Bias, sampling error and coverage: The preliminary validation of the Samples of Anonymised Records from the 1991 Census

Census Microdata Unit Occasional Paper 2

Steve Simpson, Ed Fieldhouse, and Amarjit Sandhu

The samples of anonymised records have been provided through the CMU with the support of the ESRC/JISC/DENI (ESRC grant H50726502601)

Bias, sampling error and coverage: the preliminary validation of the Samples of Anonymised Records from the 1991 Census

Census Microdata Unit Occasional Paper 2

Steve Simpson, Ed Fieldhouse, and Amarjit Sandhu
November 1993

Contents

1 Characteristics of the SARs

2 Seven aspects of pre-release validation

3 Bias: comparison with the sampled population

4 Reliability: design factors

5 Coverage: boost factors

Introduction

This paper describes the procedures used to validate the Samples of Anonymised Records from the 1991 Census (SARs) before their release to the academic community in October 1993. Section 2 summarises the range of validation procedures employed by the Census Offices of OPCS and by the Census Microdata Unit (CMU). The CMU is responsible for the disseminating the SARs and promoting their effective use by the academic community in the UK.

Section 2 refers to written descriptions of the validation procedures that have been carried out This report, in sections 3, 4, and 5, concentrates on three aspects of the validation carried out by the Census Microdata Unit: estimation of the bias, reliability, and coverage of the SARs.

Readers should initially refer to the User Guide to the SARs (CMU, 1993). Section 2 of the User Guide gives a summary and overview of the approaches described in more detail in this paper.

1 Characteristics of the SARs

It is worth emphasising some characteristics of the Samples of Anonymised Records that are important in this paper.

Great Britain SARs. The procedures described here refer to the SARs for Great Britain, at a stage when the Northern Ireland SARs and the UK harmonised SARs were planned but not yet available.

Two samples. The individual SAR is a 2% sample of persons, while the household SAR is a 1% sample of households and all persons enumerated in them (but see below for large households).

Sampling procedure. The SARs were drawn in a two stage sampling procedure, operated in an exactly equivalent way throughout England, Wales and Scotland. For both SARs the first stage was the selection of a 10% sample of the full enumerated population including visitors, present residents, and absent residents (this 10% sample has already been used in the published tables of Local Base Statistics (LBS) and Small Area Statistics). The second stage was to take two independent sub-samples of households and individuals. Both stages of the sampling procedure are described in Marsh (1993). Because of the complex sampling procedure involving stratification and clustering, the SARs are not a simple random sample, although as will be shown below, they closely approximate to one.

Population bases. There are three population bases particularly relevant to the sampling procedure for the SARs, and to this report:

(a) *Individuals in households.*
Available from both the Individual SAR and Household SAR.

(b) *Individuals in communal establishments.*
Available from the Individual SAR only.

(c) *Households.*
Available from the Household SAR only.

Person records in the individual file also contains some summary information on household characteristics. Thus, for example, the number and characteristics of crowded households cannot be estimated from the individual SAR, but the total number and the characteristics of residents who live in such households can.

No modification. Records in the SARs are exactly as on the full census database. In the Local Base and Small Area Statistics counts are subject to modification or 'blurring'.

Suppression of person records in large households. In the household SAR, details of individuals in the twenty-eight households with twelve or more persons recorded (whether visitors or residents) are suppressed. Only housing characteristics (such as amenities) are available for these households. There is no other suppression of complete records on either SAR, though certain information is not entered into files by OPCS (e.g. addresses) whilst others are top coded (e.g. age) or grouped (e.g. occupation), in order to protect confidentiality.

2 Seven aspects of validation of SARs data

This section lists the procedures that were carried out on the Great Britain SARs before their release in October 1993, and refers to publications that contain further details.

(i) National Census Office procedures relating to the complete census database:

a *Editing procedures* to remove inconsistent, missing and invalid values from the complete census database. These are described in Mills and Teague (1991). Logical inconsistencies such as married persons under 16 years of age were removed at this stage.

b *Census Validation Survey.* A sample survey carried out by the OPCS Social Survey Division 4-12 weeks after the census in 1991, to check the accuracy and coverage of census data. Reports from the survey are due in 1994.

(ii) University of Manchester Census Microdata Unit procedures relating specifically to the SARs:

c *Checks for logical and other consistency.* Checks concentrated on SAR-specific issues, for example ensuring that households of twelve or more did not include person records. During the checks of this and the next item, useful experience of particular SAR variables was gained and is included in the SARs codebook (CMU 1993, Appendix F).

d *Importing of data to software platforms supported by CMU, and example applications.* The SARs were supplied as Ascii text files by the census offices to the CMU, where they have been read into SPSS, QUANVERT, QUICKTAB, SIR, and SAS packages. A dedicated SARs tabulation package, USAR, is also being developed at the University of Leeds. A number of example applications were run before the release of the SARs, to check the versatility of the two SARs and the supported packages. These and other example applications have been published as CMU Occasional Paper 3 (Middleton, 1993).

e *Bias: comparison with the sampled population.* While the sampling procedure was unbiased, the characteristics of the sample inevitably differ slightly from the full set of records for the enumerated census population due to random sampling error. The difference is likely to be relatively greater for smaller subsets of the population. Some of these biases can be precisely measured, as below in section 3.

f *Reliability: design factors for the individual SAR.* Where biases of the SARs cannot be measured directly, the reliability of estimates from the SARs can be measured. This reliability (which is not 100% because of sampling error) is expressed as a comparison with the reliability to be expected from a simple random sample. The ratio of the two is known as the design factor. Preliminary work with the SARs before their release provided estimates of design factors for a limited number of characteristics of the individual SAR. The method and results are described in section 4 below.

g *Coverage: boost factors.* The SARs are drawn only from records of the enumerated census population. 3.8% of residents of Britain are not included in this population. Some of the characteristics of this non-response are known, as described in section 5 below.

