Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

AN INVESTIGATION INTO THE POSSIBLE USE OF MULTILEVEL MODELS BASED ON SURVEY DATA TO UPDATE CENSUS ESTIMATES FOR SMALL AREAS

P. Heady, V. Ruddock and H. Goldstein[1]

ABSTRACT

Data from the British census in 1991 are used by central government to allocate financial resources to small areas. Six years later there is concern that the characteristics of the populations of small areas may have changed and that the resulting allocation of resources is inequitable. One option is to use Structure Preserving Estimation (Purcell and Kish 1980) to update the local census statistics using trends estimated from more recent national survey data. This paper investigates the potential for estimating trends in census statistics using both a survey estimate of a national trend and survey estimates of local trends. A weighted linear combination of an estimate of the national trend and an estimate of the local trend is used to discuss the amount of survey data needed to produce accurate estimates of local trends. A multilevel model is used to estimate the variance of the local trends which is compared with the variance of the observed local trends in the census statistic between 1981 and 1991. Finally we discuss the potential of auxiliary variables to explain between area variation and improve the precision of small area estimates.

KEY WORDS: Multilevel models; Small area estimation; Census and survey data.

1. BACKGROUND

In Great Britain, as in many other countries, central government allocates some financial resources between local authorities on the basis of census statistics. However, the British census is only carried out once every ten years (in 1981, 1991, 2001 *etc.*), and during those 10 years the elevant characteristics of local populations may change. here is concern that, as the intercensal period progresses, census figures may become progressively less adequate icators of the current situation of local authorities – and , as result, financial allocations based on census istics may become progressively less fair. We therefore d to find ways of updating census statistics during the rcensal period.

How we might do so depends on the information we e available and the assumptions that we are willing to e. Simplifying slightly, it is possible to identify three d approaches.

he first approach resembles the Structure Preserving ation proposed by Purcell and Kish (1980) in that we information – provided by a survey – about how a nal distribution has changed (in this case the national rtion of individuals or households in certain categond distribute this change between local authorities by the assumption that their relative position – accor- some appropriate definition – has not changed. ntly, one can say that a standard up rating factor – on some appropriate scale – is applied to the results authority.

The second approach supplements these data with survey data for each local authority. The optimum estimate for a particular authority is then a weighted combination of the estimate that assumes a standard up rating factor, and the estimate derived from the data collected in the authority itself. The choice of optimum weighting depends on the relative magnitude of between-authority and within-authority variance – and therefore leads naturally to a formulation in terms of multilevel modelling (see Ghosh and Rao for a review of the range of methods available). In multilevel modelling terms, the variability of local trends around the national average is seen as a random effect.

In the third approach, auxiliary variables are introduced to help predict the local trends. The difference between local and national trends is no longer assumed to be purely random – but is ascribed in part to fixed effects associated with relevant covariates. These might include related variables for which administrative time series are available. They might also include local statistics taken at the time of the last census, if these define area-types in which the subsequent trends turned out to be different. If fixed effects of either kind can be identified, they allow us to reduce the amount of random variability associated with our estimates.

Important though the third approach is, in this paper we restrict our quantitative analysis to the first two. We will evaluate the performance of these approaches in estimating local trends in a particular census statistic, the proportion of households containing only one adult, between the 1981 and 1991 censuses. The criterion we will use is the estimated average weighted mean square error. This is defined as follows. For any estimator of local trends $\hat{\beta}_i^{any}$,

[1] Heady and Vera Ruddock, Methods and Quality Division, Office for National Statistics, 1 Drummond Gate, London, SW1 V2QQ, United Kingdom, vey Goldstein, Department of Mathematics, Statistics and Computing, Institute of Education, 20 Bedford Way, London, WC1H 0AL, United Kingdom.

$$\text{average weighted MSE}(\hat{\beta}_i^{\text{any}}) = \sum_i k_i \text{MSE}(\hat{\beta}_i^{\text{any}})$$

$$= \sum_i k_i \underset{s_i \in U_i}{E} \left(\hat{\beta}_i^{\text{any}} - \beta_i\right)^2 \tag{1}$$

where k_i is the weight for each local authority, proportionate to the number of households in the local authority in the 1981 census scaled so the weights for all 127 authorities sum to 1.

