Bringing the missing million home: correcting the 1991 small area statistics for undercount

Richard Mitchell
Research Unit in Health, Behaviour and Change, Department of Community Health Sciences, University of Edinburgh Medical School, Teviot Place, Edinburgh EH8 9AG, Scotland; e-mail: Richard.Mitchell@ed.ac.uk
Danny Dorling
School of Geography, University of Leeds, Leeds LS2 9JT, England; e-mail: d.dorling@geog.leeds.ac.uk
David Martin
Department of Geography, University of Southampton, Southampton SO17 1BJ, England; e-mail: D.J.Martin@soton.ac.uk
Ludi Simpson
Centre for Census and Survey Research, University of Manchester, Manchester M13 9PL, England; e-mail: Ludi.Simpson@man.ac.uk
Received 6 August 2001; in revised form 11 January 2002

Abstract. The 1991 UK Decennial Census missed about 1.2 million people. These missing individuals present a serious challenge to any census user interested in measuring intercensal change, particularly amongst the most marginalised groups in society who were prominent amongst the missing population. Recently, a web-based system for accessing census data from the 1971, 1981, and 1991 censuses was launched (www.census.ac.uk/cdu/lct). The 'LCT' package also provides access to a set of 1991 small area statistics (SAS) which have been corrected to compensate for the missing million. The authors explain the methods used for adjusting the SAS counts, provide examples of the differences between analysis with the adjusted and unadjusted data, and recommend the use of the new data set to all those interested in intercensal change.

Introduction

Change in the social and spatial distribution of people and their characteristics is both a product and creator of change in wider society and is therefore of interest to social scientists, policymakers, and the general public. Repeated survey of a population at fixed points through time is a common means of creating data with which to describe such societal change. However, only one social survey gets near to a complete coverage of everyone who lives in Britain—the decennial census (Openshaw, 1995). This is one reason why data from censuses are such a potentially rich source of information. They provide perhaps the best resource for exploring changes in the nature of Britain; who lives where, and what has happened or might happen to them.

Although the quality of census coverage always falls short of the desired 100%, the magnitude of that shortfall was especially marked in the 1991 Census process with 1.2 million people, whom the census attempted to count, estimated as 'missing' (Martin et al, forthcoming; Simpson and Dorling, 1994). Two strategies have been employed by researchers to cope with this shortfall in coverage. These can be referred to as 'the ostrich' and the 'correct your margins' strategies. The ostrich strategy, as you might expect, is to ignore the problem altogether and carry out analysis which treats the census counts as unproblematic. The ostrich strategy was broadly encouraged by a statement from the Census Office "that the effects of under-coverage in 1991 are likely to be unimportant for most purposes" (OPCS, 1995). Indeed, for some (perhaps many) purposes, the 'missing million' probably had little impact on research findings. Clearly though, where a risk

of error exists, it is sensible to try and minimise that risk. As the susceptibility of the analysis to undercount induced error increases, so does the strength of the case for using a corrected data set.

In the years following the 1991 Census a considerable amount has been learnt about who was missed on census night and from where they were missed. There is general agreement that the best way to characterise the distribution of absence from the census is to consider groups defined by age, sex, and location. Following validation work, the census offices released estimates of the number of people missed from the 1991 Census, broken down by age and sex. The "Estimating With Confidence" (EWC) project (Simpson et al, 1997) then enhanced this information by producing estimates of the undercount by age, sex, and enumeration district (or output area). Whatever the methods employed to derive them, estimates of undercount are widely available and utilised by academics, local authorities, and other interested parties. However, these adjustments to the census counts are really only useful if one is simply interested in knowing how many people there actually were in a particular area on census night rather than any detail about their individual characteristics (other than age and sex). If one considers a typical small area statistics (SAS) cross-tabulation (Cole, 1993), such estimates facilitate greater confidence in the population totals (table *margins*), but do not readily permit inference (substantive, or statistical) about table *cells* which count people with specific combinations of characteristics. This 'correct your margins' strategy thus has limited utility.

If your interest lies in groups of people not defined by age, sex, or location alone and not properly enumerated in 1991, these two strategies for dealing with undercount are not much help. Ideally, a set of 1991 SAS would exist in which all appropriate cell counts had been corrected for undercount. The census user could then choose whether to use the standard 1991 SAS, or those corrected for estimates of the undercount, as appropriate. The rest of this paper describes our approach to producing corrected 1991 SAS in theoretical and then practical terms. A small number of examples are then presented through which we seek to demonstrate the utility of the new data and to make the case for their widespread adoption. The undercount problem is one of a set which our ESRC-funded project has tackled (Martin et al, forthcoming) with the resulting solutions, including a set of corrected 1991 SAS, available at www.census.ac.uk/cdu/lct. It should be noted that corrected data are available only for the SAS and not the more detailed local base statistics (LBS).

Correcting the 1991 small area statistics—theory and design
Options for an approach to correcting each SAS count were limited by the tools available. The existing EWC work could be utilised together with the various census data sources. There was no scope for retrospective survey work, or access to individual-level data beyond those already in the academic domain. A 12-month project time scale and just one full-time research post were also significant problems. In recognition of those limitations a simple plan of attack was devised.