The remainder of this report deals with the last three of these validation procedures. A common format is used: an introduction to the problem; a description of the method used to address the problem; and a summary of results.

3 Bias: comparison with the sampled population

3.1 Introduction

The aim here is to compare the SARs with the population from which they were drawn, where this comparison can be made directly.

The population from which the SARs were drawn comprises:

- persons in communal establishments (individual SAR only);
- enumerated households and persons in enumerated households, ie excluding 'imputed households'.

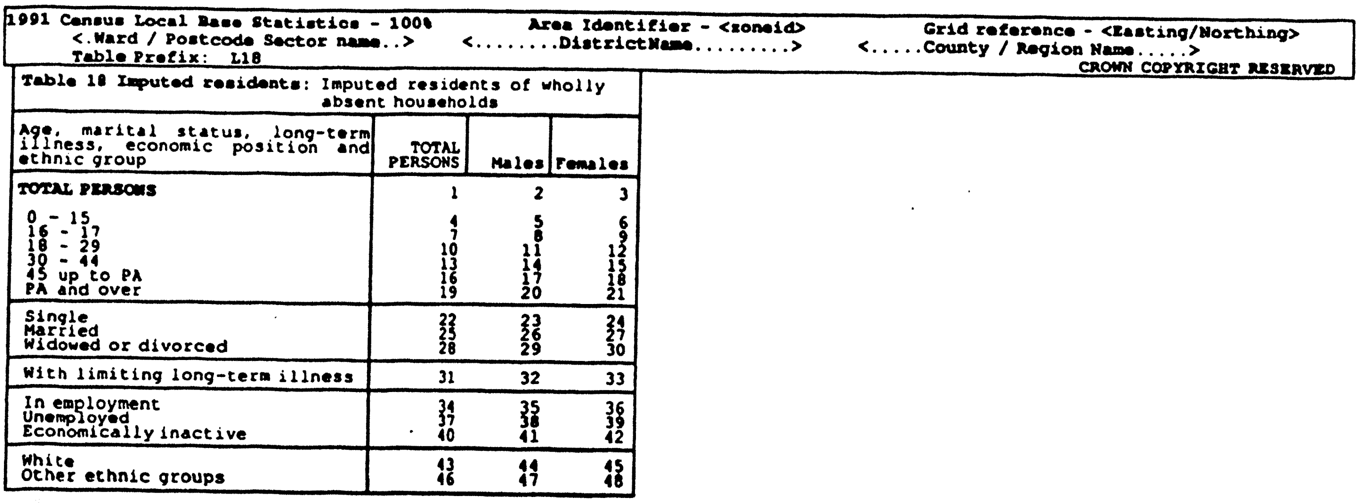
There are a great many published statistics for persons in communal establishments, for each geographical unit of the SARs. These are found primarily in the Local Base Statistics (and in printed form in the County Reports).

However, the characteristics of enumerated households and persons in them can be derived only when there are statistics for both the full census database and the imputed households, so that the latter can be deducted from the former. This in effect limits the comparisons that can be made to the characteristics of imputed records provided in summary form in Local Base Statistics Tables 1, 18 and 19 (also in the County Reports with the same numbering). The assessment of the SARs' bias reported here thus focuses on the characteristics provided in these tables.

3.2 Method

The characteristics of imputed households and person records in them are provided in Tables 18 and 19 of the Local Base Statistics (see table 1).

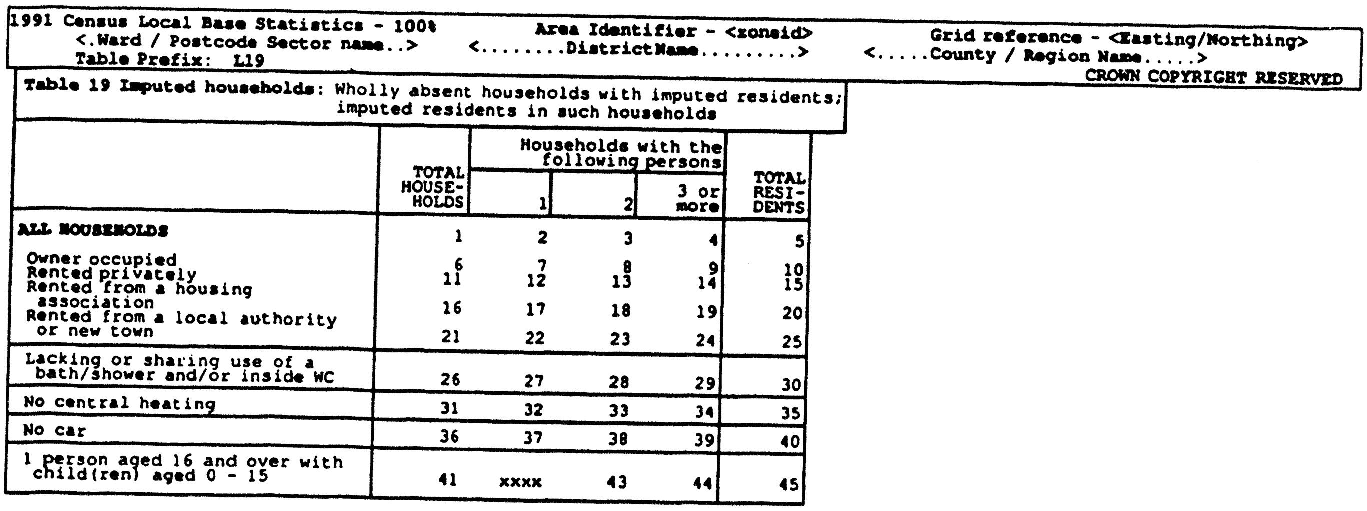
Table 1: **The census characteristics of imputed households (and residents in them): Tables 18 and 19 of the Local Base Statistics**



