GEOGRAPHIC ESTIMATES OF UNDERENUMERATION

David Steel and Jennifer Poulton, Australian Bureau of Statistics
David Steel, PO Box 10, Belconnen, ACT 2616, Australia

Introduction

The Australian population census provides the basic information from which estimates are made of the population of the nation, each of the six States and two mainland Territories, and sub-State local government areas. These population estimates are required for the determination of the number of seats each State will have in the Federal House of Representatives, the allocation of funds to each State, and the funding of local government authorities (LGAs). Population estimates are also used in their own right as indicators of population growth and distribution and as denominators for various demographic, social and economic indicators.

In Australia, population estimates have been obtained from census counts, incorporating an adjustment for underenumeration in 1976, 1981 and 1986. The adjustments are based on the results of a Post Enumeration Survey (PES) and demographic analysis.

Data for the assessment of the level of under-enumeration are primarily derived from a census PES. Results of the PES are validated by comparing them with estimates based on demographic statistics and other independent data such as statistics on school enrolment on children whose parents receive government family allowances, and on persons registered with the government Medicare insurance system. In Australia, school enrolments for children aged 6-15 years are compulsory and, until means-testing was introduced in November 1987, family allowances had been universally paid to mothers of all children of ages less than 17. Medicare insurance is also compulsory and universal for all residents. These independent statistics are helpful as a check of the PES results and demographic estimates.

Although population estimates include an adjustment for under-enumeration, no adjustment is made for other census data. Census counts are published without adjustment.

In its five yearly population census, the Australian Bureau of Statistics (ABS) employs census collectors for the delivery of forms to each household and for the collection of completed forms from each household. The census is conducted on the basis of enumerating people where they are located on census night. This collector-based field system allows the census collection phase to be completed within two weeks of the census date and allows the census PES to be conducted reasonably close to the census date - in 1986 within 4-5 weeks of census night.

The 1986 PES involved interviews with a sample of the population from about 35,000 private dwellings (2/3 of one percent of dwellings) across Australia involving about 100,000 persons. The sampling fraction varied between States and Territories, with the smaller States and Territories having higher sampling fractions. Personal data on name, age, sex, marital status and birthplace were obtained by interviewers for matching with information on the census form. For each person in the survey, information was sought on their place of usual residence, where they spent census night, their address before and after census night and any other address where they might have been included on a census form. At each given address, the personal information was matched to census forms to establish whether a person was missed, counted once or the number of times counted if counted more than once. Conducting the PES close to the census date minimises recall error and also reduces the number of exclusions due to deaths and overseas travel.

Further details on the adjustment of the 1986 Census and the PES are given in Choi et al (1988). The general consistency of the PES results and other data sources for the last three censuses has given some confidence in the basic reliability of the PES.

The PES is used to obtain estimates of underenumeration for age, sex categories at the National level with reasonable standard errors. Population estimates at the Part of State (Capital City/rest of State) and Territory level by age and sex, and at the local government area level were not derived directly from the PES. Sampling errors at the Part of State/Territory level by age and sex are high, many unacceptably high, relative to the amounts of adjustment for underenumeration which need to be made. An alternative indirect method, using an iterative proportional fitting (IPF) procedure, was used to produce Part of State/Territory estimates by age and sex from those higher level PES estimates with a low sampling error. This procedure involved taking the national population estimates by age and sex and the Part of State/Territory estimates within each sex and adjusting the census age by Part of State/ Territory counts to these two margins.

For estimates for local government areas, the problem with high sampling error is more acute and results of the PES are not sufficiently reliable to make direct estimates of underenumeration for each local government area.

2. Methods Used to Estimate LGA Underenumeration Rates

In 1976 it was decided to group LGAs and use the PES estimate of the underenumeration rate for the group which was then to be applied to all LGAs within that group. Initially the groups were to be formed so that the group underenumeration rate had a relative standard error of less than 25% and the difference between any two LGA underenumeration rates within the group was 30% or less. In the event considerations such as geography and statistical significance of differences within and between groups were also used as well as a degree of subjectivity.

The method used in 1976 was not considered very satisfactory because it was felt that it had a considerable degree of arbitrariness about it. In particular the procedure had discontinuity associated with it, in that geographically adjacent LGAs of apparently similar nature had different underenumeration rates assigned.

In 1981 a synthetic estimation approach was adopted in which the Part of State or Territory level underenumeration rates for age, sex categories were applied to the LGA counts. (The Part of State level estimates themselves were obtained by IPF.) In 1986 a similar procedure was adopted with the addition of a basic birthplace split (Australian born/foreign born) in the synthetic estimation procedures.

The approach taken by the ABS so far is to adjust for a few major demographic variables and geographic effects for which there is clear evidence of differences in underenumeration rates.

3. Investigations of Alternative Approaches to Estimating LGA Underenumeration Rates

The approach the ABS is using is deliberately conservative in using a small number of key demographic variables in the synthetic estimation procedures. However, the ABS is investigating a number of other approaches to producing estimates of LGA underenumeration rates and population estimates some of which are discussed in the remainder of this paper.

3.1 Person Level Underenumeration Rates

The synthetic estimators used in 1981 and 1986 employ estimates of the probability of people being underenumerated within age, sex, birthplace categories which is assumed to be the same for all LGAs within a Part of State. That is the population estimate for LGA k is

$$\hat{N}_k = \sum_a \frac{N_{ka}}{1-P_a}$$

where N_{ka} is the census count in age, sex, birthplace cell a and P_a is the estimate of the probability, calculated at the Part of State level, that a person in such a cell is underenumerated. In theory the number of variables used to define the cells in the adjustment could be increased. However, in practice this could cause some problems, since the estimates of these probabilities at the Part of State level would have high standard errors. Alternatively the Part of State population estimates and associated underenumeration rates could be based on synthetic estimates. (This was in fact done in 1986 where the Part of State by age by sex by birthplace population estimates were obtained by forcing census counts to add to PES estimate of the National age sex margin and birthplace Part of State margin using IPF.)