We chose to correct SAS at ward level (part postcode sector in Scotland), because this was the smallest level of geographical detail the delivery mechanism would offer (see Dorling et al, 2001; Martin et al, forthcoming). The LCT interface offers larger scale geographies but these are all derived through aggregations of ward-level information. Data from the EWC project gave an estimate of the number of people of each age and sex missing from the SAS in each ward. In theory, if it were also possible to know how many people of each age and sex possessed a particular characteristic tabulated in a SAS cell, a combination of these two items of information could produce an estimate of how many people might be missing from that particular cell count. For example, if a

SAS cell typically counts a younger male population, and it is known that a high proportion of younger men were missed from that ward, it is reasonable to assume that cell is undercounted and to then adjust it appropriately.

It is important to realise that the Office for National Statistics (ONS) does not believe any households were missed in 1991—only individuals within households [despite some evidence to the contrary from the census validation survey (Heady et al, 1994)]. Because the counts of households are regarded by ONS as being as accurate as possible, the correction process was applied only to SAS cells carrying information about individuals. This made the task more manageable and eliminated the need to distinguish between individuals missing from enumerated households and missed individuals comprising missed households.

In more detail then, each SAS cell count was conceptualised as potentially comprised of people from up to forty different age–sex categories (men aged 0, 1, 2–4, 5–9, 10–14, ..., 85+, women aged 0, 1, 2–4, 5–9, 10–14, etc). The first task was to identify the typical proportional distribution of individuals in every SAS cell amongst these forty age–sex categories (at the national scale). Getting this information required a disaggregate data source from the census which could be freely tabulated and analysed. The household-level Sample of Anonymised Records (SAR) was thus an appropriate data source for this purpose (Marsh, 1993). The SARs differ from other census outputs in that they are abstracts of individual census returns rather than aggregated counts. Information which might lead to the identification of an individual or household (such as name and address) has been removed from the SAR records. For each household in the SAR sample, information about all household members is given (Marsh and Teague, 1992). This hierarchical structure (individuals within households) was vital because some SAS cells count individuals whose own characteristics are defined with reference to other members of the household (children within specific family structures, for example). Because the SAR data stem from the same census process as the SAS, it is probable that the SAR do not precisely represent the British population as it *really* was in 1991. Although it can be assumed that the SAR is representative of numbers and characteristic features of households, some members of households that were missed by the census process itself will therefore also be missing from the SAR. There was no practical means of dealing with this potential source of error, although its impact is limited by the fact that in this research the SAR provides a sample distribution of population and household characteristics rather than a count. As such, the impact of all but the most severe and systematic undercounts will be minimised.

The SAR data were thus to be manipulated and cross-tabulated in such a way as to mimic the SAS tables, but with the addition of an age–sex breakdown for each SAS cell count. When actually calculating the correction for a 1991 ward-level SAS cell count, account would also be taken of how the age and sex of the ward's population differed from the national distribution represented in the SAR data. The actual correction calculation was made using the formulae below.

Variables
P   are probability estimates (defined below),
C is the SAS count of people,
E are EWC counts of people (including those missing),
S is the SAR count of people,
X is the count of people adjusted for the undercount (using age, sex, and location information).

Subscripts and superscripts
a   age group,
g   sex,
w   ward,
c   SAS cell,
G   Great Britain.

Problem: *To estimate a count for cell c in the SAS, adjusted for undercount.*

Solution

$$X_c^w = \sum_{ag} X_{agc}^w, \tag{1}$$

$$X_{agc}^w = C_c^w \frac{P^w(a, g)}{P^G(a, g)} P^G(a, g \,|\, c) \frac{E_{ag}^w}{C_{ag}^w}, \tag{2}$$

where
C_c^w is the SAS count in ward w of people in cell c,
$P^w(a, g)$ is the joint probability for ward w of age a and gender g,
$P^G(a, g)$ is the joint probability for Great Britain (SAR data) of age a gender g,
$P^G(a, g \,|\, c)$ is the conditional probability for Great Britain (SAR data) of age a and
                  gender g of given cell c,
E_{ag}^w is the EWC count of people of age a and gender g (including the missing
                  people) in this ward,
C_{ag}^w is the SAS count of people of age a and gender g in this ward.

Probability estimates

$$P^w(a, g) = \frac{C^w(a, g)}{\sum_{ag} C^w(a, g)}, \tag{3}$$

$$P^G(a, g) = \frac{S^G(a, g)}{\sum_{ag} S^G(a, g)}, \tag{4}$$

$$P^G(a, g \,|\, c) = \frac{S^G(a, g, c)}{\sum_{ag} G^G(a, g, c)}. \tag{5}$$

In simple nonalgebraic terms, for a SAS cell of interest these formulae perform the following five tasks. (a) Work out how many people of a given age and sex live in the ward and what proportion of each age and sex group are estimated to be missing. (b) Calculate what proportion of a SAS cell might belong to each age and sex group, using the information from cross-tabulation of the SAR data but adjusting this to take account of differences between SAR-based age – sex distributions based on a national sample and the local population in the ward in question. (c) Calculate whether the portion of the SAS cell count which belongs to each age and sex group needs to be adjusted or not, and if so, by how much. (d) Repeat this process for every age and sex group and add the results together to get an adjustment factor for the SAS cell as a whole. (e) Adjust the SAS cell. This process had to be done once for all appropriate SAS cells, for every ward in the country.