1991 Census Local Base Statistics - 100% Area Identifier - <zoneid> Grid reference - <Easting/Northing>
<.Ward / Postcode Sector name..> <........DistrictName.........> <.....County / Region Name.....>
Table Prefix: L18
CROWN COPYRIGHT RESERVED

Table 18 Imputed residents: Imputed residents of wholly absent households

Age, marital status, long-term illness, economic position and ethnic group	TOTAL PERSONS	Males	Females
TOTAL PERSONS	1	2	3
0 - 15	4	5	6
16 - 17	7	8	9
18 - 29	10	11	12
30 - 44	13	14	15
45 up to PA	16	17	18
PA and over	19	20	21
Single	22	23	24
Married	25	26	27
Widowed or divorced	28	29	30
With limiting long-term illness	31	32	33
In employment	34	35	36
Unemployed	37	38	39
Economically inactive	40	41	42
White	43	44	45
Other ethnic groups	46	47	48



1991 Census Local Base Statistics - 100% Area Identifier - <zoneid> Grid reference - <Easting/Northing>
<.Ward / Postcode Sector name..> <........DistrictName.........> <.....County / Region Name.....>
Table Prefix: L19
CROWN COPYRIGHT RESERVED

Table 19 Imputed households: Wholly absent households with imputed residents; imputed residents in such households

	TOTAL HOUSE-HOLDS	Households with the following persons			TOTAL RESI-DENTS
		1	2	3 or more	
ALL HOUSEHOLDS	1	2	3	4	5
Owner occupied	6	7	8	9	10
Rented privately	11	12	13	14	15
Rented from a housing association	16	17	18	19	20
Rented from a local authority or new town	21	22	23	24	25
Lacking or sharing use of a bath/shower and/or inside WC	26	27	28	29	30
No central heating	31	32	33	34	35
No car	36	37	38	39	40
1 person aged 16 and over with child(ren) aged 0 - 15	41	xxxx	43	44	45

Source: SASPAC User Manual part 2

Copyright: Local Government Management Board, and London Research Centre,1992.

By deducting these imputed households and imputed residents from the equivalent counts in other tables for all households and all residents, the characteristics of the enumerated population were derived. Table 2 simply provides a reference to the comparisons that are then possible between the SARs and the population from which they were drawn. The possible comparisons differ between the two SARs, because each SAR refers to different population bases as described earlier.

Table 2: Possible comparisons between the SARs and the 100% enumerated population from which they were drawn

Population base	Characteristics	Individual SAR	Household SAR
All residents	As in table 18	yes	no
Residents in communal establishments	As in table 18	yes	no
Residents in households	As in tables 18 & 19	yes	yes
Households	As in table 19	no	yes

This report limits itself to the comparisons based on the summary characteristics described by tables 18 and 19 of the LBS and their equivalents for other population bases. Further comparisons for households and residents in them are possible from Table 1 of the LBS (for visitors by sex), and as already mentioned for communal establishments where the deduction of imputed residents is not necessary.

3.3 Results

Some of these comparisons, for Great Britain, are contained in table 3 below, which is reproduced from the SARs User Guide (CMU 1993: 12). As one would expect, for Great Britain each SAR is so large that it very closely resembles the population from which it was drawn. However, one should not expect such a close representation for very small sub-populations, as sampling error will be considerably greater. To illustrate this point, the difference in the percentage of residents in households with no car between the 100% enumerated population and individual SAR are shown for all 278 SAR areas in Figure 1.

Table 3: Reproduced from SARs User Guide Table 2.1:
Characteristics of the SARs and the population from which they were drawn
Great Britain, percent.