An alternative being investigated is to obtain estimates of the probability a person is underenumerated by constructing a person level model based on the PES.

If the resulting model involves variables x_1, ... x_p and b indicates the cells of the p-way cross tabulation defined by these variables and P(b) = Prob (unit in cell b is missed) obtained from the person level model then

$$\hat{N}_k = \sum_b \frac{N_{kb}}{1-P(b)}$$

Table 1 gives the result of fitting a logistic regression model to the 1986 PES using a forward selection process. The model includes all the variables used in the analysis except sex. Using this model would only necessitate production of 6 way tables instead of the 4 way tables used at present.

This approach needs to be extended to account for people who were falsely included in the census or double counted.

3.2 LGA Regression Models

The idea of using a regression analysis at the small area level has been suggested by Ericksen and Kadane (1985) and criticised by Freedman and Navidi (1986). Isaki et al (1987) report on progress in using regression methods at the US Bureau of the Census. Ericksen and Kadane (1987) investigated some aspects of the sensitivity of the methods they propose to variations in the assumptions on which the methods are based.

We have conducted a regression analysis using the 157 LGAs in the major cities of Sydney, Melbourne, Adelaide and Perth which account for approximately half of the Australian population. (The other two State capitals of Brisbane and Hobart were excluded from the analysis because of an unusual geographic-administrative set up and the small number of LGAs respectively.) The dependent variable was the LGA underenumeration rate estimated directly from the PES. We let y_i denote the estimated underenumeration rate in the i^{th} LGA which is based on a sample of size n_i and Y_i denotes the actual underenumeration rate. The potential explanatory variables included in the regression analysis are given in Appendix 1. Unweighted linear regression models were fitted and weighted regression models with weights n_i.

Weighted regression has the intuitively appealing property of reducing the influence of LGAs with small sample size but, as will be discussed below, can also be justified as close to the appropriate weighting. Sample sizes varied between 45 and 1288 with an average of 312. The estimated LGA under-enumeration rates vary from -0.013 to 0.104.

An analysis of covariance was performed to see if there were any significant differences between the cities and it was found that a common regression could be fitted across the cities for both weighted and unweighted analyses.

A number of models were tried but attention focussed on the four models given in Table 2.

Model 1 was obtained from a stepwise selection procedure. Model 2 was chosen from a range of models using basic demographic and socio-economic variables. Models 2 and 3 include the variables used in the synthetic estimation procedures in 1986 and 1981 respectively. $\bar{R}^2_y$ is the adjusted R^2 and S_y is the estimated residual standard error. These results show that the age, sex, birthplace model can be significantly improved upon. Furthermore, inclusion of marital status and the proportion of persons residing at the same address in 1981 results in the birthplace variables becoming non-significant. Adding the labour force participation rate and proportion of persons with low income did not improve model 2.

The resulting estimated models from the unweighted analysis are as follows.

Model 1:

$$\begin{array}{llll} y = & -0.20 & + 0.45\ \text{MALE} & + 0.32\ \text{CELSE} \\ & (0.05) & (0.10) & (0.08) \\ & & + 0.05\ \text{SEMDET} + 0.26\ \text{(OTHER)} & - 0.09\ \text{ONEFAM} \\ & & (0.02) \qquad\qquad (0.18) & (0.02) \\ & & - 1.21\ \text{ABOR} + 0.01\ \text{OUTER} & + 0.07\ \text{PRIV} \\ & & (0.30) \qquad\qquad (0.004) & (0.03) \end{array}$$

Model 2:

$$\begin{array}{lllll} y = & -0.19 & + 0.36\ \text{MALE} & - 0.15\ \text{AGE1} & - 0.23\ \text{AGE2} \\ & (0.11) & (0.11) & (0.05) & (0.08) \\ & + 0.15\ \text{MST1} & + 0.29\ \text{MST2} & - 0.12\ \text{RES81} & \\ & (0.12) & (0.13) & (0.03) & \end{array}$$

Model 3:

$$\begin{array}{lllll} y = & -0.20 & + 0.45\ \text{MALE} & - 0.14\ \text{AGE1} & - 0.04\ \text{AGE2} \\ & (0.07) & (0.13) & (0.04) & (0.06) \\ & + 0.04\ \text{BPL1} & + 0.50\ \text{BPL2} & - 0.04\ \text{BPL3} & \\ & (0.03) & (0.23) & (0.04) & \end{array}$$

Model 4:

$$\begin{array}{lllll} y = & -0.13 & + 0.45\ \text{MALE} & - 0.17\ \text{AGE1} & + 0.02\ \text{AGE2} \\ & (0.05 & (0.12) & (0.04) & (0.06) \end{array}$$

The figures in brackets are the estimated standard errors on the parameter estimates but do not include any adjustment for the clustered design of the PES in which an average of 8 dwellings are selected from each selected Collection District.

All the models reflect the strong influence of the proportion of males. Model 1 includes the variable reflecting the proportion of people included in the census in an LGA different from the LGA of usual residence and probably reflects the higher underenumeration rate of people away from their place of usual residence at census night. Most of the remaining explanatory variables in Model 1 reflect areas with a lower than usual proportion of separate houses, or flats, occupied by one family. It is interesting to note that model 1 only includes the geographic variable for LGAs in the outer areas and then with a positive coefficient, despite the fact that there is a tendancy of LGAs in the inner and high density areas to have higher underenumeration rates. This feature is presumably due to the other explanatory variables accounting for these observed differences between area types.

Model 2 includes the basic demographic variables of age, sex and marital status and the variable indicating areas with relatively stable populations. An area with a relatively lower proportion of people in the same address as 1981 may be one with a mobile population or a newly developed area. Adding the variable indicating the proportion of people counted elsewhere did not improve this model.