Practical steps for correcting the 1991 small area statistics
The procedure for correcting the 1991 SAS counts was clear in theoretical and algebraic terms. The first step to operationalising these ideas was to cross-tabulate the SAR data in

such a way that cells precisely mimicked the individual-level SAS cell definitions. A C++ program called SASGEN was written to do this. SASGEN was able to mimic all but a very few cells where differences between the two data sets prevented it working exactly. These are documented in the appendix. Figure 1 illustrates how SASGEN worked.

SASGEN took approximately five hours to run on a Pentium II machine (under NT), but the code itself was not complex. Logical rules defined each cell in each SAS
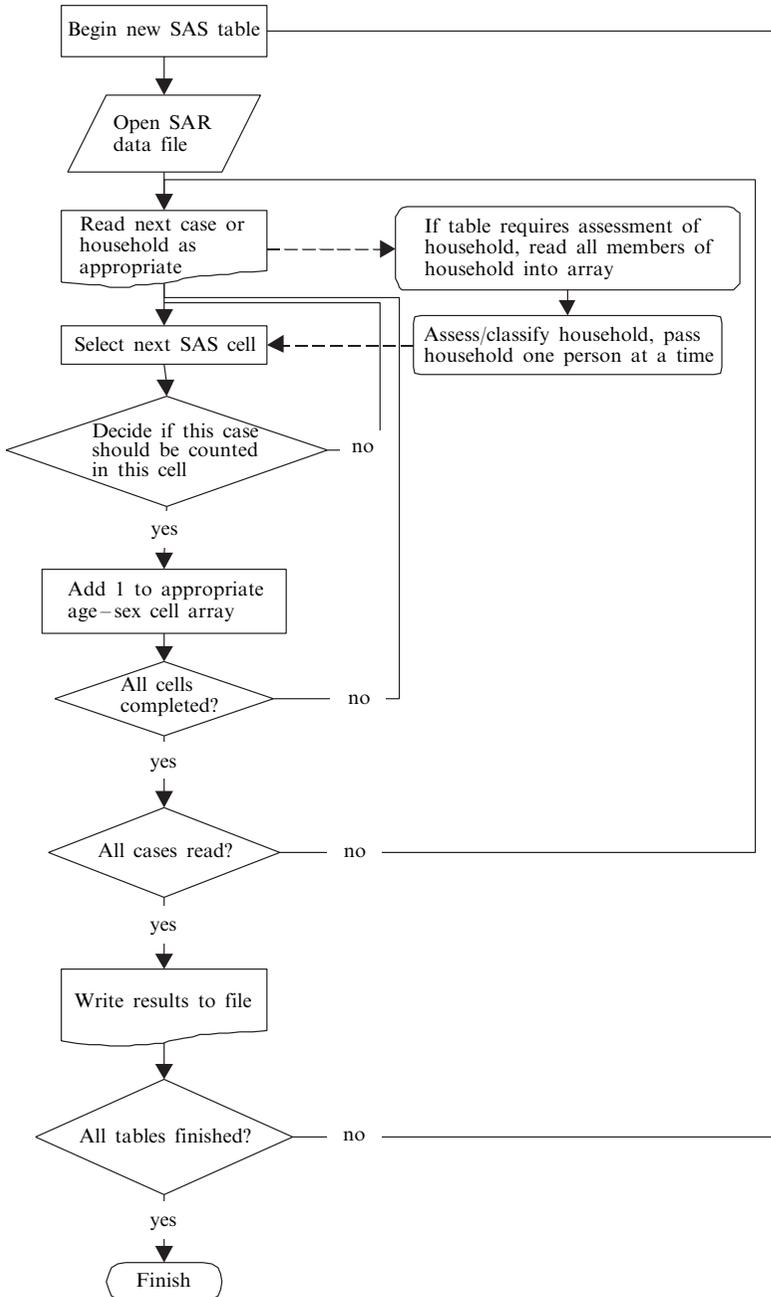

Figure 1. A flow chart describing the sequence of steps by which SASGEN produced a national age – sex breakdown of each appropriate SAS cell.

table and every case in the SAR was examined to see if that person matched the
critiera to belong in the SAS cell. For example, if a SAS cell counts people living in
owner-occupied houses who are also economically active, SASGEN checked every
individual in the SAR to see if they lived in an owner-occupied dwelling and were
economically active. If they matched those criteria, 1 was added to the total for the
appropriate age–sex group for that SAS cell. In practice, the design of the SAS tables
allowed many of the cells to be dealt with using loops within the program, removing
the need to precisely define every SAS cell with a unique line of code. As noted already,
some SAS cells counting individuals are defined by characteristics of the household
rather than the individual. In appropriate cases, data about all household members
were read into memory to help determine which SAS cells individual householders
belonged to. In practice, the complexity of the task lay in creating logical rules to
ensure that SAR members with the right characteristics were counted into the right
SAS cells.

SASGEN's first complete run created a set of data with which to validate that the
correct SAR members had been counted in the correct SAS cells. Because no other
data set breaks down every cell in the SAS by age and sex there was no directly
comparable data set which to test SASGEN's output. The best means of testing
that SASGEN was working properly was to compare its results with the SAS them-
selves. SASGEN's separate age–sex counts for each SAS cell were aggregated to give a
single value which was then compared directly with the SAS count for Britain. Because
the household-level SAR is a 1% sample of the UK census, in theory the aggregated
age–sex counts should have summed to about 1% of that for Britain as a whole (or
about 10% for those SAS counts which are a 10% sample rather than a 100% count).

