

Working Paper No. 13
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES (EUROSTAT)**

Joint ECE/Eurostat work session on statistical data confidentiality
(Luxembourg, 7-9 April 2003)

Topic (iv): Confidentiality issues for small areas

**NEIGHBOURHOOD STATISTICS IN ENGLAND AND WALES:
DISCLOSURE CONTROL PROBLEMS AND SOLUTIONS**

Invited paper

Submitted by the Office for National Statistics, United Kingdom¹

¹ Prepared by Phil Armitage (phil.armitage@ons.gov.uk) and David Brown (david.brown@ons.gov.uk),
Methodology Group.

Abstract: The Neighbourhood Statistics project has the ambitious aim of providing a detailed profile of every neighbourhood throughout England and Wales. It will collect together information derived from Census and Surveys conducted by the Office for National Statistics and other government departments, and also derived from administrative sources throughout central and local government. The information provided, all on a common geographic basis, will include basic demographic data on births and deaths, various indicators of the health and wealth of the local population, levels of government support (e.g. welfare benefits) provided, educational performance of the area's children, local access to key services, and many other descriptions. Virtually all the outputs provided will be aggregate statistics of some kind. Nevertheless, there are substantial disclosure risks arising from the very small geographical scales involved, and the fact that multiple sets of government data covering very many facets of the local population and environment will be brought together and available for public perusal from a single web site. Many of the datasets will be from administrative sources and will therefore be effectively complete censuses, further increasing the disclosure risk.

There are also potential disclosure control problems arising from the continually changing administrative units for which descriptions are required. In the United Kingdom, the Boundary Commission is continually revising administrative boundaries - which are also often electoral boundaries - to adjust for changes in population. The initial solution proposed for Neighbourhood Statistics is to provide all outputs for standard building blocks related to administrative areas fixed at one date, and only to change the geographical definition of these building blocks very occasionally. All outputs for these building blocks will have adequate disclosure control applied, so that confidentiality of individuals contained within the areas is appropriately protected. If information is required to be aggregated to different boundaries for administrative purposes, then this will be done using some synthetic apportionment method.

Many of the tabulations for which some disclosure risk occurs are frequency tables, and more attention has so far been given to improving disclosure control methods for these types of outputs. Because of the many tables describing the same population with similar classifications (e.g. gender and broad age group), there are many opportunities for linkage between tables that could substantially increase the disclosure risk. This paper will describe methods we have devised for systematically assessing the weaknesses of one standard method - random rounding - for frequency tables arising because of table linkage. Potential problems with this method have been known for some time, but, so far as we know, few systematic procedures are available as practical tools for systematically monitoring and remedying these weaknesses. The paper will also briefly describe the main features of the overall approach to the management of disclosure control in this complex project.

Keywords: Neighbourhoods, frequency tables, random rounding, linked tables

I. INTRODUCTION

1. Within the UK, as in many other countries, there are substantial differences throughout the country in the quality of the areas in which people live. Such differences arise in the physical environment and infrastructure; in the services (medical, educational, etc) provided by public and other bodies; in the opportunities for advancement; in the availability of leisure facilities; and in myriad other aspects e.g. pollution. Often these differences occur from one neighbourhood to another within cities, as well as between cities and regions. The Neighbourhood Statistics project for England and Wales is a government-sponsored attempt to provide detailed information on areas containing quite small populations throughout both countries that can be used to build evidence-based policies to support urban and rural renewal, with the eventual objective of reducing these differences.

2. Substantial disclosure control problems are associated with this project, related partly to the small populations involved, but also to other issues such as the fact that many aspects of each area will be presented in the statistics. One disclosure control tool advocated for use in neighbourhood statistics is random rounding of frequency tables. The paper will describe methods we have devised for

systematically assessing and correcting problems associated with random rounding of frequency tables arising because of table linkage. Potential weaknesses in this methodology have been known for some time, but, so far as we know, few systematic methods are available as practical tools for systematically monitoring and remedying these weaknesses. The longer-term goal is to use controlled rounding with in-built protection levels, as advocated by Fischetti and Salazar (1998, *Journal of Official Statistics*, 14, 553-565). In the meantime, some other methods to reduce disclosure risks are required and these will be described. The paper will also briefly describe the main features of the overall approach to the management of disclosure control in this complex project.

