043	5.8	0.011	0.29
Degree	0.07	18.8	0.040	14.5	0.030	0.77
Born UK	0.14	14.7	0.031	12.6	0.026	0.86
Emp other	0.28	11.7	0.024	4.1	0.007	0.35
Born Aust	0.72	11.3	0.023	10.5	0.021	0.92
Manual occup	0.10	10.6	0.022	7.2	0.014	0.68
Age 20–29	0.21	10.3	0.011	6.0	0.011	0.58
Age 45–59	0.19	9.2	0.018	1.0	0.000	—
Unemployed	0.07	8.4	0.017	4.6	0.008	0.55
Age 15–19	0.10	6.3	0.012	4.2	0.007	0.66
Female	0.52	2.6	0.004	1.9	0.002	0.70

but larger than the corresponding individual level correlation. In most cases the change is sufficient to affect the substantive interpretation. There a number of cases in which the correlations have different signs at the two levels. The effect of aggregation on the regression coefficents are similar.

5.3. Mutivariate aggregation effects. 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 [8] 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 $d_1^{(2)},\dots,d_p^{(2)}$ of $(S^{(1)})^{-1}S^{(2)}$ with associated eigenvalues $\theta_1^{(2)},\dots,\theta_p^{(2)}$. Let $D^{(2)} = [d_1^{(2)},\dots,d_p^{(2)}]$; then the CGVs are defined by $U = D'Y$ and have covariance matrix $\operatorname{diag}(\theta_l^{(2)})$ at the CD level and I_p at the individual level. Subject to the constraints of being mutually uncorrelated at the individual 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_a and the eigenvalues give the aggregation effect of each of the mutually orthogonal grouping dimensions in the set of variables being considered. Summary measures of the aggregation effects in multivariate data are given by $\overline{\theta} = \sum_l \theta_l^{(2)}/p = \operatorname{trace}[(S^{(1)})^{-1}S^{(2)}]/p$ and

Table 5.2. Summary of aggregation effects on covariances between income and other variables.

		Before adjustment		After adjustment	
Variable	Ind level	CD level	Aggregation effect	CD level	Aggregation effect
Renting Govt	−540	−141414	261.7	−839	1.6
Housing type	309	−1804	−5.8	−4	−0.0
Owner occupied	699	121671	174.1	602	0.9
Age 60+	−1167	−35023	30.0	−1169	1.0
Emp other	903	37059	41.0	−171	−0.2
Manager prof	2762	110152	39.9	71416	25.9
Married	1226	41566	33.9	8519	6.9
Degree	1321	64250	48.6	48330	36.6
Born UK	125	−10204	81.8	−8021	−64.3
Born Aust	182	38752	212.7	36264	199.0
Manual occup	166	−38640	−233.3	−30572	−184.6
Age 20–29	−46	−17291	376.3	−14744	320.8
Age 45–59	753	2534	16.6	860	1.1
Unemployed	−687	−30274	44.1	−14081	20.5
Age 15–19	−1087	5939	−5.5	−300	0.3
Female	−2087	−2730	1.3	7082	−3.4

Table 5.3. Summary of aggregation effects on conditions between income and other variables.

		Before adjustment		After adjustment	
Variable	Ind level	CD level	Aggregation effect	CD level	Aggregation effect
Renting Govt	−0.129	−0.533	4.1	−0.050	0.4
Housing type	0.069	−0.008	−0.1	−0.000	−0.0
Owner occupied	0.106	0.404	3.8	0.028	0.2
Age 60+	0.194	−0.203	1.0	0.049	0.3
Emp other	0.136	0.337	2.5	−0.003	−0.0
Manager prof	0.510	0.916	1.8	0.880	1.7
Married	0.169	0.265	1.6	0.123	0.7
Degree	0.339	0.786	2.3	0.817	2.4
Born UK	0.024	−0.105	−4.4	−0.108	−4.6
Born Aust	0.027	0.351	13.0	0.415	15.4
Manual occup	0.037	−0.549	−14.8	−0.641	−17.3
Age 20–29	−0.008	−0.188	24.1	−0.256	32.9
Age 45–59	0.135	0.153	1.1	0.039	0.3
Unemployed	−0.181	−0.569	3.1	−0.435	2.4
Age 15–19	−0.255	0.115	−0.5	−0.009	0.0
Female	−0.287	−0.048	0.2	0.179	−0.6

Table 5.4. Summary of aggregation effects on regression between income and other variables.

