Proceedings of Statistics Canada Symposium 2002
Modelling Survey Data for Social and Economic Research

AN INVESTIGATION INTO THE DEVELOPMENT AND TESTING OF A METHODOLOGY FOR UPDATING CENSUS INDICATORS

Mohammed Yar, Neil Higgins, Philip Clarke and Patrick Heady[1]

ABSTRACT

This paper looks at the use of logistic hierarchical models to update small area census data, and considers ways of combining two sources of information namely the data from the census itself and the data from a successively repeated survey. The issues considered are the characterisation of local trends and whether local trends are dominated by the national trend. A procedure for producing postcensal estimates is proposed. The procedure is applied to two examples, where repeat survey data sets are independent over time. It is found that in these examples national trend can be adequately represented by a series of annual intercepts and there is no evidence of distinct linear local trends. However, there is some evidence that modelling the residuals as an AR process would improve the estimates.

1. INTRODUCTION

A basic requirement of efficient planning and distribution of resources, especially at local level, is the availability of up to date information at small area level. In the UK, like in many other countries, the census data is widely used in allocation of financial resources. However, the population census in the UK is conducted every ten years, the most recent being in 2001. As the time diverges away from the census date, the census data becomes out dated and potentially less relevant and the uncertainty increases as the time moves away from the census date. This is particularly so for the areas undergoing rapid socio-economic change. This may make census based distribution of resources inefficient and gives rise to concerns regarding equability of distribution of resources for inter-censal years. Therefore, official statisticians in the UK are interested in investigating alternative methods of producing estimates for census variables for the years in between the censuses, referred to as the postcensal estimates or alternatively as the census updates.

In the UK context, our goal is to produce postcensal estimates for the selected 1991 census variables relating to the years up to and including 2001 using a combination of cross sectional (the 1991 Census) and successively repeated survey data sets (the LFS data 1993 to 2001). Over the last 5 years or so the UK Office for National Statistics have developed a methodology for producing small area estimates, referred to as SAEP methodology. A limitation of the SAEP methodology is to use data sets at one time point only. In contrast, the census updates problem requires us to use repeated survey data sets and , therefore, naturally poses the question of whether to treat each area separately or attempt to borrow strength across the years. This paper represents an effort to extend the SAEP methodology to repeated data sets. By representing local trends as random variations around the national trend a logistical hierarchical mixed model for the temporal evolution of census characteristics is developed. Using the census data for small areas and the estimated local trends, the model produces postcensal estimates

The structure of the rest of the paper is as follows. Section 2 presents the extension of the ONS Small Area Estimation Programme (SAEP) methodology for the 1991 census updates. Section 3 tests the validity of the methodology through application to a simulation conducted using the 100% Finnish Population Register data for the years 1987-1996, where the actual values are known. Section 4 reports the results of its application to the 1991 UK Census and presents results for a typical census variable namely ‘proportion of 1-person households’. Finally, Section 5 describes the work currently in progress and highlights areas for further research.

[1] Office for National Statistics, UK

2. THE DEVELOPMENT OF METHODOLOGY FOR PRODUCING CENSUS UPDATES

The centrepiece of the SAEP approach is to develop a model establishing a relationship between the survey variable of interest (survey data) and auxiliary variables (administrative data sources, censuses) and to use the model to produce small area estimates. For the case of binary variables, it is the probability of an individual/household having the attribute that is modelled and the model used is:

$$\begin{aligned} & y_{ij} \sim B(1,\pi_j) \\ & y_{ij} = \pi_{ij} + e_{ij} \\ & \text{logit}(\pi_j) = \alpha + \beta \overline{X}_j + u_j \end{aligned} \qquad (1)$$

where u_j is a random variable with mean 0 and variance σ_u^2. The '$\bar{}$'above the letter denotes area level covariates, specifically means or proportions available for all the areas js and are usually centred. The small area estimate and the 95% confidence interval for the jth area proportion π_j is given by:

$$\hat{\pi}_j = (1 + \exp[-(\hat{\alpha} + \hat{\beta}\overline{X}_j)])^{-1}. \qquad (2)$$

$$(1 + \exp[-(\hat{\alpha} + \hat{\beta}\overline{X}_j \pm 2\hat{\sigma})])^{-1}$$
$$\hat{\sigma}^2 = VAR(\hat{\alpha} + \hat{\beta}\overline{X}_j + \hat{u}_j)$$

The models are fitted using the package MlWin. For further details of the SAEP methodology, refer to Heady et al (2001).