Comparison between SASGEN's aggregated output and the real 1991 SAS counts
for Britain took place in a spreadsheet. Cells for which the SAR data had not provided
an 'appropriate' sample were investigated. An 'appropriate' sample varied according to
the nature of the cell. Cells with large numbers in them (for example, cells tabulating
people of a certain age, tenure, or occupation group) were expected to provide almost
precisely a 1% or 10% sample. Those where the numbers were rather smaller, perhaps
tabulating combinations of less common characteristics, were allowed to match to
within 0.2% (or 2% for the 10% SAS cells). Cells not reaching these criteria were
investigated further and the code corrected as necessary. SASGEN was run again to
produce a final set of age–sex counts for each SAS cell and these were then converted
to a proportional distribution.

The next step was to combine the age–sex breakdown for each SAS cell with the
EWC and SAS counts for each ward in accordance with the formula given above, again
in a spreadsheet environment. The decision was made to create corrected SAS values
just once and then store the results within the census data delivery package for users to
request or ignore as they wished. The data delivery package (LCT) holds census data
at ward level and aggregates 'on the fly' to the user-specified areal units (see Martin
et al, forthcoming).

Further validation and justification

Two questions remained unanswered: were the corrected 1991 cell values accurate and
is it really worth using the corrected 1991 data in preference to the standard 1991 data?
In the last two sections of the paper we address these questions.

How accurate are the corrected cell counts?

Having made much of the need to correct the SAS for undercount in 1991, and offered
these new values as a solution to the undercount problem, we must provide an

exploration of their accuracy. As with all estimation in situations where the true values cannot be known, the best strategy for testing is to make a comparison with other, similar, estimates. As noted, various attempts at correcting the undercount have been made, but no other attempts to correct for the missing million have covered all individual SAS cells in the way that the LCT project has done. Some estimates exist for broad sociodemographic groups, defined by ethnicity or employment status, for example, and these may be compared with our LCT counts. Table 1 (see over) presents a comparison between counts from the corrected SAS available through the LCT package and other attempts to correct for underenumeration. This analysis is based on data for England and Wales only.

One source of error in the LCT corrections can be readily identified. In table 1, the nonresponse rates have been calculated as a percentage of those enumerated. The LCT project has its roots in data at enumeration district level, (see Martin et al, forthcoming). These were partially suppressed for some very small areas within England and Wales to ensure confidentiality of census respondents. The number of those enumerated according to the LCT correction was thus somewhat smaller than the published total because it excludes the 34 000 people who were resident in enumeration districts for which SAS were suppressed. The 34 000 people missing from the LCT's corrected SAS tabulations is trivial when compared with the 1.2 million missing in the standard SAS.

The census offices advised census users to apply their estimates of age – sex-specific nonresponse in each local authority district (OPCS and GROS, 1994). The resulting population (column 2 in table 1) shows a smaller total number of residents than the LCT corrections (column 1) because LCT has included two adjustments extra to nonresponse. First, by putting students at their term-time address the LCT adds a net gain of 54 000 students whose vacation address was outside England and Wales. Second, by shifting the date forward ten weeks from census day to mid-year, a further 43 000 residents are gained from migration and the excess of births over deaths. Thus the LCT estimates are the only set that is consistent with the government's mid-1991 population estimates.

The estimates based on census officers' advice show a more even impact of undercount on each social group because they do not recognise the likelihood of higher undercount in inner-city and other poor areas. For example, each ethnic group has an adjustment under 4%, whereas the LCT's adjustment for each group other than White is over 4%. This is not based on any assumption of greater likelihood of underenumeration of groups other than White, but simply a result of their living in areas that were less well enumerated.

The EWC project's own adjustments (column 3) are most similar to the LCT, applying age, sex, and ward-specific nonresponse rates to each ward's census counts (Simpson et al, 1997). This similarity is to be expected, because the LCT correction procedure is based heavily on the EWC project's outputs. They do not agree precisely because of the LCT's addition of students and the ten-week demographic shift described above. Lacking other information, the LCT gives students the characteristics of the local population, which may partly account for the excess of Black Caribbean residents in comparison with the EWC project's own adjustments. Students do tend to live nearer to this ethnic minority group than does the wider population. They also tend to be White, Indian, or Chinese rather than Black Caribbean.

All the estimates discussed to date assume that census undercount was evenly spread within any local area and age – sex group. The final set of estimates in table 1 (column 4) incorporate findings that the unemployed, students, tenants with private landlords, migrants, and ethnic groups other than White are more likely to be missed

Table 1. A comparison between the 'LCT' corrected data and other attempts to correct for undercount.

	Census without adjustment as published	Adjustments to census								
		as LCT[a]	LCT[b]		Statistical agencies[c]		EWC[d]		The census data system[e]	
			no. (1)	%	no. (2)	%	no. (3)	%	no. (4)	%
Total of all residents	49 890 277	49 856 211	51 098 507	2.5	51 002 454	2.2	51 002 605	2.2	51 002 599	2.2
White	46 937 861	49 908 158	47 995 982	2.3	47 952 088	2.2	47 936 382	2.1	47 749 127	1.7
Black groups[f]	884 374	883 104	936 648	6.1	917 748	3.8	923 794	4.5	1 034 789	17.0
South Asian[g]	1 447 269	1 446 404	1 513 559	4.6	1 490 317	3.0	1 498 381	3.5	1 540 295	6.4
Other	620 773	619 106	651 022	5.2	642 302	3.5	644 049	3.7	678 388	9.3
Employed[h]	22 134 273	22 116 512	22 719 463	2.7	22 708 685	2.6	22 698 178	2.5	22 446 708	1.4
Unemployed	2 235 341	2 233 934	2 351 414	5.3	2 325 696	4.0	2 333 679	4.4	2 561 780	14.6
Student	1 531 744	1 525 262	1 621 827	6.3	1 590 208	3.8	1 588 913	3.7	1 636 705	6.9
Other inactive[i]	23 988 919	23 980 503	24 405 803	1.8	24 377 865	1.6	24 381 835	1.6	24 357 406	1.5