II. THE NEIGHBOURHOOD STATISTICS PROJECT

3. The Neighbourhood Statistics project organised from the Office for National Statistics (ONS) in London grew out of the report of a Government Policy Action Team to provide information on which to build evidence-based policies to remedy deprivation in neighbourhoods throughout the UK. This paper deals with Neighbourhood Statistics in England and Wales; separate but related similar projects are underway in Scotland and Northern Ireland. This project has been driven by the needs of neighbourhood renewal: a concerted effort on the part of central government to identify and remedy inequalities throughout the country at a local level in social and physical environment, resulting in great differences in opportunity for economic advancement, and in quality of life. The project has the ambitious aim of providing a detailed profile of every neighbourhood throughout the country. It will collect together information derived from Census and Surveys conducted by the Office for National Statistics and other government departments, and also derived from administrative sources throughout central and local government. The information provided, all on a common geographic basis, will include basic demographic data on births and deaths, various indicators of the health and wealth of the local population, levels of government support (e.g. welfare benefits) provided, educational performance of the area's children, local access to key services, and many other descriptions. The aim therefore is to obtain a detailed picture at local levels of deprivation of various kinds and inequalities in opportunity and performance. The aim is then also to obtain indicators of resources currently available in each neighbourhood to remedy any deficiencies found.

4. As deprivation and many other relevant characteristics of the population are often very localised, these new statistical descriptions were required for small areas with relatively low populations: the smallest units within the project have target populations of 200-250 people. The basic units with this level of population are so termed *census output areas*. These areas are defined by an algorithm that uses information from the 2001 Census of Population to ensure that, as far as possible, populations within census output areas (i) are as uniform as possible, and (ii) contain approximately the target population size, while (iii) the areas themselves are kept reasonably compact.

5. The heterogeneity of the distribution of population is illustrated in Figure 1, demonstrating the need for information aggregated to a very low level. This example - involving the number of primary school children per sq km - illustrates the typical heterogeneity found for many variables in many areas. It also illustrates that the standard administrative units previously used for many statistical descriptions that include much larger populations, related to electoral and administrative boundaries can sometimes present very misleading pictures of the spatial distributions of some important variables.

6. A further motivation of Neighbourhood Statistics is to monitor local changes on an ongoing basis, so that the success or otherwise of policies to remedy deprivation and inequality can be continuously assessed. This requirement means that the boundaries used should not change from year to year, so that assessments of change can be arrived at that are not confused by changes of boundaries. Obviously boundaries will have to change from time to time when population grows or falls substantially. The current intention is to restrict changes in boundaries to splitting old census output areas into two new ones when population has risen substantially, and to combining two or more old output areas into one, when population has fallen. In this way the damage to a consistent time-series will be minimised.

III. DISCLOSURE RISKS ASSOCIATED WITH NEIGHBOURHOOD STATISTICS

7. Virtually all the outputs provided will be aggregate statistics of some kind. Nevertheless, there are substantial disclosure risks arising from the very small geographical scales involved, and the fact that multiple sets of government data covering very many facets of the local population and environment will be brought together and available for public perusal from a single web site. Many of the datasets will be from administrative sources and will therefore be effectively complete censuses, further increasing the disclosure risk. The risk from this source is not perhaps as serious as it at first sight seems, because of the biases through incomplete coverage, and the different population frames and methods used for some administrative sources. Nevertheless the risks are much greater than when reporting the results of sample surveys.