2.1 An extension of the SAEP methodology for census updates

We now describe an extension of this model to cover the census-update situation. It is assumed that initially at time t=1 both the survey variable y and the corresponding census variable Y measure the same entity. The survey is subsequently periodically repeated at t=2,…..,T. The problem is to produce small area estimates for the variable Y at t=T using the census data at t=1 and the survey data at t=1,2…..T and possibly data on auxiliary variable(s) X at t=1,2…..T, if available.
Assuming that the model (1) holds for the survey variable y_{ijt} with the covariate Y, more specifically the covariate $\log it(\Pi_{jt})$,for each of the individual time points t=….., T one gets a system of T equations:

$$\begin{aligned} & y_{ijt} \sim B(1,\pi_{jt}) \\ & \text{logit}(\pi_{jt}) = \alpha_t + \gamma\ \text{logit}(\Pi_{j1}) + u_{jt} \end{aligned}$$

Using indicator functions $ind\_k(t)$, which takes the value 1 when t equals k and 0 otherwise, after some reprametrisation the above system of equation can be written as denote an indicator function.:

$$\begin{aligned} & y_{ijt} \sim B(1,\pi_{jt}) \\ & \text{logit}(\pi_{jt}) = \beta_{j0} + \gamma\ \text{logit}(\Pi_{j1}) + \beta_{j2}\text{ind\_2(t)} + \ldots\ldots\ldots\ldots + \beta_{jT}\text{ind\_T(t)} \end{aligned}$$

$$\begin{aligned} & \beta_{j0} = \beta_0 + u_{j0} \\ & \beta_{j2} = \beta_2 + u_{j2} \\ & \ldots\ldots\ldots\ldots\ldots\ldots \\ & \ldots\ldots\ldots\ldots\ldots\ldots \\ & \beta_{jT} = \beta_T + u_{jT} \end{aligned} \qquad (3)$$

The variables π_{jt} and Π_{j1} respectively are survey and census variables and are probabilities for an individual to have a characteristic. Although in principle the area level residuals $(u_{j0},u_{j2},.....,u_{jT})$ can be assumed to be correlated over time, in this paper it will be assumed that $(u_{j0},u_{j2},.....,u_{jT})$ are independent and have multivariate normal distribution with zero mean (0,0,……..,0) and variance and covariance a diagonal matrix with ii-th element σ_{ii} ie u_{jt}. Thus for the j-th area, the marginal population mean (ignoring covariate) for the year t on the logit scale is given by

$$\beta_{j0} + u_{j0}$$
$$\beta_{j0} + u_{j0} + \beta_{jt} + u_{jt}, t = 2,.....T$$

This results in the variance and covariance matrix for the marginal π_{jt} (on logit scale) with the ik-th element

$$\sigma_{00}, i = k = 1$$
$$\sigma_{00} + \sigma_{ii}' i = k = 2,...,T$$
$$\sigma_{00}, i \neq k$$

This variance matrix is of the type compound symmetry with arbitrary diagonal elements. It assumes constant correlation for different years.

For the case of fixed coefficients for individual years, the coefficient β_t measures the change in the national proportion π during the time period (0, t) on logit scale. Thus the coefficient β_{jt} may be interpreted as measuring change in the proportion π during t=0 to t=t on logit scale for the area j. An important characteristic of the model (3) is that it does not make explicit assumption about the shape of the underlying principle trend.

The model (3) has the potential to identify local deviations from the national trends in relation to the national trend at the small area level But assumes that these deviations consist only of (a) differences that affect all years equally (effectively differences between census and survey data and (b) random terms affecting the particular year in question. No allowance is made for possible temporal autocorrelation between successive years. As a result, the estimator derived from this model (formula shown below on logit scale) does not borrow strength from previous years.

$$\text{logit}(\hat{\pi}_{jt}) = \hat{\beta}_{j0} + \hat{\gamma}\ \text{logit}(\Pi_{j1}) + \hat{\beta}_{j2}\text{ind\_2(t)} + ... + \hat{\beta}_{jT}\text{ind\_T(t)}$$

$$\hat{\beta}_{j0} = \hat{\beta}_0 + \hat{u}_{j0}$$
$$\hat{\beta}_{j2} = \hat{\beta}_2 + \hat{u}_{j2}$$
$$.....$$
$$.....$$
$$\hat{\beta}_{jT} = \hat{\beta}_T + \hat{u}_{jT} \qquad (3A)$$

Effectively this estimator differs from the SAEP estimator merely by the inclusion of an estimated the random area-level variable for the year in question. It does not borrow strength over time.