2. THE DATA AND SOME BASIC ESTIMATES

A scatterplot of the proportion of households containing only one adult as measured by the 1991 census versus the proportion for 1981 (Figure 1) showed that this proportion increased in all local authorities over the period 1981-1991. There was also some change in the relative position of different local authorities, although this variation in the amount of increase was small relative to the overall range of values in either census year – and relative to the overall national trend.

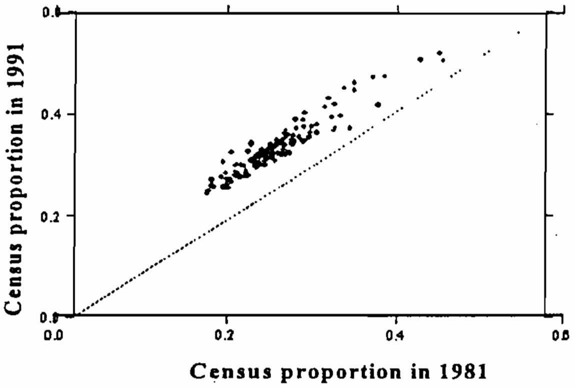

Figure 1. Relationship between the proportion of households containing only one adult in a local authority in 1981 and 1991.

Since we planned to estimate local trends using logistic models we calculated the true annual trend as one tenth of the difference in the logics of the proportions in 1981 and 1991, and calculated local trends in the same way. The national trend was 0.036, the variance was 0.000055 – equivalent to a standard deviation of 0.0074. In the rest of this paper, when we discuss trends or apply the weighted mean square error criterion, the trends in question will refer to the logics – not to the proportions themselves. We restrict our attention to the average local trend for a local authority for the whole period 1981-1991, though in reality we believe the true local trends would probably have varied during that time.

It would be convenient if we had survey data to cover the whole 1981-91 intercensal period, but unfortunately the earliest useable desegregated Labour Force Survey (LFS) data set only goes back to 1985. However we can supplement this with survey data collected after 1991, so that our full survey data set covers the period from 1985 to 1994. It simplifies the modelling process if the survey data are unclustered within each local authority – and so we have restricted this paper to the 127 urban local authorities in which the LFS sample was unclustered for all the years between 1985 and 1994[2]. For the years 1985-1991 the data were obtained from household interviews between March and May whereas for 1992-1994 the data included interviews carried out throughout the year. The mean number of households per year in each local authority was 165, but in some local authorities the achieved sample had less than 100 households per year.

We can use this survey data on its own – both to estimate both the national trend and the local trends for each of the local authorities.

We use survey data to estimate $\bar{\beta}^{\text{surv}}$ – the average trend between 1985 and 1994 – by fitting a model of the form

$$\begin{aligned} \log_e\left(\pi_{t+1985,i} \text{ or } 1 - \pi_{t+1985,i}\right) &= \alpha_i + \bar{\beta}^{\text{surv}} t \\ y_{t+1985,i} &\sim \text{Binomial}(n_{t+1985,i}, \pi_{t+1985,i}) \end{aligned} \tag{2}$$

where

$n_{t+1985,i}$ is the number of households in our survey sample in local authority i at time $t + 1985$.

$y_{t+1985,i}$ is the number of single adult households in our survey sample in local authority i at time $t + 1985$.

Since we used survey data from 1985-1994 to estimate trends over the period 1981-1991 we knew in advance that our survey based estimate of the national trend would be biassed. The survey based estimate of the national trend was
0.0384, with an estimated standard error of 0.0016 (*i.e.,* variance 0.0000026). The difference between the survey based estimate of the national trend and the national trend as measured by the census was 0.0024 – which is not statistically significant. Considering that the two estimates refer to different time periods, the two estimates are amazingly close. We conclude that there is no reason to suppose that survey and census data taken over the same period would provide different estimates of the national trend – apart from the (small) effect of sampling error on the national survey estimate.

