Metodološki zvezki, Vol. 9, No. 1, 2012, 61-79

Statistical Disclosure Control using Random Rounding and Quadratic Programming

Neeraj Tiwari[1]

Abstract

The most common method of providing data to the public is through statistical tables. The problem of protecting confidentiality in statistical tables containing sensitive information has been of great concern during the recent years. Rounding methods are perturbation techniques widely used by statistical agencies for protecting the confidential data. Random rounding is one of these methods. In this paper, using the technique of random rounding and quadratic programming, we introduce a new methodology for protecting the confidential information of tabular data with minimum loss of information. The tables obtained through the proposed method consist of unbiasedly rounded values, are additive and have specified level of confidentiality protection. Some numerical examples are also discussed to demonstrate the superiority of the proposed procedure over the existing procedures.

1 Introduction

Statistical offices collect information about society. The most common method of providing data to the public is through statistical tables. Statistical agencies throughout the world are practicing the methods of maintaining confidentiality of sensitive information. In some situations, it is required that the statistical offices do not disclose in any way the information provided by the individual respondent. The release of statistical data inevitably reveals some information about individual data subject.

When confidential information is revealed, disclosure occurs. Thus statistical offices need to protect the confidentiality of data it collects. Not all the data collected and published by the statistical offices are confidential. The statistical offices have to protect only confidential data. The cells in a table containing confidential data are termed as “Sensitive cells” and all other cells are termed as

[1] Department of Statistics, Kumaun University, SSJ Campus, Almora-263601, Uttarakhand, India; kumarn_amo@yahoo.com

"non-sensitive cells". Before publishing any information, statistical offices face two problems. The first problem is of identifying the sensitive cells in a table. Identification of sensitive cells is carried out through several rules such as threshold rule, linear sensitivity rule, p percent rule, p-q percent rule, etc. This problem has been discussed in details by Cox (1980, 1981), Willenborg and Waal (2000) and Merola (2003a). The second problem is of protecting the confidential information contained in sensitive cells, while minimizing the loss of information. This problem is generally termed as "Disclosure control". The confidential information can be protected by the application of statistical disclosure limitation methods, which ensure that the risk of disclosing confidential information is very low, while minimizing the loss of information. The rising concerns of privacy, give rise to problems of disclosure including the issues of disclosing micro data. Several disclosure control techniques are used in the literature to achieve the required protection of confidential information. Two widely used techniques of disclosure control are "Controlled rounding" and "Cell suppression".

Rounding techniques involve the replacement of the original data by multiples of a given rounding base. Controlled rounding problem is the problem of optimally rounding real valued entries in a tabular array to adjacent integer values in a manner that preserves the tabular structure of the array. Rounding methods are used for many purposes, such as for improving the readability of data values, to control statistical disclosure in tables, to solve the problem of iterative proportional fitting (or raking) in two-way tables and controlled selection. Statistical disclosure control is one of the area in which rounding methods are widely used. Fellegi (1975) proposed a technique for random rounding which unbiasedly rounds the cell values and also maintains the additivity of the rounded table. The drawback of the random rounding procedure proposed by Fellegi (1975) is that it is applicable to one-dimensional tables only. Cox and Ernst (1982) used the transportation theory in linear programming to obtain an optimal controlled rounding of a two way tabular array. Using the general theory of transportation problems they demonstrated that solutions always exist to the controlled rounding problems. Causey, Cox and Ernst (1985) summarized the idea of Cox and Ernst (1982) and used the transportation theory to solve the controlled rounding problem. They discussed several statistical applications in which controlled rounding can be used and applied the concept of controlled rounding to solve the controlled selection problem. Cox (1987) presented a constructive algorithm for achieving unbiased controlled rounding which is simple to implement by hand. He also discussed a controlled rounding problem in three dimensions and provided a counter example to the existence of unbiased controlled rounding in three dimensions. Tiwari and Nigam (1988) improved the method of Cox (1987) to terminate in fewer steps. Salazar (2005) proposed a technique, termed as cell perturbation, which allows reducing the data loss from controlled rounding. This method is closely related to the classical controlled rounding methods and has the advantage that it also ensures the protection of sensitive cells to a specified level,

while minimizing the loss of information. Glover, Cox, Kelly and Patil (2008) applied a single mixed integer linear program to protect the sensitive information in tabular data using the method of controlled tabular adjustment.

