Working Paper No. 14
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

DIFFERENT APPROACHES TO DISCLOSURE CONTROL PROBLEMS ASSOCIATED WITH GEOGRAPHY

Invited paper

Submitted by the Office for National Statistics, United Kingdom[1]

[1] Prepared by David Brown (david.brown@ons.gov.uk).

Abstract: Geography is frequently assigned a special status in disclosure risk evaluations and during development of disclosure control treatments. There are two main reasons for this in the United Kingdom, which apply to other countries to varying extents, depending on how administrative areas are defined, and whether statistics for small areas are routinely produced.

First, there are frequent changes in the administrative boundaries used for production of government statistics in the UK, especially at 'ward' level (including populations of between one and a few thousand people). Administrative authorities always require tabular statistics for the current boundaries for each set of data. But they also need to assess changes in time, so they usually require tabular statistics derived from the same microdata aggregated to both the old and the new boundaries. As the boundaries change gradually from year to year, this frequently results in many cases of disclosure by differencing the statistics aggregated to the two closely related boundary sets. In particular when the area defined by one boundary, say A, is a subset of the other, say A + dA, either for the whole population or a subclass, tabular statistics for the difference dA can be derived. If the disclosure control treatment is one that focusses particularly on low frequencies, then small frequencies (often resulting in potential disclosure) for the small area dA can often be obtained by differencing the two tables.

Secondly, statistical tables are routinely produced in the UK for areas with target populations as low as 200-250 people, what in the UK are called 'census output areas'. This raises problems because low frequencies are frequently obtained, resulting in potential for disclosure if precautions are not taken. It is more difficult to adequately protect such tables using standard disclosure control methods, especially without doing substantial damage to the tables. So there will be more instances of potential disclosure occurring for this reason. Once a potential disclosure has occurred, this disclosure can be more easily converted into an actual disclosure since the area to be searched in in order to identify the potentially disclosive case is frequently a small compact area. It will frequently happen that detailed information on most of the population contained with such a small area will be within a single person's knowledge. Furthermore, in cases where the intruder does not already posses the knowledge, it is a relatively easy matter to collect information about the residents of such an area, merely by visiting the area, and conducting an investigation. So the risks of converting the potential disclosure into an actual disclosure are much greater. Therefore in order to manage the total risk adequately, we need to reduce the incidence of potential disclosure to a much lower level.

There are various possible approaches to solving these disclosure problems relating to geography, two of which are discussed in this paper. One involves providing adequately disclosure protected tabular outputs for a standard set of building blocks, and obtaining *estimates* of the outputs for other areas by synthetic means (using methodology related to small area estimation techniques). The second involves a pretabulation method involving perturbations of the geographical position of sets of households, similar to record swapping applied within the US and UK Censuses. Initial investigations of the method's properties are briefly discussed.

Keywords: record swapping, local geography, small area statistics, pretabulation disclosure control

I. INTRODUCTION: IS GEOGRAPHY A SPECIAL VARIABLE IN DISCLOSURE CONTROL?

1. Over many years, geographical location has been considered an important variable in disclosure control, for various reasons. (a) Knowledge that might be used by an intruder is often geographically clustered, and this is particularly risky when small area statistics are published. (b) Some disclosure control techniques e.g. record swapping, operate entirely by creating uncertainty over location. (c) In some other techniques there are substantial controls exercised over geography (e.g. involving population thresholds in microdata and tabular data release). (d) Confidentiality problems often occur because of the requirement to release statistical outputs for closely related geographical boundaries.

2. Perhaps the most important reason for geography being considered in a special manner in disclosure control is that the knowledge of potential intruders is frequently geographically clustered. Typically people know about the area they live in. They are more likely to have detailed knowledge of all the people who live close to them than all the people who share their occupation, or age group. In towns and villages particularly, it quite often occurs that people have a good knowledge of almost everyone that lives in their neighbourhood. This danger is only realised if statistics for areas containing small populations are released, and so this is a much greater issue in the context of projects like the UK Neighbourhood Statistics that publish statistics for areas containing as few as 200 people.

