THE DESIGN OF A ONE NUMBER CENSUS IN THE UK

O Abbott[1], J Brown[2], L Buckner[1], R Chambers[2], M Cruddas[1], I Diamond[2] and J Woolford[1]

[1]Office for National Statistics
Segensworth Road
Titchfield
Fareham
Hampshire
PO15 5RR
UK

Tel: +44 (0) 1329 813512
Fax: +44 (0) 1329 813407
Email: marie.cruddas@ons.gov.uk

[2]University of Southampton
Highfield
Southampton
Hampshire
SO17 1BJ
UK

Tel: +44 (0) 1703 592518
Fax: +44 (0) 1703 593846
Email: idd@socsci.soton.ac.uk

This paper describes the development of a 'One Number Census' in the UK. This research project aims to estimate the extent of underenumeration in the 2001 Census and adjust census outputs down to the lowest level. All census outputs will then add to the national estimate of the population on Census Day, the 'one number'.

The One Number Census process has six stages:

1. The Census Coverage Survey (CCS) will re-enumerate a sample of postcodes (geographical units of around 15 households).
2. The CCS data will be matched to the Census data.
3. Estimates of the population based on the Census and CCS will be produced by age and sex for each area of a broad regional stratification of the UK.
4. Estimates of the population will be produced for Local Authority Districts (LADs), important units of resource allocation in the UK.
5. Estimates produced at 3 and 4 will be quality assured using demographic techniques and estimates.
6. Individual and household level records will be imputed for those estimated to be missed by the Census.

The paper expands on each of the above stages, paying particular attention to areas where there has been development of new methodology as opposed to the application of existing methods.

Background

One of the major uses of the decennial UK census is in providing figures on which to rebase the annual population estimates. This base needs to take into account the level of underenumeration in the census, which has traditionally been measured from data collected in a post-enumeration survey (PES) and (at the national level) through comparison with the estimate of the population based on the previous census. In the 1991 Census, although the level of underenumeration was not high (estimated at 2.2 per cent), it did not occur uniformly across all socio-demographic groups and parts of the country. There was also a significant difference between the survey-based estimate and that rolled forward from the previous census. Further investigation showed that the PES had failed to measure the level of underenumeration and its degree of variability adequately.

Maximising coverage in the 2001 Census is a priority. A number of initiatives have been introduced to help achieve this, for example:

- the Census forms have been redesigned to make them easier to complete;

- population definitions for the Census have been reviewed;
- postback of Census forms will be allowed for the first time; and
- resources will be concentrated in areas where response rates are lowest.

Despite efforts to maximise coverage in the 2001 Census, it is only realistic to expect there will be some degree of underenumeration. The One Number Census (ONC) project aims to measure this underenumeration, provide a clear link between the Census counts and the population estimates, and adjust all Census counts (which means the individual level database itself) for underenumeration.