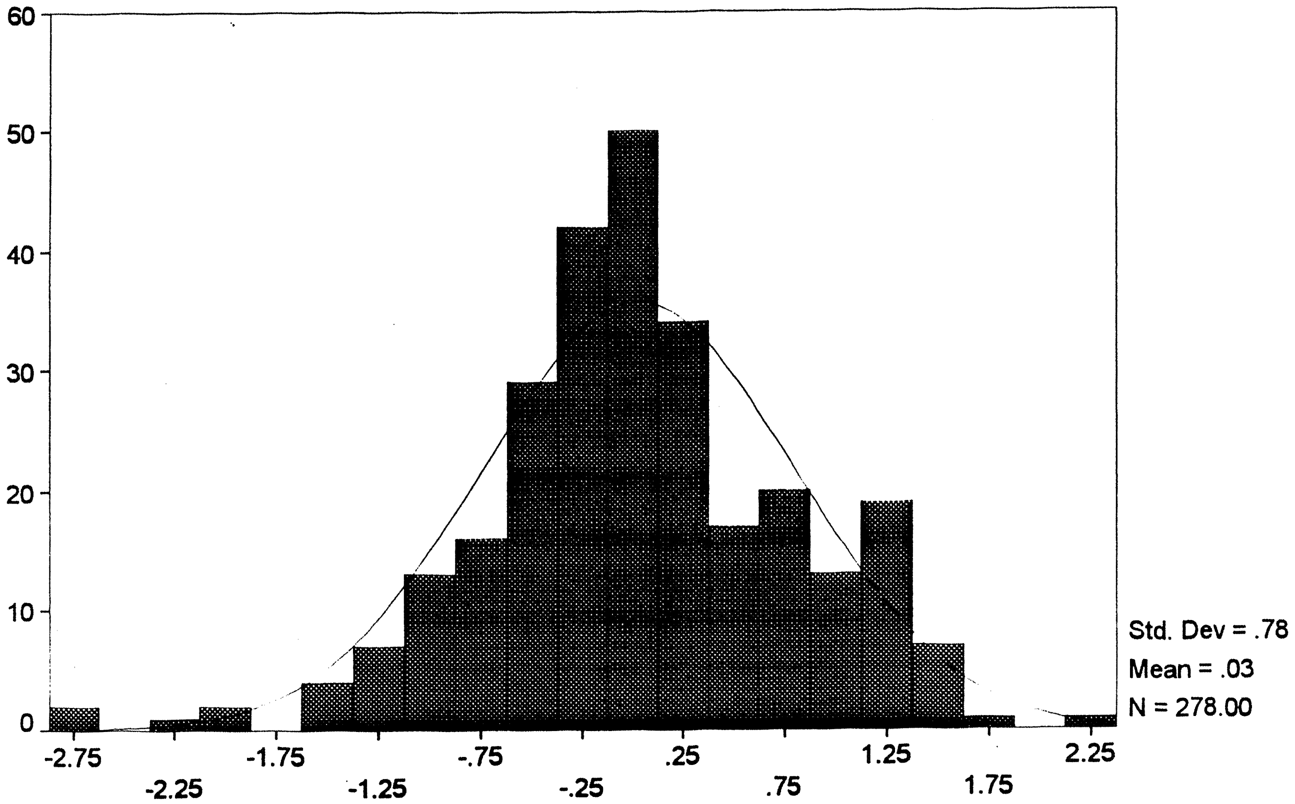
INDIVIDUAL CHARACTERISTICS	% OF ALL RESIDENTS		% OF RESIDENTS IN COMMUNAL ESTABLISHMENTS	
	Individual SAR	Census population	Individual SAR	Census population
Male	48.4	48.4	41.7	41.2
Female	51.6	51.6	58.3	58.8
Age 0-15	20.2	20.2	3.4	3.7
16-17	2.5	2.5	1.2	1.3
18-29	18.1	18.1	21.6	21.3
30-44	21.3	21.2	9.9	10.1
45 up to pensionable age	19.3	19.3	8.6	8.9
Pensionable age	18.7	18.7	55.2	54.7
Single	41.0	41.1	47.3	48.0
Married	46.9	46.8	12.5	12.7
Widowed/Divorced	12.1	12.1	40.1	39.3
With limiting long-term illness	13.1	13.1	63.5	63.3
In employment	44.1	44.3	21.6	22.1
Unemployed	4.6	4.5	3.7	3.6
Economically inactive	31.2	31.1	71.3	70.6
White	94.6	94.6	94.3	94.5
Other ethnic groups	5.4	5.4	5.7	5.5

HOUSEHOLD CHARACTERISTICS	% OF RESIDENTS IN HOUSEHOLDS		% OF HOUSEHOLDS	
	Individual SAR	Census population	Household SAR	Census population
One person in household	10.6	10.6	26.3	26.3
Owner occupied	69.9	70.0	66.4	66.7
Rented privately (excl. 'with job')	5.5	5.5	7.2	6.9
Rented from a housing ass.	2.4	2.4	3.2	3.1
Rented from a local authority, new town, or Scottish Homes	20.0	20.0	21.3	21.4
Lacking or sharing use of a bath/shower and/or inside WC	0.74	0.75	1.3	1.2
No central heating	16.8	16.8	18.8	18.8
No car	24.9	24.9	33.3	33.1
Lone parent	n/a	4.15	3.7	3.7

Sources. Individual SAR, Household SAR, LBS (Tables 18 and 19 for imputed households, deducted from equivalent cells for 100% data in other LBS tables). The base in each case excludes imputed households and residents in them.
Crown Copyright.

Figure 1.

Distribution of differences between percentageof persons in households with no car in individual SAR and 100% enumerated population.