3. Some pretabulation disclosure control techniques consist solely of providing uncertainty about geographical location, e.g. classical untargetted record swapping as adopted in the UK 2001 Census (supplemented in this case with some post-tabulation disclosure control methods). Untargetted record swapping often involves the following stages: (a) selecting a proportion of households at random from the population (household set A); (b) matching them with other households on two sets of variables: **set 1**: those that are thought very important in themselves whose spatial distributions ideally should not be damaged by the record swapping (e.g. broad age and sex structure of the household) and **set 2**: other variables that might be known to be geographically clustered (household tenure in the UK); (c) exchanging each of the households in set A, with a randomly selected household from those which have been matched using the variables in the two lists above. Alternatively, the method can be thought of as changing the geographical location of the two selected households. The technique operates on the microdata before tabulation, and then, in theory, tabulations can be performed on the modified microdataset without the need for any further disclosure control. In practice, depending on the nature of the tabulations released, sometimes other procedures are required on the outputs after tabulation. Another version, implemented for the US 2000 Census involves targetted record swapping (Zayatz, 2001, 2003). The details are unimportant for our present argument, sufficient to say that it involves differentially selecting for geographical perturbation those records that appear *a priori* likely to be particularly disclosive. The disclosure protection provided in both these versions of record swapping rests on providing some uncertainty about the location of any potentially disclosed households or individuals.

4. Thirdly, in microdata releases (e.g. samples of anonymised microdata from the Census), it is conventional not to provide small scale geographical information. One rule of thumb is not to provide any geographical information about units containing fewer than 100,000 population, i.e. any geographical information is restricted to a relatively high level. In other surveys (e.g. Labour Force Survey microdata), geographical information released is typically restricted to much higher levels of geography, even under special access agreements. In tabular releases, a frequent requirement is that the area for which the aggregate statistics are being disseminated should contain a minimum population e.g. in the UK, any tables for wards containing fewer than 1000 people are suppressed or combined with tables for adjacent wards.

5. There have, however, been trends operating in the opposite direction, i.e. to provide statistics containing more detailed geographical information. One such development is Neighbourhood Statistics in the UK, where the intention is to provide detailed information on many characteristics of small areas so as to provide information that can be used to develop evidence-based policies for the relief of various kinds of deprivation and under-performance. The general policy favoured at present is to strongly target assistance to those areas that specifically require it, rather than using a more general assistance regime that is thought wasteful and less effective overall. The areas of deprivation sometimes involve very small areas and correspondingly low populations in cities, or very low populations in larger areas in more rural environments. That is to say, the pattern of deprivation and relative wealth is a patchy mosaic. If information collected and released were restricted to larger geographical areas, frequently such areas would be heterogeneous. In some areas of London, for example, the very wealthy often live just a street or two away from the very poor. In other cases, the areas of deprivation do not follow the administrative boundaries to which statistics are usually released, but are irregular shaped areas. Furthermore, the areas in which one type of deprivation largely occurs are frequently not the same as those in which other types of deprivation or underperformance occur.

6. A second case when statistics aggregated to closely related but different boundaries are required arises as a result of the continually changing administrative boundaries in the UK, particularly at what is termed ward level (corresponding to populations typically of between one and a few thousands). Administrative authorities always require tabular statistics for the current boundaries for each set of data. But they also need to assess changes in time, so they usually require tabular statistics derived from the same microdata aggregated to both the old and the new boundaries. As the boundaries change gradually from year to year, this frequently results in many cases of disclosure by differencing the statistics aggregated to the two closely related boundary sets.

7. Typically national statistical institutes do not release outputs for areas corresponding to such multiple sets of boundaries without some further treatment to reduce the risk of disclosure by differencing. For example, when the area defined by one boundary, say A, is a subset of the other, say A + dA, either for the whole population or a subclass, tabular statistics for the difference dA can be derived (see Figure 1). If the disclosure control treatment is one that focusses particularly on low frequencies, then small frequencies (often resulting in potential disclosure) for the small area dA can often be accurately provided by differencing the two tables.