Figure 1: A Schematic overview of the One Number Census Process

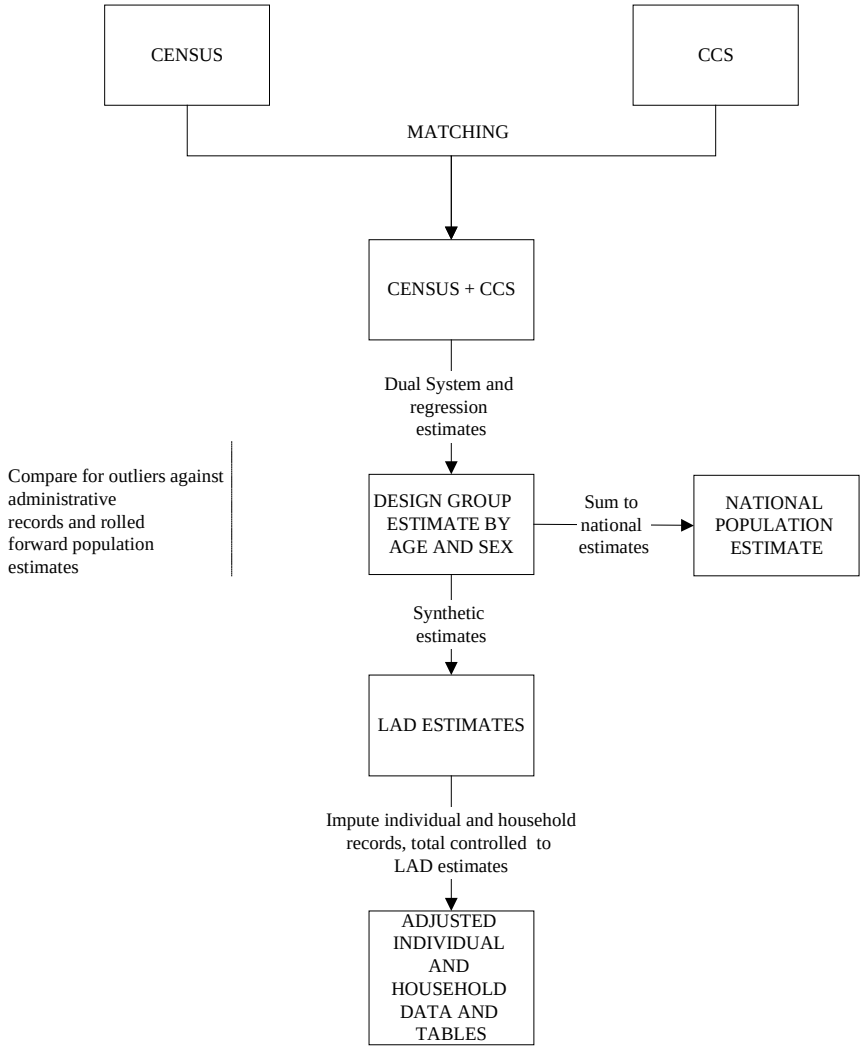

Figure 1 illustrates the One Number Census process, which comprises six stages:

1. A Census Coverage Survey (CCS) will re-enumerate a sample of postcodes (geographical units of around 15 households). The survey will collect data on a small number of key variables central to measuring underenumeration.
2. The CCS data will be matched, using a probability based matching procedure, against individual Census records.

3. Combined regression and dual system estimation will be used to produce estimates of the population based on the Census and CCS, by age and sex, for each area of a broad regional stratification of the UK. These regions, each with a population of around 0.5 million, are referred to as 'Design Groups' and are large Local Authority Districts (LADs) or groups of smaller LADs. The size of the Design Groups was selected to ensure a high efficiency of the design, based on a simulation study. LADs are important units of resource allocation in the UK. There are over 400 LADs of varying population sizes.
4. LAD estimates will be derived from the Design Group estimates using synthetic estimation.
5. National, Design Group and LAD estimates will be compared with a set of 1991 based estimates to assess their plausibility. In the event that any estimate is implausible a contingency strategy will be used.
6. Individual and household level records will be imputed for those estimated to have been missed by the Census.

The Design of the Census Coverage Survey

Following the 1991 Census, a Census Validation Survey (CVS) was carried out in England, Scotland, and Wales. This survey aimed to estimate net underenumeration and to validate the quality of Census data (Heady *et al*., 1994). The second of these aims required a complete re-interview of a sample of households that had previously been enumerated in the Census. This requirement was costly, due to the time required to fill out the complete census form, resulting in a small sample size. It also meant that the ability of the CVS to find missed households was compromised, since no independent listing of households was carried out.

An alternative strategy was required for 2001. Administrative records were found not to be accurate enough to measure Census quality to the required precision. It was therefore concluded that a PES was needed with a clear objective and different design. The CCS (as the PES will be known in 2001) will address coverage exclusively. Focusing on coverage allows for a shorter, doorstep questionnaire. Savings in time can be translated into a larger sample size. Information on question response error in the Census data will be obtained from other sources, particularly the question testing programme, the 1997 Census Test and through a separate quality survey carried out in 1999.

The CCS will be a postcode-unit based survey, re-enumerating a sample of postcode units rather than households. It is technically feasible to design a household-based CCS by sampling delivery points on the UK Postal Address File, but the incomplete coverage of this sample frame makes it unsuitable for checking coverage in the Census. Consequently, an area-based sampling design has been chosen for the CCS, with census Enumeration Districts (EDs) as primary sampling units and postcodes within EDs as secondary sampling units. Sub-sampling of households within postcodes was not considered since coverage data from all households in a sampled postcode is necessary for estimation of small area effects in the multilevel models proposed for stage six of the ONC.