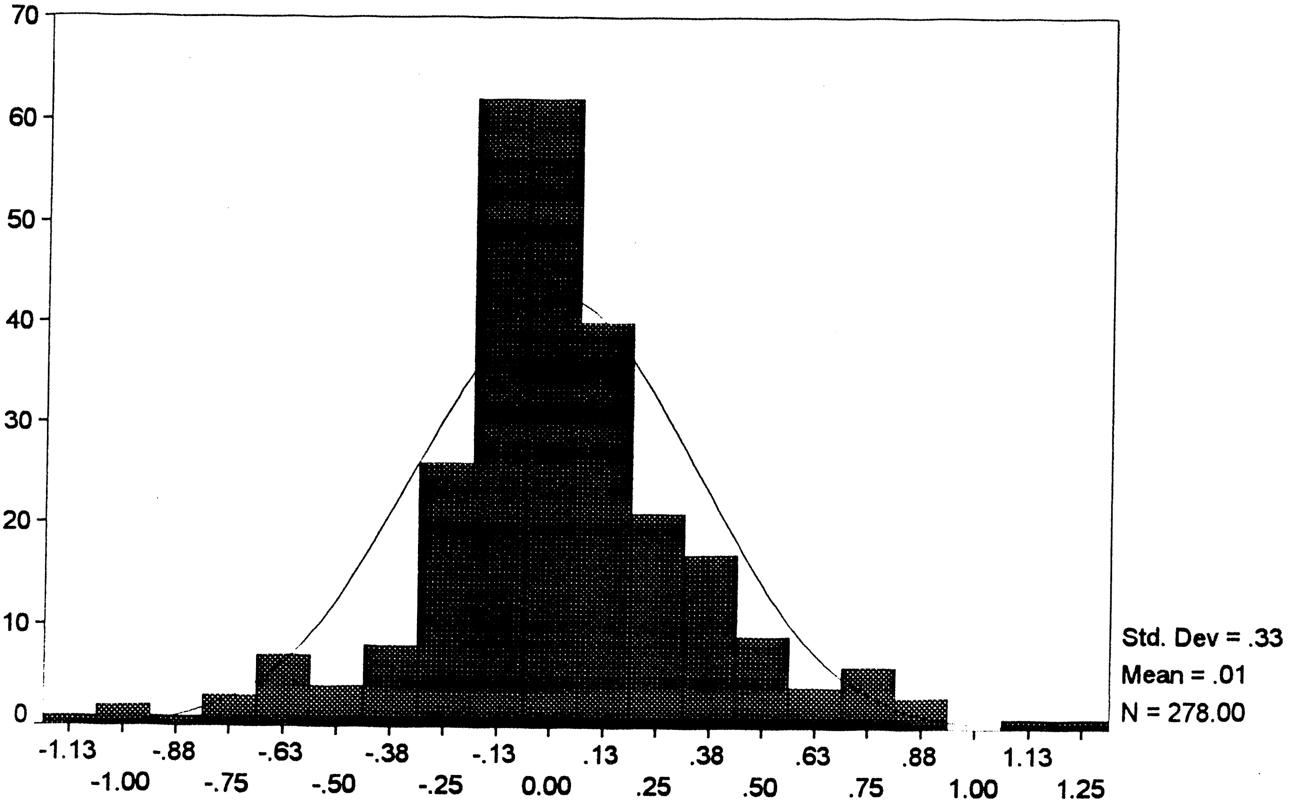
Note: Shows the distribution of raw percentage differences between the individual sar and the total enumerated population across 278 sar areas.

These differences each arise from sampling error, but the sample size in each case depends on the size of the SAR area. The degree to which the differences match what one might expect from a simple random sampling scheme may be quantified and is used to calculate design factors in section 4 below. The mean of the differences shown in Figure 1 over the 278 SAR areas is 0.03% demonstrating there is no substantial systematic bias in either direction. The standard deviation of the errors is 0.78%; all but 10 SAR areas (3.6%) falling within 2 standard deviations of the mean.

Sampling theory would suggest that larger errors be found in areas with smaller sample sizes and where the percentage in question approaches 50%. In this example, as expected, there is a significant correlation of -0.19 between the size of the error (ignoring the sign) and the total number of residents sampled within the SAR area. However, there is no linear relationship between the error and the percentage of residents with no car, because car ownership varies either side of 50%.

Figure 2.

Distribution of differences between percentage non-white population in individual SAR and 100% enumerated population.



Note: Shows the distribution of raw percentage differences between the individual sar and the total enumerated population across 278 sar areas.

Figure 2 shows a similar distribution of errors for the percentage of the population belonging to ethnic minorities (non-white). Again the overall mean error is very nearly zero and most cases fall within two standard deviations of the mean. There are 19 exceptions to this (6.8%) the largest (with errors of just over 1%) being in Newham and Hounslow (over represented) and Greenwich and Harrow (under represented). These are all areas of relatively high non-white populations. As expected this is reflected in a significant positive correlation (+0.5) between the proportion of non-whites in the sample and the size of the error. Perhaps surprisingly, however, there is no significant relationship between size of sample and error on this variable. This is probably because the benefit of large samples is offset by a higher proportion of non-whites (and thus higher sampling error) in the same areas (eg. Birmingham, Leeds, Sheffield).

CMU can provide registered users of the SARs with the equivalent of Table 3 for other national, or sub-national, areas within the SAR geography.

4 Reliability: design factors for the individual SAR

4.1 Introduction

The differences discussed above between the individual SAR and the population it was drawn from, also provide estimates of design factors for the individual SAR. These were reported in the SAR User Guide in section 2.2.