II. SOLUTIONS TO THE DISCLOSURE-BY DIFFERENCING PROBLEM FOR RELATED GEOGRAPHICAL BOUNDARIES

8. The most obvious approach to this problem is to examine the difference in the two boundary systems to assess whether disclosure problems are likely to arise. A recent investigation involved comparing two very different boundary sets (the usual administrative boundaries, and a grid square set of boundaries) and concluded that there was little disclosure risk as a result of differencing between the two boundary sets. However, this is rather a special pair of boundary sets, since the boundaries only very occasionally coincide, or almost coincide. The more usual case in the UK involves small changes to administrative boundaries, as just described. In many of these cases, there is no problem as typically the areas defined by the new boundaries do not form strict subsets or supersets of the areas defined by the old boundaries. Usually, a ward will gain some population at one end, and lose it at another. So in most cases, it will not be possible to infer statistics for a small compact area from the difference between the two releases. In some cases though, where population is increasing sharply, the new boundaries will frequently lie entirely within the old, or in the case of rapidly falling population, the old will lie entirely within the new boundaries. But comparison of boundary sets is typically tedious and expensive, and without such checks, there can never be any certainty that no disclosure-by-differencing problem will occur.

9. Another solution is to provide information for low population building blocks that can be assembled flexibly to suit whatever purpose is required. The tables for each building block will have disclosure control treatment applied before release. This solution provides protection against disclosure by differencing by not releasing statistics aggregated to other boundaries, except by synthetic estimation from the standard building blocks. Problems arise when statistics for areas that do not correspond to an integer assembly of these building blocks are required. Various methods of synthetic estimation, with varying degrees of accuracy, are then possible. This is the standard solution adopted for UK Neighbourhood Statistics in the first place, discussed further in section III.

10. An alternative approach to the demands of producing statistics to many different boundaries for different purposes, that we are at present assessing as a potential disclosure control tool, is to make some pretabulation perturbation to the geography, in such a way that release of data to several sets of closely related boundaries is facilitated. The study of this technique has acquired greater focus with the development of geographical tools that make it feasible to convert each postcode at any specified time, or even address, into accurate geographical grid reference positions. The ONS has invested in this methodology for the purposes of improving accuracy, and obtaining greater flexibility in geographical aggregation, subject to overcoming the disclosure control problems. Some initial explorations of techniques based on this principle - still very much in the development phase - are described in section IV.

III. USE OF BUILDING BLOCKS AND SYNTHETIC ESTIMATION

11. The primary geographical system adopted by Neighbourhood Statistics for releases from 2001 onwards will be based on administrative boundaries as defined on the last day of 2002. The reason for this is that frequently it takes some time after collection of data, whether administrative, or resulting from a census or survey, to process the data and check for biases of various kinds, and to produce the disclosure controlled outputs. For example, Population Census data collected in April 2001, and aggregated to local authority level (with typical populations greater than 30,000) has just appeared in mid-February 2003. The decision was therefore taken to use, within the Neighbourhood Statistics system, the last feasible set of available boundaries before the release of the decennial Census of population data, and to release all post-2001, administratively-collected and other data falling within the Neighbourhood Statistics umbrella to these boundaries. All subsequently collected data for some time (i.e. for a number of years) would be released using these end-2002 boundaries to allow for comparisons across time. Any data required for other boundaries - and local authorities will typically require information about their area, and sub-areas (wards, as described above, and at lower levels) - will be obtained synthetically from the outputs for end-2002 boundaries. Thus the standard outputs for local authority, wards and lower areas, each protected to the require disclosure control standard, will effectively become building blocks for the construction of statistics aggregated to all other boundaries as required (Armitage, 2003).