8. The disclosure risk arising from the small geographical areas involved occurs for two interacting reasons. First of all, because the total populations are small, any tabulations that have useful breakdowns of the main associated variables (age, sex, ethnicity, type of household etc) will frequently contain cells that have zero or very low counts. Thus there will be many instances of potential disclosure, of the types commonly described as group (attribute) disclosure or identification disclosure. This of course could constitute a high disclosure risk, but if the populations covered by the tabulation were only vaguely defined or were dispersed in a manner that it is difficult for a potential intruder to gather information on them, then the risk is lower. In these circumstances, it is difficult to convert a potential disclosure into an actual disclosure.

9. The second reason is that if the tabulation is for a very small population in a geographically compact area, then many individuals will have a sufficiently complete knowledge to convert the potential disclosure into actual disclosure. Also if someone does not have such a knowledge, it is relatively easy to acquire the knowledge by collecting together whatever information about the area is in the public domain, and then supplementing this by visiting the area and making discreet enquiries about the people living there. So by various means, it is much easier for statistics relating to such small populations in compact areas to convert the potential disclosure into an actual disclosure. This means that the disclosure control has to be particularly effective. How to achieve powerful disclosure control using one of the standard techniques for managing group and identification disclosure risk in frequency tables - random rounding - is the subject of the second half of this paper.

10. There are also potential disclosure control problems arising from the continually changing administrative units for which descriptions are required. In England, the Boundary Committee for England and the Boundary Commission are continually revising administrative boundaries - which are also often electoral boundaries - to adjust for changes in population to ensure that different areas have comparable electoral representation. We have alluded above to the need to provide adequate data to monitor changes in time, which could also be seriously damaged by using changing boundaries for the areas covered by the Neighbourhood Statistics tabulations. The initial solution of the changing boundaries problem proposed for Neighbourhood Statistics is to provide all outputs for standard building blocks related to administrative areas fixed at one date, and only to change the geographical definition of these building blocks very occasionally (as discussed above, by splitting one unit into two, or combining units). All outputs for these building blocks will have adequate disclosure control applied, so that confidentiality of individuals contained within the areas is appropriately protected. If information is required to be aggregated to different boundaries for administrative purposes, then this will be done using some synthetic apportionment method. The original data will not be aggregated to different boundaries, and so there will be little opportunity for disclosure by differencing the aggregate statistics for two closely related boundary sets.

11. Many of the tabulations for which some disclosure risk occurs are of frequency tables, and more attention has so far been given to improving disclosure control methods for these types of outputs. Because of the many tables describing the same population with similar classifications (e.g. gender and

broad age group), there are many opportunities for linkage between tables that could substantially increase the disclosure risk.

IV. ASSESSING THE SECURITY OF RANDOM ROUNDING

12. We now describe the approach we have developed for systematically assessing the weaknesses of one standard method - random rounding - for frequency tables arising because of table linkage. Cells of the table are usually rounded to an adjacent multiple of the rounding base using a probabilistic algorithm. Taking as an example random rounding to base 3, the classical version rounds a frequency of one to zero with probability $2/3$, and to 3 with probability $1/3$. 2 would be rounded to zero with probability $1/3$, and to 3 with probability $2/3$. Any figure originally at an integer multiple of 3, including zero, would not change. The pattern is repeated for higher values. Other probabilities could be used, but they are usually chosen so that the perturbations applied in the rounding are unbiased.

13. Potential weaknesses in this method have been known for some time. Interior cells of tables are rounded independently of the margins - a feature not liked by users, that also contributes to a lack of security, in the sense that for some tables the original frequencies can be determined from the rounded ones. The possibility of unpicking the random rounding is simply illustrated by a row of a table consisting of two zeroes that sum to 6 (Table 1a). The intruder is told that conventional random rounding using a rounding base of 3 has been applied. The intruder would therefore know that the original frequencies that could have given rise to the rounded frequencies could have been as given in brackets in Table 1b. We know that the true frequencies in a and b must add up to the total given in the third column. The only way that this can happen is for the true value in a and in b to be 2 and the total to be 4 (Table 1c). Examples of this form abound in the literature, but, so far as we know, no systematic methods are available as practical tools for systematically monitoring and remedying these weaknesses.