3. COMPARING DIFFERENT ESTIMATORS OF LOCAL TRENDS

3.1 A SPREE Estimator

Since the SPREE estimator of the annual trend in every authority is effectively $\bar{\beta}^{\text{surv}}$, the MSE of the trend for local authority i is $(\bar{\beta}^{\text{surv}} - \beta_i)^2$. Averaging nationally, and relying on the fact that sampling variation is independent of the varying values of the areas themselves, we get

[2] All the analyses and figures in this paper are restricted to data from these 127 local authorities.

$$E(\text{MSE}) = \sum_i k_i E(\hat{\bar{\beta}}^{\text{surv}} - \beta_i)^2 = \sum_i k_i E(\hat{\bar{\beta}} - \bar{\beta})^2 +$$

$$\sum_i k_i (\bar{\beta} - \beta_i)^2 = 0.000055 + 0.0000026 = 0.000058.$$

$\sqrt{E(\text{MSE})}$, the corresponding indicator of expected deviation, is therefore 0.0076.

3.2 An Estimator Based on Local Survey Data Alone

At the opposite extreme from an approach which looks only at national-level data, is one which relies entirely on survey data for the local authority concerned. To estimate the local trends for each local authority, we fitted a fixed effects model in SAS (SAS Institute) which allowed the annual trend to be different in different local authorities, *i.e.*,

$$\log_e\left(\frac{\pi_{t+1985,i}}{1-\pi_{t+1985,i}}\right) = \alpha_i + t\,\beta_i^{\text{surv}} \qquad (3)$$

$$y_{t+1985,i} \sim \text{Binomial}\,(n_{t+1985,i},\ \pi_{t+1985,i})$$

where $y_{t,i}$ = number of single adult households in local authority *i* in the survey sample at time *t*;
$n_{t,i}$ = number of households in local authority *i* in the survey sample at time *t*.

The average variance of the local trend estimators $\hat{\beta}_i^{\text{surv}}$ was estimated as 0.000337, over six times the variance of the true local trends about the national trend. This is equivalent to a standard error of 0.018, giving an average coefficient of variation of about 0.5 – which is clearly unacceptable. As would be expected from such an unstable estimator, the correlation between the estimated and true values of the local trends – $\hat{\beta}_i^{\text{surv}}$ and β_i – was very small: $r = 0.23$. This is low, even allowing for the fact that the estimate $\hat{\beta}_i^{\text{surv}}$ and the actual β_i refer to different periods.

3.3 An Estimator that Combines National and Local Survey Data

We clearly do not have enough survey data to accurately estimate trends for each local authority using data from that authority alone. However we can incorporate estimates of local trends into an estimator,

$$\hat{\beta}_i^{\text{comb}} = w\hat{\bar{\beta}}^{\text{surv}} + (1-w)\hat{\beta}_i^{\text{surv}},$$

which is a weighted linear combination of the survey based estimators of the national trend and the local trends, and has a lower average mean squared error than either.

We choose the value of w which minimises the average mean squared error of $\hat{\beta}_i^{\text{comb}}$ *i.e.*, we minimise

$$M = \sum_i k_i \underset{s_i \in U_i}{E}\left((w\hat{\bar{\beta}}^{\text{surv}} + (1-w)\,\hat{\beta}_i^{\text{surv}}) - \beta_i\right)^2 \qquad (4)$$

where U_i is the population of all households within local authority i, k_i is the scaled weight and s_i is the sample of households drawn from local authority i.

If we ignore the small covariance between $\hat{\bar{\beta}}^{\text{surv}}$ and any particular $\hat{\beta}_i^{\text{surv}}$, this gives an expression which is simply a particular example of a well-known result

$$w = \frac{\sum_i k_i \underset{s_i \in U_i}{E}(\beta_i - \hat{\beta}_i^{\text{surv}})^2}{\sum_i k_i\, E(\hat{\bar{\beta}}^{\text{surv}} - \beta_i)^2 + \sum_i k_i \underset{s_i \in U_i}{E}(\beta_i - \hat{\beta}_i^{\text{surv}})^2} \qquad (5)$$

Substituting in values we already know, we obtain

$$w = \frac{0.000337}{0.000058 + 0.000337} = 0.85$$

i.e., the combined estimator of β_i with minimum average mean squared error is

$$0.85\,\hat{\bar{\beta}}^{\text{surv}} + 0.15\,\hat{\beta}_i^{\text{surv}}.$$

So, in the event, our survey based estimates of local trends has not made much impact on our combined estimator – which is appropriate given the imprecision of the local survey figures.