12. A serious problem arises as a result of the mismatch between the new boundary for which statistics are required, and the best approximation consisting of whole building blocks (see Figure 2). How do we then proceed? Various answers can be provided. The statistical agency might release the statistics for the closest approximation it can find utilising just its standard set of building blocks, and refuse to release anything else. If the agency tries to produce outputs nearer to the users' demands, a series of options are available. For example, suppose the aggregate statistic for half a ward is required for the new purpose, how might we obtain it from the whole ward figures. One possibility is to use pro-rata allocation of the whole ward to the two halves, allocating proportionally to population, or number of households or geographical area. However, these approaches do not cope very effectively with heterogeneity within the ward. An alternative approach is to use the types of regression estimator used within the small area statistics field, which also allocates in a pro-rata way, but in effect uses several covariates (population, number of households, area, number of small/large businesses, etc etc) for the pro-rata allocation. This might be developed so as to provide acceptable accuracy for many types of table for most areas. A major disadvantage is that it might require substantial development time, and further resources for implementation of the estimation for any set of boundaries. The reason is that you have to select, estimate and validate the model that will be used for the estimates from a suitable dataset in the first place, then check whether its use in any given case is warranted. However, this might be a feasible solution for a restricted set of datasets and boundaries. It is unlikely that it could be automated so as to operate without substantial human intervention for each new type of data and boundaries.

IV. PERTURBING LOCATION IN THE MICRODATA, INCLUDING LOCAL RECORD SWAPPING

13. The introduction of tools for converting each postcode or household address into a detailed geographical grid reference, so that in theory all households could be accurately plotted as points on a map, presents new challenges and new opportunities in disclosure control. These tools were introduced to allow accurate aggregation to each and every of the myriad boundary sets requested by users. However, as discussed above, unless some method is found for managing the disclosure-by-differencing risk, such use of geo-referencing tools could be severely curtailed. We have already discussed one solution to the disclosure-by-differencing problem - the use of a standard set of building blocks - that does not utilise the full power of the geo-referencing tools. Access to geo-referencing tools also permits new types of analysis - so called spatial analysis in which spatial relationships between events are

modelled and explored - but again without some tools to manage the disclosure risk, such possibilities could only be explored by someone allowed privileged access to the disclosive microdata.

14. However, the new geo-referencing tools present new disclosure control possibilities, two of which we discuss next.

Local perturbation. One possible technique is to randomly perturb the geographical location of a subset of the individual household microdata (for the moment we restrict ourselves to data collected from households, as in a census of population, or in some administrative statistics). This could be done by the application of a random perturbation from a specified distribution (e.g. a bivariate normal distribution, with zero mean and specified variance) to the location of each *selected* household. The perturbed microdataset is then aggregated in the normal way, and the uncertainty in locations if well enough organised can cast enough uncertainty on the aggregate statistics that disclosure risk is much reduced.

Local swapping: swapping with nearest neighbour households. A more interesting alternative is to adopt a variant of local record swapping: that is to exchange the household records of a subset of the data with households nearby. Various swapping schemes could be defined, but we just discuss one for the moment. We initially consider untargetted selection of households for swapping. We first select at random a proportion, p, of households for swapping. For each randomly selected household, we swap it at random with a household randomly selected from its n nearest neighbouring households. To prevent the possibility of households migrating substantial distances as a result of continued swapping, it seems sensible to impose a rule that once a household has been swapped it becomes ineligible for further swaps. This method does not change the spatial distribution of household locations, although of course, it changes the spatial distributions of the characteristics of the households, unless some matching of swapped households is carried out (see 19 below for further discussion).

15. *Concentration of disclosure control protection at the edges of the aggregation area.* We now discuss an interesting feature of any local swapping or perturbation procedure: the disclosure control protection could be said to be focussed particularly on the households at the the edges of the area for which the aggregate statistics are produced. For any household map, and for any specified local swapping rule (that restricts swapping within a given number of nearest neighbours), it is possible to draw a boundary round those households that can only be swapped with other households within the area (area 1, in the interior of the specified area, see Figure 3). By definition, households in the remainder of the area (in area 2, adjacent to the edges of the area) could possibly be swapped with households outside the area. So there is some uncertainty about whether households to be found within area 2 after record swapping actually were within the total area before swapping, whereas those within area 1 after swapping could only have been inside the total area before swapping. This feature of local swapping/perturbation techniques provides protection against disclosure by differencing. Recall that the disclosure by differencing problem occurs because we are differencing two tables for geographically similar areas, the A and A+dA area mentioned above, A being entirely contained, as you would expect from the notation, within A+dA. There is some uncertainty about the original location of households at the edge of each area, then differencing these two tables will provide estimates for the area dA with a relatively large random error affecting it. By contrast, for the area A or A+dA, the random error will be small since only the edge cells will be damaged by potential error, provided the areas are reasonably large and compact.

