Working Paper
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Ottawa, Canada, 28-30 October 2013)

Topic (ii): New methods for protection of microdata

Analytical tests of controlled shuffling to protect statistical confidentiality and privacy of a ten per cent household sample of the 2011 census of Ireland for the IPUMS-International database

Prepared by Robert McCaa*, Krish Muralidhar**, Rathindra Sarathy***, Michael Comerford****, and Albert Esteve*****

*Minnesota Population Center, University of Minnesota, Minneapolis, MN 55455 USA, rmccaa@umn.edu
**University of Kentucky, Lexington, KY 40506. krish.muralidhar@uky.edu
***Oklahoma State University. Stillwater, OK 74078 rathin.sarathy@okstate.edu
****School of Computing Science, University of Glasgow, comerm@dcs.gla.ac.uk
*****Centre d'Estudis Demogràfics, Autonomous University of Barcelona, Bellaterra 08193, Spain. aesteve@ced.uab.es

Analytical tests of controlled shuffling to protect statistical confidentiality and privacy of a ten per cent household sample of the 2011 census of Ireland for the IPUMS-International database.

Robert McCaa*, Krish Muralidhar**, Rathindra Sarathy***, Michael Comerford****, and Albert Esteve*****

*Minnesota Population Center, University of Minnesota, Minneapolis, MN 55455 USA, rmccaa@umn.edu
**University of Kentucky, Lexington, KY 40506. krish.muralidhar@uky.edu
***Oklahoma State University. Stillwater, OK 74078 rathin.sarathy@okstate.edu
****School of Computing Science, University of Glasgow, comerm@dcs.gla.ac.uk
*****Centre d'Estudis Demogràfics, Autonomous University of Barcelona, Bellaterra 08193, Spain. aesteve@ced.uab.es

Abstract. IPUMS-International disseminates more than two hundred integrated, confidentialized census microdata samples to thousands of researchers world-wide at no cost. The number of samples is increasing at the rate of several dozen per year, as the process of integrating metadata and microdata is completed. Protecting the statistical confidentiality and privacy of individuals represented in the microdata is a sine qua non of the IPUMS project. For the 2010 round of censuses, even greater protections are required, while researchers are demanding ever higher precision and greater utility. This paper describes a tripartite collaborative experiment using a ten percent household sample of the 2011 census of Ireland to estimate risk, mask the data using controlled shuffling, and assess analytical utility by comparing the masked data against the unprotected source microdata. Controlled shuffling exploits hierarchically ordered coding schemes to protect privacy and enhance utility. With controlled shuffling, the lesson seems to be more detail means less risk and greater utility. Overall, despite substantial perturbations of the masked dataset, we find that data utility is very high and information loss is slight, almost imperceptible even for fairly complex analytical problems..

Acknowledgement. The authors greatly appreciate the cooperation of the Central Statistics Office of Ireland in providing a ten per cent household sample of the 2011 census for this experiment. The authors alone are solely responsible for the contents of this paper. The dataset described here-in is solely for experimentation and, as of this writing, the CSO has not approved its release to third parties.

1 Introduction.

IPUMS-International disseminates integrated, confidentialized census microdata samples to researchers world-wide at no cost. Currently, 238 samples representing 74 countries (544 million person records) are available to more than 7,000 registered users, representing more than one hundred nationalities. Each year the database expands with the addition of samples for the 2010 round of censuses and for more countries, as the task of integrating the microdata and metadata is completed.

Protecting the confidentiality and privacy of individuals represented in the microdata is a sine qua non for the IPUMS project. Access is restricted by means of a rigorous vetting process. Researchers must demonstrate their bona fides, agree to stringent conditions of the user license, and demonstrate a specific research need to be granted access. The microdata are further protected by the fact that researchers do not obtain complete copies of samples, but instead must submit an individual ("extract") request, specifying not only the sample or samples but also the precise variables and even sub-populations required. In other words, each extract is unique, and none is complete. This process of dissemination provides additional safe-guards against researchers sharing data with unauthorized persons.