In a feasibility study conducted earlier Heady, Ruddock and Goldstein (1997) proposed the following time series generalisation of the model (1) as a way of allowing for temporal autocorrelation in the context of the census-updates problem.

$$
\begin{aligned}
&y_{ijt} \sim B(1,\pi_{jt}) \\
&\log it(\pi_{jt}) = \beta_{0j} + \beta_{1j}t + \gamma \log it(\Pi_{j1}) \\
&\beta_{0j} = \beta_0 + u_{j0} \\
&\beta_{1j} = \beta_1 + u_{j1} \\
&u_{j0} \sim N(0,\sigma_0^2), u_{j1} \sim N(0,\sigma_1^2), Cov(u_{j0},u_{j1}) = 0
\end{aligned}
\qquad (4)
$$

The formulae for the estimates of proportions analogous to the formula (3A) follow in an obvious way by substituting the estimates of fixed and random effects in (4).

The rationale for the model (4) is that reliable trend estimates for larger geographical areas can be computed using time series survey data and by making coefficients random at small area and time levels, local trend estimates can be computed. A key feature of the model is that it assumes that the underlying trend (national) is known, linear in time in this case. The authors noted that the performance of the model (4) in estimating locally specific time-trends was not satisfactory.

3. VALIDATION OF THE CENSUS UPDATES METHODOLOGY

The validation of the census updates methodology outlined in the preceding section is very important. In the UK, there is hardly any up to date good quality source of census type information at local authority level for the variables studied. Therefore, independent external assessment of the quality of the census updates is a difficult problem. However, during the project the ONS had access to the 100% aggregated Finnish Register data on a number of socio-demographic variables at NUTS 5 area level (average size in 1987: 11,000 people) for the period 1987-1996. This enabled us to use a simulation based approach for assessing the quality of the estimates. The approach involved drawing samples from the same data sets, applying the methodology, producing the estimates and comparing the estimates against the actual values.

For each year 1987-1996, separate random samples of 10,000 individuals were drawn of the variable 'employment status'. The simulated time series data sets were constructed by combining (concatenating) samples for individual years. For example, the i-th time series simulated data sets comprised i-th samples for individual years linked over time and in total it consisted of 100,000 records, with 10,000 records for each year between 1987 and 1996. In the context of Finnish data, the variables were:

(i) y_{ijt} = ith individual in jth NUTS 5 area at time t employed or not employed (sample binary variable)

(ii) Y_{j0} = proportion of resident population aged 15-74 employed in NUTS 5 in 1987 (1987 Finnish Register)

To see the suitability of the model (3) and (4) the bivariate correlations between sample estimates of the proportion of people employed (linear scale) are shown in Table 1.

Table 1: The sample estimates of bivariate correlations between proportion of people employed (aged 15-74) for different years, 1987-1996

	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996
1987	1	.979	.965	.954	.920	.895	.837	.867	.890	.893
1988	.979	1	.982	.971	.939	.906	.844	.870	.891	.890
1989	.965	.982	1	.979	.937	.900	.841	.865	.884	.877
1990	.954	.971	.979	1	.961	.934	.883	.900	.914	.904
1991	.920	.939	.937	.961	1	.971	.932	.934	.937	.925
1992	.895	.906	.900	.934	.971	1	.967	.964	.959	.945
1993	.837	.844	.841	.883	.932	.967	1	.974	.960	.942
1994	.867	.870	.865	.900	.934	.964	.974	1	.984	.971
1995	.890	.891	.884	.914	.937	.959	.960	.984	1	.986
1996	.893	.890	.877	.904	.925	.945	.942	.971	.986	1

From these correlations it can be seen that the correlations among the estimates more than 4 years apart roughly settle to a constant value. The correlations less than 4 years apart gradually decay as the time lag increases. After allowing for constant correlations among estimates for different years, the correlations less than 4 years apart are seen to be weak.

