

Geographically Intelligent Disclosure Control for Census Data

Production

Young, Caroline

University of Southampton, Division of Social Statistics

Highfield, Southampton, SO17 1BJ

E-mail: cjy@soton.ac.uk

Martin, David

University of Southampton, Department of Geography

Highfield, Southampton, SO17 1BJ

E-mail: d.j.martin@soton.ac.uk

Skinner, Chris

University of Southampton, Division of Social Statistics

Highfield, Southampton, SO17 1BJ

E-mail: cjs@socsci.soton.ac.uk

1. Introduction

Disclosure Control of census data is a complex and ongoing concern for many statistical offices. Advances in technology and capabilities for disseminating, processing and analyzing data have led to a new demand on the part of census users for more detailed data at multiple user-defined small area geographies. The growth of digital geographical information allows for the possibility of automatically generating census geographies as required (Openshaw and Rao 1995, Martin 2000). However tables relating to small areas can be highly disclosive due to the increased likelihood of individuals with *unique* characteristics. Moreover tables produced for similar geographies, although independently safe, may be compared with one another in order to reveal new, previously unpublished information, a situation described by Duke-Williams and Rees (1998) as *disclosure by differencing*. The response of the UK statistical organizations has been to publish counts only for hierarchical aggregations of the smallest 2001 census output areas. However there is pressure for the development of a flexible tabulation system in, for example, the UK, Australia and the US. New innovative methods of disclosure control are necessary to meet these disclosure challenges.

2. Statistical Disclosure Control (SDC)

Tables of counts are produced by cross-classification of subsets of the attribute variables. Geographical coordinates (X, Y) are associated with each unit (individual or household), typically by matching to a master address list. The data available for tabulation are compiled into a microdata file, represented by an $N \times (K + 2)$ matrix, Z , where N is the number of units across the whole country and the rows of the matrix contain the values of K attribute variables denoted A_1, A_2, \dots, A_K as well as values of (X, Y) for the different units. We refer to the N units as the population. A basic notion of disclosure is identification, which arises if it is possible for the intruder to establish a one-to-one correspondence between an element of a table and a known unit in the population (Bethlehem et al., 1990). Focusing on cell uniques provides an overall indicator of disclosure risk.

Methods of SDC that are applied to the data before aggregation (pre-tabular methods) can provide protection against the problem of disclosure-by-differencing if the base data, Z is made safe, since then any resulting outputs must also be safe. However, post-tabular methods of SDC (Shlomo, 2005) suffer the problem that each table must be protected individually and checked against all other existing outputs for potential disclosure. This article introduces a new class of pre-tabular methods for disclosure control called *geographical perturbation*. In particular we focus on a new geographical perturbation method called *Local Density Swapping* and compare it to an established method of *Random Record Swapping*. The objective is to provide a new SDC method to minimize the disclosure risk from the above challenges whilst retaining the usefulness and benefits of the data. A summary of a preliminary empirical

analysis is described.

3. Geographical Perturbation

Geographical Perturbation refers to the class of SDC methods that modify the true spatial locations of some or all of the units in the microdata file. Such methods reduce disclosure risk in the aggregated table since it will not be known whether an observed cell unique in a table corresponds to a unit which genuinely falls in that cell or indeed whether the true cell count is one.

Geographical perturbation could be implemented in a number of ways including the established approach of random record swapping (RRS) employed in practice by several census agencies (Shlomo, 2005 and Zayatz, 2006). RRS is *zone-dependent* in the sense that the outcome is determined by the size and shape of the geographical zones (see section 4). Another form of geographical perturbation is *displacement* where the coordinates of each unit are displaced either according to some deterministic rule, for example using rotation or change of scale, or according to a random procedure, for example combining randomly determined distances and directions (Armstrong et al., 1999). This idea is commonly used for protecting confidentiality of locational data (e.g. maps) in the field of geoprivacy but can also be applied to census data. Such an approach may have several advantages in reducing the disclosure risk of census data. However in this article we introduce and concentrate on a new method of geographical perturbation called Local Density Swapping which is zone-independent and a swapping rather than displacement approach. An empirical analysis will illustrate the risk-utility outcome of this new approach to RRS.

4. Random Record Swapping (RRS)

Random record swapping results in the geographical variables of two units i and j being swapped. Thus, geographical coordinates (X_i, Y_i) are exchanged with (X_j, Y_j) but the remaining attribute variables remain unchanged. Different approaches may be adopted for deciding which pairs of units are swapped. Often, pairs of units are matched in some way to limit the potential damage to resulting analyses. Following Willenborg and de Waal (2001) the i th record of the microdata may be partitioned into three sub-vectors: M_i, A_i^* and (X_i, Y_i) . Match variables M_i consist of a subset of the attribute variables to be controlled when seeking recipient households for swapping. Thus, only records j with $M_i = M_j$ are considered for swapping with record i . The specific details of record swapping, such as the proportion of records selected in the sample for swapping, are kept confidential so as not to aid any malicious intruder. The UK 2001 censuses were disclosure-protected by RRS in combination with a post-tabular SDC method. A sample of households was swapped between OAs (output areas) within LADs (local authority districts) (Boyd and Vickers, 1999), that is each record in the sample was paired with a matching record from a different OA within the same LAD. Thus, LAD is effectively a match variable also and this method is zone-dependent, being based on specific LADs and OAs. A key drawback of such RRS is that membership of the same LAD is the only control over the distance and direction of perturbation, which may result in significant damage to statistical analyses involving spatial elements finer than LADs.

