2009 International Symposium on Computing, Communication, and Control (ISCCC 2009)
Proc .of CSIT vol.1 (2011) © (2011) IACSIT Press, Singapore

Statistical Disclosure Control Methods for Microdata

Oleg Chertov [1+] and Anastasiya Pilipyuk [1]

[1] National Technical University "Kyiv Polytechnic Institute", Kyiv, Ukraine

Abstract. In this paper we formulate three basic tasks of statistical disclosure control for microdata, analyze existent methods for achieving optimal ratio between minimal disclosure risk and minimal information loss, and substantiate an availability of masking methods interconnected with microdata wavelet transform.

Keywords: Statistical Disclosure Control, Microdata, Microfile, Masking Methods, Wavelet Transform.

1. Introduction

Last two decades active researches are conducted in an area of data mining, i.e. extracting hidden patterns from data. Let us draw attention on an opposite direction – "disclosure limitation". Classical methods of information encryption or security (organizational, technical, with usual or electronic keys, steganography etc.) are used for making data access hard or impossible. In this paper we consider such transformation of a given to *any* user data sample that saves all main features of this sample, while ensuring a disclosure control of confidential information.

A necessity of such data transformations emerges when users get access not only to analytical results (in form of tables, plots, diagrams etc.), but also to original data. Lately such situation has become more widespread when providing results of various statistical and sociological investigations. A thematic example is the IPUMS-International project. While accomplishing this project data have been gathering from 130 censuses in 44 countries. The project database contains 279 millions personal records [1].

The usual approach for protecting the released data is to distort or to mask them in some way before publication. The methods that attempt to perform such distortion are named as statistical disclosure control methods.

We mark out three actual tasks of statistical disclosure control. The first one is an ensuring of *individual respondent anonymity*. For example, suppose somebody knows only the range of ages, exact amount of children and belonging to top-ranking officers. Then in the dataset it is possible to identify the record which is related to the current President of Ukraine, because he has 5 kids.

The second task is to provide an ensuring of *group anonymity*. For example, it is necessary to exclude the geographical location of cantonments from disclosure by calculating a concentration of military-aged youth.

The third task is an ensuring of *reverse transformation*, i.e. transformation from masking data to original data. It is needed when the reconstruction of primary sample is impossible, for example, because some information has lost its actuality.

2. Problem Definition

[+] Tel.: + 38067-3094309; fax: +38044-2419658
E-mail address: chertov@i.ua.

Microdata are information about an individual respondent, for example, person, household, company. Microfile is a set of microdata reduced to one file, which contains the attributive record for each individual respondent.

By *X* denote an original microdata set. The aim is to release a new (protected) microdata set *X'* in such a way that:

- Disclosure risk is low or at least is adequate to the importance of protected information,
- Results of original and protected microdata set analyses are the same or at least similar,
- Cost of microdata transformation is acceptable.

These requirements are equivalent to an assertion that the information and pecuniary loss during microdata transformation must be acceptable and the level of disclosure risk must be adequate. In other words, the microdata distortion should be small enough to preserve data utility, but it should be sufficient to prevent private information about an individual from being deduced or estimated from the released data. The possible aims of microdata analysis by user are not considered in this paper.

Methods of mapping from the set *X* into the set *X'* have to ensure a solution of statistical disclosure control tasks that were considered in § 1.

3. A Classification of Microdata Attributes

An each record of microdata set contains the same attributes. An attribute value is individual for each respondent. The attributes from an original microdata set can be classified in four categories [2, p. 55-56] which are not necessarily disjoint. See table 1.

TABLE I. A CLASSIFICATION OF MICRODATA ATTRIBUTES

Category	Definition/Characteristic	Disclosure risk	Example
identifier	to identify the respondent unambiguously	unambiguously identification	serial and number of passport, full name, Social Security Number
quasi-identifier	to identify the respondent ambiguously (but in combination it may provide unambiguous identification)	high probability	age, gender, position, home address
confidential attribute	sensitive information on the respondent	high or medium probability	salary, confession, ethnic origin
non-confidential attribute	general information	low probability	language skills

4. Statistical Disclosure Control Concepts

Removing and encrypting are only acceptable ways for disclosure control of identifier values. For another types of attributes we can apply masking methods. They can in turn be divided on two groups depending on their effect on the original data [3].