Despite the lower adjusted R^2 value we will concentrate on the evaluation on model 2, mainly because it is a relatively simple model which can be relatively easily interpreted. We feel there is a danger in Model 1, because of its being the result of a stepwise selection, including variables by chance and it is not as easy to interpret. The predicted values from Model 2 differ from those from Model 1 by less than .01 in over 90% of cases with an average absolute difference of 0.0048.

For ease of discussion we will concentrate on the results obtained for Melbourne. Figures 1 and 2 show the estimated LGA underenumeration rates for Melbourne obtained from the regression Models 1 and 2 respectively. The main feature to note is that the range of the predicted values is less than the original estimates and an associated tendency for small rates to be made larger and large rates to be made smaller, as should be expected from a regression approach.

The model behind the regression approach needs to recognise the variance introduced by the estimation of $\underset{\sim}{Y}$ by $\underset{\sim}{y}$ based on the PES. A simple formulation is

$$\underset{\sim}{y} = \underset{\sim}{Y} + \underset{\sim}{e} \qquad (1)$$

where $E[\underset{\sim}{e}|\underset{\sim}{Y}] = 0$, $V(\underset{\sim}{e}|\underset{\sim}{Y}) = \Delta = \text{diag}(\delta_i^2)$ and δ_i^2 is the sampling variance of y_i, and

$$\underset{\sim}{Y} \sim N(X\beta, \sigma^2 I) \qquad (2)$$

where X is the matrix of the values of the p explanatory variables.

In assessing the adequacy of the fit of any of the regression models it must be remembered that the sampling errors will severely restrict the ability of any model to achieve a good fit. It is easy to show that if

$$S_y^2 = \frac{1}{n-1}\ \underset{\sim}{y}'(I - P_X)\underset{\sim}{y}$$

is the usual estimate of variance obtained from a regression, where $P_X = X(X'X)^{-1}X'$, then under the model given by (1) and (2)

$$E[S_y^2] = \sigma^2 + \frac{1}{n-p}\ \text{tr}[(I-P_X)\Delta]$$

If Δ is known or, as in practice, estimated unbiasedly from the PES, then

$$S_Y^2 = S_y^2 - \frac{1}{n-p}\ \text{tr}[(I-P_X)\Delta] \qquad (3)$$

is unbiased for σ^2. Estimates of S_Y obtained in this way are given in Table 2. Similarly it is easy to show

$$E[\underset{\sim}{y}'(I-P_0)\underset{\sim}{y}] = \underset{\sim}{Y}'(I-P_0)\underset{\sim}{Y} + tr[(I-P_0)\Delta]$$

where $P_0 = \underset{\sim}{1}\ (\underset{\sim}{1}'\underset{\sim}{1})^{-1}\ \underset{\sim}{1}'$. Using this result it is possible to correct for the sampling error and get an estimate $\bar{R}_Y^2$ of the adjusted R^2 associated with a regression of $\underset{\sim}{Y}$ on X.

These estimates are given in Table 2 and show that models that appear to fit only moderately well may correspond to good fits in terms of actual underenumeration rates.

If we think of Y_i as the result of some superpopulation model, as is done in (2), the assumption $V(\underset{\sim}{Y}) = \sigma^2 I$ may not be realistic. A more realistic assumption would be $V(Y_i)\ \alpha\ \frac{\mu_i}{N_i}$ where $\mu_i = E(Y_i)$ and N_i is the population size in the i^{th} LGA. Moreover if any design effect associated with the PES is approximately constant across SLAs δ_i^2 is approximately proportional to $\frac{Y_i}{n_i}$. Now

$$\begin{aligned} V(y_i) &= E[V(y_i|Y_i)] + V(E[y_i|Y_i]) \\ &= a\ E[\frac{Y_i}{n_i}] + V(Y_i) \quad \text{for some constant a} \\ &= a\ \frac{\mu_i}{n_i} + b\ \frac{\mu_i}{N_i} \quad \text{for some constant b} \end{aligned}$$

The PES uses equal probabilities of selection (within States) so $E[n_i] = f\ N_i$ where f is the sampling fraction. Hence

$$V(y_i)\ \alpha\ \frac{\mu_i}{n_i}$$

which gives some justification in weighting by n_i in the regression analysis. Weighted regression analysis produced very similar models with slightly better $\bar{R}y^2$ values.

Having estimated β we could consider estimating $\underset{\sim}{Y}$ by the fitted values $X\hat{\beta}$. Ericksen and Kadane (1985) adopt a hierarchical Bayesian approach in which (1) and (2) are assumed to hold and it is also assumed $\beta \sim N\ (\gamma, \Omega)$. The methods described by Lindley and Smith (1972) can then be applied.

Ericksen and Kadane (1985) and Isaki et al (1987) both concentrate on the case where $\Omega^{-1} = 0$ in which case the posterior mean of $\underset{\sim}{Y}$ is

$$\hat{Y}_1 = [\ \Delta^{-1} + \sigma^{-2}\ (I-P_X)]^{-1}\ \underset{\sim}{y}$$

which can be shown to be equal to

$$\hat{Y}_2 = [\Delta^{-1} + \sigma^{-2}\ I]^{-1}\ [\Delta^{-1}\ y + \sigma^{-1}\ X\hat{\beta}]$$

where $\hat{\beta} = X(X'\ \Sigma^{-1}\ X)^{-1}\ X'\Sigma^{-1}\underset{\sim}{y}$ is the GLS estimate of β calculated assuming $\Sigma = \Delta + 6^2\ I = V(\underset{\sim}{y})$ is known. In practice Δ and σ^2 must be estimated. These results assume the sampling errors have a Normal distribution.

As noted by Freedman and Navidi (1986) $\hat{Y}_2$ is also the minimum variance unbiased linear estimate of $\underset{\sim}{Y}$.

