

Secretary, Census Advisory Group
Office for National Statistics
Segensworth Road
Titchfield, Fareham
Hants PO15 5RR

From: Ludi Simpson
Centre for Census and
Survey Research,
and Bradford Council.

May 17th 2002

AG(02)03: 2001 census disclosure control in England and Wales

I enclose my letter to Len Cook of 9th April 2002 where I expressed serious concerns about the proposals then circulating. Advisory Group paper AG(02)03 does not alter the proposals addressed in my letter. Indeed an internal ONS note of Feb 8th from Paul Vickers to Graham Jones confirms my detailed description of option 1 (rounding to 3). So please consider all the points of the enclosed letter as part of this, my response to the Advisory Group paper.

I believe that neither option is justified by data disclosure risks, when balanced against the damage to the utility of the data that rounding or adjustments to small counts will create. I would urge you to consider returning to the position on data disclosure measures prior to November 2001.

In particular, there is no assessment in AG(02)03 of the damage that the proposals will cause to the utility of the output data. The lack of such an assessment is a major reason not to implement the proposed changes. The paper simply restates that there will be no impact on statistical conclusions, a statement which is untrue.

Please consider the following points in addition to the enclosed letter.

1. The ONS Output Working Party early in May agreed that the ONS Neighbourhood Statistics site would not be the vehicle to distribute a complete set of census tabulations. Such restriction is sensible to enable on-line users to explore the broad statistics, perhaps as part of considering their needs for more detailed statistics. It lessens the risk of uninformed use of statistics for small areas. Such intelligent routing of users to proper use of the statistics, together with education on the quality of the data and the license restrictions against attempts to identify individuals, are alternative means of protection from claims of and real data disclosure, as I argue in my letter attached.
2. Migration analysts at the Centre for Interaction Data at Leeds University have shown that over 90% of values in the special migration statistics are within the range 0-3, and that this figure is over 97% for those aged 45-64. They feel that the migration interaction data will be little better than random insertion of 0s and 3s for many of these data. This is a real concern for users of these very important migration data, including myself.

3. A re-analysis of data by Robin Flowerdew after rounding to three showed all the expected impacts of adding random noise to statistical analysis: the model results he had published were no longer a satisfactory fit to the data, most estimates of regression parameters reduced in significance, some to below the significance level, and one interaction changed in its direction. The noise altered the interpretation of the data as well as lessened its power. I would urge you to assess the real impact of the proposed adjustments.

Complete prevention of data disclosure is impossible¹, and thus there is a balance to be made between measures to prevent it and their impact on the utility of the data. It is the assessment of that balance that is missing from AG(02)03.

Yours sincerely,

Ludi Simpson

¹ In both options, it is possible – albeit not straightforwardly – to reconstruct some individual cell values of ‘1’s and ‘2’s. In option 1 this is possible where a table has sub-totals. In option 2 it is possible where tables have slightly differing definitions, for example of age bands, such that knowledge of 20 people of age 65 or over and 19 people aged 65-84 would allow the deduction of 1 person aged 85 or over. Zeros also present a problem: a classification with only one non-zero gives that information about all people. More generally, any zero entry shows information – that people do *not* have that characteristic.

[copy]

Len Cook, National Statistician
Office for National Statistics
Drummond Gate
London SW1V 2QQ

From: Ludi Simpson
CCSR and
Bradford Council

April 9th 2002

Dear Len Cook

Measures to prevent disclosure of confidential information in outputs from the 2001 Census

I am very pleased that ONS have reconsidered measures to control disclosure risk, as reported in Advisory Group paper AG(02)02. I note that the decisions in paper AG(01)08 last November – to double and combine thresholds for release of data for geographical areas, and to round all output to multiples of 3 – are still on the table, and that they will be sent for consultation in a further paper later this month only if alternatives prove feasible. My comments are therefore addressed to the November decisions as well as to the possible relaxed version of rounding.

This issue has been a huge distraction to users and to Census officers. Below I summarise the measures proposed before November 2001, and comment in three sections on the new measures added then. The final section offers alternatives. I conclude that all the decisions made last November could be withdrawn with gain to ONS, census users, and the public whom the statistics ultimately serve. I do hope that you will consider the arguments and make a decision to return to the positive track that Census dissemination was on in the Autumn.

The situation before November 2001

1. Users agree that no risks can be taken with regards to confidentiality. Until November last year, the measures proposed by ONS to avoid data disclosure in outputs were to swap the area identity of a small proportion of records, to not identify the items and records that had been imputed in the One Number Census procedures, and to restrict some variables to broader categories than had been collected.