Perturbative methods. The original microfile is distorted, but in such a way that a difference between values of statistical rates of masking and original data would be acceptable.

Non-perturbative methods. They protect data without altering them. Non-perturbative methods base on principles of suppression and generalization (recoding). Suppression is a removing some data from original set. Recoding is a data enlargement. In case of continuous data (for example, age) non-perturbative methods can transform it to interval, fuzzy or categorical datatype.

Sometimes methods generating synthetic data [4] or combined methods are also applied. Such methods are not acceptable at microdata preparation, because if we don't know the aim of the follow-up analysis, we can not generate adequate set of synthetic data.

5. Perturbative Masking Methods

Table 2 lists perturbative masking methods and gives a short description of them [2, p. 58-63; 5].

TABLE II. MAIN PERTURBATIVE MASKING METHODS

Name	Short description
Data (or rank) swapping	To transform a microfile by exchanging values of confidential attributes among individual records.
Additive noise	To add to the vector for each attribute of the original dataset *X* a vector of normally distributed errors drawn from a random variable with zero mathematical expectation.
Resampling	To build the masked attribute as a *n*-dimensional vector, where *n* is the number of records and each element of the vector is the average value of the corresponding ranked values from some independent samples.
Rounding	To replace original values of attributes with rounded values from a rounding set.
Microaggregation	The set of original records is partitioned into several groups in such way that records in the same group are similar to each other and so that the number of records in each group is at least *k*. For each attribute, the average value over each group is computed and is used to replace each of the original averaged values.
Post-**RA**ndomization **M**ethod (PRAM)	To replace original scores of some categorical attributes for certain records in the original dataset *X* with the scores according to a Markov matrix.
Micro **A**gglomeration, **S**ubstitution, **S**ubsampling and **C**alibration (MASSC)	Combined method with four steps: micro agglomeration that applies to partition the original dataset *X* into groups of records which are at a similar risk of disclosure, optimal probabilistic substitution, optimal probabilistic subsampling, and optimal sampling weight calibration.

6. Non-perturbative masking methods

Table 3 lists main non-perturbative masking methods and gives a short description of them.

TABLE III. MAIN NON-PERTURBATIVE MASKING METHODS

Name	Short description
Global recoding	To form a new (less specific) attribute by combining several (more specific) existing attributes (for categorical attribute). In other words, global recoding is a vertical aggregation. To replace an existing attribute by its discretized version (for continuous attribute).
Top (bottom) coding	To form a new category by gathering values those above (top coding) or below (bottom coding) a certain threshold. In other words, top and/or bottom coding is a horizontal aggregation.
Local suppression	To suppress some values of individual attributes for increasing the set of records agreeing on quasi-identifiers.

7. A comparison of masking methods

Only non-perturbative masking methods can guarantee full individual anonymity, but they do not provide, in general, group anonymity and reverse transformation.

Lately the microaggregation and different variants of this method [6] are applied into practice for providing a group anonymity. These methods guarantee so-called *k*-anonymity, it means that each combination of values of quasi-identifier attributes will correspond to at least *k* records existing in the dataset sharing that combination. In case of one-parameter microaggregation, i.e. for univariate data, a polynomial-time optimal algorithm is given in [7]. But for multivariate records, optimal microaggregation is an NP-hard problem, i.e. a combinatorial problem with non-polynomial estimation of number of iterations [8]. So in practice multivariate microaggregation is used only as an heuristic method. An implementation of the reverse transformation for existent perturbative methods is practically impossible, because requires saving a high volume of temporary values.