16. *Discriminate locally unique/rares and wide-area unique/rares***.** Disclosure usually occurs as a result of households that are quite rare with respect to combinations of spanning variables defining the cells of the table, within the area for which statistics are produced. It is useful to distinguish those that would be unique/rare within a large area - much larger than the area for which statistics are being produced - since uncertainty about exact geographical location would provide less disclosure protection for these individuals. Such individuals would be less well protected by local swapping. They might require either swapping over longer distances, or the application of some other perturbation technique, to adequately protect them. On the other hand, local unique/rares are likely to be sufficiently protected by local record swapping. This raises the possibility of using a combination of disclosure control techniques - one geared to the local unique/rares, the other to the wide-area unique/rares. These disclosure control

techniques would typically be applied in addition to initial techniques such as global broad banding, and use of total population thresholds.

17. *Record swapping tends to homogenise spatial distribution*. One major problem with record swapping of any form is that it tends to reduce heterogeneity in spatial distribution of any variable that is not matched in the swapping method. Thus pockets of deprivation or riches will tend to be geographically dispersed to some extent, if variables related to wealth and deprivation are not included in the matching variables. The differences between the extremes will tend to be reduced. Local record swapping will of course induce a smaller amount of spread, that might be more acceptable in some applications.

18. *Other advantages and disadvantages of nearest neighbour swapping.* Nearest neighbour swapping has the advantage that it adapts to the local density of households. Households in thinly populated rural areas will be swapped longer distances on average than those in densely populated urban areas (Figure 4). This is appropriate since, to provide the same degree of protection against disclosure by differencing, it will be necessary to swap rural households over longer distances. There is a potential disadvantage, in that households on the edges of urban areas will be more likely to be swapped with households on the urban side, since there would generally be a greater density of households there, and the majority of the *n* nearest neighbours are therefore more likely to be on that side (Figure 4). If the household in question were genuinely an urban household on the edge of the urban area, this would be useful in that less damage would be done to the outputs if the household were swapped with one with similar characteristics. However, if the household were a rural one just outside the conurbation, this tendency to be preferentially swapped with an urban household might be considered a disadvantage, since more damage would result.

19. *Swapping only with matched households.* One possible answer is to attempt to swap the household only with households that are similar with respect to some variables, one of which could be its urban or rural status. Other variables conventionally used in matched swapping schemes are household size, broad age and sex structure, and household tenure. Household tenure is often controlled in this way, since it is sometimes geographically clustered. For example, moving a household in public sector rented accommodation into an area consisting entirely of privately owned housing could be potentially risky, as such households if detected in a tabulation might thereby be recognised as having been swapped. It is also possible to vary the strictness with which matching is carried out. A balance has to be struck between in effect having very little matching, in which case substantial damage might be done to the outputs, and having too strict a matching process when different problems might arise because matching households cannot be found within a desirable distance. You are then faced with either accepting longer distance moves - and the resulting damage - or not being able to find a household to swap with. Both could be inappropriate.

20. *Targetted swapping.* So far we have mainly discussed untargetted swapping. However, swapping particularly targetted on those households that are likely to be particularly disclosive is likely to provide much greater disclosure control benefit for a given level of damage done to the outputs (see also Zayatz, 2001, 2003). Disclosure often occurs in tabular outputs as a result of households that are quite rare with respect to combinations of spanning variables defining the cells of the table. If these households could be preferentially selected for swapping, then two benefits ensue. The first is that the more-likely-to-be-disclosive households will be moved around, and hence preferentially given disclosure protection. (Although it has also to be said that moving these households around also provides some protection to the households not selected for movement.) The second benefit is that the damage due to disclosure control will be focussed more on those types of households that are relatively rare, and about which the information in the dataset is likely to be poorer quality initially because of their low frequencies. So we are in a sense concentrating the damage on the least reliable data.