A plot of the actual trends in the proportion of people in employment (not shown here) revealed that local trends (NUTS 5 areas) were approximately similar to national trend. Furthermore, the national trend does not follow a regular pattern of monotonically increasing or decreasing over time. This can be further confirmed by looking at the national trend shown on Figure 3. This suggests that the model (3) may be more appropriate for this data set than the model (4).

Since the actual population values for the variable y are known for each year and for each NUTS5 area, the predictive performance of the models (3) and (4) can be evaluated by comparing the model based estimated values against the actual values. A plot of the actual values and model based estimates for the year 1996 is shown in Figure 1. A similar plot for the changes during 1995-1996 is shown in Figure 2.

Figure 1: A comparison of the actual values and the model based estimates of the proportions of residents in employment for NUTS 5 in Finland, 1996.

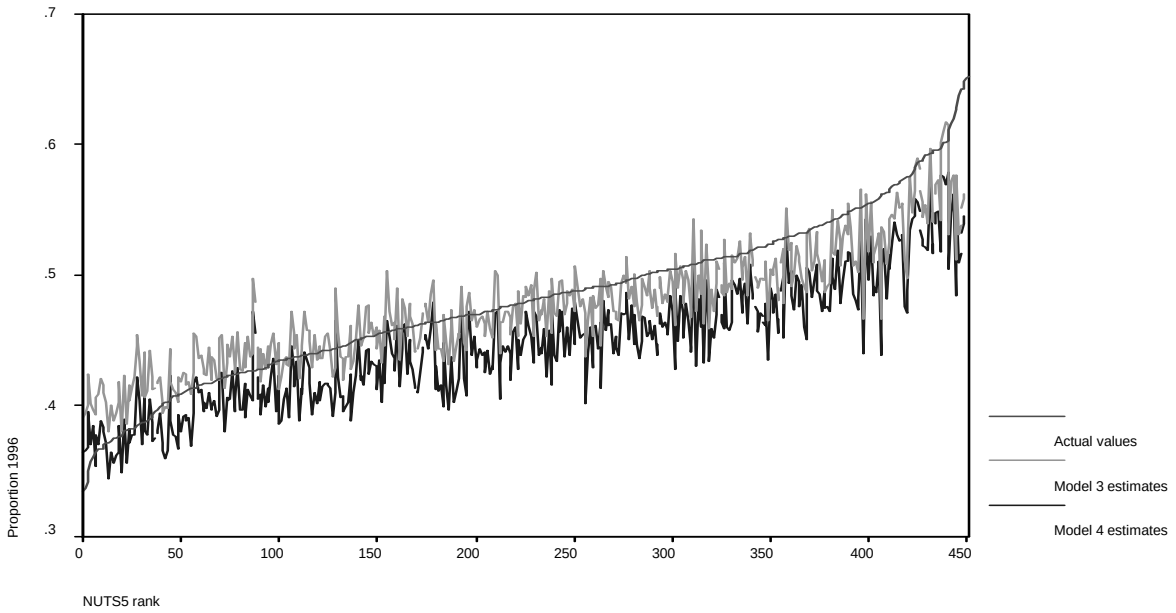


Figure 2: A comparison of the actual values and the model based estimates of changes in the proportion of residents in employment for NUTS 5 in Finland, 1987-1996.