Technical measures, such as sampling of households, suppression of variables and codes, and swapping of records, are also used to protect the microdata. For the 2010 round of censuses, even greater protections are required due to the explosion of big data, the development of ingenuous techniques of data mining and matching, and the erosion of responsible behavior on the internet. Honesty, trust and professional responsibility continue to be held in highest esteem by all but the tiniest minority of researchers. Nonetheless, census microdata must be protected such that the slightest allegation of violation of confidentiality may be immediately and credibly dealt with.

The threat of de-anonymization in the age of "Big Data" is real. Despite the fact that to gain access to the IPUMS-International database the conditions of use license endorsed by each user expressly prohibits any attempt to identify individuals in the census microdata, strong technical measures must be applied to protect the microdata against even a remote likelihood of re-identification. At the same time we must assure researchers that the microdata are of the highest precision and utility.

The paper describes a tripartite collaborative experiment to estimate risk (Comerford), protect the data using controlled shuffling (Muralidhar and Sarathy), and assess the analytical utility (McCaa and Esteve). The challenge is to assure statistical confidentiality, yet disseminate data of the highest precision and analytical utility. Thanks to the cooperation of the Central Statistical Office of Ireland, a 10% household sample of the 2011 census is being used as a test case. The sample is richly detailed with 474,535 person records, 117,945 families, and 79,785 couples described by 43 variables and more than 1,400 attributes. Person records include variables for single year of age (0-85+), occupation. (number of categories=90), industry (110), country of birth (92), nationality (75), relationship to reference person (12), educational level (7), etc. Before beginning the experiment, we recoded "County of usual residence" (35) into region (8), thereby sacrificing geographical detail to facilitate analysis of social, demographic, cultural and economic attributes.

2 k-anonymity.

A standard approach to the assessment of disclosure risk addresses three key aspects in the literature: the data environment, the sensitivity of the data and the data characteristics. Examples of this type of approach can be seen in (Elliot et al. 2011; Elliot & Dale 1999). In our analysis we interpreted these three aspects in the following ways. The data environment is an attempt to capture information about the world outside of the data under consideration for release. This information is used to demonstrate the a priori knowledge of a would-be intruder and can be configured in a number of ways to simulate different intruder scenarios. In our experiments we wanted to provide a robust analysis and therefore chose a deliberately conservative re-identification key. This was based on the growing concerns about the amount of information publicly available online through social networking sites, e.g., Facebook and LinkedIn. Searching public profiles on LinkedIn using one of our author's names revealed a number of individuals that share a very detailed personal curriculum vitae, without the need for a 'friend request' style level of security.

Extrapolating the information we learned from social media we constructed our conservative key with the following variables from the census sample: sex, age, marital status, nationality, ethnicity, level of education, occupational group, industry classification, region of usual residence, region of birth, country of usual residence and country of birth. This assumes a high level of knowledge for an intruder and should be seen as a worst case scenario.

In this context, 'data sensitivity' means the extent to which the data's subjects might consider the information held in the dataset to represent a threat to their privacy. This is often considered aside from the legal obligations of the data holders. For example, projects like the Scottish Health Informatics Programme (SHIP) use this aspect of the risk assessment to build trust with the data subjects, holding focus groups with patient representatives. For our experiments the data sensitivity contributed to the selection of our test parameters as set out below, taking into account also that we are working with a sample of the population.

The data characteristics take the information gathered from the environment and the data sensitivity and seek to describe the data in an empirical analysis. For this purpose we used k-anonymity, a well-established tool for highlighting re-identification risk. K-anonymity is satisfied if a record is indistinguishable from k-1 other records for a given key. Despite certain criticisms and enhancements k-anonymity still offers a reliable test providing the results are interpreted within the test's definition. For a discussion of k-anonymity see Domingo-Ferrer & Torra (2008). Given our assessment of the data sensitivity in this case, we set the k-anonymity threshold at 3, and the key as referenced above.[1]

1 K-anonymity tests were carried out using the NIAH algorithm available from: https://sourceforge.net/projects/niahsdc/

The first pass of the data, using a k-anonymity threshold of three, flagged 78% of records as not meeting the k-anonymity criteria. This high level was to be expected given such a strong key. This allowed us to look at those records that did meet the criteria and unpick their apparent homogeneity. The results showed that at this level young people made up the bulk of our records meeting the k-anonymity criteria because they share a number of values in our key i.e. they tend not to have been married, they don't work, and they are still in school.