Together with minimum thresholds of 20 households or 50 residents, these measures would protect private information from disclosure and satisfy the user requirement that tables should be consistent both internally and in comparison with others. This had not been achieved in the 1981 and 1991 Censuses due to the random addition and subtraction of '1's to output. In 1991, the minimum threshold was 16 households and 50 residents, while in 1981 the thresholds were 8 households and 25 residents. The main change in 2001 was to be a replacement of random modification by record swapping and greater imputation.

In addition, disclosure of private information is made less likely by the levels of error in answers (5% to 40% for most questions), migration since Census day (10% per annum), and changes in circumstances since the Census (considerable for the employment and social characteristics that are most sensitive).

2. These measures in total complied with the need to maintain confidentiality of census output laid down in the Government's White Paper *The 2001 Census of Population*, and had been considered sufficient to prevent data disclosure by the Census Division of ONS. The measures were considered adequate by an independent review of disclosure issues by Dick Carter of Statistics Canada, as reported in Advisory Group paper AG(02)01.

Perception of disclosure: a problem of education

3. Only after a further review by the ONS Methods and Quality division in 2001, a new objective was laid down: the avoidance of *perceived* disclosure as well as of *real* disclosure.

The argument was that census data, being more freely disseminated than ever before, would allow claims of data disclosure that although not able to be verified would damage confidence and trust in the confidentiality of Census data and more generally in ONS. The over-riding issue was the many values of '1' that would occur in the Census Area Statistics to be released as tables for Output Areas.

It is said that if a table shows 1 person aged 60-64 with a limiting long-term illness (for example in table CAST02) and a person in that Output Area is known to have been that age in April 2001, then their declaration of illness on the Census form can be claimed to have been made public, with resulting bad press for ONS. The argument concludes that therefore there should be no '1's in tables of output from the Census.

As a result Advisory Group Paper AG(01)08 declared that the thresholds should be doubled and combined: each output area must satisfy a minimum of 40 households *and* 100 residents. Further, all values in output would be randomly rounded to a multiple of 3.

4. Undoubtedly such a claim of disclosure could and perhaps will be made, but this does not mean that disclosure has occurred. There may be others in the same Output Area of the same age while this person's record was swapped to another area or missing altogether. Alternatively, information on illness may have been missing on this person's record and therefore imputed. There can be no certainty of disclosure. The only way of being certain of the information is if the person themselves makes public their details, in which case no disclosure by the Census has taken place.
5. It is up to the Census Offices to clarify and make public the ways in which disclosure is prevented. Public understanding that the Census output database itself is not 100% accurate but a very good estimate of the situation in each local area is a gain, not a loss. ONS actions to achieve this could include a statement clearly made at all dissemination points that small adjustments to census data have been made to ensure that no disclosure of personal information is possible. This would reinforce the proposed 'point-and-click' on-line agreement that each user must neither attempt nor claim to recognise an individual in the tables, and be linked to more detailed explanation of disclosure control

measures. Such public statistical education should also be directed to media organisations and to politicians.

6. The greatest reason for dropping the requirement to hide '1's from tables is that it cannot achieve the aim of preventing claims of data disclosure. In my experience, many people wrongly believe that census forms are available to government departments; if we were to act on the objective of preventing claims of data disclosure, there would be no census at all.
7. The attempt to avoid *claims* of data disclosure even where this is no *real* data disclosure is a poor approach for a statistical agency whose job is to tell the truth with scientifically collected evidence.

Raising and combining thresholds does not reduce the number of '1's and will significantly reduce the fitness for purpose of census data.

8. The raising and combining of thresholds from 20 households or 50 residents to 40 households and 100 residents is intended by ONS to reduce the occurrence of values of '1' in output tables (before rounding).
9. There has been no evidence from ONS that the higher thresholds decided *will* reduce the occurrence of values of '1'. They will certainly reduce the occurrence of zeros, but may *increase* the occurrence of '1's. This happens in the example I have investigated. The table shows all six of the 933 1991 Census EDs in Bradford District with resident population under 100. The first part of the table shows the number of zeros, '1's and '2's in Table S02, which provides 154 cells cross-tabulating age, sex and marital status. Combining the adjacent EDs in the table produces new areas with greater population and new versions of Table S02 by adding together the cells from each pair of EDs. There are fewer zeros, but *more* '1's and *more* '2's in each table S02 for the three larger areas. This is because the '1's from each small ED were usually in cells where the other ED had zeros.