[a] Sum of enumeration districts.
[b] Adjustments specific to age–sex–ward, including students and shift to mid-1991.
[c] Adjustments specific to age–sex–district.
[d] Adjustments specific to age–sex–ward.
[e] Adjustments specific to age–sex–ward–social group.
[f] Black groups: Black African, Black Caribbean and Black Other.
[g] South Asian: Indian, Pakistani, and Bangladeshi.
[h] Employed: includes those on government schemes, and economically active students.
[i] Other inactive: includes children.

by a census (Simpson, 2002; Simpson and Middleton, 1999), within each local area. They incorporate the age–sex distribution recorded nationally on the SARs, as did the LCT estimates, but substituted the local age–sex distribution where known from census local statistics. They are produced only for 1991 electoral ward areas, which have less data modification and more social detail on which to base the adjustments. These estimates suggest a considerably larger undercount of some groups—for example, 17% among Black groups, and 15% of the unemployed.

Two further issues must be noted. First, account was not taken of the differential degrees to which imputation was applied in 1991 in obtaining data for households readily identified as missing. It seems likely that areas with a substantial degree of undercount were also those in which imputation took place to a greater extent, a factor which might reduce the accuracy of both the uncorrected and the corrected SAS for those areas. In addition, the SAR is derived from a sample of the SAS without imputed households. Second, much of the correction process was based on the EWC estimates which are just that—*estimates*. Any error in these will be transferred into the LCT corrections. No solution can be offered for these problems, only that their existence be flagged for consideration by the user.

In summary then, the LCT estimates are sensitive to the likely nonresponse rate in local areas including within local authority districts. However, social groups susceptible to undercount will still be underestimated: alternative estimates are available for a limited set of tables and only for aggregates of electoral wards. LCT remains the only source of corrected 1991 data for all SAS cells which count individuals, and from this analysis the corrections appear plausible.

Do you need to use the corrected values?
The utility of the corrected SAS data provided by the LCT package will vary according to the research question at hand. First, consider the analysis of change in Britain.

It should be noted that changes in definition of the population base between 1981 and 1991 (OPCS and GROS, 1992) mean that comparisons of counts in these years is probably best undertaken using the uncorrected 1991 SAS. The 1981 Census did not include a proportion of the population because of a restrictive definition of 'resident' population which excluded large groups of people living in Britain. It thus makes little sense to use corrected counts for 1991, when the 1981 undercount may have been as extensive and for which we have no comparable correction. Indeed, it should be carefully noted that this correction to the 1991 figures can have no impact at all on the inaccuracies brought to intercensal change analysis by the undercounts in 1981 and 1971 (although the latter are believed to be relatively small).

In contrast to 1981 and 1991, SAS from the 2001 Census will try to include all residents in Britain through a 'one number' system which adds anyone missed on census night back into the data before they are released to the public (ONS et al, 1999). In many ways the LCT corrections have tried to create a 'one number census' for 1991, and in fact the EWC methodologies used in the LCT project will also be used to prepare the 'one number' 2001 Census data set (Brown et al, 1999). The advent of the one number census means that analysis of change between 1991 and 2001 which does not use the corrected 1991 SAS is likely to be flawed. We strongly advise census users to compare counts of individuals in the 2001 data only with the LCT corrected data for 1991.

The underenumeration in 1991 missed about 1.2 million people, which is quite a small proportion of the entire British population and, if the analysis in question is focused on large groups of people (whether defined by geography or sociodemographic characteristics) and/or at one point in time rather than an analysis of change over time, the additional 1.2 million people might not make a substantive difference

to research findings. Similarly, where an analysis is focused on proportions of the population which might be placed into particular categories by virtue of their characteristics (as opposed to counts of people with those characteristics), correction for underenumeration might make little difference to research findings. Minority groups will not become majorities by correcting for undercount. However, where the analysis is focused on temporal change or smaller groups of people (defined by geography or individual characteristics, or both), the corrected values could certainly make a difference to substantive research findings.

For some types of analysis the corrected 1991 data may be more appropriate. Work using 1991 data only, or comparisons between 1991, 1981, and 1971 in terms of *proportion* or *distribution* (rather than raw counts) are examples. Working with uncorrected 1991 SAS will always leave open the possibility that observed changes are the result of undercount in 1991 rather than 'real' change. Use of the corrected data from the LCT package minimises that risk as far as possible. Perhaps the best strategy is to run an analysis using both the corrected and the noncorrected data. The LCT package makes it relatively quick and easy to extract both sets. If the results are sufficiently similar, researchers might opt to employ the standard 1991 SAS in their final results, and draw comfort from their use of the 'official' statistics. If, however, the results are substantially different, careful consideration might lead the user to opt for the most accurate 1991 counts which stem from this project.