In summary, the distribution of differences that might be expected from 278 areas of known size can be calculated on the assumption that the sampling procedure was a simple random sample. By comparing the actual sample differences with that expected from a simple random sample, one estimates the extent to which the SAR sample procedure was more or less reliable than a simple random sample. The result is expressed as a ratio, the design factor, which can then be used to multiply sampling errors (for example standard errors) provided by many statistical software packages.

The method used here to estimate design factors is an approximate one for various reasons. It is based on only 278 observations. It is based only on characteristics of the SARs that can be measured precisely in the individual SAR and in the population from which it is drawn The method uses no 'sampling point' information concerning which of the SAR records was drawn from the same strata in the SAR sampling design. This information is held by the Census Offices on the confidential census database.

The sampling point information will be available within OPCS late in 1993, and will allow the calculation of more detailed and accurate design factors for both the individual SAR and the household SAR, for many more variables than have been treated here.

4.2 Method

The comparisons described in Section 3, give for various characteristics, the SAR value and its corresponding value from the 100% enumerated population. These comparisons are used here to estimate the reliability of particular characteristics in the individual SAR file, by expressing the differences as a distribution of accuracy across 278 SAR areas.

The steps in the calculation are as follows, using the example of the proportion of all residents who are of non-white ethnic group:

(a) For each of the 278 individual SAR areas, the standard error for the statistic is calculated, assuming simple random sampling. In the case of a sample proportion of non-whites, this is (following the SAR User Guide Section 2.2(ii)):

sqrt(Pr*(1-Pr)/n)

where Pr is the proportion of non-whites found in the SAR area and n is the sample size (all residents selected in the SAR in the area).

(b) The 'actual' error or the difference between the SAR statistic (the proportion of residents who are non-white) and the corresponding value from the 100% enumerated population was also calculated (as in section 3). When this is divided by the estimated standard error calculated above this provides a new variable (a statistic for each area) which, on the assumption of random sampling, should have come from a distribution with mean 0 and standard deviation 1.

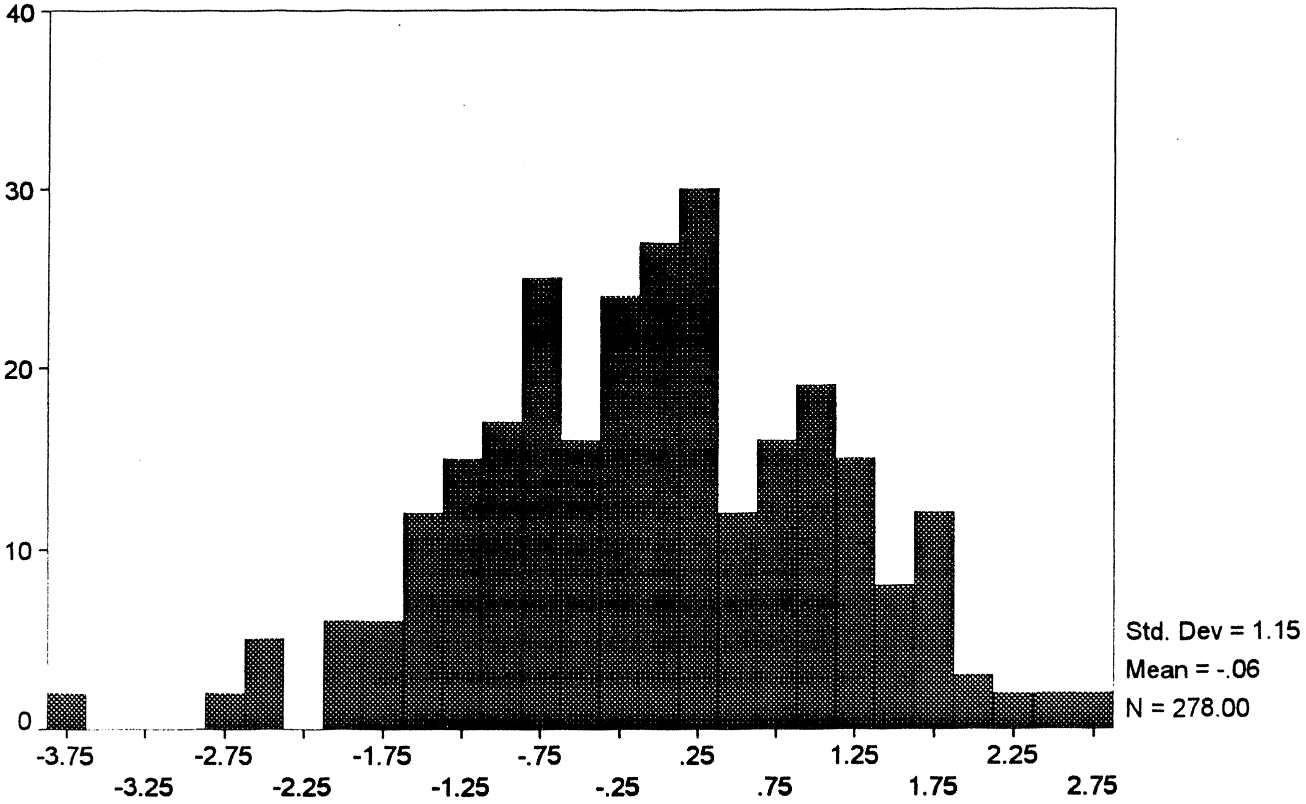
(c) The actual standard deviation of this variable is then calculated, and is the estimated design factor for the statistic. To the extent that it is more or less than one, the SAR statistic has greater or lesser sampling variability than expected under simple random sampling.