The estimate $\hat{Y}_2$ can also be written as

$$\hat{Y}_3 = \sigma^2\ \Sigma^{-1}\ \underset{\sim}{y} + \Delta\ \Sigma^{-1}\ X\ \hat{\beta}$$

this form shows clearly how $\underset{\sim}{\hat{Y}}_2$ is a weighted mean of $\underset{\sim}{y}$ and $X\hat{\beta}$. If for some reason β is estimated by something other than $\hat{\beta}$ (eg the OLS estimate) $\hat{Y}_1$ and $\hat{Y}_2$ are not identical although $\hat{Y}_2$ and $\hat{Y}_3$ will be.

The sample size in some LGAs is such that the estimated sampling variances are very unstable. For this reason it was decided to estimate Δ by "smoothed" variances obtained by estimating an average design effect in each city and applying these design effects to variances calculated from assumptions of simple random sampling. Problems still arose because of small estimated under-enumeration rates giving unrealistically small estimates of variance. To overcome this problem sampling variances were calculated initially using the underenumeration rate estimated from the appropriate model. σ^2 was estimated by S_Y^2 given by (3).

For model 2 the average weight given to the regression estimate is about 0.63 in Melbourne, but can be a lot higher for LGAs with small sample size and/or large under-enumeration rate and hence large sampling variance. Clearly LGAs with large sample sizes will have smaller weight given to the regression estimate. Figure 3 shows the plot of the Bayesian estimates associated with model 2 against the direct survey estimate. The range of the estimates is virtually identical to the estimates obtained from the regression model.

3.3 Incorporating Geographic Effects in Regression Models

There are two main ways of incorporating geographic effects:

(1) introduce geographic variables into the explanatory variables

(2) introducing geographic effects in the regression residuals.

We have already tested for differences between cities using ANOCOVA methods, however differences between different areas within cities might still exist. To examine this possibility LGAs were classified as Inner, High Density, Suburban, Outer and regression models incorporating terms reflecting these area types were fitted. A model with only these variables gave on $\bar{R}_y^2$ of 0.17. When these variables were included in a stepwise selection procedure only the variable indicating LGAs in the outer areas was chosen (see model 1). A problem with this approach is that the allocation of LGAs into area types has a degree of subjectivity. Such a procedure is open to criticisms similar to those made of the 1976 procedure.

The assumption that $V(\underset{\sim}{Y}) = \sigma^2 I$ assumes there is no relationship between errors in the regression model for geographically close LGAs. One way of incorporating geographic effects in the error in the regression model (2) is to allow for spatial autocorrelation in the residuals of the model. Hence we can replace (2) by the assumption

$$\underset{\sim}{Y} \sim N(X\beta,\ V)$$

where V is the variance matrix associated with the spatial autocorrelation. Cliff and Ord (1981) discuss two forms of spatial autocorrelation models, a simultaneous and a conditional which lead to the following forms of V

$$\text{Conditional}\quad V = 6^2\ (I-\rho W)^{-1}$$

$$\text{Simultaneous}\quad V = 6^2\ (I-\rho W')^{-1}\ (I-\rho W)^{-1}$$

W is a weight matrix, assumed known, reflecting the spatial connection of the LGAs and ρ is the parameter of the autocorrelation process.

The hierarchical Bayes approach can be applied to this case also, giving

$$\hat{Y}_1 = [\Delta^{-1} + V^{-1}(I-\tilde{P}_X)]^{-1}\ \Delta^{-1}\underset{\sim}{y}$$

where $\tilde{P}_X = X(X'\ V^{-1}\ X)^{-1}\ \ X'\ V^{-1}$

It can be shown that $\hat{Y}_2$ is equal to

$$\hat{Y}_2 = [\ \Delta^{-1} + V^{-1}]^{-1}\ [\Delta^{-1}\ \underset{\sim}{y} + V^{-1}\ X\ \hat{\beta}]$$

where $\hat{\beta} = (X^{-1}\ \Sigma^{-1}\ X)^{-1}\ X^{1}\ \Sigma^{-1}\underset{\sim}{y}$ assuming $\Sigma = \Delta + V$ is known. Again it is possible to show that $\hat{Y}_2$ is the minimum variance unbiased linear estimate of $\underset{\sim}{Y}$.

The estimate $\hat{Y}_2$ can also be written as

$$\hat{Y}_3 = V\ \Sigma^{-1}\ \underset{\sim}{y}\ +\ \Delta\ \Sigma^{-1}\ X\ \hat{\underset{\sim}{\beta}}$$

If β is not estimated by $\hat{\beta}$, $\hat{Y}_1$ and $\hat{Y}_2$ are not the same although $\hat{Y}_2$ and $\hat{Y}_3$ are.

To apply this method would require the estimation of Δ, σ^2 and ρ. Before developing a procedure to do this it is worthwhile seeing if there is any evidence of spatial autocorrelation. To do this we can apply the methods described by Cliff and Ord (1981, pp 200-203) modified for the unequal sampling variances.

For a regression model the test statistic proposed by Cliff and Ord is

$$I = \frac{m}{S_0}\ \frac{\hat{\underset{\sim}{e}}'\ W\ \hat{\underset{\sim}{e}}}{\hat{\underset{\sim}{e}}'\ \hat{\underset{\sim}{e}}}$$

where $S_0 = \sum_{i \neq j} w_{ij}$, $\hat{\underset{\sim}{e}}$ is the vector of residuals from an OLS fit, and m is the number of LGAs. The properties of this test statistic are derived by Cliff and Ord based on the model $y \sim N(X\ \beta,\ \sigma^2)$ under the null hypothesis of no spatial autocorrelation. In our problem $\underset{\sim}{y} \sim N\ (X\beta,\ \Delta + V)$ and we assume that $V(\underset{\sim}{y}_i)$ is approximately proportional to n_i^{-1}.

With this assumption the theory of Cliff and Ord can be applied but with all the formulas involving weighting by n_i.

In our application W was constructed by setting $w_{ij} = 1$ if the i^{th} and j^{th} LGAs touched at any point and rescaling so $W\underset{\sim}{1} = \underset{\sim}{1}$. The test statistic was calculated for each city for a purely spatial model where $\underset{\sim}{X} = 1$ and for the regression model 2. The results are given in table 3.