Subject to resource constraints, the CCS sample design will be optimised to produce population estimates of acceptable accuracy for the 24 age-sex groups defined by sex (male/female) and 12 age classes: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-79, 80-84, 85+. The ages 45-79 have been combined since there was no evidence of any marked underenumeration in this group in 1991. This age grouping will be reviewed prior to finalising the CCS design.

Underenumeration in the 2001 Census is expected to be higher in areas with particular characteristics. For example, people in dwellings occupied by more than one household (multi-occupancy) have a relatively high probability of not being enumerated in a census. In order to control for the differentials, EDs within each Design Group are classified by a 'Hard to Count' (HtC) score. This score was chosen to represent social, economic and demographic characteristics that were found to be important determinants of underenumeration by the Office for National Statistics (ONS) and the Estimating With Confidence Project (Simpson *et al.*, 1997). The variables making up the HtC score will be reviewed prior to finalisation of the CCS design. The prototype HtC score used in the CCS Rehearsal, which was undertaken as part of the 1999 Census Rehearsal, was based on the following variables from the 1991 Census:

- percentage of young people who migrated into the Enumeration District in the last year;
- percentage of households in multiply-occupied buildings; and
- percentage of households which were privately rented.

For sample design purposes, the HtC score has been converted to a five point HtC index, with each quintile assigned an index value from 1 (easiest to count) to 5 (hardest to count). At the design stage, the role of the HtC index is to ensure that all types of EDs are sampled. Further size stratification of EDs, within each level of the HtC index, based on 1991 Census counts improves efficiency by reducing within stratum variance.

The second stage of the CCS design consists of the random selection of a fixed number of postcodes within each selected Enumeration District.

A companion paper describes the practicalities of conducting the Census Coverage Survey (Dixie, 1999).

Matching the Census Coverage Survey and Census Records

The estimation strategy requires the identification of the number of individuals and households observed in both the Census and CCS and those observed only once. Underenumeration of around two to three percent nationally means that, although absolute numbers may be large, percentages are small. Thus the ONC process requires an accurate matching methodology.

The independent enumeration methodologies employed by the Census and CCS mean that simple matching using a unique identifier common to both lists is not possible. Furthermore, simple exact matching on the variables collected in common by both methods is out of the question as there will be errors in both sets of data caused by incorrect recording, misunderstandings, the time gap, errors introduced during processing etc. The size of the CCS also means that hand matching is not feasible. Thus a largely automated process involving probability matching is necessary.

Probability matching entails assigning a probability weight to a pair of records based on the level of agreement between them. The probability weights reflect the likelihood that the two records correspond to the same individual. A blocking variable, e.g. postcode, is used to reduce the number of comparisons required by an initial grouping of the records. Probability matching is only undertaken within blocks as defined by the blocking variables.

Matching variables such as name, type of accommodation and month of birth are compared for each pair of records within a block. Provided the variables being compared are independent of

each other, the probability weights associated with each variable can be summed to give an overall probability weight for the two records. Records are matched if, for the Census record that most closely resembles the CCS record in question, the likelihood of them relating to the same household or individual exceeds an agreed threshold.

The CCS data will be used for two purposes; to enable the data to be matched against the Census; and to identify the characteristics of underenumeration via the modelling process, so that adjustments can be applied to the whole population. In order that the second part is not biased by the first the matching and modelling variables should be as independent as possible.

The initial probability weights used in 2001 will have been calculated from the data collected during the 1999 Census Rehearsal. These weights will be refined as the 2001 matching process progresses.

As the data are structured both geographically and by individuals within households we utilise this structure within the matching strategy.

The key stages of the matching are as follows:

1. Use blocking variables to reduce the number of comparisons made
2. Match households
3. Match individuals within matched households
4. Clerically check any CCS forms left unmatched.

Estimation of Design Group Age-Sex Populations

There are two stages to the estimation of Design Group age-sex populations. First, a Dual System Estimation (DSE) method is used to estimate the number of people in different age-sex groups missed by both the Census and CCS within each CCS postcode. Second, the postcode level population counts obtained from these DSEs are used in regression estimation to obtain final counts for the Design Group as a whole.

To start, we describe the DSE component of this methodology. It is unlikely that the union count (i.e. the total of those counted in the Census and/or CCS) for an area will constitute a complete count. DSE assumes that:

(i) the Census and CCS counts are independent; and
(ii) the probability of 'capture' by one or both of these counts is the same for all individuals in the area of interest.

