Available online at www.sciencedirect.com
SCIENCE DIRECT®
European Journal of Operational Research 171 (2006) 39–52
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/ejor

Stochastics and Statistics

Minimum-distance controlled perturbation methods for large-scale tabular data protection

Jordi Castro *

Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Pau Gargallo 5, 08028 Barcelona, Spain

Received 3 July 2003; accepted 9 August 2004
Available online 2 November 2004

Abstract

National Statistical Agencies routinely release large amounts of tabular information. Prior to dissemination, tabular data needs to be processed to avoid the disclosure of individual confidential information. One widely used class of methods is based on the modification of the table cells values. However, previous approaches were not able to preserve the values of the marginal cells and the additivity relations for a general table of any dimension, size and structure. This void was recently filled by the controlled tabular adjustment and one of its variants, the quadratic minimum-distance controlled perturbation method. Although independently developed, both approaches rely on the same strategy: given a set of tables to be protected, they find the minimum-distance values to the original cells that make the released information safe. Controlled tabular adjustment uses the L_1 distance; the quadratic minimum-distance variant considers L_2. This work presents both approaches within an unified framework, and includes a new variant based on L_∞. Among other benefits, the unified framework permits the simple comparison of the three distances, and a single general result about their disclosure risk. The three distances are evaluated with the unique standard library for tabular data protection currently available. Some of the complex instances were contributed by National Statistical Agencies, and, therefore, are good representatives of theirs real needs. Unlike alternative methods, the three distances were able to solve all the instances, requiring only few seconds for each of them on a personal computer using a general purpose solver. The results show that this class of methods are an effective and promising tool for the protection of large volumes of tabular data. All the linear and quadratic problems solved in the paper are delivered to the optimization community in MPS format.

Keywords: Statistical confidentiality; Statistical disclosure control; Controlled tabular adjustment, Linear programming; Quadratic programming; Interior-point methods

* Tel.: +34 93 4015854; fax: +34 93 4015855.
E-mail address: jcastro@eio.upc.es

0377-2217/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2004.08.034

1. Introduction

The safe dissemination of data is one of the main concerns of National Statistical Agencies. The released data can be classified as disaggregated or aggregated. Disaggregated data (a.k.a. microdata or microfiles) consists of files of records, each record providing the values for a set of variables of an individual. Aggregated data (a.k.a. tabular data) is obtained from microdata crossing two or more variables, which results in sets of tables with a likely large number of cells. It must be guaranteed, for both types of data, that no individual information can be derived from the released information. The available methods for this purpose belong to the field of statistical disclosure control. Good introductions to the state-of-the-art in this field can be found in the monographs Willenborg and de Waal (2000) and Domingo-Ferrer (2002).

In this paper we focus on tabular data protection. Although each cell of the table shows aggregated information for several individuals, there is a risk of disclosing individual data. This is clearly shown in the example of Fig. 1. Table (a) of that figure gives the average salary for age interval and ZIP code, while table (b) shows the number of individuals for the same variables. If there was only one individual in ZIP code z_2 and age interval 51–55, then any external attacker would know the salary of this single person is 40,000€. For two individuals, any of them could deduce the salary of the other, becoming an internal attacker. Usually, cells showing information about few individuals are considered sensitive, although other rules

		z_1	z_2	
⋮	...	...	...	...
51–55	...	38000€	40000€	...
56–60	...	39000€	42000€	...
⋮	...	...	...	...

(a)

		z_1	z_2	
⋮	...	...	...	...
51–55	...	20	1 or 2	...
56–60	...	30	35	...
⋮	...	...	...	...

(b)

Fig. 1. Example of disclosure in tabular data: (a) average salary per age and ZIP code, (b) number of individuals per age and ZIP code. If there is only one individual in ZIP code z_2 and age interval 51–55, then any external attacker knows the salary of this single person is 40,000€. For two individuals, any of them can deduce the salary of the other, becoming an internal attacker.

can be used in practice. Methods for detecting sensitive cells are out of the scope of this work. A recent discussion about sensitivity rules can be found in Domingo-Ferrer and Torra (2002), and Robertson and Ethier (2002).

Fig. 1 shows a two-dimensional example. This can be considered the simplest case. However, in practice we must deal with more complex situations, including multidimensional, hierarchical and linked tables. Multidimensional tables are obtained crossing more than two variables, and they can be individually protected. Hierarchical tables are sets of tables whose variables have a hierarchical relation (e.g., ZIP code and city). In that case, the total or marginal cells of some tables are internal ones for the others. They have to be protected together, to avoid the disclosure of sensitive data. Finally, linked tables are a generalization of the previous situation, where several tables are made from the same microdata, thus sharing information or cells, either hierarchical or not. Again, they have to be protected together. Linked tables can deal with any table dimension, size and structure, and thus include the other situations. Dealing with linked tables is a desired feature of any tabular protection method. Eventually, the final goal would be the protection of the whole set of linked tables that can be produced from some microfiles (e.g., a population census). Clearly, the number of cells involved in that case might be of several millions, an impractical size for most current tabular protection techniques. The family of protection methods considered in this work deal with linked tables, and, as shown in the computational results, can solve real-world large instances in few seconds. All the above situations can both refer to frequency tables (i.e., cell values are integer and are usually associated to the number of individuals in that cell) or magnitude tables (i.e., cell values are real, and, for instance, they show the mean for some other variable of all the individuals in that cell). In this work we focus on tables of magnitudes. For tables of frequencies the procedures here described can also be applied followed by some heuristic post-process.

Current methods for tabular data protection can be classified as perturbative (they change the cell values) or nonperturbative (no change is per-