Differences between corrected and uncorrected census data are presented in the following examples. Clearly, we cannot provide examples of work with the 2001 data, where the true value of the corrected 1991 SAS may lie. It would also be an impossible task to provide examples which are relevant to the entire range of research interests represented in this journal's readership. Instead, some different census counts for a set of areas are presented and the impact of correction on these counts is demonstrated. Finally, we present one very much smaller scale example which seems pertinent to the census user community in Britain.

Figure 2 illustrates the impact of using corrected values on the change in unemployment rate 1981–91 for a set of British local authority districts (as defined in 1991), selected to represent a variety of local authority sizes and characters. Note that this example uses *rates* and not counts of unemployed people. Although other versions of unemployment rates abound in the United Kingdom, the census is the only data source in which everyone has a chance to say whether *they* think they are unemployed or not. The impact of correction is clear for this selection of districts. The rise in unemployment rate between 1981 and 1991 is greater using the corrected values in almost all cases and this is because unemployed people are amongst those most likely to have been missed by the enumeration process. The correction process adds them back in because they also tend to be younger. In one sense, the impact of correction shown in this graph might not be of tremendous significance. The urban districts still have very high unemployment rates, and the direction of the trend remains the same in all example districts. However, when these differences are converted to counts of individuals, they are far more dramatic. In Manchester, for example, 6658 more unemployed people are found in the corrected SAS data than in the uncorrected data. This change does not alter the fact that Manchester had high unemployment rates in 1991, or that they were growing at this point in time, but for anyone working at smaller spatial scales, constructing a sample frame, or analysing mortality rates amongst the unemployed in this area, an extra 6658 unemployed people would make a tremendous difference.

Table 2 illustrates the difference between corrected and standard 1991 Census counts of Asian and Black people in the example districts. Note the system's ability to distinguish between places with concentrations of Asian and Black people and those without.
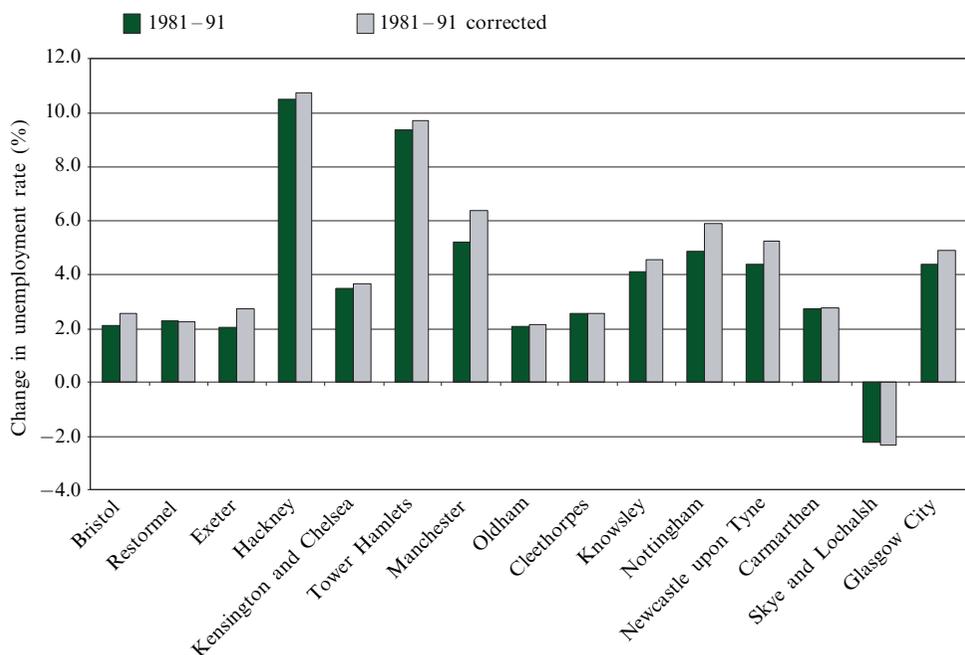
Figure 2. Change in unemployment rate using standard and corrected 1991 Census data.

Table 2. Changes to counts of ethnic minorities from correcting data.

District	Additional Asian population		Additional Black population	
	number	% increase	number	% increase
Bristol	476	7	706	8
Restormel	0	0	1	1
Exeter	71	15	26	14
Hackney	590	4	2 100	5
Kensington and Chelsea	450	7	553	7
Tower Hamlets	1 944	5	661	6
Manchester	3 202	14	3 283	17
Oldham	429	3	40	2
Cleethorpes	1	0	0	0
Knowsley	6	2	27	4
Nottingham	1 747	14	1 443	12
Newcastle upon Tyne	996	13	0	0
Carmarthen	0	1	0	0
Skye and Lochalsh	0	0	0	0
Glasgow City	1 101	7	122	8

Although considerable differences between the corrected and noncorrected 'univariate' SAS counts (such as unemployment or ethnic group) are easy to demonstrate, the true value of the corrected 1991 SAS available from the LCT system lies in the correction of multivariate, cross-tabulated counts. Table 09 in the SAS, for example, cross-tabulates gender, economic position, and ethnic group. Here, table 3 (see over) illustrates the difference in counts between corrected and noncorrected versions of the data, for men in Glasgow.

Table 3. Percentage difference between corrected and noncorrected counts for a section of cells from SAS table 09 (figures expressed as a percentage increase from noncorrected to corrected).