For the second pass of the data we experimented by removing variables from the key to see what effect this would have on the k-anonymity rate. After each k-anonymity test we analysed the remaining risky records to inform the order in which variables could be removed from the key. Once an order was chosen those records that flipped from 'not satisfying' to 'satisfying' k-anonymity were flagged with a dummy variable indicating which variable had affected the change.

We concluded that the variables age, education, occupational group and industry classification followed by the geographical variables should be considered for our later data shuffling experiments.

3 Controlled Data Shuffling to Prevent Disclosure and Preserve Quality.

The purpose of disclosure risk assessment is to identify the extent to which the unmodified release of the data could result in potential re-identification of the records and, possibly, the subsequent disclosure of sensitive information regarding individuals. If the risk of such disclosure is deemed low, then it may be appropriate to allow users to analyse the original data resulting in the highest level of analytical utility. When the risk of disclosure is high, then it may be necessary to modify the data prior to analysis so as to prevent re-identification and disclosure of confidential information. The process of modifying the data prior to allowing access is often referred to as data masking.

There are a wide variety of data masking solutions that are available. At the broadest level, they can be classified as input or output masking. In input masking, the original data is masked and all analyses are performed on the masked data. In output masking, the analyses are performed on the original data and the results of the analyses are masked prior to release. For static data, which includes all the samples integrated into the IPUMS-International database, input masking is generally preferred since it provides the assurance that the results of the same analysis on the same data performed at any point in time will *always* yield the same results. Maintaining consistency at this basic level is crucial to maintain users trust in the validity of the data. Unfortunately, it is extremely difficult (if not practically impossible) to ensure that output masking provides consistent results. Hence, in the remainder of this paper, we limit our discussion to input masking.

There are many input masking techniques that are available. Hundepool et al (2012) provide an excellent discussion of these techniques. Given that we have used k-anonymity to identify risky records, it seems reasonable that input masking through aggregation, simple aggregation for categorical data (Sweeney 2002) and micro-aggregation for numerical data (Domingo-Ferrer amd Mateo-Sanz 2002), would be relevant. Unfortunately, given that close to 80% of the records were identified as being at risk, the level of aggregation that is required in order to prevent disclosure is so high the types of analyses that can be performed on the aggregated data would be severely limited. In order to provide users with greater flexibility in analysing the data, we have to investigate alternative procedures.

Input masking through data perturbation is one approach that can be used in these situations. There are many data perturbation techniques that are available (see Hundepool et al 2012). Most of these techniques rely on modifying the original data through random noise, and the values in the masked data are different from those in the original data. This would be perfectly acceptable for traditional numerical data. The treatment of nominal data is a more difficult problem for data perturbation approaches, and only a few select techniques are capable of perturbing nominal data (see Hundepool et al 2012 for a comprehensive discussion).

Recently Domingo-Ferrer et al (2012) identified the specific problem of taxonomic data, that is, data whose value are nominal but also have a hierarchical structure such as medical diagnosis coded using the International Classification of Diseases (2008). In the Irish data, there are two variables that fall under the category of taxonomic data (Industry classification with 110 hierarchical categories and Occupation group with 90 hierarchical categories). For example, the 90 3-digit occupation groups are divided into 9 1-digit groups. Group 1, “Managers, Directors and Senior Officials”, contains 12 3-digit occupations, while Group 9, “Elementary Occupations”, has only 9. By controlling the shuffling to take into account the hierarchical codes, the perturbed data are more likely to preserve associations with other variables, such as education, industry, and even age.

One approach to handling taxonomic data is to convert them to purely nominal data (by representing every unique code within the taxonomy as a nominal variable). The problem with this approach is that it results in a very large number of nominal variables making it extremely difficult to carry out the perturbation. More importantly, this transformation ignores the inherent taxonomy that is an integral part of the variable. Hence, in the presence of taxonomic data, perturbation approaches that “generate new values” for the original values are not appropriate.