Table 1a

a	b	Total
0	0	6

Table 1b

a	b	Total
0	0	6
=	=	=
(0,1,2)	(0,1,2)	(4,5,6,7,8)

Table 1c

a	b	Total
0	0	6
=	=	=
(2)	(2)	(4)

V. A GENERAL METHOD OF AUDITING RANDOM ROUNDING

14. We illustrate a general method using a simple example, involving two one way tables and a common grand total, that have been randomly rounded using base 5, as in Table 2a. The ranges of the original frequencies that could have given rise to these rounded frequencies are given in Table 2b. A maximum and minimum is specified for each frequency.

Table 2a

Under 30	30-60	Over 60	Total
15	15	0	20

Male	Female	Total
10	5	20

Table 2b

Under 30	30-60	Over 60	Total
15	15	0	20
(11 ... 19)	(11 ... 19)	(0 ... 4)	(16 ... 24)

Male	Female	Total
10	5	20
(6 ... 14)	(1 ... 9)	(16 ... 24)

Table 2c

Under 30	30-60	Over 60	Total
15	15	0	20
(11 ... 19)	(11 ... 19)	(0 ... 4)	(22 ... 23)

Male	Female	Total
10	5	20
(6 ... 14)	(1 ... 9)	(22 ... 23)

We first make use of the fact that the original frequencies for each of the one-way tables must add up to the grand total. In particular, a total cannot exceed the sum of the contributing maxima and cannot be less than the sum of the contributing minima. Using the first table we see that the range of the total can be reduced to (22 ... 24), since the sums of the minima and maxima are 22 and 38 respectively. Using the second table in a similar manner we can narrow the range on the total to (22 ... 23). We call this process the *max-min rule*, that we are applying working down from the two 1-way tables to a 0-way marginal.

15. Next we consider what we can infer about the interior cells of the one way tables from the reduced interval on the total. Suppose the three frequencies are a, b, c, and the total is t.

So

$$a = t - b - c$$

Then

$$\max(a) = \max(t - b - c) = \max(t) + \max(-b) + \max(-c) = \max(t) - \min(b) - \min(c).$$

i.e. the maximum of a particular cell equals the maximum of the total minus the sum of the minima of the other cells contributing to the total. So applying this to the 'under 30' cell of Table 2c, we find that the maximum of this cell = $23 - 0 - 11$ i.e. 12. So we can replace the range 11...19 by the new range 11...12 in this cell. We call the rule for this part of the process the *squeeze rule*: we are, in effect, squeezing the intervals on interior cells using the tighter intervals on the margins. This can be applied using the information on any marginal cell to reduce the intervals on cells in any row or column to which the marginal cell is marginal. Similarly the maximum of the '30-60' cell in Table 2c is also 12, and of the 'Over 60' cell is $23 - 11 - 11 = 1$. There is a similar relationship between the minima

$$\min(a) = \min(t) - \max(b) - \max(c).$$

Applying this to the three cells in the age distribution, produces minima of $22-19-4 = -1$ for the first two cells and $22-19-19 = -16$ for the third. Thus they do not improve on what we already knew. So we obtain the revised intervals as in Table 2d. In this case, no further improvement is possible by repeated application of the squeeze rule. In many cases however, the squeeze rule can be repeatedly applied to any row or column, since when one interval narrows, this frequently means that other intervals can be narrowed. This is then continued until no further change occurs in the row or column.

16. The same rules can be applied to the second table, with the results also in Table 2d.

Table 2d

Under 30	30-60	Over 60	Total
15	15	0	20
(11 ... 12)	(11 ... 12)	(0 ... 1)	(22 ... 23)

Male	Female	Total
10	5	20
(13... 14)	(8 9)	(2223)

In this manner, we have reduced the widths of the intervals substantially, although we have not determined any frequency exactly.