When these assumptions hold, DSE gives an unbiased estimate of the total population. Hogan (1993) describes the implementation of DSE for the 1990 US Census. In this case assumption (i) was approximated through the operational independence of the Census and PES data capture processes, and assumption (ii) was approximated by forming post strata based on characteristics believed to be related to heterogeneity in the capture probabilities.

In the context of the ONC, DSE will be used with the Census and CCS data as a method of improving the population count for a sampled postcode, rather than as a method of estimation in itself. That is, given matched Census and CCS data for a CCS postcode, DSE is used to define a new count which is the union count plus an adjustment for people missed by both the Census and

the CCS in that postcode. This DSE count for the sampled postcode is then used as the dependent variable in a regression model, which links this count with the Census count for that postcode.

The regression model is based on the assumption that the 2001 Census count and the dual system adjusted count within each postcode satisfy a linear regression relationship with a zero intercept (a simple ratio model). However, for some age-sex groups there is the possibility of a non-zero intercept as in some postcodes the Census can miss all the people. For such age-sex groups an intercept term α_d will be added to the ratio model described below. This issue is currently being researched.

It is known from the 1991 Census that undercount varies by age and sex as well as by local characteristics, therefore a separate regression model within each age-sex group for each HtC category within each Design Group is used. Let Y_{id} denote the DSE count for a particular age-sex group in postcode i in HtC group d in a particular Design Group, with X_{id} denoting the corresponding 2001 Census count. Estimation will be based on the simple zero intercept regression model:

$$\begin{aligned} &E\left\{Y_{id} \middle| X_{id}\right\} = \beta_d X_{id} \\ &Var\left\{Y_{id} \middle| X_{id}\right\} = \sigma_d^2 X_{id} \\ &Cov\left\{Y_{id}, Y_{jf} \middle| X_{id}, X_{jf}\right\} = 0 \text{ for all } i \neq j;\ d, f = 1,\ldots,5 \end{aligned} \qquad (1)$$

Substituting the Ordinary Least Squares estimator for β_d into (1), it is straightforward to show (Royall, 1970) that under this model the Best Linear Unbiased Predictor (BLUP) for the total count T of the age-sex group in the Design Group is the stratified ratio estimator of this total given by:

$$\hat{T} = \sum_{d=1}^{5} \left\{ T_{Sd} + \sum_{i \in R_d} \left(\hat{\beta}_d X_{id} \right) \right\} = \sum_{d=1}^{5} \hat{T}_d \qquad (2)$$

where T_{Sd} is the total DSE count for the age-sex group for CCS sampled postcodes in category d of the HtC index in the Design Group; and R_d is the set of non-sampled postcodes in category d of the HtC index in the Design Group. Strictly speaking, the model specified by (1) is known to be wrong as the zero covariance assumption ignores correlation between postcode counts within a ED. However, the simple OLS estimator (2) remains unbiased under this mis-specification, and the OLS estimator is only marginally inefficient under a non-zero covariance structure (Scott and Holt, 1982).

The variance of $\hat{T} - T$, the estimation error associated with (2), can be estimated using model (1). Unlike (2), this is sensitive to mis-specification of the variance structure (Royall and Cumberland, 1978). Consequently, as the postcodes are clustered within EDs, the conservative ultimate cluster variance estimator will be used. This is given by:

$$\hat{V}(\hat{T} - T) = \sum_{d=1}^{5} \frac{1}{m_d (m_d - 1)} \sum_{e=1}^{m_d} \left(\hat{T}_d^{(e)} - \hat{T}_d \right)^2 \qquad (3)$$

where $\hat{T}_d^{(e)}$ denotes the BLUP for the population total of category *d* of the HtC index based only on the sample data from ED *e* and m_d is the number of EDs in HtC group *d*.

The above estimation strategy represents a regression generalisation of the Horvitz-Thompson DSE estimator proposed in Alho (1994). As a postcode is a small population in a generally small geographic area, and with the counts split by age and sex, the DSE homogeneity assumption should not be seriously violated. In the situation where people missed by the Census have a higher chance of being missed by the CCS than those counted by the Census, one would expect the regression estimator based on the DSE count to underestimate, but to a lesser extent than the regression estimator based on the union count. When the reverse happens and the CCS is very good at finding the missed people (the requirement for getting unbiased estimates when using the union count in the regression estimator) one would expect the DSE count regression estimator to overestimate. However, unless these dependencies are extremely high, one would not expect a gross error.