5. Local Density Swapping (LDS)

The proposed local density swapping method is designed to overcome this key drawback, by applying pre-tabular perturbation to the microdata in a way which is not dependent upon the choice of output zones. The idea behind LDS is to sample swapping distances from a probability distribution. This allows greater control because the type of distribution can be selected, as can its parameters such as mean and standard deviation. The same distribution could be used uniformly for all records or could differ by region or area type. In addition, a key idea of LDS is to sample household distance rather than Euclidean distance from the distribution. The perturbation distance would therefore represent the number of households between swapped households. This takes into account population density which can be considered a fundamental predictor of disclosure risk; the risk of identifying an individual by differencing depends on the proximity of other individuals with similar characteristics. In general, the more people there are in the area, the more likely it is that someone else will share the same characteristics. In an urban area which is densely populated, less noise is generally needed to disguise the true location of a household. A similar concept has been identified in the geoprivacy literature (Armstrong et al. 1999, VanWey et al, 2005), masking locations of disease or crime on a map in proportion to population density. Density is not taken into account by conventional approaches to

record swapping. Finally we propose to further reduce the risk from geographical differencing by swapping much shorter distances but potentially more records. In practice this could be implemented by swapping broadly at the same scale as the small area geography (for the UK this might be adjacent postcodes for example, or blocks in the US) with variation depending on population density.

6. Evaluating Effectiveness in a Risk-Utility Framework

The effectiveness of the swapping methods can be assessed in a risk-utility framework (Duncan et al: 2001). RRS will be used as a benchmark against which LDS can be compared. We introduce a measure of disclosure risk as the percentage of true uniques at postcode level (representing highly disclosive ‘differenced’ areas). Let z_i^* be a $1 \times R$ subvector of the original $1 \times K$ vector z_i for R selected variable values. Let c be any $1 \times R$ vector of values taken by z_i^* , representing a cell in the table T for a given area formed by the cross-classification of the R variables. Then the frequency for cell c in table T may be expressed as $T(c) = \sum I(z_i^* = c)$, where I is the indicator function. Let $T(c)^o$ denote the cell value in table T before perturbation and $T(c)^p$ the cell value after perturbation. Let n_T denote the number of cells in the table T . Furthermore, let $match = 1$ if $T(c)^o = T(c)^p = 1$ and the same unit appears in cell c before and after perturbation. The probability of finding a true unique is then:

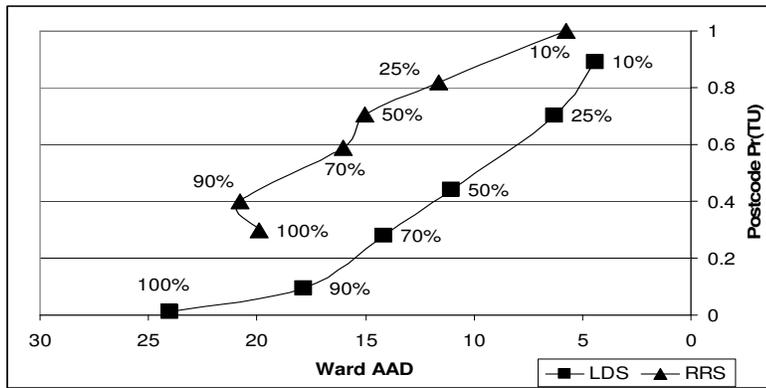
$$\Pr(\text{TU}) = \frac{\sum^{n_T} I(T(c)^o = 1 \ \& \ match = 1)}{\sum^{n_T} I(T(c)^p = 1)}$$

Utility is often measured in terms of distortion to the data. We introduce a measure of ‘dis-utility’ as the absolute average deviation per cell (or AAD) averaged across all tables at ward level.

$$\text{AAD} = \frac{\sum^{n_T} |T(c)^p - T(c)^o|}{n_T}$$

Implementation using a synthetic census-like population (53,680 households) relating to a region in Hampshire produces the results shown in figure 1. An exponential distribution was used for the LDS with mean equivalent to a ‘local’ swap. For both RRS and LDS, similar match variables to those used in the 2001 UK census were used, and the same initial samples for all swapping sizes were used to form one half of the pair of swapped records.

Figure 1: Mapping the Risk-Utility of RRS and LDS for different sampling fractions



Our preliminary analysis in figure 1 indicates that for samples of the same size, LDS results in higher utility at ward level as well as lower levels of risk for differenced areas when compared to RRS. Tests at higher levels of geography (ED, LSOA, ward) showed comparable levels of disclosure risk but it is at postcode level where there are significant improvements from the LDS method.