4.3 Results

Figure 3 shows the distribution across 278 areas of this variable. (i.e. the difference between the SAR and the 100% census proportion of residents who are non-white, after standardising by the estimated standard error in each SAR area.) As a consequence of standardisation, unlike the raw differences (reported in section 3.3), the standardised differences are not influenced by the size of the ethnic population of the SAR area or the sample size. The standard deviation of the distribution in Figure 3 is 1.15. This is the estimated design factor, shown in table 4 below.

Figure 3.

Distribution of standardised differences between percentage non-white population on individual SAR and 100% enumerated population.



Note: Shows the distribution of standardised percentage differences between the individual sar and the total enumerated population across 278 sar areas. The raw percentage difference has been divided by the standard error expected under simple random sampling.

Table 4: **Estimated Design Factors For The Individual SAR; Base: all residents**

Age, marital status, long term illness, economic position and ethnic group	Total Person	Males	Females
Total Persons			
		0.99	0.99
0 - 15	1.03	1.05	1.11
16 - 17	0.95	0.98	1.02
18 - 29	0.97	1.03	1.05
30 - 44	0.97	0.98	0.43
45 up to PA	1.04	1.01	1.03
PA and over	1.01	0.99	1.00
Single	1.00	1.00	0.97
Married	0.97	1.01	0.93
Widowed or Divorced	1.01	1.05	1.00
With Limiting long-term illness	0.98	0.96	0.97
In employment	1.05	1.04	1.03
Unemployed	1.03	1.03	0.97
Economically inactive	1.04	0.98	1.05
White	1.15	1.09	1.09
Other ethnic groups	1.15	1.09	1.09

Table 4 (continued)
Base: Residents in households

	Residents in 1 person households	Total Residents
All Households	1.00	
Owner occupied	0.92	0.90
Rented privately	1.01	1.05
Rented from a housing association		
Rented from a local authority	0.99	0.86
or new town		
	0.96	0.82
Lacking or sharing use of a bath/shower and/or inside w.c	1.04	1.09
No central heating	0.91	0.96
No car	1.00	1.10
1 person aged 16 and over with child(ren) aged 0 - 15		

Average design factors for each group of variables were given in the SARs user guide. So for example, the design factor for 'employment' is the mean of the design factors for each employment status category in Table 4. Note: In the User Guide (first edition) the figures provided were, incorrectly, averages of the square of the calculated design effects.

5 Coverage: boost factors

5.1 Introduction

The User Guide section 2.4 describes how the SARs lack complete coverage of the population in two particular ways that can be measured.

Firstly, like all census output, the SARs exclude those who were missed by the census. The count of usual residents in the 100% census output misses 2.2% of the complete resident population of Great Britain as estimated by OPCS/GRO(S) (1993).

Secondly, like other sample census products including the 10% tabular output and the Longitudinal Survey, the SARs exclude the 1.6% of records on the census database that were imputed. These records were created in order to represent residents that an enumerator believed existed in a household from which no census form had been obtained. While arguably improving the tabular output for small areas, their inclusion in the SARs was deemed to be inappropriate since such records were not fully coded.

This section treats only the second of these two aspects of incomplete coverage: the difference between the SARs and the 100% census database.

The Local Base Statistics provide SARs users with exact characteristics of the imputed records excluded from the SARs' population base; these are to be found principally in Tables 1, 18 and 19, which have also been essential to sections 3 and 4 above.

The SARs User Guide and this section attempt to provide some measure of the extent to which the SARs would have to be boosted in order to mirror the 100% census output. These boost factors reflect not only the imputed residents not included in the SARs, but also any bias in the sample selected that makes it over- or under-representative of the census population it was drawn from, as already discussed above in section 3.

5.2 Method

A very great number of boost factors can be calculated directly, in fact for any variable or cross-classification that exists in both the 100% census output and the SAR in question. The boost factor is simply the ratio of the former to the latter, when the latter has been multiplied by the inverse of its sampling fraction:

Individual SAR: 100% output / (50 * SAR output)

Household SAR: 100% output / (100 * SAR output)

Here we restrict attention to five-year agegroup and sex categories, simply to be able to set results side by side with those of OPCS/GRO(S) in their description of under-coverage in the census already referred to.

5.3 Results

The main results, already presented in the SARs User Guide, are reproduced below.

Panel (a) in Tables 5 and 6 shows the ratios of full census counts to those in the 2% individual SAR (grossed up by factor of 50). Table 5 for Great Britain shows that the records imputed for residents in wholly absent households are spread across all ages, and do not show the same dominance of men as in those not covered by the census (shown in panel (b)). Perhaps this is partly because the imputed household are simply records copied from nearby households of the same size that did return a late census form (see Mills and Teague 1991).

The purpose of Tables 5 and 6 is simply to provide a factor by which the SARs can be grossed to the full Census data base characteristics. It is not to claim that the inclusion of imputed households in the latter provides a higher quality output (beyond the enumerators' evidence of the existence of residents, it may not do so). More information on the quality of imputed data will be contained in the Census Validation Survey reports, published by HMSO.