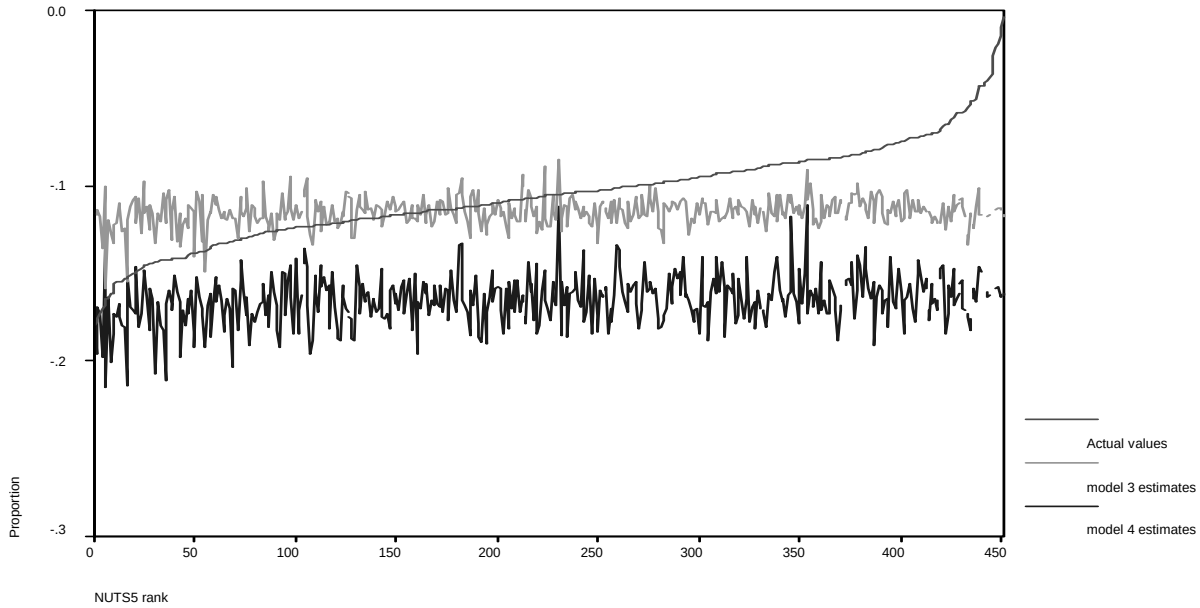
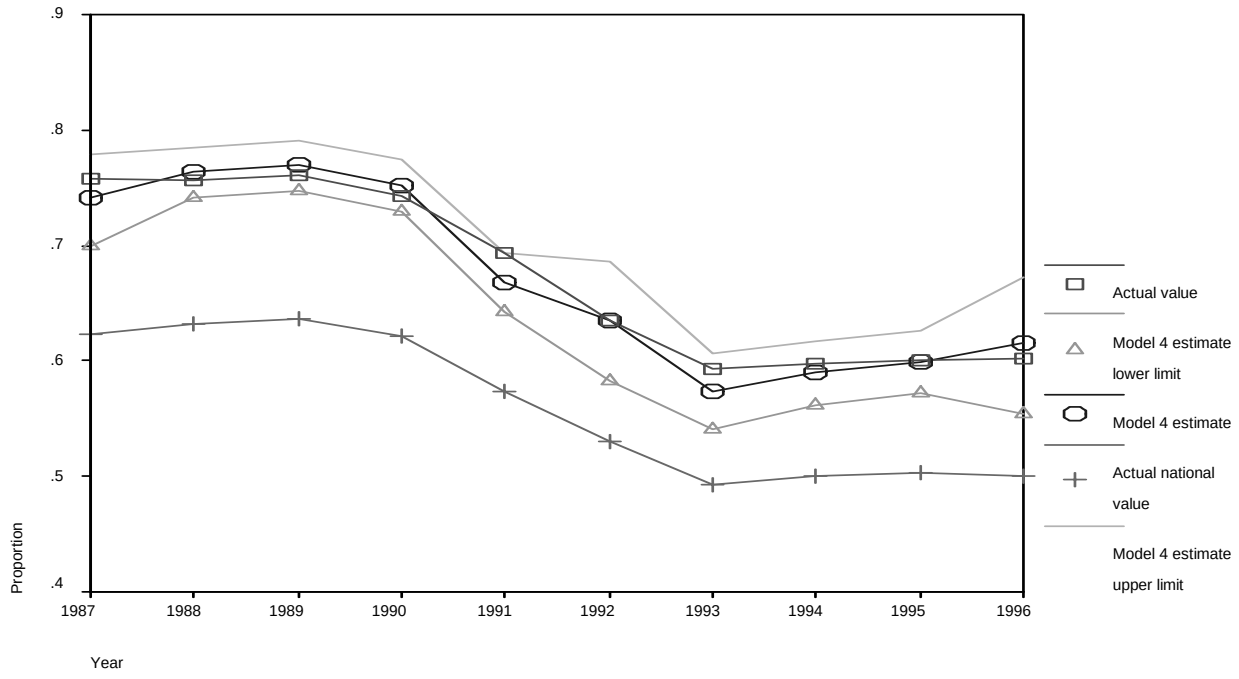


Figure 3: Actual values and the 95% confidence intervals for the model based estimates of the proportion of residents in employment for the NUTS area 257 in Finland, 1987-1996.



