WP.24
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 UNECE/Eurostat work session on statistical data confidentiality
(Manchester, United Kingdom, 17-19 December 2007)

Topic (ii): Tabular data protection

INTEGER ROUNDING VERSUS CONTINUOUS ADJUSTMENT FOR TABULAR DATA

Supporting Paper

Prepared by Juan-José Salazar-González, University of La Laguna, Spain

Integer Rounding versus Continuous Adjustment for Tabular Data

Juan-José Salazar-González*

* DEIOC, University of La Laguna, Tenerife, Spain. (jjsalaza@ull.es)

Abstract. Controlled Rounding and Adjustment are two perturbation techniques in fashion as an alternative procedure for guaranteeing confidentiality during tabular data publication. To apply each technique on large data, automatic tools based on modern optimization procedures are necessary. In this paper we discuss the advantage and disadvantage inherent to mathematical models for both techniques.

1 Introduction

Statistical agencies collect data to make reliable information available to the public. This information is typically made available in the form of tabular data (i.e., a table), defined by cross-classification of a small number of variables. A fundamental characteristic of all tables is the existence of mathematical equations. Each equation says that a cell value (marginal cell) is identical to the sum of other values (internal cells). Depending on the size and structure of the table, the set of equations may create a complex linear system of equations, which may have a negative impact when applying some methodologies to protect private information. See Salazar [4] for a survey of articles concerning approaches for protecting tables.

Controlled Rounding consists of replacing each cell value by a multiple of a prespecified base number (e.g. 5). There are several variants of the methods, but the better accepted is the one where

1. original cell values which are multiple of the base number must remain unchanged;
2. other cell values must be replaced either by the minimum multiple which is larger or equal to the original value, or the maximum multiple which is smaller or equal to the original value;
3. the modified table must satisfy the same system of linear equations as the original table.

Figure 1 shows an example of unrounded table, which in Figure 2 has been rounded using base number 5. When the table structure satisfies some conditions (e.g., the

Unrounded data	total	male	female	young	adult	thin	fat
North East	60593	29225	31368	13856	46737	34565	26028
North West	174414	78129	96285	25673	148741	3432	170982
Yorkshire and Humberside	108769	46119	62650	2342	106427	32223	76546
East Midlands	93346	43201	50145	23443	69903	23434	69912
West Midlands	131817	61046	70771	23878	107939	432	131385
East	107060	47376	59684	24532	82528	34233	72827
London	110811	49053	61758	17635	93176	3423	107388
South East	123359	50949	72410	34223	89136	4567	118792
South West	119863	44718	75145	35980	83883	56356	63507
England	1030032	449816	580216	201562	828470	192665	837367
Wales	95388	49579	45809	34989	60399	6454	88934
Scotland	124678	61327	63351	36789	87889	5643	119035
Great Britain	1250098	560722	689376	273340	976758	204762	1045336

Figure 1: Original (unprotected) table.

Rounded data (base=5)	total	male	female	young	adult	thin	fat
North East	60595	29225	31370	13855	46740	34565	26030
North West	174415	78130	96285	25675	148740	3430	170985
Yorkshire and Humberside	108770	46120	62650	2340	106430	32225	76545
East Midlands	93345	43200	50145	23445	69900	23435	69910
West Midlands	131815	61045	70770	23875	107940	430	131385
East	107060	47375	59685	24530	82530	34235	72825
London	110810	49055	61755	17635	93175	3420	107390
South East	123360	50950	72410	34225	89135	4570	118790
South West	119860	44715	75145	35980	83880	56355	63505
England	1030030	449815	580215	201560	828470	192665	837365
Wales	95390	49580	45810	34990	60400	6455	88935
Scotland	124675	61325	63350	36790	87885	5640	119035
Great Britain	1250095	560720	689375	273340	976755	204760	1045335

Figure 2: Modified (protected) table.

cells can be represented by arcs in a network) a modified table exists, and it is known how to find a closest one to the original table with an efficient approach (e.g., a min-cost flow algorithm). However, for a general structure the problem of finding a modified table may be infeasible and some variants have been proposed in the literature (see, e.g., Salazar [3]). The better accepted variant relax the conditions (1) and (2), so a modified value is not necessary an adjacent multiple to the original value. Finding a solution to this extended model implies solving an Integer Linear Programming model, which is known to be be (in general) a complex mathematical problem (e.g., $\mathcal{NP}$-hard in Complexity Theory). This classification means that there are examples of tables where it is very difficult to find a modified table, and this has motivated the research of alternative methodologies.

Tabular Adjustment is an alternative approach to Controlled Rounding. It was originally proposed by Dandekar and Cox [1], and it consists of

- deciding whether each sensitive cell value should be rounded up or down;
- determining the continuous value for each non-sensitive cell value.

A mathematical formulation to find a solution also contains integer variables, but only for the sensitive cells. The non-sensitive cells are associated to continuous variables, which leads to a Mixed Integer Programming model. The problem of finding a Tabular Adjustment solution is again $\mathcal{NP}$-hard, but in practice it is much easier than a problem of finding a Controlled Rounding solution because the number of integer variables is smaller. Note that solving a mathematical problem with only continuous variables is easy ($\mathcal{P}$ in Complexity Theory). Other similar methods have also been proposed in the literature (see, e.g., Cell Perturbation in Salazar [3]), based on a different understanding of protection, but exploiting the advantage of simplifying the problem resolution by having continuous mathematical variables instead of integer mathematical variables.