Occurrence of low values in Table S02: age by sex by marital status

1991 ED code	Residents	'0's	'1's	'2's
08CXGB32	56	83	19	11
08CXGD37	56	89	14	18
08CXGD38	63	68	34	18
08CXGD56	73	103	10	4
08CXFW23	78	63	26	13
08CXFS40	79	103	17	5
Average	67.5	84.8	20.0	11.5

Combine 'adjacent' EDs

GB32+GD37	112	57	17	11
GD38+GD56	136	56	30	19
FW23+FS40	157	46	31	16
Average	135.0	53.0	26.0	15.3

10. The new stipulation that both people and household thresholds must be met will mean that all data for institutions, even the largest ones, will be diluted with residential areas. This will reduce the value of the data significantly.
11. Increased thresholds will make less accurate the creation of data for new areas by aggregation and apportionment of Output Areas. The widespread use of such area aggregation is a key way in which analysts and processors of census data add value to the ONS standard products. This is addressed again below in the context of rounding.

Rounding to three will significantly reduce the fitness of census output for a variety of purposes

12. ONS papers have stated that true values of 0 and multiples of 3 will not be modified, but all other values will independently and randomly be rounded up or down to a multiple of 3, and that there will be no bias in the results. ‘No bias’ means that *on average* the modification due to rounding will be zero.

A rounded value will be modified either by one to the nearest multiple of 3, or by two to the next nearest. For example each occurrence of 2 will be rounded up by 1 or down by 2. If the probability of rounding to the nearest 3 is p , then ‘no bias’ means that $1 \cdot p = 2 \cdot (1-p)$, from which p is derived as $2/3$:

Rounding to the nearest 3 takes place with probability of $2/3$.

Rounding to the next nearest 3 takes place with probability of $1/3$

Although ONS may not wish to confirm or deny these probabilities, these are clearly what is intended, unless ONS is proposing to modify to a multiple of 3 further away than either of the two closest values. In that case all the calculations that follow in this paper are severe underestimates of the impact of rounding.

13. Under the decision made in November, where all output will be rounded, but totals will be independently rounded to avoid the further effect of adding rounded cells, a table may appear as follows (taken from an ONS paper sent to Census user representatives in February 2002, ‘Draft: 2001 Census disclosure control in England and Wales’):

Females	Total	Single, widowed or divorced	Married
20-24	6	3	3
25-29	-	3	-
30-34	6	-	6
35-39	12	-	9

Each value of 3 may hide a true count on the Census output database of a 1, a 2, a 3, a 4, or a 5. Totals will usually not be the same as the sum of their component cells, and may be less as in the case of women aged 25-29 in this example. The calculation of percentages will baffle users, and software will not cope with it without careful reprogramming. Frankly, many users will be baffled by such a table well before they consider calculating percentages.

14. From the rounding probabilities stated earlier, the probability distribution of the impact of rounding for each variable count is: -2 (with probability 1/9), -1 (2/9), 0 (1/3), +1 (2/9), and +2 (1/9). The sum of n counts subject to rounding (including unrounded multiples of three) will be in error by a quantity with variance $4n/3$. By the central limit theorem, as n gets larger the probability of modification in the sum of n counts approaches a gaussian normal distribution with mean 0 and standard deviation $\sqrt{4n/3}$. As the distribution for a single count is already triangular, the approximation is good already with $n = 10$. It can be calculated exactly when n is small by multiplying the probabilities of each possible outcome.

For example, with exact calculations for the sum of 10 counts subject to rounding:

The 95% confidence interval is +/-7.

A modification of 5 or more happens 22% of the time.

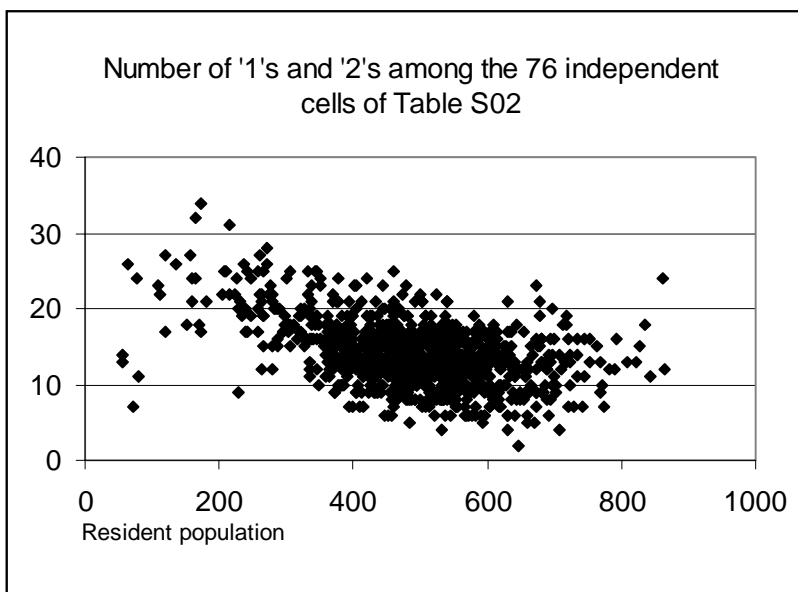
A modification of 8 or more happens 4% of the time.