Local Authority District Estimation

Direct estimation using the CCS only produces estimates by age and sex for each Design Group. In the case of a LAD with a population of approximately 500,000 or above this will give a direct estimate of the LAD population by age and sex. However, for the smaller LADs clustered to form Design Groups, this will not be the case; although all LADs will be sampled in the CCS. For these LADs it will be necessary to carry out further estimation, and allocate the Design Group estimate to the constituent LADs.

Standard small area synthetic estimation techniques are used for this purpose. These techniques are based on the idea that a statistical model fitted to data from a large area (in our case the CCS Design Group) can be applied to a smaller area to produce a synthetic estimate for that area. The problem with this approach is that, while the estimators based on the large area model have small variance, they are usually biased for any particular small area. A compromise, introduced in the 1980s, involves the introduction of random effects for the small areas into the large area model. These allow the estimates for each small area to vary around the synthetic estimates for those areas. This helps reduce the bias in the estimate for a small area at the cost of a slight increase in its variance (Gosh and Rao, 1994).

As described in the previous section, direct Design Group estimation is based on the linear regression model (1) linking the 2001 Census count for each postcode with the DSE-adjusted CCS count for the postcode. This model can be extended to allow for the multiple LADs within a Design Group by writing it in the form

$$Y_{idl} = \beta_d X_{idl} + \delta_{dl} + \varepsilon_{idl}$$

where the extra index $l = 1 \ldots L$ denotes the LADs in a Design Group, δ_{dl} represents an LAD 'effect' common to all postcodes with HtC index d, and ε_{idl} represents a postcode specific error term. The addition of the δ_d term above represents differences between LADs that have been grouped to form a Design Group.

This regression model can be fitted to the CCS data for a Design Group, and the LAD effects δ_{dl} estimated. For consistency, LAD population totals obtained in this way will be adjusted so that they sum to the original CCS Design Group totals, and they are always at least as large as the 2001 Census counts for the LAD.

Imputation of Missed Household and Individuals

This final stage of the ONC process starts by modelling the probability of being counted in the Census in terms of the characteristics of individuals and households. This is possible in CCS areas where there are two independent counts of the population. These models are applied to all individuals and households counted by the Census in order to calculate their coverage weights. The coverage weights are calibrated to agree with the total population estimates by age-sex group and by household size for each LAD.

The imputation procedure is based on the fact that there are two processes that cause individuals to be missed by the Census. First, when there is no form received from the household and therefore all household members are missed. Second, when contact with the household fails to enumerate all household members and therefore some individuals are omitted from the form. These two processes are treated separately by the methodology.

Creating Household Coverage Weights

After the Census and CCS it can be assumed that all households within CCS areas fit into one of the following categories:

1) Counted in the Census, but missed by the CCS;
2) Counted in the CCS, but missed by the Census; and
3) Counted in both the Census and the CCS.

Underlying this is the assumption that no household is missed by both. While this is an unrealistic assumption, the calibration process accounts for such households. The final imputed database is constrained to the population estimates at the Design Group level. Categories (1) - (3) above define a multinomial outcome variable that can be modelled for each LAD using a logistic specification. Based on this model, the probability $\theta_{jidl}^{(t)}$ that household j in postcode i in HtC group d in LAD l has outcome t can be estimated. For outcomes $t = 1$ and $t = 3$ this estimated probability will be a function of the characteristics of the household as measured by the Census. This model can therefore be extrapolated to non-CCS areas to obtain estimated coverage probabilities for all households. Consequently, for each household j counted in the Census a household (h/h) coverage weight

$$w_{jidl}^{h/h} = \frac{1}{\theta_{jidl}^{(1)} + \theta_{jidl}^{(3)}}$$

can be calculated. In general, the weighted sums of households of different sizes computed using these weights will not agree with the corresponding estimates for the LAD. Consequently, these weights are calibrated (via an iterative scaling procedure) so these constraints are satisfied.