Among data perturbation techniques, there are two techniques that differ from all others in the fact they do not replace the original values with newly generated values, but reassign the original values between records. These two techniques are data swapping (Dalenius and Reiss 1982) and data shuffling (Muralidhar and Sarathy 2006). In data swapping, the values of a variable are exchanged between two records

within a specified proximity. The process will then have to be repeated for every variable that is to be masked. The problem with this approach is that the swapping is performed on a univariate basis and it is difficult to maintain consistent levels of swapping across many variables. Swapping also results in attenuation of the relationship both between the swapped variables and between the swapped and unswapped variables.

Data shuffling, by contrast, is a multivariate procedure where the values of the individual records are reassigned to other records in the data set based on the rank order correlation of the entire data set. One of the key features of data shuffling is that the rank order correlation of the masked data is asymptotically the same as that of the original data. This ensures that all monotonic relationships between the variables are preserved by the shuffling process. When compared to data swapping, data shuffling provides a higher level of utility and lower level of disclosure risk (Muralidhar et al 2006). Data shuffling is capable of handling all types of data. Numerical and ordinal data inherently lend themselves to data shuffling. Nominal data are converted to binary data prior to shuffling. And for taxonomic data, the numerical mapping proposed by Domingo-Ferrer et al (2012) is used.

Data shuffling can be briefly described as follows. Let X represent the set of confidential variables and let S represent the set of non-confidential variables. Let Y represent the masked confidential variables. Data shuffling models the joint distribution of {X, S, Y} as a multivariate normal (Gaussian) copula. Let {X*, S*} represent the normalized values of the {X,S}. The perturbed normalized values Y* are created using the conditional distribution {X*,S*}. Once the values of Y* have been generated in this manner, the original values of X are reverse mapped to Y* to result in the perturbed values Y. For a complete description of data shuffling please refer to Muralidhar and Sarathy (2006).

Data shuffling offers the following advantages:

1. The shuffled values Y have the same marginal distribution as the original values X. Hence, the results of all univariate analyses using Y provide exactly the same results as that using X.
2. The rank order correlation matrix of {Y, S} is asymptotically the same as the rank order correlation matrix of {X, S}. Hence, the results of most multivariate analysis using {Y, S} should asymptotically provide the same results as using {X, S}.

One of the key features of data shuffling is that the process of shuffling is based on joint rank order correlation matrix of *all variables* {X, S, Y}. This provides the data administrator with the ability to control for disclosure risk by specifying the appropriate relationship between the original (X) and masked (Y) variables. This specification can range anywhere from no protection (no shuffling), to maximum protection (where X and Y are conditionally independent given S), and any level in

between. Prior illustrations of data shuffling have used the maximum level of protection. We use the term *controlled data shuffling* to indicate that the desired level of disclosure protection has been specified by the data administrator. This new approach provides a much higher level of flexibility in implementing data shuffling. We now provide the results of implementing data shuffling for the Irish data.

4 Confidentiality protection.

We consider confidentiality protection, taken as a whole, to be strong. Almost 3 of every 4 person records were modified. Age was masked for 50.1%, sex for 13.6% (49% of children aged 0-19), educational attainment 8.1%, industry 13.7%, and occupation 12.4%. One-fifth of the records were modified on two or more variables; one-fourth for those with ages 20 or more. For 80% of adults (aged 20+ years) codes for at least one masked variable are no longer identical to the source microdata. For 25% this was true for two or more variables.

Masking at the individual level is additive for households. Thus for a family with two children, there is a 50% chance that the sex of one was shuffled.

For couples (excluding same-sex unions, which are too few in number to successfully shuffle), joint attributes were taken into account to maintain the associations between characteristics of husbands and wives. Ages of both husbands and wives were perturbed for 50.1% of the 79,785 pairs. Ages of husbands only were masked for 20.7%, and of the wives 20.5%. In sum, age was masked for at least one of the pair in 91.3% of the cases.