8. The microdata wavelet transform

The wavelet transform is a synthesis of ideas emerging over many years from different fields, notably mathematics, physics and engineering. The main area of wavelet transform usage is an analysis and processing of signals that are heterogenous in space and non-stationary in time [9]. A multiresolution analysis is the design method of most of the practically relevant discrete wavelet transforms. A signal can be viewed as composed of a smooth background and fluctuations (details) of it. The distinction between the smooth part and the details is determined by the resolution, that is, by the scale below which the details of a signal cannot be discerned. It is possible to research global data features in large scale representation and to select local features in smaller scales.

In the area of disclosure control of confidential information it is possible to make a data redistribution by the use of the wavelet transform. For example, territorial belonging (or value of any other quasi-identifier or confidential attribute) of respondent can be changed. It provides an ensuring of group anonymity and (partially) of respondent individual anonymity.

The wavelet transform can be referenced to «data exchange» methods of perturbative masking methods. But as opposite, for example, to data swapping method which was applied to protect data for 2001 UK Census [10] or to rank swapping which are applied in μ-Argus [11, p. 16, 61-62] the wavelet transform allows to fulfill the exact renewal of information (perfect reconstruction), i.e. to solve the third selected task without saving any temporary values.

The wavelet transform interconnecting with non-perturbative methods (first of all with global recoding and top/bottom coding) will make it possible to guarantee 100% of individual respondent anonymity.

9. Resume

None of the existing statistical disclosure control methods in microfiles makes it possible to solve all of three tasks selected in the article. It is necessary to apply a number of mutually complementary methods. The interconnecting of wavelet transform with non-perturbative masking methods seems the most perspective one. It is planned to accomplish this interconnecting while preparing microfile with data about the last All-Ukrainian population census.

10.Acknowledgements

We are grateful to Lubov Stelmakh (the State Statistics Committee of Ukraine) for constant attention to this work.

11.References

[1] Minnesota Population Center. Integrated Public Use Microdata Series — International: Version 5.0. Minneapolis: University of Minnesota, 2009 (https://international.ipums.org/international/).

[2] J. Domingo-Ferrer. A Survey of Inference Control Methods for Privacy-Preserving Data Mining. In: C.C. Aggarwal, and P.S. Yu (eds.). *Privacy-Preserving Data Mining: Models and Algorithms*. NY: Springer. 2008, pp. 53-80.

[3] L. Willenborg, and T. de Waal. *Elements of Statistical Disclosure Control. Series: Lecture Notes in Statistics*. V. 155. NY: Springer-Verlag. 2001, 261 p.

[4] J.P. Reiter. Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*. 2002, **18** (4): 531-544.

[5] R. Brand. Microdata protection through noise addition. In: J. Domingo-Ferrer (ed.). *Inference Control in Statistical Databases. Series: Lecture Notes in Computer Science*. V. 2316. NY: Springer. 2002, pp. 97-116.

[6] J. Domingo-Ferrer, F. Sebé, A. Solanas. A polynomial-time approximation to optimal multivariate microaggregation. *Computers and Mathematics with Applications*. 2008, V. 55: 714-732.

[7] S. L. Hansen, and S. Mukherjee. A polynomial algorithm for optimal univariate microaggregation. *IEEE Transactions on Knowledge and Data Engineering*. 2003, **15** (4): 1043-1044.

[8] A. Oganian, and J. Domingo-Ferrer. On the complexity of optimal microaggregation for statistical disclosure control. *Statistical Journal of the United Nations Economic Commission for Europe*. 2001, **18** (4): 345-354.

[9] S. Mallat. *A wavelet tour of signal processing*. NY: Academic Press. 1999, 851p.

[10] M. Boyd, and P. Vickers. Record Swapping – A possible disclosure control approach for the 2001 UK Census. In: *Joint ECE/Eurostat work session on statistical data confidentiality. Working Paper № 26*. Thessaloniki. 1999, 13 p.

[11] *μ-ARGUS version 4.2 Software and User's Manual* / Hundepool A., Van de Wetering A., Ramaswamy R., et al. Voorburg: Statistics Netherlands. 2008, 82 p.