It can be seen from Figures 1-2 that the model (3) estimates of proportions and changes are closer to the actual values than the model (4) estimates. An examination of the fitted models (3) and (4) (not presented here) showed that:

- The model 3 estimates of variances for the random terms were very small apart from those for the intercept term and the 1996. In fact variances of the random terms for the years 1989, 1990, 1991, 1994, 1995 were estimated to be zero

- The model (4) estimates of the variance of random term for slope was nearly zero

This suggests that the model (3) and (4) estimates for small areas are predominately determined by the 1987 actual values and the assumed form of the underlying trend. While the model (3) assumes non-linear trend, the model (4) assumes linear trend. The better modelling of the underlying trend in the model (3) is thought to mainly account for the superior performance of the model (3) estimates as against the model (4). It is the fact that most changes apply roughly equally to all areas in the country that explains the surprising fact that an estimator that doesn't borrow strength over time can nevertheless provide a reasonable procedure for updating local census estimates. For this reason we have provisionally adopted model three for the purpose of constructing provisional updated estimates of local census data – as will be described further in the next part of this paper.

Before moving on to British data, it is worth making a couple of further points. For a typical NUTS 5 area, how well the model (3) traces the actual trend is shown in Figure 3. In particular, note that the model based estimates successfully follow the actual trend well and the 95% confidence intervals are satisfactory.

However, it is clear that there is still room for improvement in model 3 – since it is inherently unable to provide an estimate of locally specific trends because of its assumption of that the random terms $\{u_{jt}, t = 1,.....T\}$ are independent. It was noted earlier that there is some evidence of weak and decaying correlations between the sample estimates less than 4 years apart. This suggests assuming an autocorrelation structure for the terms $\{u_{jt}, t = 1,.....T\}$ such as suggested by AR(1). It is likely that in future some further improvement in the estimates could be obtained by modelling the residuals in this way.

4. APPLICATION TO THE UK 1991 CENSUS DATA

The methodology developed in the preceding sections was applied to produce postcensal estimates for three 1991 UK census variables (derived) for the year 2001 relating to 1998 local authority districts or unitary authorities (LAD/UA), (average size in 1991: 133,000 people) specifically:

PHHSNG = the 1991 census proportion of single person households
PPLA = the 1991 census proportion of pensioners living alone
PETHNIC = the 1991 census proportion of people from ethnic minorities

For space limitations, only results for one of the variables PHHSNG will be presented here. For details of the results on other variables refer to Yar et al (2002). The data sets employed were:

the 1991 census data
the LFS data 1993-2001

It would have been desirable to use the LFS data from 1991 and onwards although it is not clear how much difference it would have made. Due to a major redesign of LFS in 1992 and inadequate geocoding of the data for the year 1992, only LFS data from 1993 onwards could be used. Furthermore, the LFS employs a rotating panel design with each sampled household interviewed in five successive quarters making the LFS less than 5 quarters apart non-independent. Therefore, to ensure independence and avoid any complication of seasonal effect, it was decided only to use waves 1-4 of Spring Quarter (March-May) for each year 1993-2001. The individual LFS data was linked to the 1991 census covariates at UA/LAD level.

Table 2. Two level logistic model for the LFS response variable HHSNG.

Covariate	Parameter Estimate	Std. Error	T Value
Fixed part			
INTERCEPT	0.118	0.054	2.181
LGT(PHHSNG)	1.031	0.049	20.911
IND94	0.012	0.079	0.157
IND95	0.034	0.081	0.423
IND96	-0.135	0.079	-1.709
IND97	-0.268	0.081	-3.329
IND98	-0.220	0.079	-2.779
IND99	-0.208	0.077	-2.699
IND00	-0.174	0.079	-2.211
IND01	-0.232	0.078	-2.988
LGT(PHHSNG)*IND94	-0.019	0.072	-0.262
LGT(PHHSNG)*IND95	0.008	0.074	0.105
LGT(PHHSNG)*IND96	-0.110	0.072	-1.534
LGT(PHHSNG)*IND97	-0.248	0.073	-3.403
LGT(PHHSNG)*IND98	-0.199	0.072	-2.776
LGT(PHHSNG)*IND99	-0.216	0.070	-3.091
LGT(PHHSNG)*IND00	-0.195	0.071	-2.724
LGT(PHHSNG)*IND01	-0.258	0.070	-3.672
Random part			
Level 2 variances			
INTERCEPT	0.001	0.000	2.545
IND94	0.006	0.004	1.758
IND95	0.002	0.004	0.547
IND96	0.005	0.004	1.421
IND97	0.007	0.004	1.812
IND98	0.003	0.003	0.997
IND99	0.000	0.000	
IND00	0.003	0.003	0.804
IND01	0.000	0.000	
Level 1variance	1.000	0.000	

