

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 23<sup>rd</sup> 2002

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

Please add the following comments and analyses to those you have received during consultation on Advisory Group Paper AG(02)03.

I understand that there is a balance to be made between the risks of data disclosure and the damage to the utility of the data. No assessment of the latter risks has been made available by ONS, and the analyses below go some way to fill the gap. They extend to the whole of England and Wales the analyses in my letter to Len Cook of April 9<sup>th</sup> which I have already sent to you in response to your consultation. They also extend the analyses to significance tests. I hope they are useful to your deliberations.

I strongly recommend that ONS replicate these and similar analyses relevant to policy. I expect that the analyses most vulnerable to rounding are those that use migration and workplace data and any analyses treating the Output Area data sets as building blocks for larger areas and discerning the geographical patterns of social change.

Section 1 below shows that raising the population threshold for Output Areas from 50 to 100 will have a perverse impact, *raising* the number of '1's and '2's in tabular output (before rounding or other adjustment).

Section 2 quantifies the impact on a table total of the two options proposed by ONS for eliminating '1's and '2's from Census tabular output. Under Option 2, the inconsistency between equivalent population totals from two typical tables will average more than 14 and frequently exceed 30, when several EDs are summed. For tables with unusually large numbers of cells, the inconsistencies will be greater.

Section 3 shows that the statistical significance of economic and social relationships is seriously changed after rounding. In the example, the difference between White and Other economic activity changes between significance and non-significance for one in six of all the areas compared. This is the case under either of the proposed options. Robin Flowerdew has reported that the fit of a published regression analysis is no longer adequate after random rounding to the nearest three. Other colleagues are undertaking further analyses, but it seems clear that the impact on the utility of census data is serious.

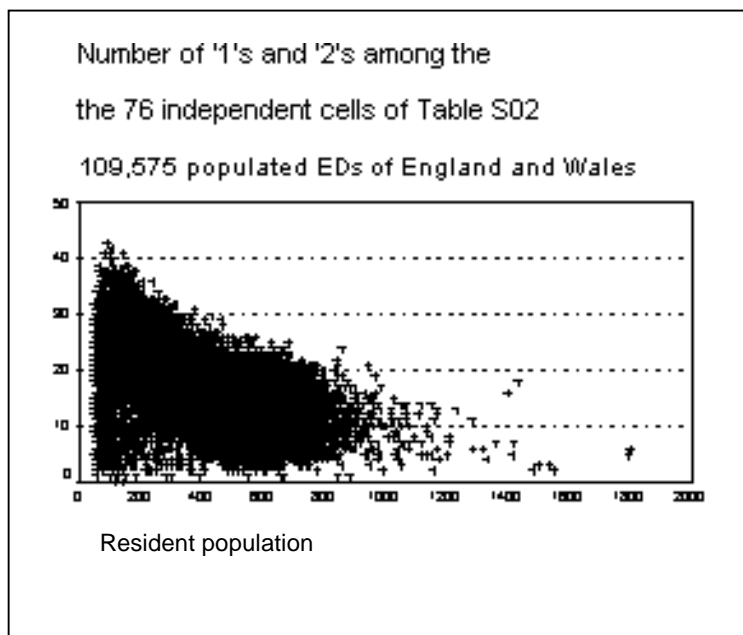
The analyses use two 1991 Census tables for all EDs in England and Wales. Table S02 provides standard data for residents by age, sex and marital status. It has 154 cells, of which 76 are not subtotals or totals, and is used in Sections 1 and 2 below. Table S09 provides employment status counts for each broad ethnic group. It has 54 cells, and is used in section 3.

In sections 2 and 3 I have applied the two options proposed by ONS to 1991 Census data as published for EDs in England and Wales. Option 1 has been confirmed to be as described in my April 9<sup>th</sup> letter. It is a random adjustment of figures that are not multiples of three, the adjustment being to the nearest three with probability 2/3 and to the next nearest three with probability 1/3. For Option 2, I have implemented the minimum adjustment that eliminates '1's and '2's and maintains no bias. This minimum adjustment restricts Option 1 to cell values of 1 and 2. Any other adjustment will induce bias, or a greater disturbance to the data. The impact discussed below is therefore the minimum impact of Option 2.

### 1. The number of '1's and '2's is not reduced by increasing the population threshold for Output Areas

The table and the chart below show the number of '1's and '2's among the 76 independent cells of Table S02 (ie., excluding sub-totals and totals), as it varies with the resident population of the EDs in England in Wales. Zero-population EDs are excluded. The chart has the shape of a shoe that rises at the heel before falling slowly as the ED population rises. EDs with between 100 and 200 residents have *more* '1's and '2's in the basic table of age, sex, and marital status than either EDs with 50-100 residents or larger EDs.

Even with a threshold of several hundred population, there would remain on average between 10 and 20 '1's and '2's among these 76 cells, that would be rounded in either Option. The impact of rounding is not reduced by increasing the population thresholds.



Resident population	No of EDs	Mean number of 1s and 2s
50-	1,735	21.1
100-	2,317	24.2
150-	3,029	23.1
200-	3,691	21.5
250-	4,847	19.8
300-	7,103	17.7
350-	10,158	15.8
400-	14,311	14.2
450-	17,557	13.2
500-	16,833	12.5
550-	13,172	12.1
600-	8,018	11.9
650-	4,007	11.8
700-	1,696	11.8
750-	639	11.6
800-	233	11.5
850-	91	11.0
900-	45	10.5
950-	28	10.2
1000-	17	9.9

## 2. Inconsistency between tables

ONS Option 2, while affecting only small numbers will have a larger impact on table subtotals and totals than option 1 in which the totals are rounded independently from their component cells. I have again used 1991 table S02 on age, sex and marital status as an example. The mean impact on the table total is a maximum of 2 in Option 1, and a maximum of 26 and an average of 4 in Option 2. See the first section of the table below.