17. The procedure can be applied to any linked set of tables. The steps involve the application of the max-min rule building up from the most interior cells of the set of tables, and working to progressively more marginal cells. At each stage it is important to apply the max-min rule to all the cells that add up to each marginal cell, starting with the least marginal cells. In this way, the process can lead to the greatest narrowing on the overall total. Then the squeeze rule is applied progressively in a reverse direction to all less marginal rows and columns. It is important to iterate the whole process until no further changes in the intervals occur, because improvements in one place might make improvements elsewhere possible. Other forms of linkage such as through hierarchical geographies can also of course be included in the process. The squeeze rule is in effect the same as the Buzzigoli-Giusti shuttle algorithm used in determining the interior cells of a table from the margins (Buzzigoli and Giusti, 1999, Cox, 2001).

18. We now use this process routinely in Neighbourhood Statistics to audit any randomly rounded tables to ensure that they are secure against this kind of unpicking. This is sometimes used to check whether further tables can be included in a series of linked tables without compromising the security of the disclosure control. For example, a series of tabulations involving one way tables (O1) age group, (O2) gender and (O3) tenure, with the two way tables (T1) age group by gender might have been proposed for release, and the random rounding audited using the procedure above had determined that these were safe. The agency releasing the tables might then suggest that it would be useful also to release a further two way table (T2) tenure by gender. We would then re-run the random rounding audit to check whether the revised set of linked tables was still safe. In some cases of this kind, we have advised against including new tabulations, because the security of the whole set might be compromised in some areas as a result.

VI. OTHER CONFIDENTIALITY ISSUES IN NEIGHBOURHOOD STATISTICS

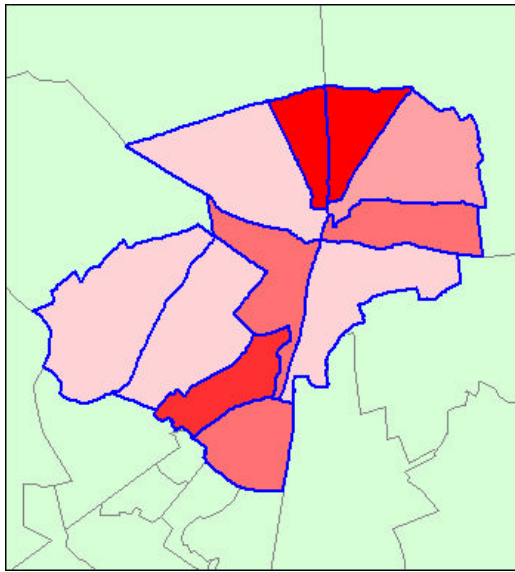
19. Many of the datasets that can be reported at a local geographical level provide information about households and individuals. Business information is generally restricted to small businesses, and as yet, there is little business information reported on the Neighbourhood Statistics web site. Some variates (e.g. mean or median individual income) planned for inclusion on the web site will be essentially small area estimates, and as they involve complex models of the response being reported, will carry little disclosure risk.

20. Reverting to frequency tables, the longer term goal in Neighbourhood Statistics is to use controlled rounding with in-built protection levels, as advocated by Fischetti and Salazar (1998, *Journal of Official Statistics*, 14, 553-565), but this might take some time to implement. This will clearly obviate the need for any auditing of the rounding procedure. We are also investigating methods of disclosure control that might provide some security against disclosure by differencing due to aggregating the same dataset to different boundaries. These methods, involving a pre-tabulation approach, would enable us to depart from the current practice of using synthetic estimation for areas that do not closely match the standard fixed building blocks used for the standard outputs. This work is still in its early stages, and is reported in another paper in this work session (Brown, 2003).