3.4 Estimating Local Trends Using a Multilevel Model

In practise we would not know the variance of the true local trends, but we can estimate this variance by fitting a multilevel model, using the package MLn. We estimate the local authority district trends β_i using a multilevel model where households are clustered within local authority districts.

$$\log_e\left(\frac{\pi_{t+1985,i}}{1-\pi_{t+1985,i}}\right) = \alpha + u_i + t\,(\bar{\beta} + \varphi_i) \qquad (6)$$

$$y_{t,ij} \sim B(1, \pi_{t+1985,i}),$$

$$u_i \sim N(0, \sigma_u^2),\ \varphi_i \sim N(0, \sigma_\varphi^2),\ \text{cov}(u_i,\ \varphi_i) = 0$$

where *i* stands for local authority and *j* for household.

In our previous notation β_i is estimated by $\hat{\beta}_i^{\text{mult}} = \hat{\bar{\beta}} + \hat{\varphi}_i^{\text{mult}}$.

The estimated deviation $\hat{\varphi}_i^{\text{mult}}$ of a particular local trend from the estimated national trend $\hat{\bar{\beta}}$ is shrunk towards the estimated national trend, the amount of shrinkage relating to the number of households in the local authority. Therefore local survey estimates based on a large number of households make a greater contribution to the combined estimator $\hat{\beta}_i^{\text{mult}}$ than local survey estimates based on a smaller number of households. This is similar to allowing different values of w for different small areas in the combined estimator outlined earlier.

We fitted this 2 level logistic model using the package MLn (Rasbash, Yang, Woodhouse and Goldstein 1995) using Restricted Iterative Generalised Least Squares estimation and a second order Taylor expansion of the non linear term about the level 2 (local authority) residuals (Goldstein and Rasbash 1996).

The correlation between the estimated $\hat{\beta}_i^{\text{mult}}$ and the true values β_i is still small, $r = 0.15$[3] (Figure 2).

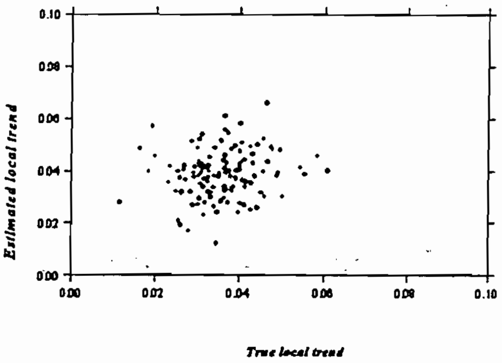

Figure 2. Relationship between survey trends estimated from survey data using MLn and true local trends.

Table 1 shows the estimated fixed and random effects from our model.

Table 1

Fixed effects			
Parameter		Estimate	Standard Error
α		-0.9529	0.0236
$\bar{\beta}^{\text{mult}}$		0.0387	0.0021
Random effects			
Parameter	Level	Estimate	Standard Error
σ_u^2	2	0.0588	0.0082
σ_φ^2	2	0.0002	0.000065
var / bin – var	1	0.9991	0.0031

There is considerable clustering of single adult households within local authority districts as measured by the estimate of σ_u^2 – which is consistent with the wide range of proportions for local authorities shown in Figure 1. The estimated variance of the deviations of the local trends φ_i from the national trend is significantly different from zero (by about three times its own standard error). Unfortunately this estimated variance (0.0002) is higher than the true value of 0.000055 obtained from the census data. One explanation for this might be that the variance of the true local trends over the period 1985-1994 was greater than the variance of the local trends over the period 1981-1991. It is also possible that differences between survey and census data collection techniques mean that their trend estimates are subject to different biases at the local level – over and above any national level differences. These issues would need to be resolved before we went ahead with local trend estimates based on multi-level models, but we would not expect them to provide fundamental difficulties if the more serious problems due to limited survey sample sizes could be resolved.