21. *What protection is provided?* Once aggregate statistics are calculated from such perturbed/swapped data, the chances of potential disclosures as a result of releasing the table in question will be reduced. Disclosure often occurs in *frequency tables* as a result of some frequencies being known to be exactly zero. Following the geographical perturbation/swapping, it will not be known for certain if a zero in the

table does actually correspond to a real zero in the population within the area specified. A record that was originally within the now-zero cell might have been perturbed outside the area, i.e. the cell might not originally have contained a frequency of zero. The degree of protection depends on the probability that any zero appearing in the table after local perturbation of the microdata does actually correspond to a zero in the original data. More protection than indicated by this is provided however. Even if the zero in the row or column in question is correct, it is possible that the individuals whose responses appear in other elements of the row or column have been swapped into the area from outside, and so do not relate to someone we might identify within the area. Evaluating the disclosure risk in frequency tables is therefore complex, requiring assessment of the effects of geographical perturbation on a number of elements of the frequency tables.

22. Protection of *magnitude tables* is a little easier to analyse. If low non-zero frequencies occur in a cell of a magnitude table after local perturbation, resulting in potential disclosure, we just have to consider the probability that the individuals in these cells have been swapped and therefore do not relate to the area for which statistics are published. Analytical or computer simulation techniques that could be used to obtain answers to these questions will be discussed in a later paper. It is evident however that targetting swapping on individuals in cells with very low counts will substantially improve disclosure protection while minimising the damage.

References

Armitage, P. (2003) 'Neighbourhood Statistics in England and Wales: disclosure control problems and solutions'. Proceedings Joint UNECE/Eurostat WorkSession on Statistical Data Confidentiality (Luxembourg, 7-9 April 2003

Zayatz, L. (2001) 'SDC in the 2000 Decennial Census', Proceedings AMRADS Workshop 'SDC: From Theory to Practice', Luxembourg, 13-14 December, 2001.

Zayatz, L. (2003) 'Disclosure Limitation for Census 2000 Tabular Data', Proceedings Joint UNECE/Eurostat WorkSession on Statistical Data Confidentiality (Luxembourg, 7-9 April 2003

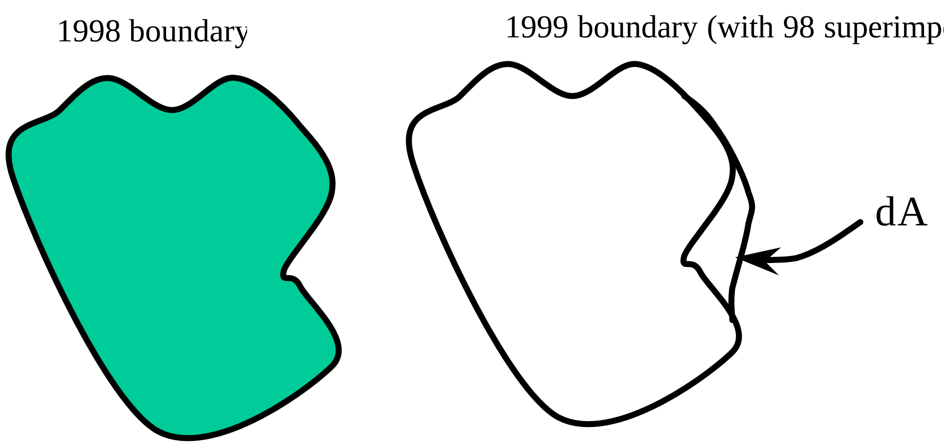

Figure 1. **Disclosure by differencing**. When the same microdata are aggregated to two closely related boundaries, A and A+dA, where one is completely contained within the other, then aggregate statistics can be obtained for the small area dA by differencing the two tables. If disclosure control focusses only on low frequencies, then often tables unaffected by disclosure control treatment can be obtained for the area dA. If these tables then contain low frequencies, disclosure can occur.