While there is some evidence of spatial autocorrelation in the LGA rates in Melbourne and perhaps Sydney, the residuals from the regression model exhibit no spatial autocorrelation. Hence this approach only appears worthwile pursuing if we are considering using the city mean as the model estimate.

3.4 Spatial Smoothing

The regression based procedures discussed in 3.3 result in the estimate for a particular LGA being a weighted average of the PES estimates of all the LGAs in the analysis. This of course is what regression does. An implied criticism in the Freedman Navidi (1986) paper is this fact.

An objection to the approach used in 1976 was the discontinuities inherent in it. Even a regression approach can give geographically contiguous LGAs very different underenumeration rates unless the predictor variables have a reasonably smooth geographic distribution.

There is no reason to suppose underenumeration is smoothly distributed. However, the procedures adopted in 1981 and 1986 of applying Part of State underenumeration rates as estimated from the PES amount to using crude geographic averages within the relevant categories. A less extreme approach would be to take some sort of spatial moving averages.

Consider the i^{th} LGA and assume that the underenumeration rates of the LGAs in the "neighbourhood" of this LGA are generated from a superpopulation with a constant mean. Hence if we let ${}_i\underset{\sim}{Y}$ be the vector of LGA underenumeration rates within the neighbourhood (including Y_i) we are assuming

$${}_i\underset{\sim}{y} = {}_i\underset{\sim}{Y} + {}_i\underset{\sim}{e}$$

$${}_i\underset{\sim}{Y} \sim N\,({}_i\mu\underset{\sim}{1},\ {}_i\sigma^2\ I)$$

where ${}_i\underset{\sim}{y}$, ${}_i\underset{\sim}{e}$ have been similarly defined.

For this model the hierarchical Bayes or minimum variance linear unbiased estimate of Y_i is

$$\hat{Y}_i = \frac{{}_i\sigma^2}{{}_i\sigma^2 + \delta_i^2}\, y_i + \frac{\delta_i^2}{{}_i\sigma^2 + \delta_i^2}\, {}_i\hat{\mu} \qquad (4)$$

where ${}_i\hat{\mu}$ is the weighted mean of the neighbourhood rates calculated using weights $(\delta_i^2 + {}_i\sigma^2)^{-1}$

The suggestion is to use $\hat{Y}_i$ as an estimate of Y_i.

As with all applications of the approach an important problem is the estimation of σ_i^2. A straightforward approach is to calculate the ${}_iS^2$ variance of the LGA rates in the neighbourhood and use

$${}_i\hat{\sigma}^2 = {}_iS^2 - {}_i\bar{\delta}^2$$

where ${}_i\bar{\delta}^2$ is the average of the sampling variances of the LGA underenumeration rates in the neighbourhood. Since ${}_iS^2$ may be based as only a few LGAs it is possible that $\hat{\sigma}_i^2$ is negative in which case we set it equal to 0. The variability of the rates in the neighbourhood is less than that expected due to sampling error and so setting $\hat{Y}_i = {}_i\mu$ seems reasonable.

The suggested procedure is as follows: for each LGA

1) define the neighbourhood
   - this can be based on distance or common boundaries
2) calculate ${}_i\bar{\delta}^2$ and ${}_iS^2$
3) estimate ${}_i\hat{\sigma}^2 = \max\,({}_iS^2 - {}_i\bar{\delta}^2,\ 0)$
4) calculate $\hat{Y}_i$ according to (4).

In our application we have taken the neighbourhoods to be defined in a way similar to that used in the test of spatial autocorrelation, so that the neighbourhood of a particular LGA is the set of all LGAs touching that LGA at some point and itself.

This procedure is trying to avoid the difficulties in constructing a regression model and casting the neighbourhood mean in the role of the fitted values from the regression model. The neighbourhood mean will be a badly biased estimate of Y_i if the actual underenumeration rates differ greatly within a neighbourhood. However in this case we would expect ${}_iS^2$ to be large relative to the sampling variance and would lead to most of the weight going to y_i, which to some extent puts us back where we started. A lot of weight will also be given to y_i when the sample size is reasonable. In 14 out of 53 cases in Melbourne the neighbourhood mean received a weight of less than 0.5, with the average weight being 0.70.

For the moment take the fitted value from regression model 2 as Y_i, then figure 4 gives a plot of the neighbourhood means against Y_i. The average absolute difference is 0.0061, with 9 cases the difference exceeding .01. When we use $\hat{Y}_i$ the average difference increases to 0.0088 (Figure 5). Comparing $\hat{Y}_i$ with the Bayesian estimate obtained from Model 2 gives an average absolute difference of 0.0061, with the difference exceeding .01 in 8 cases (Figure 6). The average absolute difference between the neighbourhood mean and the PES estimate is 0.0127 and is smaller than the corresponding figure (0.0136) for the fitted values from regression model 2 (Figure 7). These results suggest that methods based on the neighbourhood means will not be significantly worse than those based on regression models, although they may not necessarily be any better.

The use of the neighbourhood mean presupposes that the underenumeration rates vary within a city; if this is not the case then they are inefficient estimates of the overall city mean. An alternative is to use the city mean instead of the neighbourhood mean. As would be expected the weights given the city mean are less variable than those given the neighbourhood mean, but have an average of 0.52 compared with an average of 0.70. The Bayesian estimate utilising the city mean has an average absolute difference from the Bayesian estimate using the neighbourhood mean of .0059, with the difference exceeding .01 in 8 cases (Figure 8). The range of the Bayesian estimate using the city mean is smaller than the range of Bayesian estimate using the neighbourhood mean.

4. Concluding Comments

The results obtained so far encourage us to pursue the methods described further, with the exception of the spatial autocorrelation model. Clearly there is a need for better evaluation of the methods. So far we have only compared the estimates of underenumeration rates for different methods. We intend to also examine the resulting estimates of population and compare them with those obtained from the synthetic estimation methods currently used.