Users are particularly concerned about the impact of rounding on aggregates of small areas, which are often used to produce neighbourhood statistics and their analysis. I have simulated such aggregates by adding ED data for 1991 within the 9,512 wards in which they lie. There are between 1 and 63 EDs in each ward of England and Wales, which suitably simulates the variety of aggregates that many users are concerned with.

The second section of the table below shows that the impact of rounding on these aggregates of EDs is to move the total of Table S02 an average of 3 from its true value under Option 1, and an average of 14 from its true value under Option 2. For 10% of aggregates, the impact of rounding will be to change the total population by 30 or more. Under both Options, no bias is introduced.

Table: Impact of rounding on the total population in Table S02

	Min	Max	Mean Bias	Mean impact	90 <sup>th</sup> percentile
(a) 109,575 EDs in England and Wales					
Option 1 Rounding	-2	2	0.00	0.89	2
Option 2 Adjust small numbers	-26	25	0.01	4.28	9
(b) 9,512 aggregates of EDs					
Option 1 Rounding	-19	23	0.02	2.97	6
Option 2 Adjust small numbers	-85	83	0.12	13.76	30

This examination of 1991 Census tables to illustrate the consequences of rounding as proposed for 2001 cannot be directly extended to examine the inconsistency between different tables, because the 1991 tables are already inconsistent due to data modification used then.

However, the inconsistency between two tables which have been independently rounded as proposed will tend to be *larger* than the impact of rounding on either table, as the variance of rounding errors for each table must be added. Tables with unusually large numbers of cells will also have larger errors due to rounding under option 2.

The results above can therefore be taken as the minimum inconsistency between tables.

## 3. Statistical significance would be frequently affected by both rounding Options

Adding random error through rounding or adjusting small numbers will have some impact on statistical analyses. As an example, I have compared the male economic activity of the White ethnic group compared with all other ethnic groups, taken from table S09, and done so using the aggregates of EDs as above, has in the past been done for towns or city communities.

The first table below gives an example to explain the analysis. It shows the data for a single aggregate of Enumeration Districts: the 21 EDs within Eastbrook ward of the London Borough of

Barking and Dagenham. From the 1991 data as released, the economic activity rate is 79%, higher by five percentage points than for other groups combined. Both Options proposed by ONS alter the counts by a little, but only for Option 2 in this example do the changed counts have an impact on the activity rates, slightly decreasing the difference between ethnic groups.

The lower part of the table works through the calculations for the Chi-squared statistic and shows the 'p value', the probability that the observed difference in activity rates could have come by chance from a sample of this size, if the rates were really equal. That probability is only 0.011 for the raw figures. It is slightly higher at 0.016 for the figures after random rounding to three (ONS Option 1), and still higher at 0.069 after adjustment of small numbers and consistent totalling within tables (ONS Option 2).

In this particular case, both Options have altered the figures such that an analysis after the adjustments would find the differences less significant. Option 2 would lead to an assessment that the different activity rates are *not* statistically significant at the 95% confidence level, although the original data would lead to an assessment of statistically significant differences. On other occasions the changes would work in the other direction.

Table: 21 EDs aggregated within ward ABFE, London Borough of Barking and Dagenham

Male economic activity rates	White		Other		White Activity rate	Other Activity rate
	Ec Active	Inactive	Ec Active	Inactive		
Counts from Census table						
As 1991	1,859	490	344	122	79%	74%
Option 1: Rounded	1,854	486	345	120	79%	74%
Option 2: Adjust small numbers	1,859	490	355	116	79%	75%
Expected values if ethnic group and activity are independent						
As 1991	1,838	511	365	101		
Option 1: Rounded	1,834	506	365	100		
Option 2: Adjust small numbers	1,844	505	370	101		
Contributions to chi-squared statistic					Chisq	p
As 1991	0.23	0.84	1.17	4.22	6.469	0.011
Option 1: Rounded	0.21	0.76	1.05	3.80	5.811	0.016
Option 2: Adjust small numbers	0.12	0.43	0.59	2.16	3.302	0.069

The final table repeats the same analysis for all the aggregates of EDs which have expected counts of five or more, sufficient for this statistical test. For one sixth of these four thousand areas, the statistical significance of the difference in activity rates is changed, either to make it significant at the 95% level when it was not before rounding or adjustment, or to make the difference no longer statistically significant when it was previously.

The change is not predominantly in one direction: there is no overall bias. The comparison of economic activity rates involves both subtotals which are affected more by Option 2 than by Option 1, and the summation of ED values of three or more, which are affected more by Option 1 than Option 2.

In this case the impact on statistical significance for comparison of rates is large and similar for either option proposed by ONS. The comparison is not unusual and a similar impact may be expected for other analyses in 2001, should the proposals be implemented.

Table: mean impact on significance of compared male activity rates, aggregate areas with expected counts of 5 or greater.

	Number of aggregates with expected values all >=5	Bias in p value	% in which p value is increased	Mean impact on p value	% in which p value crosses .05
Option 1: Rounded	3953	-0.02	0.442	0.172	0.168
Option 2: Adjust small numbers	3978	-0.02	0.432	0.167	0.165

I can find nothing wrong with my workings and have enclosed with this letter a spreadsheet relevant to Section 3 above. You may be able to check it or do similar calculations with the same or other tables, for which I do not have time at the moment. Duncan Smith at CCSR is attempting to take it further, using Table S09 to look at the impact on different comparisons between male and female activity and unemployment rates.

Yours sincerely

Ludi Simpson