		Before adjustment		After adjustment	
Variable	Ind level	CD level	Aggregation effect	CD level	Aggregation effect
Renting Govt	−6523	−9968	1.5	−10122	1.6
Housing type	3303	−176	−0.1	−44	−0.0
Owner occupied	3394	6651	2.0	2920	0.9
Age 60+	−6812	−5344	0.9	−6822	1.0
Emp other	4336	15177	3.5	−200	−0.0
Manager prof	20004	37784	1.9	36476	1.8
Married	4959	8384	1.7	5939	1.2
Degree	18467	47725	2.6	46478	2.5
Born UK	962	−5357	−5.6	−4916	−5.1
Born Aust	840	15744	18.7	15941	19.0
Manual occup	1756	−38658	−22.0	−45197	−25.7
Age 20–29	−278	−10172	36.5	−15000	53.9
Age 45–59	5132	9259	1.8	5856	1.1
Unemployed	−10120	−53096	5.2	−45233	4.5
Age 15–19	−12664	10990	−0.9	−836	0.1
Female	−8349	−4123	0.5	15198	−1.8

$\overline{Q} = \sum_a Q_{aa}/p$. 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, that is, (y,x,z), gave $\overline{\theta} = 27.7$ and $\overline{Q} = 33.7$. The first four CGVs accounted for 83 percent of the total aggregation effect. The coefficients for the first four CGVs, showed 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 $\overline{\theta} = 13.6$ and $\overline{Q} = 14.5$. The first five CGVs accounted for 82 percent of the total aggregation effect. The coefficients for the first five CGVs, showed 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.

5.4. Aggregation effects on multiple regression. Multiple regression models were estimated using the HSF data and the CD data, weighted by CD population size. The results are summarized in Table 5.5. 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 with the exceptions being

Table 5.5. Comparison of individual CD level and adjusted CD regression equations.

	Individual level		CD level		Adjusted CD level	
Variable	Coefficient	SE	Coefficient	SE	Coefficient	SE
Intercept	11876	496	4854	834	1573	1021
Married	−9	274	4716	430	7770	564
Female	−6019	245	−3067	896	2195	915
Degree	8472	489	21700	1285	23501	1269
Unemp	−963	522	−391	1288	570	1327
Manual	9192	460	1457	1101	2705	1092
Man prof	20679	433	23682	1016	23037	1024
Empl other	11738	348	6383	675	7690	742
Born UK	1146	425	2691	508	2275	506
Born Aust	1874	337	2428	465	2899	492
Age 15–19	−9861	495	−482	1162	58	1141
Age 20–29	−3530	358	2027	770	1962	758
Age 45–59	586	361	434	610	1385	1589
Age 60+	255	400	1958	625	2280	1561
R^2	0.496		0.880		0.831	

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 $\sum_{i\in s}(y_i - \hat{y}_i^{(1)})^2/(n-p)$, which was 10351^2 and this can be compared with $\sum_{i\in s}(y_i - \hat{y}_i^{(2)})^2/(n-p)$, 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. [2]). There are likely to be variables with some explanatory power omitted from the model, but this reflects 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 so is the assessment of their statistical significance. Using a ten percent significance level the coefficients for married, aged 45–59 and aged 60+ were nonsignificant in the individual level equation. In the CD level equation the coefficients for unemployed, manual occupation, aged 15–19 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}.

6. An evaluation of the adjusted CD level regression method

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, aged 45–59 and aged 60+.