A binary LFS variable HHSNG, which takes the value 1 if the household (h/h) is a single person h/h and the value 0 otherwise, was defined. The model (3) was fitted to the response variable HHSNG using the software package MLwiN. Again due to the limitation of MLwiN the variance and covariance matrix of area level residuals u_{jt} was constrained to be to be a diagonal matrix ie:

$$\text{diag}(\sigma_{00}, \sigma_{33}, \sigma_{44}, \sigma_{55}, \sigma_{66}, \sigma_{77}, \sigma_{88}, \sigma_{99}, \sigma_{1010})$$

The only covariates available were from the 1991 census and 'time'. The model fitted to the response variable HHSNG is shown in Table 2. The labels for the variables in Table 2 are:

IND94 denotes indicator function for 1994 and it takes value 1 for 1994 and zero otherwise.
PHHSNG denotes the 1991 census proportion of single person households.
IND94* PHHSNG denotes interaction between **IND94** and **PHHSNG**.

Other indicator functions and interactions appearing in Table 2 are defined similarly.

For a typical LAD, the figures 5 shows the trend in model and LFS direct estimates with their confidence intervals for the years 1993 and 2001. For benchmarking, the LFS national estimate for each year is also shown (It can be seen that, in this instance the national trend is virtually static). The model based estimates have higher efficiency over direct survey estimates as measured by the width of confidence intervals. This is an example where the study variable 'proportion of 1-person households' has experienced slow changes over time. For this particular example, the assumption of linear trend can not be discounted either.

Model diagnostics

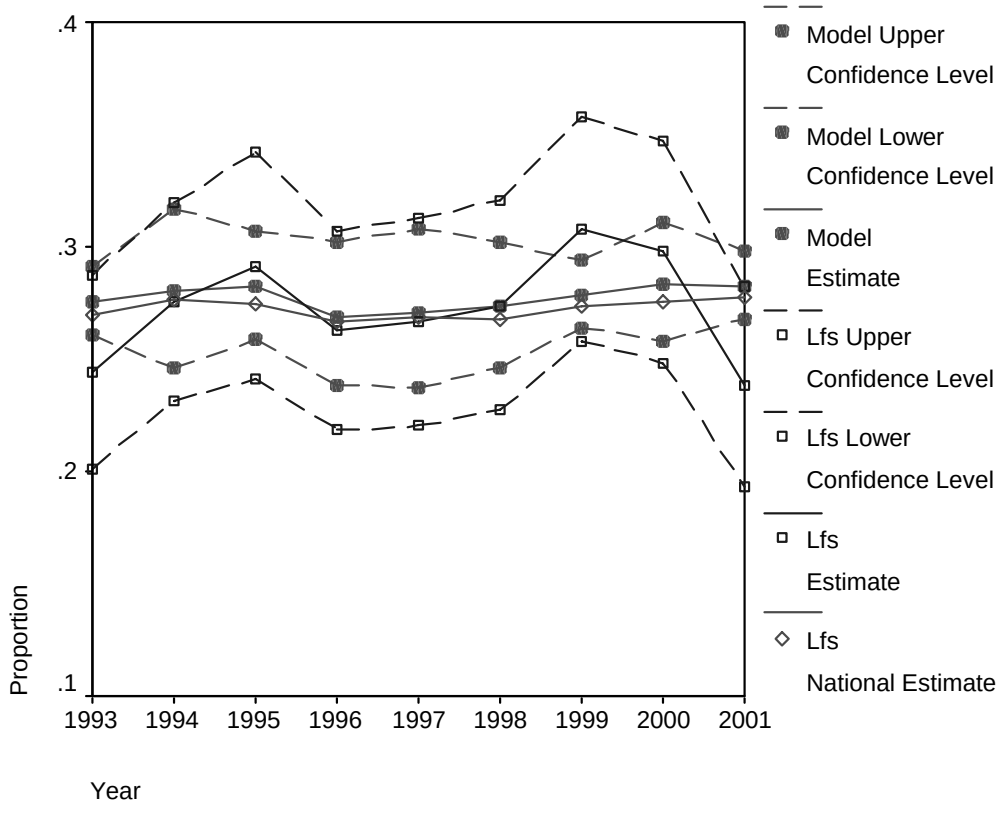
The testing and validation approach employed in Section 3 can not be employed in this particular case, as the true population values are unknown. However, in the context of small area estimation, Brown et al (2001) have developed 5 diagnostics to validate the model and check the quality of the estimates.