The problem remains of how to objectively evaluate the different estimates of LGA underenumeration rate or population given that we do not have reliable estimates of them. One method is to form "regions" within each city with a sample size sufficient to give reasonable sampling errors. The PES estimates from these regions could then be compared with the estimates obtained by summing the LGA estimates over the LGA in the region for the different estimation methods being considered. This method is in the same spirit as that used by Isaki et al (1987) in which the PEP State estimates are used as the basis of the evaluation. Our main concern in using this approach is that since the regions are formed geographically the evaluation may be unduly favourable to geographically based methods such as spatial smoothing. We are considering other methods of evaluation.

References

Choi, C.Y., Steel, D.G and Skinner, T. (1988). Adjusting the 1986 Australian Census Count for Underenumeration. Paper presented at 4th USBC Annual Research Conference.

Cliff, A.D. and Ord, J.K. (1981). Spatial Processes, Models and Application, Pion Ltd, London.

Ericksen, E.P. and Kadane, J.B. (1985). Estimating the Population in a Census Year. J. Amer. Statis Assoc. V80 pp 98-109.

Ericksen, E.P. and Kadane, J.B. (1987). Sensitivity Analysis of Local Estimates of Undercount in the 1980 US Census. In "Small Area Statistics : An International Symposium" ed by R. Platek, J.N.K. Roa, C.E. Sarndel and M.P. Singh. John Wiley and Sons, NY.

Freedman, D.A. and Navidi, W.C. (1986). Regression Models for Adjusting the 1980 Census. Statistical Science Vol 1 p 3-39.

Isaki, C.T., Schultz, L.K., Smith, P.J. and Diffendal, G.J. (1987). Small Area Estimation Research for Census Undercount - Progress Report. In "Small Area Statistics : An International Symposium: ed by R. Platek, J.K.N. Roa, C.E. Sarndell and M.P. Singh. John Wiley and Sons, NY.

APPENDIX 1

Variables Used In LGA Regression Analysis

Variable Name	Definition
TOTPER	total number of persons
MALE	proportion male
AGE1	proportion aged 0-19
AGE2	proportion aged 20-29
MST1	proportion of persons aged 15 and over married or widowed
MST2	proportion of persons aged 15 and over never married
CELSE	proportion counted elsewhere, ie not at usual residence
RES81	proportion of persons counted at home, residing at the same address in 1981
BPL1	proportion born in Australia
BPL2	proportion born in New Zealand
BPL3	proportion born in UK/Ireland, Italy, Greece, Holland/Netherlands, Germany, Vietnam
QUAL	proportion of persons aged 15 and over with some qualification
UERATE	unemployment rate (of persons aged 15 and over)
PTRATE	labour force participation rate (of persons aged 15 and over)
LOWIN	proportion of persons aged 15 and over in the low income group (income < $12,000)
NPD	proportion of persons in non-private dwellings and caravan parks
SEPHSE	proportion of private dwellings that are a separate house
SEMDET	proportion of private dwellings that are semi-detached or row/terrace houses
OTHER	Proportion of private dwellings classed as caravan, houseboat, improvised dwelling, or house/flat attached to shop/office
ONEFAM	proportion of occupied private dwellings occupied by one family household
ABOR	proportion of persons Aboriginal/Torres Strait Islander
PRIV	proportion of employed labour force employed in private sector
INNER	LGAs in the CBD of the city
HIGH DENSITY	LGAs in the high density inner city area of the city
OUTER	LGAs in the outer regions of the city
RENT	proportion of private dwellings rented
PERBEDRM	average number of persons per bedroom in private dwellings
NONENG	proportion of people aged 5 years and over, born in a non-English speaking country and using languages other than English

TABLE 1 : Person Level Model for Underenumeration

SOURCE	DF	CHI-SQUARE	PROB
INTERCEPT	1	20.08	0.0001
STATEPOS (Part of State)	4	25.46	0.0001
LOC (visitor at PES dwelling)	1	26.97	0.0001
RACE (Aboriginal/Non-Aboriginal)	2	8.50	0.0143
BPL (Birthplace)	3	11.13	0.0110
MST (Marital Status)	2	86.18	0.0001
AGE (20-39, other)	1	31.53	0.0001
MST*AGE	2	33.05	0.0001
STATEPOS*LOC	4	16.08	0.0029
STATEPOS*BPL	12	30.39	0.0024
BPL*AGE	3	6.46	0.0912
STATEPOS*RACE	7	11.40	0.1222
STATEPOS*MST	8	14.05	0.0805
LOC*BPL*MST	6	19.13	0.0040
LOC*BPL*AGE	3	10.02	0.0184
LIKELIHOOD RATIO	198	207.84	0.3015

TABLE 2 : Summary of Regression Models (Unweighted)

Model	Variables Included	$\bar{R}^2_y$	s_y	C_p	$\bar{R}^2_Y$	s_Y
1	MALE, CELSE, SEMDET OTHER, ONEFAM, ABOR OUTER, PRIV	.40	.017	-3.7	.82	.0064
2	MALE, AGE1, AGE2 MST1, MST2, RES81	.33	.018	11.2	.66	.0089
3	MALE, AGE1, AGE2 BPL1, BPL2, BPL3	.20	.020	39.5	.39	.0124
4	MALE, AGE1, AGE2	.19	.020	40.0	.36	.0127

TABLE 3: Tests of Spatial Autocorrelation

	LGA Rates		Residuals from Regression Model 2	
City	I	t	I	t
Sydney	0.217	1.76	-0.129	-0.55
Melbourne	0.184	1.94	-0.0176	-0.87
Adelaide	-0.062	-0.82	-0.071	0.29
Perth	0.077	0.34	0.018	0.83

Definition of Variables in Figures

UNDENT	PES estimate of underenumeration rate
MOD1	Fitted value from regression model 2
MOD2	Fitted value from regression model 1
NGHB	Neighbourhood mean
B_MOD2	Bayesian estimate based on regression model 2
B_NGHB	Bayesian estimate based on neighbourhood means
B_CITY	Bayesian estimate based on city means