Although the details of modification in 1991 have not been published by ONS (as they have not for the proposed 2001 rounding), one can take the following as close to the truth from an examination of modified totals with unmodified totals in 1991 output: -1 (1/10), 0 (4/5), and +1 (1/10). In this case, rounding to 3 in 2001 doubles the 1991 confidence interval for the summation of 10 counts, multiplies 100-fold the probability of a modification of 5 or more, and 10,000-fold the probability of a modification of 8 or more. The impact of the proposed rounding is very considerable and will affect the many applications of census data which rely on adding cells within tables and adding areas to make larger areas.

15. Inconsistency *within* tables arises if the totals of tables are rounded independently from their component cells, as planned according to AG(01)08. If instead the totals of tables were not to be rounded but to be derived as the sum of their rounded component cells, then those totals will be subject to the accumulating error described in the previous paragraph. The published data, while consistent within each table, would not be consistent *between* tables. The very first two standard tables for 2001 wards, which have more than 100 cells in each table, would *half* the time differ in their total by more than 16 people. Again, users would be baffled by such large differences, and distrust the data within the tables.
16. The use of OAs as building blocks is a staple of many new geographical systems making excellent use of Census data in central government, local government and commerce. In a similar but less obvious way, migration and commuting data for small areas are also used as building blocks during analysis. This is the case when larger origin-destination areas are specified around the main focus of smaller areas, say wards of a district or County. The data about flows between the larger and smaller areas must be aggregated from the more geographically detailed data.

In recognition that census Output Areas are frequently used as building blocks and that rounding error will be compounded as counts are summed, ONS suggested in their February document that they could consider a service to provide aggregated areas, applying the rounding only after the unrounded counts have been summed. Such a service would have to be free, online, and have a programmable interface for automatic return of results. This I expect would be quite unmanageable for ONS and users.

17. ONS has stated that “rounding will have no impact on the statistical conclusions to be drawn from the data”, in the February paper already referred to. In fact, standard statistics textbooks show that adding random error increases the variation of each variable, and reduces the correlation between variables, with unpredictable effects on statistical modelling.
18. Paper AG(02)02 of March 2002 suggests that ONS is considering rounding only small numbers, perhaps just ‘1’s and ‘2’s, and then deriving sub-totals and totals from the rounded independent cells (those that are not the sum of other cells). Using again the example of Table S02 for the 933 EDs of Bradford, there are on average 14 ‘1’s and ‘2’s, or 18% of the 76 independent cells. The chart below shows that the number of ‘1’s and ‘2’s does not vary much. Even among EDs with large population there are usually between 10 and 20 independent cells that would be rounded in this new plan.



In the 2001 Census output, very many tables have similarly large numbers of independent cells like Table S02 in 1991. The number of ‘1’s and ‘2’s would depend on the distribution of residents across the categories described by each table, but is likely to be similar to that of Table S02. The totals on these tables would be the sum of cells, 10-20 of which had been rounded and thus suffering the errors described above. In fact the impact would be rather more than the error described above, as that included unrounded multiples of three:

For a table in which 10 counts of ‘1’ or ‘2’ were rounded:

A modification of 5 or more happens 32% of the time.

A modification of 8 or more happens 9% of the time.

Two such tables referring to the same population would more often than not be inconsistent by *more* than 4 residents.

Output for migration and commuting statistics, which tend to be more populated by ‘1’s and ‘2’s, would suffer particularly from the impact of rounding.

Thus the new possibility in AG(02)02 of rounding only low value, still considerably damages the data before analyses, and leads to considerable inconsistencies between tables.