To assess how well these variables perform in removing aggregation effects $\hat{\Sigma}_{ww}(z)$ was calculated. The resulting adjusted aggregation effects $\hat{Q}_a(z) = \hat{\Sigma}_{aa}(z)/S_{aa}^{(1)}$ are given in column five of Table 5.1. The ratio $\hat{Q}_a(z)/Q_a$ is given in the last column of Table 5.1 and indicates that these adjustment variables remove between 9 and 75 percent of the aggregation effect. For income the reduction is 32 percent and the average reduction across the variables is 52 percent. These values tell us the effect of the adjustment for each variable separately. A CGV analyses based on $(S_{ww}^{(1)})^{-1}\hat{\Sigma}_{ww}(z)$ 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 aged 45–59 and aged 60+ were not included in this CGV analysis. The reduction in $\overline{\theta}$ 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 remaining aggregation effects. An analysis of the CGVs based on $(S_{xx}^{(1)})^{-1}\hat{\Sigma}_{xx}(z)$ 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 effects. 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 5.2, 5.3, and 5.4 show the adjusted covariances and associated correlation and regression 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 and in some cases there is an over-adjustment. There is over-adjustment for the correlations of income with the auxiliary variables and no improvement for the correlations involving the explanatory variables. Similar results apply for the bivariate regressions coefficients. The relationships between the explanatory and auxiliary variables is a factor in the ecological bias. Analysis of the relationships between all the variables in the analysis showed that for the covariances there are 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 and bivariate regression coefficients. Those involving the auxiliary variables are improved considerably, but those between income and the explanatory variables and those between the explanatory variables experience some small improvement, but not enough to give reasonable estimates of the individual level coefficients.

The estimates of the regression equation obtained from $\widehat{\Sigma}_{ww}(z)$ are given in Table 5.5. In general the adjusted CD regression coefficients are no closer to the individual level coefficients than the original CD level regression coefficients. The resulting adjustment of R^2 is still considerably higher 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 the adjustment has had a small detrimental effect. The average absolute difference between the CD and individual level coefficients has 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 CD regression coefficients into elements that can be ascribed to the adjustment variables and those that cannot. Equation (4.6) 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 6.1 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 6.2 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 6.3. This shows that with

Table 6.1. Comparison of individual and CD level regression equations including auxiliary variables.

	Individual level		CD level	
Variable	Coefficient	SE	Coefficient	SE
Intercept	11773	598	3873	827
Married	−178	283	7892	572
Female	−6015	245	−2170	916
Degree	8419	489	23456	1270
Unemp	−907	522	606	1329
Manual	9167	460	2753	1093
Man prof	20503	436	23017	1024
Emp other	11583	350	7735	743
Born UK	1199	425	2273	507
Born Aust	1887	337	2928	493
Age 15–19	−10000	496	73	1141
Age 20–29	−3558	359	1942	759
Age 45–59	552	361	1641	624
Age 60+	224	402	2250	652
Owner occupied	66	349	−1220	517
Renting Govt	−1501	495	22	460
Housing type	422	443	−1798	379
R^2	0.497		0.886	

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 $\beta_{yx|z}$. The adjustment has adjusted the component of bias it is designed to reduce. The remaining biases mean that the overall effect is largely the same. 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 MLwiN 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 suggest that to eliminate 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. Indicators that there might be difficulties in obtaining a good adjustment are the aggregation effects remaining after the adjustment, which are given in Table 5.1. While many have been reduced considerably, several are still quite large.

Table 6.2. Comparison of individual, CD and adjusted CD level estimate of β_{3x}.

	Owner occupied			Renting government			Housing type		
Variable	Ind'	CD	Adj CD	Ind'	CD	Adj CD	Ind'	CD	Adj CD
Intercept	0.497	−0.365	0.586	0.178	−0.036	0.103	0.801	−0.298	0.883
Married	0.235	1.106	0.066	−0.064	−0.094	−0.005	0.135	1.015	0.023
Female	0.009	−0.239	0.004	0.002	0.899	0.013	−0.003	0.672	0.011
Degree	0.007	0.585	0.016	−0.035	−0.332	−0.007	−0.001	0.625	0.014
Unemp	−0.088	−0.452	0.010	0.032	1.495	0.021	−0.003	0.879	0.014
Manual	0.027	0.739	0.029	−0.013	−0.608	−0.011	0.009	0.214	0.007
Man prof	0.111	0.284	−0.003	−0.105	−0.896	−0.013	0.029	−0.573	−0.009
Empl other	0.115	1.075	0.029	−0.090	−1.318	−0.022	0.029	0.006	0.005
Born UK	−0.026	−0.203	0.003	0.038	0.339	0.006	0.013	−0.090	−0.003
Born Aust	0.008	−0.057	0.016	0.012	0.586	0.008	0.010	0.324	0.006
Age 15–19	0.131	0.399	0.009	−0.066	−0.505	−0.008	0.074	0.032	0.002
Age 20–29	−0.062	−0.059	−0.015	−0.018	−0.188	−0.003	0.012	−0.013	−0.001
Age 45–59	0.071	0.397	0.154	−0.017	0.156	−0.020	0.006	0.411	0.038
Age 60+	0.074	0.516	0.072	−0.033	−0.895	0.006	−0.055	−0.199	−0.065
R^2	0.104	0.746	0.053	0.043	0.637	0.013	0.054	0.547	0.020