Table 5: Reproduced from SARs User Guide Table 2.3
Age-gender boost factors for the SARs: Great Britain

	(a) 100% Census data divided by (Individual SAR times 50): all residents			(b) Full population divided by 100% Census data: all residents		
	Persons	Men	Women	Persons	Men	Women
All ages	1.016	1.017	1.015	1.02	1.03	1.01
0- 4	1.019	1.018	1.019	1.03	1.04	1.03
5- 9	1.011	1.014	1.007	1.03	1.03	1.02
10-14	1.007	1.006	1.007	1.02	1.02	1.01
15-19	1.011	1.018	1.002	1.02	1.03	1.01
20-24	1.020	1.017	1.023	1.06	1.10	1.03
25-29	1.021	1.017	1.024	1.07	1.10	1.03
30-34	1.024	1.027	1.021	1.03	1.05	1.02
35-39	1.012	1.020	1.005	1.01	1.02	1.00
40-44	1.011	1.011	1.011	1.01	1.02	1.02
45-49	1.017	1.017	1.018	1.00	1.00	1.00
50-54	1.014	1.016	1.013	1.00	1.00	1.00
55-59	1.015	1.017	1.012	1.00	1.00	1.00
60-64	1.014	1.015	1.014	1.00	1.00	1.00
65-69	1.023	1.028	1.018	1.00	1.00	1.00
70-74	1.019	1.016	1.021	1.00	1.00	1.00
75-79	1.015	1.011	1.017	1.00	1.00	1.00
80-84	1.025	1.034	1.022	1.02	1.01	1.03
85+	1.012	1.014	1.011	1.04	1.01	1.05

Sources and notes: see foot of Table 6.

Table 6: Reproduced from SARs User Guide Table 2.4
Area type boost factors for the SARs

AREA TYPE	(a) 100% Census data divided by (Individual SAR times 50): all residents	(b) Full population divided by 100% Census data: all residents
Great Britain	1.016	1.02
Inner London	1.082	1.04
Outer London	1.024	1.02
Metroplitan areas:		
Main	1.021	1.04
Other	1.009	1.02
Non-metropolitan areas of England and Wales:		
Cities	1.016	1.03
Other	1.009	1.02
Scotland	1.015	1.02

Sources. Panel (a): individual SAR and LBS census data. Panel (b): derived from the OPCS/GRO(S) (1993). In each case the population base excludes visitors. Crown copyright.
Notes: Panel (a) is the ratio of all residents in LBS census data to 50 times all residents in the individual SAR, for the specified age-sex-group, and reflects the residents in imputed households which are included only in the former. One should be aware that this ratio also reflects any age-sex bias in the SARs due to the sampling process, but this is minimal for Great Britain as shown in Table 2.1 above.
Panel (b) is the ratio of all residents in the government population estimates for census day using the census definition of students, to all residents in LBS census data, and reflects the residents missed by the census but included in the former.

Panel (a) of Table 6 is an approximation for the following reason: Table 6 purports to use SAR characteristics for the six types of District used in panel b to summarise census under-enumeration, as in OPCS/GRO(S) 1993. In fact the categories of non-metropolitan cities and other non-metropolitan areas in panel (a) of Table 6, are approximations because the same areas cannot be reproduced in the SAR. In the table non-metropolitan cities include all individual SAR areas that cover non-metropolitan city districts, and in so doing include fourteen other non-metropolitan areas, as follows: Kingswood, Wansdyke (with Bath); South Cambridgeshire (with Cambridge); Teignbridge (with Exeter); Chester-le-Street (with Durham); Cotswold (with Cheltenham); Tewksbury (with Gloucester); Malvern Hills (with Worcester); East Lindsey, West Lindsey (with Lincoln); Broadland (with Norwich); Selby (with York); Vale of White Horse, West Oxfordshire (with Oxford).

Despite this approximation, the results in Table 6 are consistent with other sources, and show that the problems with enumerating households that led to imputation were much greater in some areas, mainly urban areas and particularly Inner London, than in others. CMU Occasional Paper 1 (Sandhu 1993) describes the geographical variation of imputation in more detail. The message for users of Table 6 is that for some areas the SARs are considerably affected by non-response.

CMU can provide the eqivalent of panel (a) of Table 5, the ratios of 100% census population to SARs for each age-sex group, for any region in the household file or any SAR area in the individual file.

References:

Census Microdata Unit (1993) *A User Guide to the SARS*. First edition. Manchester: University of Manchester Census Microdata Unit.

Marsh, Catherine (1993) *The sample of Anonymised Records* in Dale and Marsh (eds) *The 1991 Census User's Guide*. London: HMSO.

Mills, Ian and Teague, Andy (1991) Editing and imputing data for the 1991 Census. *Population Trends 64*: 30-37.

Middleton, Liz (1993) *An introductory guide to analysing the SARs*. Occasional Paper 3. Manchester: University of Manchester Census Microdata Unit.

OPCS/GRO(S) (1993) *Under-Coverage in Great Britain* Census User Guide 58.

Sandhu, Amarjit (1993) *Issues of imputation in the 1991 census*. Occasional Paper 1. Manchester: University of Manchester Census Microdata Unit.

(Census Microdata Unit Occasional Papers 1-3 are available for £3 each from the Unit. Please enquire about others published or due to be published)