4. STRATEGIES FOR IMPROVING SURVEY – BASED ESTIMATES OF LOCAL TRENDS

The small contribution of local survey data to our combined estimator of local trends suggests that our data as it stands is not sufficient to produce estimates of local trends which are accurate enough to improve on a SPREE type estimator of local proportions. We conclude this paper by considering two alternative strategies for seeking to increase the contribution of local data to survey based estimates of local trends.

4.1 Increasing the Contribution of Survey Data to Estimates of Local Trends

Our survey data set was not large enough to have a large impact on our combined estimates of local trends. Since the contribution of the survey data is calculated from the estimated variances of the $\hat{\beta}_i^{\text{surv}}$, and these variances are proportional to the inverse of the sample size in a particular local authority we can examine the impact of increasing the survey sample size on the expected value of the average mean squared error of the combined estimator[4]. In particular how much larger would the survey data need to be to reduce the expected value of the average mean squared error of the combined estimator by fifty percent? To obtain this reduction we have to increase the sample size of the survey by a factor f, such that

$$f = \frac{2ab - bc}{ac},$$

where a is the variance of the true local trends about the national trend, b is the sampling variance of the local trends estimated from the survey data and c is the original expected value of the average weighted mean squared error. For our Labour Force Survey data we would need to multiply the achieved sample size by 8.2 to obtain a fifty percent reduction in the average mean squared error of the combined estimator. With this enlarged sample the weight w would be 0.43 *i.e.*, a greater weight would be given to the local area estimators than the national estimator.

4.2 Using Appropriate Auxiliary Variables to Model Differences Between Local Trends

A more feasible approach involves incorporating auxiliary variables into a survey based estimator of the national trend. The differences in the values of these variables between the local authorities will absorb some of the variation between the local authority trends. Since the coefficients of these variables will be estimated using the

[3] The difference between this correlation and the correlation between the raw local survey trends and true values ($r = 0.23$) is due to the differential shrinkage of the estimated deviations of the local survey trends from the national trend in the multilevel model.

[4] This follows by applying a well known result on the properties of the inverse of a matrix (Healy 1986) to the matrix of second partial derivatives of the log likelihood for our logistic model (Hosmer and Lemeshow 1989).

sample data set their estimated standard error will be small and there will be only a small increase in the average mean squared error of the estimator.

We identified two possible auxiliary variables for the proportion of households containing only one adult, but in neither case were data easily available. The first possibility is to use data from local authority records on Council Tax payments. In the UK all households pay a council tax to their local authority. Households containing only one adult can claim a reduction on their annual bill so there must be a record of the proportion of households in each local authority where this reduction applies which could be used as an auxiliary variable.

Another possible source of auxiliary data is the Electoral Register. Each local authority has a register with a record for each individual at an address who is eligible to vote. The proportion of addresses containing only one voter could therefore be used as an auxiliary variable. Although addresses are not quite the same as households, and the register is incomplete, it might nevertheless be a useful auxiliary variable.

If it is important to estimate local trends in census statistics then we need to maintain series of appropriate auxiliary variables recorded at the local authority level. The availability of auxiliary variables has increased in recent years so we should be able to evaluate the gains which may be made using auxiliary variables and survey data to update statistics from the 1991 census when we have data from the UK census in 2001.

REFERENCES

Ghosh, M., and Rao, J.N.K. (1994). Small Area Estimation: An Appraisal. *Statistical Science*, 9, 1, 55-93.

Goldstein, H., and Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society Series A*, 159, Part 3, 505-514.

Healy, M.J.R. (1986). *Matrices for Statistics*. Oxford Science Publications, Clarendon Press. Oxford.

Hosmer, D.W., and Lemeshow, S. (1989). *Applied logistic regression*. Wiley Series in Probability and Mathematical Statistics, Wiley.

Purcell, N., and Kish, L. (1980). Postcensal Estimates for Local Areas (or Domains), *International Statistical Review*, 48, 3-18.

Rasbash, J., Yang, M., Woodhouse, G., and Goldstein, H. (1995). *MLn: command reference guide*. London: Institute of Education.

SAS Institute Inc. (1989). *SAS/STAT User's Guide, Version 6, Fourth Edition*, 2, Cary, NC.