One of the proposed measures of the performance of the model based estimates to the actual values is the Coverage Diagnostic. Since the direct survey estimates are considered to be unbiased, 95% confidence intervals for direct survey estimates will contain the true values 95% of the times. So should the confidence intervals for model based estimates. The coverage rates of the confidence intervals for direct survey estimates in terms of model based estimates for the period 1993-2001 are shown in Table 3. For one of the local authorities, a comparison of the confidence intervals for sample estimates and model based estimates is shown in Figure 4.

Table 3. Coverage diagnostic for the proportion of single person households in England and Wales.

Year	% model based estimates outside 95% confidence interval for direct survey estimates
1993	4.9
1994	4.6
1995	4.6
1996	3.0
1997	3.5
1998	3.2
1999	4.9
2000	4.6
2001	6.2

Figure 4. Labour force survey and model estimates of the proportion of single person households in Bradford between 1993 and 2001.



5. CONCLUSIONS AND POSSIBLE AREAS FOR FURTHER RESEARCH

Based on the empirical work reported here and the other, it is concluded that it is feasible to produce census updates using a very slightly adjusted version of the SAEP methodology. However, there are a number of ways that the quality of the postcensal estimates can be improved. For the 1-person h/h example, the only covariates employed were 'time' and the 1991 census covariates. It is thought that the availability of other time varying covariates at LAD/UA level related to the variable under study will improve the quality of the estimates. For the 1-person h/h example these could be Council Tax data and Electoral Register. However, at the time of investigation these data sets were not available to the project.

The model used to produce postcensal estimates assumed independence of area level residuals over time. The resulting analysis in this case is essentially equivalent to analysing data for individual time points separately using the SAEP model. Therefore, it does not borrow strength over time and fails to exploit the full potential of the data. It is recognised that an optimal estimation strategy for the census updates case will need to borrow strength both over space and time. For that one may consider using more general models for residuals such as autoregressive models, random walk models or the ones appearing in Rao and Yu (1994) and Pfefferman (2002). It remains to be seen how far the estimates improve after allowing for autocorrelation over time. In a related context, the use of time series models in small area estimation is being investigated as part of our European project EURAREA. Once those results become available, it would be possible to apply them to produce postcensal estimates.

REFERENCES

Brown, G; Chambers, R; Heady, P and Heasman, D (2001). Evaluation of small area estimation methods – an application to unemployment estimates from the UK LFS. Proceedings of Statistics Canada Symposium 2001: Achieving Data Quality in a Statistical Agency: a Methodological Perspective.

Ghosh, M; Nangia, N and Kim, D, H (1996). Estimation of median income of 4-person families: A Bayesian time series approach. Journal of American Statistical Association, Vol 91, no 436, 1423-1431.

Heady, P; Clarke, P et al (2002). Small Area Estimation Project Report. Model Based Small area Estimation Series No 2, Office for National statistics, UK.

Heady, P; Ruddock, V and Goldstein, H (1997). An investigation into the possible use of multilevel models based on survey data to update census estimates for small areas. Proc. of Statistics Canada Symposium 1997, 199-203.

Pfeffermann, D (2002): Small area estimation-new developments and directions. International Statistical Review, 70, no 1, 125-144.

Purcell, N and Kish, L (1980). Postcensal estimates for local areas (or domains). International Statistical Review, vol48, 3-18.

Rao, J.N.K and Yu, M (1994). Small area estimation by combining time series and cross sectional data. Canadian Journal of Statistics, 22, 511-528.

Saei, A and Chambers, R (2002). Small area estimation: a review of methods based on the application of mixed models. FP5 project EURAREA report.

Tiler, R.B (1992). Time series modelling of sample survey data from the US Current Population Survey. Journal of Official Statistics, 8, 149-166.

Yar, M and Higgins, N (2002). An investigation into the development and testing of methodology for updating the 1991 UK Census indicators. Internal ONS report.