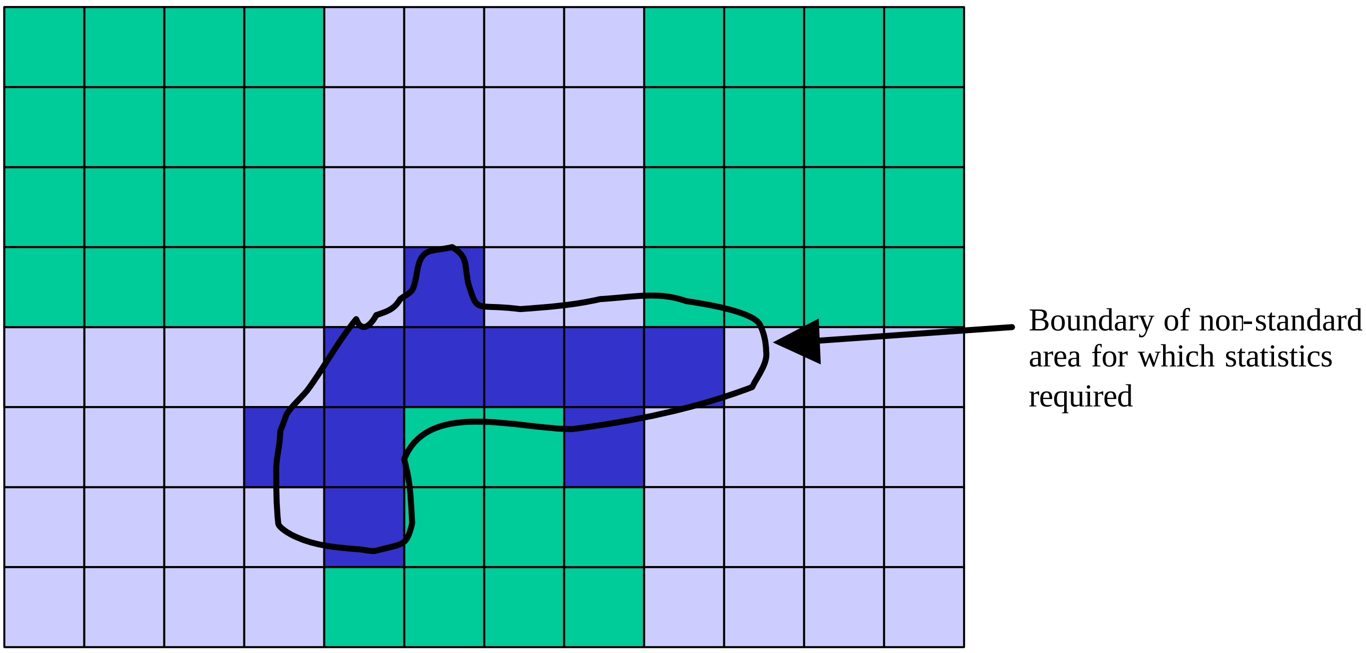

Figure 2. **Synthetic estimation for non-standard boundaries**. The building blocks used for standard releases are displayed schematically in this figure as squares. Three approaches to estimation of the outputs, using only the releases for the building blocks are possible (i) just using the closest approximation using outputs from standard building blocks only (i.e. for the dark blue squares); (ii) apportioning the building block outputs for those parts of the area round the edge of the non-standard area in proportion to population or area or some other single variable, and adding or subtracting these from the aggregate of the building blocks as appropriate; (iii) using regression estimators that effectively allocate in proportion to a number of covariates simultaneously (population, area, no of households...), as used in small area estimation methods, and again adding or subtracting these from the aggregate of the building blocks.