FIGURE 1 : PLOT OF MOD1 VS UNDENMT

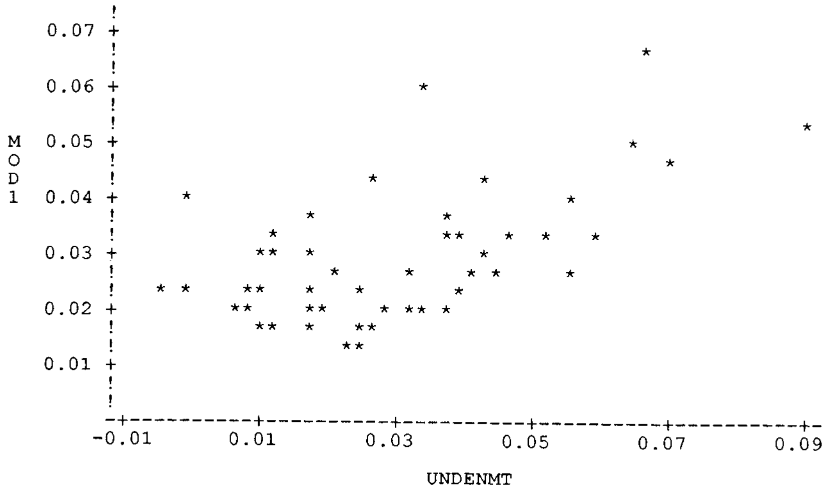


FIGURE 2 : PLOT OF MOD2 VS UNDENMT

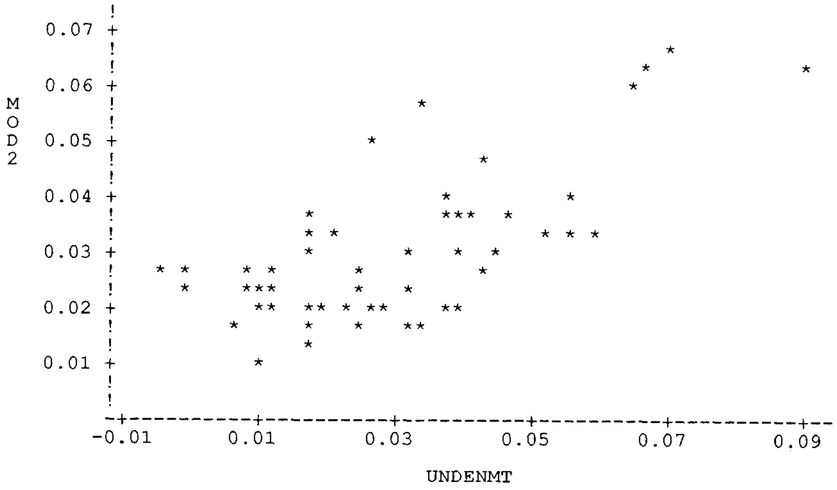


FIGURE 3 : PLOT OF B_MOD2 VS UNDENMT

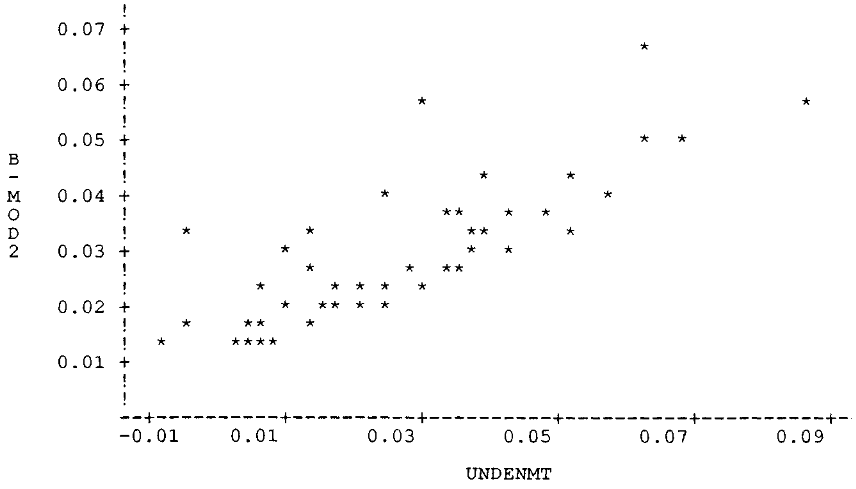


FIGURE 4 : PLOT OF NGHB VS MOD1

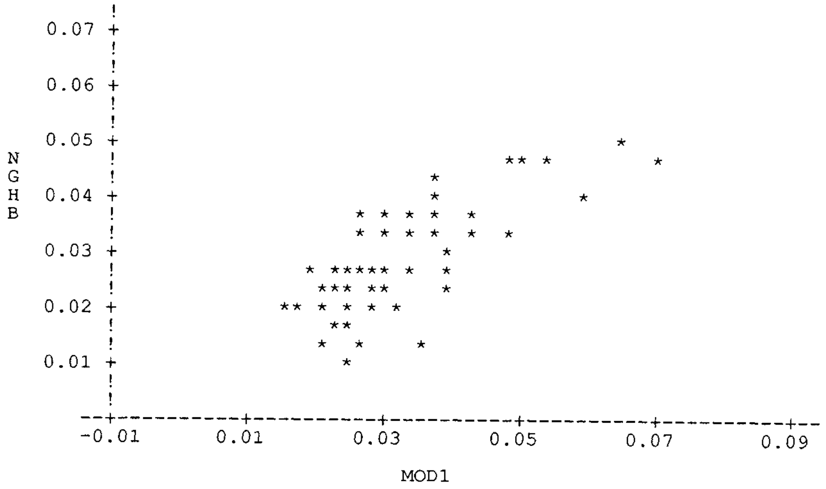