Table 6.3. Comparison of individual, CD and adjusted CD level estimate of $\beta_{zx}\beta_{yz|x}$.

	Level		
Variable	Individual	CD	Adjusted CD
Intercept	103	981	−2300
Married	169	−3177	−122
Female	−4	−897	−24
Degree	52	−1846	−45
Unemp	55	−997	−37
Manual	26	−1296	−49
Man prof	176	665	20
Empl other	155	−1351	−45
Born UK	−53	418	1
Born Aust	−14	−500	−30
Age 15–19	140	−555	−15
Age 20–29	28	85	20
Age 45–59	34	−1207	−256
Age 60+	32	−292	29

7. Discussion

The multi-level model which incorporates grouping variables and random effects provides a general framework through which the causes of ecological biases can be explained.

In the example considered, 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 enables the aggregation effects due to these variables to 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 were associated with indicators of higher socio-economic status. We could 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. However, there are always likely to be some residual group level effects and so we need methods that can satisfactorily account for them.

The 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 reflect 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 suggest that a significant amount will always remain. Hence, methods to account for the remaining homogeneity due to group level effects need to be developed.

Acknowledgment

This work was supported by the Australian Research Council.

References

[1] H. M. Blalock, *Causal Inference in Nonexperimental Research*, University of North Carolina Press, North Carolina, 1964.

[2] H. Davies, H. Joshi, and L. Clarke, *Is it cash that the deprived are short of?*, Journal of the Royal Statistical Society. Series A. Statistics in Society **160** (1997), no. 1, 107–126.

[3] H. Goldstein, *Multilevel Models in Educational and Sociological Research*, Griffin, London, 1995.

[4] D Holt, D. G. Steel, and M. Tranmer, *Area homogeneity and the modifiable areal unit problem*, Geographical Systems **3** (1996), 181–200.

[5] K. Smith, *Another look at the clustering perspective on aggregation problems*, Sociological Methods & Research **5** (1977), 289–310.

[6] T. M. F. Smith, *The Pearson adjustment for multivariate normal model*, Analysis of Complex Surveys (C. J. Skinner, D. Holt, and T. M. F. Smith, eds.), chapter 6, John Wiley & Sons, London, 1989.

[7] D. G. Steel, *Statistical analysis of populations with group structure*, Ph.D. thesis, Department of Social Statistics, University of Southampton, Southampton, 1985.

[8] D. G. Steel and D. Holt, *Analysing and adjusting aggregation effects: The ecological fallacy revisited*, International Statistical Review **64** (1996), no. 1, 39–60.

[9] ———, *Rules for random aggregation*, Environment and Planning A **28** (1996), no. 6, 957–978.

[10] D. G. Steel, M. Tranmer, and D. Holt, *Analysis combining survey and geographically aggregated data*, Analysis of Survey Data (Southampton, 1999) (R. L. Chambers and C. J. Skinner, eds.), chapter 20, Wiley Ser. Surv. Methodol., John Wiley & Sons, Chichester, 2003, pp. 323–343.

D. G. Steel: School of Mathematics and Applied Statistics, University of Wollongong,
Wollongong NSW 2522, Australia
E-mail address: dsteel@uow.edu.au

M. Tranmer: Cathie Marsh Centre for Census and Survey Research, University of Manchester,
Manchester, M13 9PL, UK
E-mail address: mark.tranmer@manchester.ac.uk

D. Holt: Department of Social Statistics, University of Southampton, Southampton SO17 1BJ, UK
E-mail address: tholt@socsci.soton.ac.uk