Glasgow City	Total persons (%)	White (%)	Black groups (%)	Indian, Pakistani, and Bangladeshi (%)	Chinese and other (%)
Total persons	4	4	11	9	12
Males 16+	6	6	15	11	15
Economically active	7	7	16	10	11
Unemployed	10	9	14	11	15
Economically inactive	3	2	16	15	23

Table 3 shows that correction for undercount adds an extra 14% to the count of unemployed Black people, but that this is dwarfed by the extra 23% of economically inactive men from Chinese or other ethnic groups. Such corrections are likely to make a substantive difference to research on these groups, and to policies focused on them. They are also essential when it comes to retrospective review of the effects of policy following release of the 2001 Census data. Even if there has been no 'real' change in economic inactivity rates amongst the Chinese and other ethnic minorities in Glasgow, the 2001 'one number census' may suggest a 23% increase on 1991 because of flawed data from 1991. This would be avoided by using the corrected data.

The magnitude of the differences between corrected and standard counts are such that in the corrected counts, the Black male unemployment *rate* in Manchester is 2.25% higher than in the standard counts. The equivalent difference for the White population is 1%. Again, the true value of correction will be revealed in analysis of change when the data from 2001 are available.

Table 4 provides one further example of the differences between corrected and noncorrected counts, this time for people living in overcrowded accommodation. At first glance, the table suggests significantly higher numbers of overcrowded people in the corrected data. However, careful interpretation of these figures is needed because

Table 4. Numbers in overcrowded households (percentage differences between corrected and noncorrected counts).

District	Persons per room		
	<0.5 (%)	1–1.5 (%)	>1.5 (%)
Bristol	5	5	9
Restormel	1	2	2
Exeter	6	7	19
Hackney	3	4	5
Kensington and Chelsea	5	6	8
Tower Hamlets	4	5	6
Manchester	7	9	15
Oldham	1	3	3
Cleethorpes	1	2	1
Knowsley	3	4	4
Nottingham	5	8	14
Newcastle upon Tyne	6	8	13
Carmarthen	2	2	1
Skye and Lochalsh	0	0	−1[a]
Glasgow City	3	4	6

[a] Because the effect of the corrections may be to move students from home to location of study, some comparisons of corrected with noncorrected data show a small reduction in population in those areas which are 'home' for students.

those areas in which extra numbers of overcrowded householders appear to have been found are also those in which extra numbers of students are often found. Student living arrangements tend to yield households which are technically 'overcrowded' but which might not be associated with the social problems of overcrowding amongst families and older people.

The single biggest difference between the corrected and noncorrected SAS counts is probably the location in which they place the student population. The 1991 Census attempted to locate students at their home address. The correction procedures employed in this project effectively put students back to their term-time location. Thus, big differences in the numbers of young people in key student areas (such as Oxford or Manchester) need to be treated with some caution. The corrected data are 'right' in the sense that more young people really did live in these areas for most of the year in 1991, but their identity as students needs to be borne in mind when assessing the implications of the higher population count and deciding whether to use the corrected data or not for a specific analysis.

The final example looks at ward DAGF (better known as 'University ward', in Leeds, England) and further develops the student numbers theme. For this analysis, the ward's boundaries have been held constant from 1971 to 1991 by retrospective reconstruction. The example illustrates one impact of correcting the 1991 Census statistics for a small area. According to the 1971 Census there were some 25 243 people present in this area. This had shrunk to 17 707 by 1981 (although the basis on which resident population was counted had changed, and using the 1971 definition the population in 1981 was 19 606). Whichever count definition is used, the population appeared to have declined. The number of students living in this ward had also declined over this period, from 1649 to 956 (although the definition had again changed from aged 15+ to age 16+). In 1991 the census suggested a further population decline to 17 326, though with a rise in student numbers to 1 711. Later, however, the EWC project estimated that some 21 886 people were resident in this area, including (according to the LCT-corrected 1991 counts) 3092 students. Although the definition of population base has changed from census to census, raw figures from each census give the impression of a trend *opposite* to that which appears to have actually happened.

Finally, some speculation about the results for this ward in the 2001 Census. A by-election was held in the ward on 7 June 2001 with the electorate on that day at 14 992. However, electorates are notoriously underreported in areas such as this and the estimate of the ward's population in 1998, provided by Oxford University (and available from the ONS website www.national-statistics.gov.uk) is 21 600 residents, of which 18 700 were aged 16+. Has the population really declined? Those who live and work in the ward doubt it. In theory, the 2001 Census will provide a firmer estimate and one which we can compare with the corrected 1991 Census to reveal the changes more clearly and accurately.

Conclusions

In this paper we have described a solution to the problem of undercount in the 1991 Census. We have explained how the corrected data were derived, explored their quality, and demonstrated their utility. In terms of quality, the corrected data are broadly in agreement with other attempts to account for undercount in 1991. Each method has advantages and disadvantages as discussed earlier in the paper. However, by far the greatest advantage of these corrections is that they have been applied to all SAS cells which count individuals, as opposed to just demographic or univariate variables. In addition, the LCT package provides the corrected data at a variety of spatial scales and geographies. As far as it is possible to tell, these data are certainly plausible and as

accurate as possible within the constraints of the correction methods and the limited time scale of one year's work.