FIGURE 5 : PLOT OF B_NGHB VS MOD1

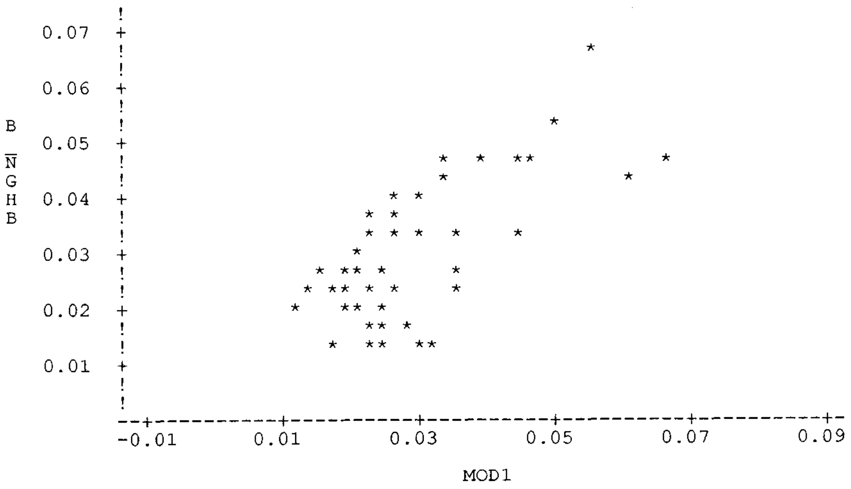


FIGURE 6 : PLOT OF B_NGHB VS B_MOD2

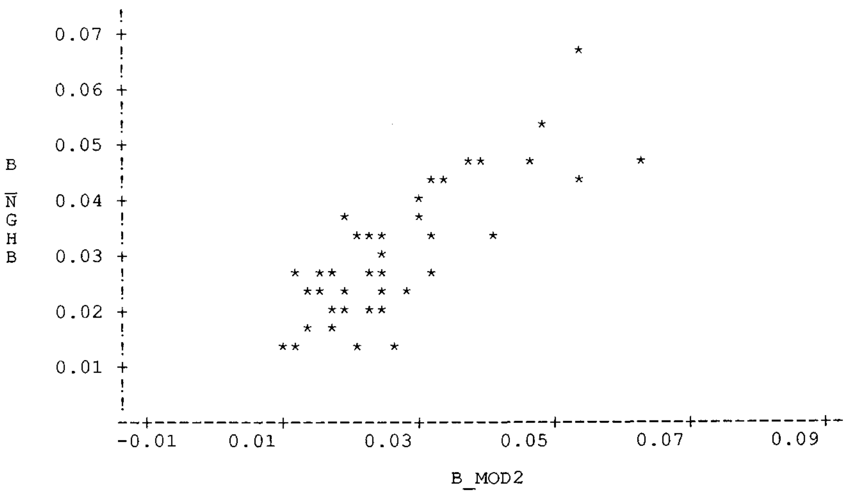


FIGURE 7 : PLOT OF NGHB VS UNDENMT

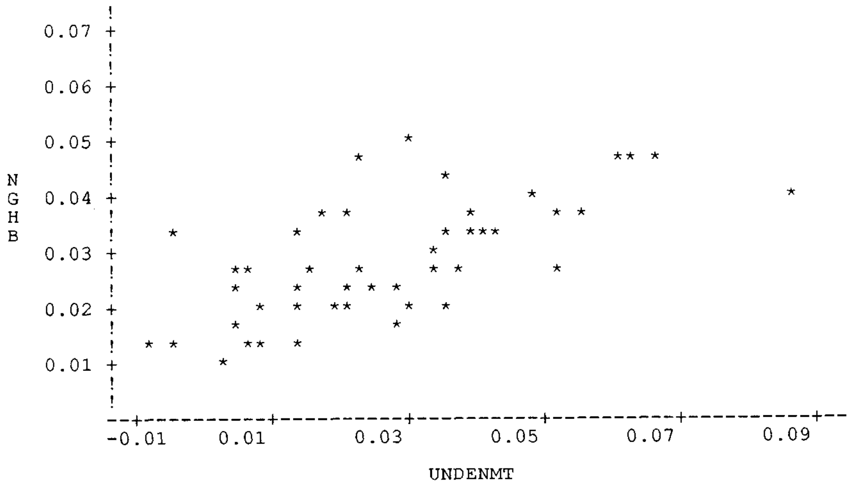
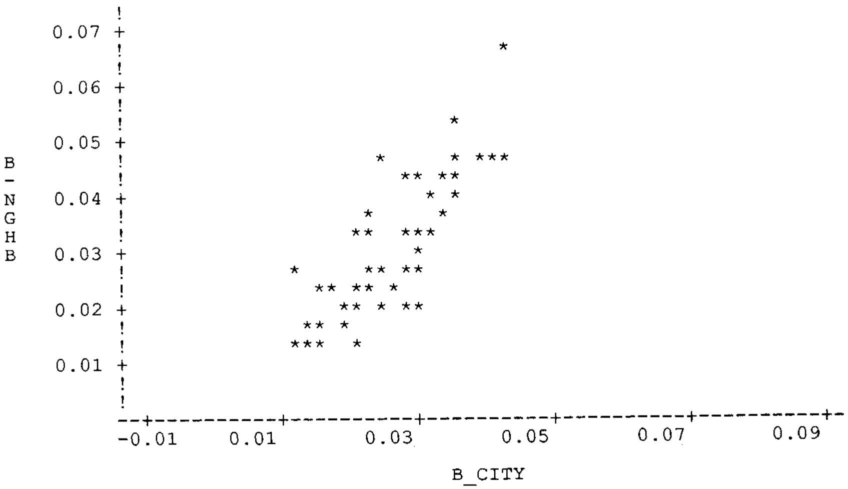


FIGURE 8 : PLOT OF B_NGHB VS B_CITY


B-NGHB
B_CITY