5 Analytical utility: 3 tests.

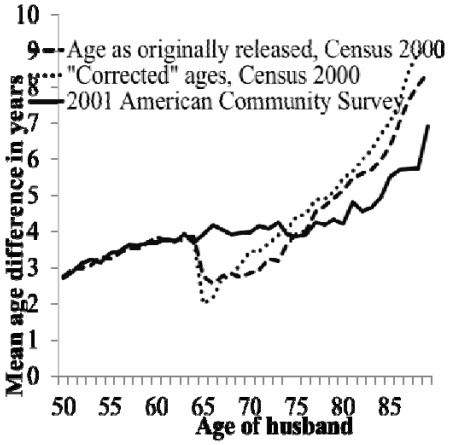
The primary purpose of IPUMS data is to provide researchers with the ability to analyse data from across the globe. Hence, a successful data protection mechanism must ensure that the masked data provides results that are similar to that using the original data. In this section, we provide the results of some ad hoc analyses conducted on the data. One important aspect of this evaluation is that the data was masked without knowledge of the subsequent analyses that will be done on the data. Hence, this evaluation provides a more general assessment of the effectiveness of the masking procedure.

5.1 Age gap between spouses.

For a first test, consider the gap in ages between spouses, a challenging correlation to retain with masked microdata. The sample of the 2000 census of the USA contains a notorious error due to masking of ages for persons 65 years of age and older. Later,

the Census Bureau "corrected" the error, but seemingly worsened the discrepancy (see left panel Figure 1).

USA 2000: perturbation gone wrong



McCaa et al 2012, p. 185.

Ireland 2011: controlled shuffling OK

Calculations by authors.

Fig 1. Masking effects on age gap between spouses: two examples

In contrast, for the 10% household sample of Ireland, comparing the unperturbed and shuffled microdata reveals surprisingly minor discrepancies throughout the age range, despite the fact that in 50% of the cases age was masked for both members of the pair and in 90% for at least one. The age gap between spouses is a strong test of data utility, a test that the shuffled Irish sample readily passes.

5.2 Own-Child fertility.

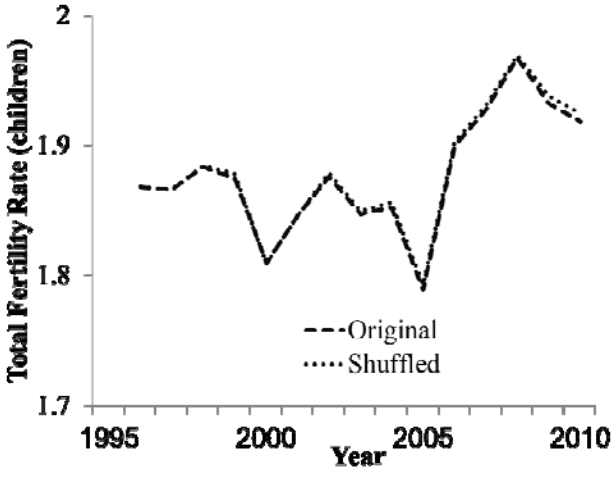


Fig 2. A 15 year series of Total Fertility Rates from a household sample of the 2011 census of Ireland: Shuffled microdata closely track the unperturbed source data.

As a second test, we focus on fertility. Fertility is fundamental for researchers, and population censuses offer valuable insights on fertility levels, trends, and differentials. Where the census lacks a fertility module, estimates can still be derived indirectly (the "own-child method") from household samples, such as the Irish sample for the 2011 census. Children aged 0-14 are matched to their mothers using relationship to head, then a 15 year fertility series is constructed based on the ages of mothers and their children. (The data are adjusted both for children who cannot be matched to mothers and for mortality). A challenging test for masked data is to replicate the age differences between mothers and their children.

Figure 2 shows that the shuffling strategy in this experiment yields astonishingly robust results in spite of the fact that the data was masked without any knowledge that it would be used for fertility analysis. Differences in total fertility rates between the original data and the shuffled are at 3 decimal places, almost imperceptible.