7. Spatial Measures of Utility

Since swapping distorts the association between geographical and attribute variables, we further assess how geographical areas change relative to one another in terms of their ranking for a particular attribute. This may be considered important to census users who wish to differentiate between proximate areas. Empirical assessment of rank changes for RRS and LDS for Lower Super Output Areas on the synthetic data (LSOAs: a geography independent of

RRS method) was carried out for two different attributes (A) % unemployment and (B) % of male head of households, aged 35-50, in a professional job with a first degree or higher. As with any large mixed urban/rural area, these attributes vary over space. Results showed LDS results in fewer rank changes, potentially indicating that distortion to underlying patterns is generally less than RRS.

Another test that can be performed to assess changes in spatial distributions is to study the effect on clustering/spatial autocorrelation (see Fotheringham et al, 2002). If we know of an attribute which exhibits spatial dependency, we can exploit this relationship to assess the effect of the methods. The global measure of spatial autocorrelation used is the Moran's I: (Moran, 1950) with spatial autocorrelation at a local level measured and mapped by zone, using the LISA statistic (Local Indicators of Spatial Association) in GeoDa. The results showed that both RRS and LDS did not alter the patterns of spatial autocorrelation for the two attributes at the 25% level. However at the 80% level for attribute (A), the original pattern begins to disappear for RRS and LDS decreasing the number of significant areas of spatial autocorrelation. At the 80% there is major damage to the autocorrelation structure for the cross-classification (B). Some elements are retained with LDS, but with RRS many of the significant LSOAs are showing incorrect directions of spatial autocorrelation; an observation also supported by Moran's I. At 25% the spatial correlation is little changed from 0.2640 for the original data for both methods. At 80% this has reduced to 0.0367 for RRS whereas 0.2336 for LDS. This divergence is more apparent for the cross-classification than a single variable.

8. Conclusion

This article has described a new method of geographical perturbation for census data. Empirical results indicate that LDS compares favourably to RRS in terms of risk-utility and has other potential benefits against new disclosure challenges. Further exploratory work will be carried with regard to utility after swapping.

REFERENCES

- Armstrong, M.P., Rushton, G. and Zimmerman, D.L., 1999, Geographically masking health data to preserve confidentiality, *Statistics in Medicine*, 18/5, 497-525
- Bethlehem, J.G., Keller, W.J. and Pannekoek, J., 1990, Disclosure Control of Microdata, *JASA*, 85, 38-45
- Boyd, M. and Vickers, P., 1999, Record swapping – a possible disclosure control approach for the 2001 UK census. UNECE/Eurostat work session on statistical data confidentiality, March 1999, Thessaloniki
- Boyd, M. and Vickers, P. (1999) Record Swapping – A possible disclosure control approach for the 2001 UK Census. Joint ECE/Eurostat work session on statistical data confidentiality, 1999.
- Duke-Williams, O. and Rees, P., 1998, Can Census Offices publish statistics for more than one small area geography? An analysis of the differencing problem in statistical disclosure. *International Journal of Geographical Information Science*, 12, 579-605
- Duncan, G., Keller-McNulty, S. & Stokes, S., 2001, Disclosure Risk vs. Data Utility: the R-U Confidentiality Map. Technical Report LA-UR-01-6428, Statistical Sciences Group, Los Alamos National Lab, N.M.
- Fotheringham, A.S. Brunson, C. and Charlton, M., 2002, *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*, New York: Wiley
- Martin, D. 2000 Towards the geographies of the 2001 UK Census of Population. *Transactions of the Institute of British Geographers*, 25, 321-332
- Openshaw, S. and Rao, L., 1995, Algorithms for reengineering 1991 Census geography, *Environment and Planning A* 27, 425-446
- Shlomo, N., 2005, Assessment of statistical disclosure control methods for the 2001 UK Census. UNECE/Eurostat work session on statistical data confidentiality, 9-11 Nov 2005, Geneva
- Moran, P.A.P., 1950, Notes on continuous stochastic phenomena, *Biometrika*, 37: 17-23
- VanWey, L.K., Rindfuss, R.R., Gutmann, M.P., Entwisle, B., and Balk, D.L., 2005, Confidentiality and spatially explicit data: concerns and challenges. *Proceedings of the National Academy of Sciences of the United States of America: Spatial Demography Special Feature*, 102 (43) Available online at: <http://www.pnas.org/cgi/reprint/102/43/15337>
- Willenborg, L. and De Waal, T., 2001, *Elements of Statistical Disclosure Control*, Lecture Notes in Statistics, 155 (Springer Verlag: New York)
- Zayatz, L., 2006, Disclosure avoidance practices and research at the US Census Bureau: An update. US Census Bureau Research Report Series. Available online at: <http://www.census.gov/srd/papers/pdf/rrs2005-05.pdf>