The corrected data will become of most use when data from the 2001 Census are released. The corrected 1991 data will be the most appropriate baseline data set with which to explore the changes that occurred during the 1990s. That decade saw tremendous social, political, and economic change in Britain, with increasing strength and influence of 'minority' groups, increasing political and financial focus on the welfare of the marginalised and impoverished communities, devolution in Scotland, advent of the Welsh assembly, and, of course, a change of government. Interestingly many of these changes may have taken place (or been hoped for most) in the places and to the people who were more likely to have been missed in the 1991 Census. A proper evaluation of the impact of politics, policies, and processes in Britain at the end of the 20th century needs the best possible baseline data set from which to draw comparisons in the future. The corrected data described here provide that baseline.

References

Brown J, Diamond I, Chambers R, Buckner L, Teague A, 1999, "A methodological strategy for a one-number census in the UK" *Journal of the Royal Statistical Society Series A* **162** 247 – 267

CMU, 1993, "A user guide to the SARs", Census Microdata Unit, University of Manchester, Manchester

Cole K, 1993, "The 1991 local base and small area statistics", in *The 1991 Census User's Guide* Eds A Dale, C Marsh, Office of Population Censuses and Surveys (HMSO, London) pp 201 – 247

Dorling D, Martin D, Mitchell R, 2001, "Linking censuses through time", working paper, School of Geography, University of Leeds, Leeds

Heady P, Smith S, Avery V, 1994, "1991 Census Validation Survey: coverage report", Social Survey Division, Office of Population Censuses and Surveys (HSMO, London)

Marsh C, 1993, "The sample of anonymised records", in *The 1991 Census User's Guide* Eds A Dale, C Marsh, Office of Population Censuses and Surveys (HMSO, London) pp 295 – 311

Marsh C, Teague A, 1992, "Samples of anonymised records from the 1991 Census" *Population Trends* **69** 17 – 26

Marsh C, Skinner C, Arber S, Penhale B, Openshaw S, Hobcraft J, Lievesley D, Walford N, 1991, "The case for Samples of Anonymised Records from the 1991 Census" *Journal of the Royal Statistical Society A* **154** 305 – 340

Martin D, Dorling D, Mitchell R, forthcoming, "Linking censuses through time: problems and solutions" *Area*

ONS, GROS, NISRA, 1999, "2001 Census: a guide to the one number census", Office for National Statistics, General Register Office, Scotland, Northern Ireland Statistics and Research Agency, http://www.statistics.gov.uk/census2001/pdfs/onc.pdf

OPCS, 1995, "Census coverage", Office of Population Censuses and Surveys *Census Newsletter* **32** 2

OPCS, GROS, 1992 *1991 Census: Definitions: Great Britain* Office of Population Censuses and Surveys, General Register Office, Scotland (HMSO, London)

OPCS, GROS, 1994, "Undercoverage in Great Britain. Census user guide 58", Office of Population Censuses and Surveys, London, and General Register Office, Scotland, Edinburgh

Openshaw S, 1995, "A quick introduction to most of what you need to know about the 1991 Census", in *Census Users' Handbook* Ed. S Openshaw (GeoInformation International, Cambridge) pp 1 – 26

Simpson L, 2002, "Dealing with census undercount", in *The Census Data System: Resources, Tools and Developments* Eds P Rees, D Martin, P Williamson (The Stationery Office, London) forthcoming

Simpson S, Dorling D, 1994, "Those missing millions: implications for social statistics of non-response to the 1991 census" *Journal of Social Policy* **23** 543 – 567

Simpson S, Middleton E, 1999, "Undercount of migration in the UK 1991 Census and its impact on counterubanization and population projections" *International Journal of Population Geography* **5** 387 – 405

Simpson S, Cossey R, Diamond D, 1997, "1991 population estimates for areas smaller than districts" *Population Trends* **90** 31 – 39

APPENDIX

Owing to differences in the information about households and individuals provided by the SAR data set and the SAS tables [often as part of the measures to protect anonymity in the SAR data (CMU, 1993; Marsh et al, 1991)], not all SAS cells could be matched perfectly by SASGEN. These are documented below.

Table affected	Reason for mismatch[a]	Likely impact on results
Table 34, cell 30	SAR do not give enough detail about student's economic position.	Because so many students were missed in 1991, the 'corrected' data will underestimate for this cell count.
Table 67	SAR do not carry the same detail as SAS about reading, writing, and speaking Welsh or Gaelic.	Corrections for these tables are likely to be too small.
Table 74	SAR do not carry a code for inadequately described SOC.	Cells listed as counting people with 'inadequately described' occupations will be undercounted in the corrected data.
Table 76	SAR do not carry a code for inadequately described SOC or use same definition of 'district' as SAS.	Cells listed as counting people with 'inadequately described' occupations will be undercounted in the corrected data; cells referring to work outside usual district of residence are based on SAR districts.
Table 78	SAR do not carry a code for inadequately described SOC.	Cells listed as counting people with 'inadequately described' occupations will be undercounted in the corrected data.
Table 79	SAR do not distinguish between agricultural and forestry or fishing occupations.	Forestry and fishing cell counts may be artificially inflated.
Table 82	SAR do not use same definition of 'district' as SAS.	Cells referring to work outside usual district of residence are based on SAR districts and may therefore be undercounted.
Table 92	SAR does not identify those in SEG (socioeconomic group) 1.1 as distinct from 1.2.	Cells referring to SEG 1.1 and 1.2 are likely to be incorrect counts.

[a] SOC = Standard Occupational Classification.

p