Drilling down to the level of age specific rates, we find that both the unperturbed and masked data reveal declining fertility for ages 15-29 and rising fertility for ages 30-49. While this is not news to experts, what is surprising is that the pattern is unmistakable even in the shuffled data

5.3 Homogamy: educational assortative mating.

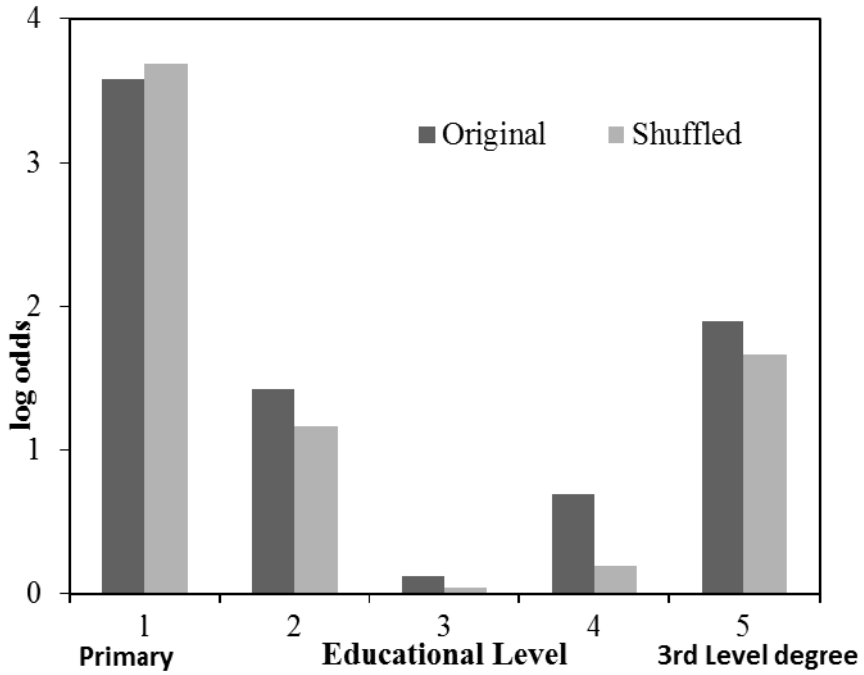


Fig 3. Shuffled microdata approximate the sampled homogamy log-odds, except for educational level 4 (third level – non-degree).

Homogamy (like marries like) is a near universal social rule, and Ireland is no exception. When out-marriage does occur, it is often a matter of hypergamy, highly educated men marrying less educated women. In recent decades, census microdata reveal a reversal—a growing trend toward hypogamy, the propensity of highly educated women marrying less educated men (Esteve et al 2012).

To analyse educational assortative mating using census microdata, the characteristics of married partners must be coupled together and analysed jointly. Indeed the sample of the 2011 census of Ireland shows that women completing the third level degree or higher are twice as likely to marry down as men: 43.5 to 22.3%. The masked data yieldd similar, but slightly dampened propensities—44.8 to 26.5%.

One important feature of the Irish marriage market is the propensity of those with third level non-degrees to marry spouses of the same educational level. We analyzed this propensity for women in the 25-34 age group. According to the original data, there are twice as many marriages of this type (log odds = 0.7), but the shuffled data shows this propensity to be only 1.2 times (log odds = 0.2, see Fig. 3, code 4; n = 122 vs. 93). We are currently investigating the reason for this difference and, if necessary, to modify the shuffling parameters to ensure high analytical validity for all analyses.

When the microdata are re-shuffled for a final round, to minimize this error, joint characteristics should be taken into account. Otherwise homogamy effects are likely to be under-estimated.

6 Conclusions.

Data shuffling is widely recognized as a robust masking procedure for confidentializing microdata. Controlled shuffling allows the data administrator greater flexibility to protect privacy and enhance utility. It also provides the ability to model hierarchically ordered coding schemes which are common in census microdata. The promising results of the experiment may be of interest not only for masking census microdata but for all types of microdata with explicit hierarchical codes, whether international standards such as ISCO, ISIC, NACE, NUTS, etc. or ex post facto integrated codes such as those developed by the IPUMS projects.