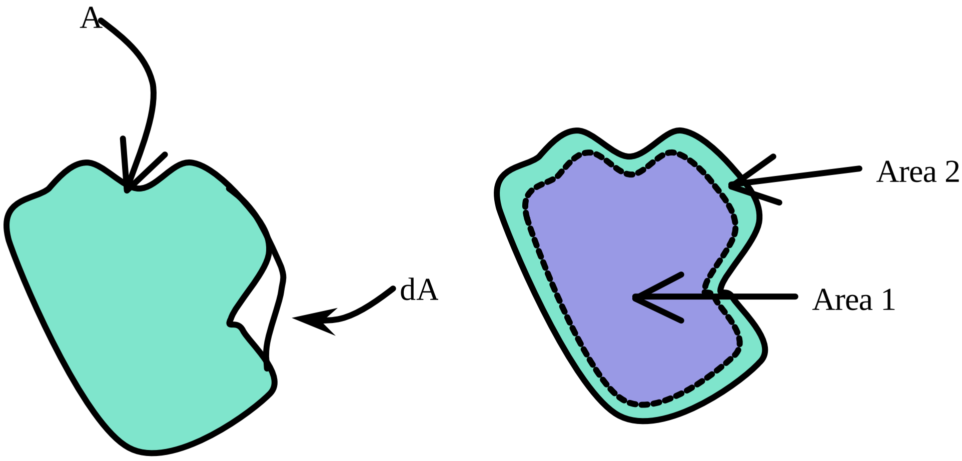

Figure 3. An illustration that the disclosure protection is concentrated around the edges of the area using local record swapping. Area 1 contains those households that can only be swapped with other households within the total area. Households in the remainder of the area (in area 2, adjacent to the edges of the area) could possibly be swapped with households outside the area. The majority of households in the areas A and A+dA will be unaffected by local record swapping, but the majority of households in area dA will be affected. In this manner, protection against disclosure by differencing for small areas comparable to dA will be provided by local record swapping.

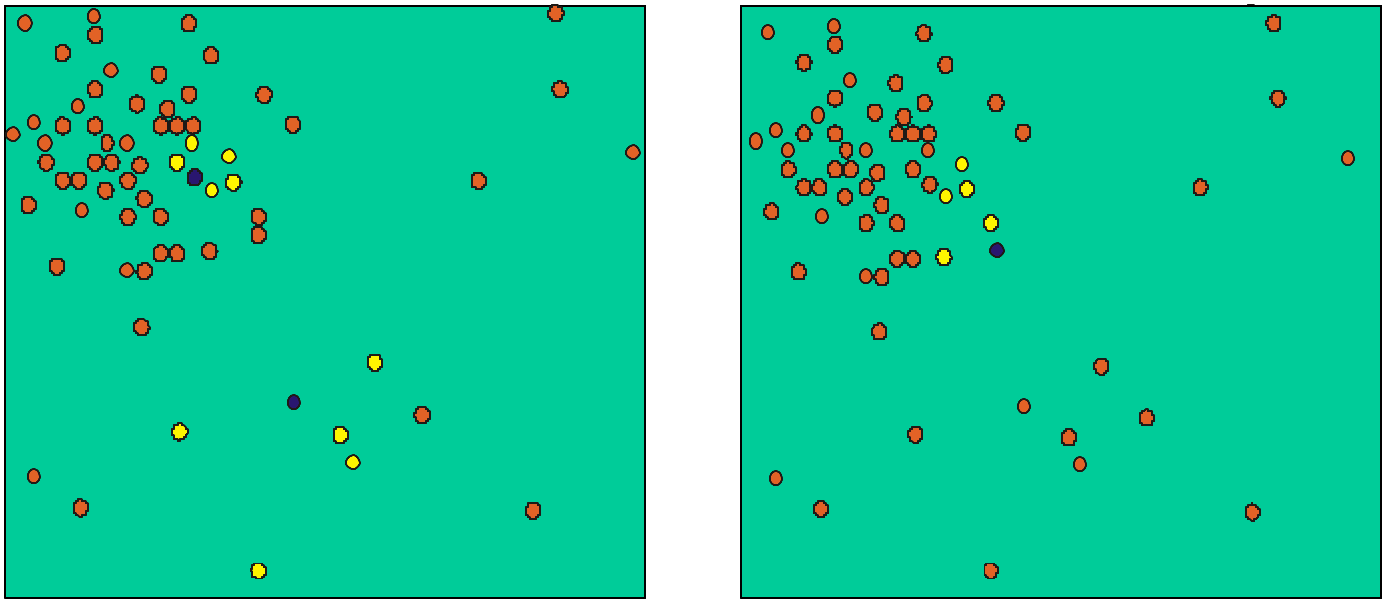

Figure 4. Illustrations of some properties of local record swapping. The blue circles indicate the household selected for swapping, the yellow ones their 5 nearest neighbours. The left plot indicates how rural households are likely to be swapped further than urban households. The right hand plot demonstrates how the nearest neighbours of the household on the edge of a cluster of households (that might constitute a village or part of a town) are more likely to be swapped with households in the cluster.