The new measures are not based on relevant literature and experience

19. In none of the ONS documents on disclosure control released since November is literature referenced that could support the decisions made then.
20. No evidence of claims of data disclosure have been referred to, to support the decisions made. While the dissemination of Census data freely on the Internet is a new departure, the measures proposed are far too radical to be made on the hunch that a problem might occur in the future. An examination of evidence of such problems in similar situations in the past is necessary.
21. The General Registrar in Scotland plans neither to raise thresholds, nor to combine thresholds, nor to round to three. I am sure he has an equal care for confidentiality of census data and the reputation of the government statistical agency.
22. Britain used to round its census migration data to the nearest 5, but has not done so in the last 20 years. Many other countries do not round their output, and none round in the way decided for England and Wales. ONS representatives refer to New Zealand and Australia, where yourself and your head of Methods and Quality have gained your experience, and to Canada. Canada is an interesting case, since it was their Associate Census Manager for Research and Testing who reviewed UK plans for disclosure control and considered them sufficient before the latest measures were taken. If other countries' experience is to be appealed to then an examination of that experience and its relevance to the UK is required.
23. As the decisions on disclosure control have not been derived scientifically from theory nor from evidence, they should be put to users and gain their acceptance before implementation. This has clearly not been achieved.
24. It does seem clear that the decision to raise thresholds and to round has been a decision based on good judgement made late in the day. Therefore it can be reversed by good judgement.

Alternatives to doubled thresholds and rounding

25. It has been suggested that if data are not rounded then it will be impossible to satisfy demands for non-standard areas, for fear that differencing similar areas will implicitly provide data for areas that do not meet the thresholds, whatever level they are set at. Research at Leeds University has questioned the feasibility of disclosure of confidential information by such differencing. Nonetheless, user representatives have indicated that they believe users would be willing to have non-standard areas computed by sensibly apportioning Output Area data. The apportionment could use the person or address count from the Census database. Alternatively, an output Census database with more severe record-swapping *within* Output Areas would suffice to ensure that small 'differenced areas' did not hold identifiable data. ONS could usefully canvass opinion on these alternatives in their paper later in April.

26. ONS have rightly drawn attention to the new and free availability of census data on the Internet, via the Neighbourhood Statistics site, which users welcome. It has been suggested that this will encourage new and inexperienced users who will be concerned to see '1's in tables for very small areas. To reduce the frequency of '1's without rounding, it is reasonable that users should have to select aggregates of Output Areas (directly, or implicitly by drawing a boundary on a map), and that these aggregates should have a minimum threshold of population considerably higher than that for Output Areas themselves. Tables available in this way on the Internet could be limited to univariate and simpler cross-tabulations. The full data for all Output Areas would then only be directly available to the traditional and new 'bulk' users of the Census output. Such users require detailed data for value-added services and analyses. They can be educated on the nature and most appropriate presentation of the census data.
27. As suggested earlier, education is a key to reducing the risk of false claims about confidentiality. ONS could state clearly at all dissemination points that small adjustments to census data have been made to ensure that no disclosure of personal information is possible. A 'point-and-click' agreement neither to attempt nor to claim to recognise an individual in the tables should precede access to the data, and links to more detailed explanation of disclosure control measures should be easily available.
28. These measures will deal with data disclosure without the damage to the quality of data that is a consequence of raised and combined thresholds and rounded output.

Best wishes

Ludi Simpson

Manchester University Cathie Marsh Centre for Census and Survey Research, Research Fellow; Bradford Council, policy officer responsible for demographic and census analysis

Copy to:

Census Organisations

Andy Teague, Census Division, ONS
 Chris Denham, Census Division, ONS
 John Pullinger, ONS
 Frank Thomas, GROS
 David Orr, GROS
 Robert Beatty, NISRA
 Norman Cavan, Registrar General
 Northern Ireland

Government and Local Government user
 representatives

Barbara Noble, DTLR
 Gillian Goddard, DoH
 Jon McGinty, ONS Neighbourhood
 Statistics
 Rogers Sykes, LGA
 John Hollis, GLA

Business User representatives

Keith Dugmore, DUG
 Barry Leventhal, MRS

Academic sector representatives

Phil Rees, ESRC Census Programme
 Director

Ian Diamond, Chair, ESRC RRB
 Mike Batty, Chair, ESRC/JISC CAC
 Tony Champion, Deputy Chair,
 ESRC/JISC CAC

Keith Cole, Director, CDU, ESRC/JISC
 Census Programme

Justin Hayes, CDU, ESRC/JISC Census
 Programme

[This version has two typographical errors
 corrected in paragraphs 4 and 14]