Before submitting the masked sample of the 2011 census of Ireland to the CSO for permission to integrate into the IPUMS-International database, the authors plan to:

1. Fine-tune controlled shuffling as follows:
    a. When shuffling sex for unmarried children aged 0-19, take into account educational level
    b. For industry, take into account 23 first level groups instead of only 10
    c. For occupation and industry, maintain the association between these variables as well as with segment, social class code and disability

   d. For educational attainment, take into account the joint characteristics of spouses, and associate with field of study
2. Apply the classic technical protections for all datasets entrusted to IPUMS:
   a. Top/bottom code sparse categories
   b. Convert large households to "group quarters" removing household identities.
   c. Swap a fraction of households across places of residence
3. Take into account criticisms and suggestions from others.

We expect to resolve these matters expeditiously so that the 2011 sample may be integrated into the IPUMS-International database for launch in 2014.

References.

Dalenius, T. and Reiss, S. P. (1982). "Data-swapping: A Technique for Disclosure Control," *Journal of Statistical Planning and Inference* 6 73-85.

Domingo-Ferrer, J. and Mateo-Sanz, J.M. (2002). "Practical data-oriented microaggregation for statistical disclosure control," *IEEE Transactions on Knowledge and Data Engineering*, 14(1):189-201, 2002.

Domingo-Ferrer, J. & Torra, V. (2008). "A critique of k-anonymity and some of its enhancements," In *Availability, Reliability and Security, 2008. ARES 08. Third International Conference*. 990–993. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4529451 [Accessed July 14, 2012].

Domingo-Ferrer, J., K. Muralidhar, K. and Ruffian-Torrell, G. (2012). "Anonymization Methods for Taxonomic Microdata," J. Domingo-Ferrer and I. Tinnirello (Eds.): *Privacy in Statistical Data (PSD 2012)*, LNCS 7556, pp. 90-102, 2012. Springer-Verlag Berlin Heidelberg 2012

Elliot, M. & Dale, A. (1999). "Scenarios of attack: the data intruder's perspective on statistical disclosure risk," *Netherlands Official Statistics*, 14, pp. 6–10, 1999.

Elliot, M. et al. (2004). "Data Environment Analysis and the Key Variable Mapping System," in *Privacy in Statistical Databases*. pp. 138–147. Available at: http://www.springerlink.com/index/6KL805434G016U15.pdf [July 13, 2012].

Esteve, A., Garcia, J., Permanyer, I. (2012). "Union formation implications of the gender gap reversal in education: The end of hypergamy?," *Population and Development Review*, 38(3).

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E., Spicer, K. and de Wolf, P.-P. (2012) *Statistical Disclosure Control*, Wiley Series in Survey Methodology, United Kingdom: John Wiley & Sons.

McCaa, R., Cleveland, L., Ruggles, S. and Sobek, M. (2012). "When Excessive Perturbation Goes Wrong and Why IPUMS-International Relies Instead on Sampling, Suppression, Swapping, and Other Minimally Harmful Methods to Protect Privacy of Census Microdata," in J. Domingo-Ferrer and I. Tinnirello (Eds.): *Privacy in Statistical Data (PSD 2012)*, LNCS 7556, 179-187.

Muralidhar, K. and Sarathy, R. (2006). "Data Shuffling- A New Masking Approach for Numerical Data," *Management Science*, 52(5), 658-670.

Muralidhar, K., Sarathy, R., and Dandekar, R. (2006). "Why Swap when you can Shuffle? A Comparison of the Proximity Swap and the Data Shuffle for Numeric Data," in Domingo-Ferrer and Franconi, Eds.: *Privacy in Statistical Databases (PSD 2006)*, LNCS 4302, 164-176, Springer Verlag, Berlin, 2006.

Sweeney, L. (2001). "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, 10:557–570.

World Health Organization. (2008). *International Classification of Diseases*, 9th Revision, Clinical Modification, Sixth Edition. http://icd9cm.chrisendres.com/