Creating Individual Coverage Weights

Coverage weights for individuals counted by the Census are obtained using similar assumptions to those described above. It is assumed that if a household is counted by the Census only, then no individuals from that household are missed by the Census. Similarly, if the household is counted by the CCS only then it is assumed that no individuals from that household are missed by the CCS. Although this assumption is violated in practice, the extra people are accounted for by constraining to estimated totals at the LAD level. Using these assumptions it is only necessary to consider individuals in households counted by both the Census and the CCS. In this case the possible categories are:

a) Counted in the Census, but missed in the CCS;
b) Counted in the CCS, but missed by the Census;
c) Counted in both the Census and the CCS.

Matched Census/CCS data and an assumed multinomial logistic model are used to estimate the probability $\pi_{kjidl}^{(r)}$ that individual *k* in household *j* in postcode *i* in HtC group *d* in LAD *l* has outcome *r*. As with the household model, the individual probabilities for outcomes $r = a$ and $r = c$ depend on individual and household characteristics as measured in the Census. Therefore, they can be extended to allow computation of coverage probabilities for all individuals counted by the Census within households also counted by the Census. For each such individual (*ind*), therefore, a coverage weight

$$w_{kjidl}^{ind} = \frac{1}{\pi_{kjidl}^{(a)} + \pi_{kjidl}^{(c)}}$$

can be calculated.

Donor Imputation for Missed Households

The next stage of the imputation process involves imputing missed households. Households are split into impute classes defined by similar household characteristics and processed sequentially in order of increasing coverage weight. When the cumulated weighted count of the households gets more than 0.5 ahead of the cumulated unweighted count a new household is imputed. The donor household is defined by the characteristics of the impute class as well as those households with the current weight and not only donates the household characteristics but all the individuals within the household as well. This process ensures that the total number of households after imputation matches the estimated LAD total. It will also correspond to totals defined by any other variables to which the household weights have been calibrated.

Donor Imputation for Missed Individuals

This is the most complex stage of the imputation since adding individuals to households changes the structure of the recipient household. This stage is best thought of in two parts. The first identifies how many individuals need to be imputed and obtains the appropriate donors. Individuals are processed sequentially, in order of coverage weight within impute class. When the cumulated weighted count exceeds the cumulated unweighted count by more than 0.5 an individual needs to be imputed. The impute class and weight define the basic characteristics of that person. A donor household is then found that contains a person of the required type. Second, the person is imputed into a ‘nearby’ recipient household. The recipient household is the

household nearest to the donor household in both space and household structure. The imputed person is added into the recipient household. The recipient household is then subject to Census edit checks to ensure internal consistency.

Pruning and Grafting of Individuals

The preceding stages of imputation add individuals to the Census database, either as part of an imputed household or as an addition to a counted household. Typically, this results in an excess of synthetic individuals on the database. The final stage of the imputation process therefore is to make sure that the totals of individuals match LAD totals by age and sex and that the resulting household size distribution is correct. A process of 'pruning off' and 'grafting on' imputed individuals from the database is then carried out until these key LAD totals are achieved.

Eventually, an individual level database will be created which will represent the best estimate of what would have been collected had the 2001 Census not been subject to underenumeration. Tabulations derived from this database will automatically include compensation for underenumeration and therefore all add to the 'One Number'.

References

Alho, J. M. (1994) Analysis of sample-based capture-recapture experiments. *Journal of Official Statistics,* **10**, 245 - 256.

Dixie, J. (1999) Planning for the 2001 Census Coverage Survey in England and Wales. *Paper to be presented at the 1999 Research Conference of the Federal Committee on Statistical Methodology.*

Ghosh, M. and Rao, J.N.K. (1994) Small area estimation: An appraisal. *Statistical Science,* **9,** 55-93.

Heady, P., Smith, S. and Avery, V. (1994) *1991 Census Validation Survey: Coverage Report*, London: HMSO.

Hogan, H. (1993) The 1990 post-enumeration survey: operations and results. *J.A.S.A.,* **88**, 1047-1060.

Royall, R. M. (1970) On finite population sampling under certain linear regression models. *Biometrika,* **57**, 377-387.

Royall, R. M. and Cumberland, W. G. (1978) Variance estimation in finite population sampling. *J.A.S.A.,* **73**, 351-361.

Scott, A. J. and Holt, D. (1982) The effect of two-stage sampling on ordinary least squares methods. *J.A.S.A.,* **77**, 848-854.

Simpson, S., Cossey, R. and Diamond, I. (1997) 1991 population estimates for areas smaller than districts. *Population Trends,* **90**, 31-39.