Although replacing some integer variables by continuous variables may help to solve in practice a model, this paper points out some disadvantages that one should have in mind when replacing Controlled Rounding by a Continuous Adjustment.

2 Linear-programming relaxations

To illustrate some negative consequences of using continuous variables instead of integer variables, let us analyze the following mathematical problem:

$$\begin{aligned} \min\ x_0 & \\ 75000\, x_0 &= 75001\, x_1 + 75002\, x_2 \\ & x_0 \geq 1, x_1 \geq 0, x_2 \geq 0 \\ & x_0 \in \mathbb{Z}, x_1 \in \mathbb{Z}, x_2 \in \mathbb{Z} \end{aligned}$$

This is an artificial simple example with three cells and one linear equation. It does not correspond to any table in practice, but the small size will help us to make clear the main observation of this paper.

Solving the integer mathematical model is difficult in practice. Indeed, using the best commercial solver (like Cplex or Xpress) will take more than one hour on a modern personal computer before finding an optimal solution. The difficulty of this problem is clearly not on the size, but on the integrability of the variables. If the integer variables are replaced by continuous variables then the problem becomes trivial:

$$x_0 = 1 \qquad , \qquad x_1 = \frac{75000}{75001} \qquad , \qquad x_2 = 0.$$

There is no need of a sophisticated solver for finding this trivial solution. However, when the variables must be integer, then a sophisticated solver is fundamental, and using this solver we will find:

$$x_0 = 37502 \qquad , \qquad x_1 = 2 \qquad , \qquad x_2 = 37499.$$

The immediate conclusion when comparing the two solutions is that both can be very far one from the other. Indeed, in theory, for any large number M, it is possible to design an instance where the integer and the continuous solutions are farther than M. The above example shows that this situation may happen in practice, even with tiny numbers of cells and equations.

3 Conclusion

The previous section has shown that the solution of the Linear Programming relaxation of an integer program may be very different from its integer solution. Then, using alternative methodologies where integer variables are replaced by continuous variables may create easier-to-solve models but wrong-to-use solutions.

In addition, it is also obvious that continuous variables contain decimal part. This is a serious drawback for protecting tables with frequency data, but also with magnitude values due to numerical errors during the computation and displaying. In fact, when using magnitude data one does not want to public cell values like 345.0000001, and the simple task of eliminating the decimal part of the numbers (i.e., rounding) may create non-additive tables. Also, if one wants to display the continuous solution of the above optimization problem with only four decimals, number 75000/75001 = 0.9999866668... would be replaced by 1.0000, thus leading to the non-additive solution

$$x_0 = 1.0000 \qquad , \qquad x_1 = 1.0000 \qquad , \qquad x_2 = 0.0000.$$

Hence, after applying a methodology (like Tabular Adjustment), the Controlled Rounding is mandatory unless we have been *lucky* with the original table and the

modified values (continuous numbers) are suitable to be published as they come out from the methodology.

The use of continuous variables in a methodology (as in Controlled Rounding) may reduce the computational complexity of finding a solution, but depending on the table itself the found solution may contains fractional values with a significant decimal part. The finite precision of computers and the necessary truncation of decimals during the publication phase require the use of Controlled Rounding to guarantee additivity. In other words, *only Controlled Rounding guarantees that the modified table satisfies the same linear equations as the original table.*

A final remark is that the above example belongs to a class of optimization problems (with one equation, no matter the number of variables) which can be solved in a very efficient way by using dynamic programming. For solving instances of this class a general-purpose commercial software (like Cplex or Xpress) is not convenient. Indeed, it takes less than one second to solve the above instances by dynamic programming on a computer. This positive result is the outcome of a research work done on the model, and remark also the importance of analyzing mathematical models instead of using a commercial software as a black-box solver. In other words, the fact of having a mathematical model for a new disclosure limitation methodology is not the end of a research line, as it may be of interest to study ad-hoc approaches to solve it.

References

[1] Dandekar, R.A., Cox, L.H. Synthetic Tabular Data: an alternative to complementary cell suppression for disclosure limitation of tabular data. Technical report (2002).

[2] Salazar, J.J. A Unified Mathematical Programming Framework for Different Statistical Disclosure Limitation Methods. *Operations Research* **53** (2005) 819–829. http://dx.doi.org/10.1287/opre.1040.0202

[3] Salazar, J.J. Controlled Rounding and Cell Perturbation: Statistical Disclosure Limitation Methods for Tabular Data. *Mathematical Programming* **105** (2006) 583–603. http://dx.doi.org/10.1007/s10107-005-0666-4

[4] Salazar, J.J. Statistical confidentiality: Optimization techniques to protect tables. *Computers & Operations Research* **35** (2008) 1638-1651. http://dx.doi.org/10.1016/j.cor.2005.09.009