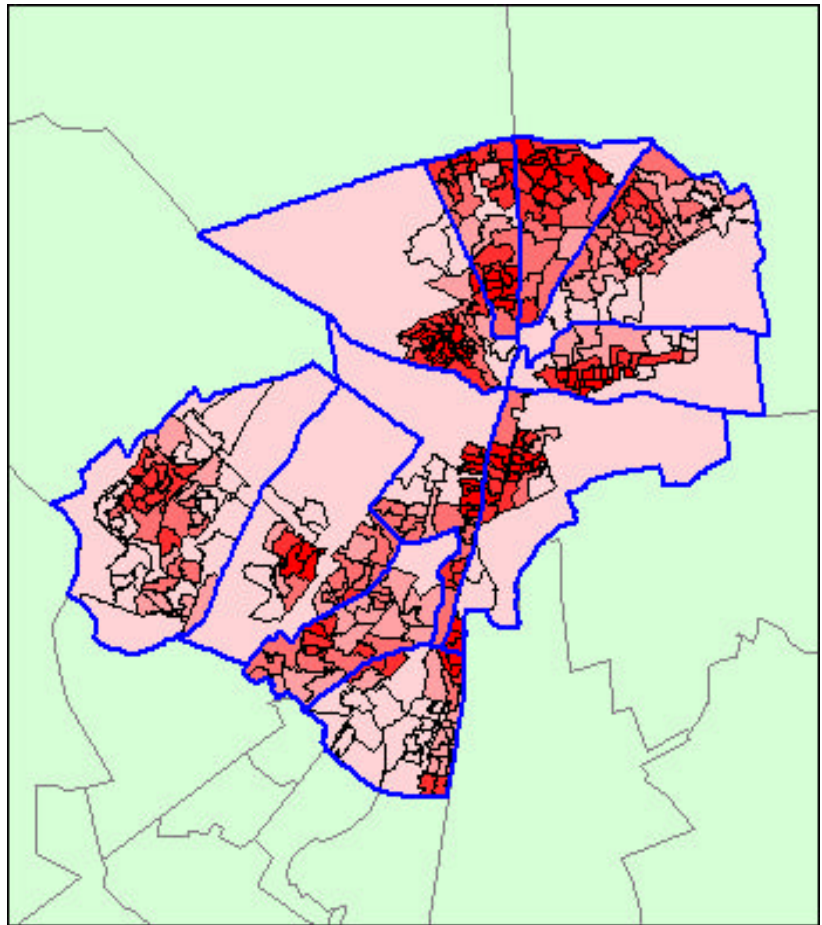
21. Other issues arise because the disclosure control will often need to be exercised at source. The government departments providing the administrative data have the primary responsibility for protecting the confidentiality of the people they report on, and in many cases they are not yet legally empowered to transfer confidential data to other government departments, even for statistical purposes. So they will generally apply disclosure control to their aggregate statistics before transfer to the Neighbourhood Statistics project staff for display on the Neighbourhood Statistics web site. This means that there are substantial coordination problems to ensure that adequate and reasonably uniform standards of data quality etc, and confidentiality protection, are applied. A further complication arises because in some cases knowledge of the parameters used in the disclosure control method sometimes helps an intruder to undo the disclosure protection. For this reason, each agency will apply slightly different confidentiality procedures from the others, using parameters to characterise each variant. Each agency will keep confidential the parameter values used in its confidentiality procedures.

REFERENCES

- Brown D (2003) Different approaches to disclosure control problems associated with geography. Proceedings Joint UNECE/Eurostat WorkSession on Statistical Data Confidentiality (Luxembourg, 7-9 April 2003).
- Buzzigoli L and Giusti A (1999) An algorithm to calculate the lower and upper bounds of an array given its marginals. Statistical Data Protection: Proceedings of the Conference, Luxembourg: Eurostat 131-147
- Cox L (2001) Bounding entries in 3-dimensional contingency tables. Proc AMRADS Workshop, SDC: *From Theory to Practice*, Luxembourg, December 2001.



(a)



(b)

Figure 1. An illustration of the need for information at a very local (census output area) level. The boundaries denoted by the thick (blue) lines in both (a) and (b) are for wards, while the thinner (black) lines within them in (b) denote census output area boundaries in an area near the city of Lincoln in England. The colour coding in both maps is related to the number of primary school children per sq km., the more intense (red) indicating a higher number. It is clear (i) that the wards are extremely heterogeneous, and that the ward picture in (a) is very unrepresentative of the distribution over this area, because of the very substantial variation within the wards. Note in (b) that communities with a high density of school children sometimes cross ward boundaries.