Another method widely used by different researchers for protecting sensitive cells in a table is the method of cell suppression; in which sensitive cells are not published i.e. they are suppressed. This problem has been widely discussed by Cox (1980, 1995), Sande (1984), Carvalho et al. (1994) and Fischetti and Salazar (1999, 2000). In cell suppression, a large amount of information is lost as in addition to suppression of sensitive cells, some non-sensitive cells are also suppressed. To reduce the loss of information, Fischetti and Salazar (2003) proposed an improved methodology, known as partial cell suppression, in which instead of wholly suppressing primary and complementary suppressed cells, some intervals obtained with the help of a mathematical model, are published for these cell entries. The loss of information in partial cell suppression is smaller in comparison to complete cell suppression. Other statistical disclosure control approaches include data swapping, random noise, collapsing and roughly comparing. For details about statistical confidentiality, the readers may refer to Duncan, Elliot and Salazar (2011).

In this article, we use the idea of random rounding and quadratic programming to propose an improved methodology for disclosure control in an array that perturbs only the sensitive cells and adjusts some non-sensitive cells to preserve the marginal values of the array. The table obtained through the proposed procedure guarantees the protection level requirement and also attempts to minimize the information loss by minimizing the distance between the original and final table.

In Section 2, we describe the basic notations, problem of attacker and the protection of sensitive cells. The proposed methodology is introduced in Section 3. In Section 4, we discuss some numerical examples to demonstrate the utility of the proposed procedure. Section 5 concludes the findings of the paper.

2 Basic notations, problem of attacker and the protection of sensitive cells

In what follows, we describe the basic notations used in this manuscript. The problem of attacker and the protection of sensitive cells are discussed using the notations of Salazar (2005). In sensitive cells, we assume the existence of individuals who may analyze the published pattern to disclose the confidential information. These individuals are referred to as “Attackers” (or “Intruders” or “Snoopers”). If there exists more than one attacker in a cell, the problem is referred to as “Multi-attacker” problem. On the other hand the problem with only one attacker in a cell is referred to as “Single-attacker” Problem. Attackers can

also be categorized as "External attacker" and "Internal attacker". External attacker knows the set of linear system $My = b$ and the information that the cell values are non-negative. Internal attacker knows the set of linear system $My = b$ and also the tighter bounds (lower and upper bounds) on cell values. In this paper, we concern ourselves with the problem of disclosure control with single internal attacker.

Let A denote the tabular array

$$\begin{matrix} (a_{pq})_{m\mathrm{X}n} & (a_{p.})_{m\mathrm{X}1} \\ (a_{.q})_{1\mathrm{X}n} & (a_{..})_{1\mathrm{X}1} \end{matrix} \quad .$$

The tabular array A can be represented with the help of a vector $\underline{a}=(a_i : i \in I)$, where $a_1 = a_{11}$, $a_2 = a_{12}$, $a_3 = a_{13}$ … are all non-negative integers and I is the set of all elements including internal, marginal and grand total, consisting of $mn+m+n+1$ elements with the structure $M\underline{a}=0$, i.e.,

$$\begin{bmatrix} 1 & 1 & \dots & 1 & -1 & 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & 1 & 1 & \dots & 1 & -1 & \dots & . & . & \dots & 0 \\ . & . & \dots & . & . & . & . & \dots & . & . & \dots & . & . & \dots & . \\ . & . & \dots & . & . & . & . & \dots & . & . & \dots & . & . & \dots & . \\ . & . & \dots & . & . & . & . & \dots & . & . & \dots & . & . & \dots & . \\ 1 & 0 & \dots & . & 1 & 0 & \dots & . & . & \dots & -1 & 0 & \dots & . & 0 \\ 0 & 1 & \dots & . & . & 1 & \dots & . & . & \dots & . & -1 & \dots & . & 0 \\ 1 & 1 & \dots & 1 & 0 & 1 & 1 & \dots & 1 & 0 & \dots & . & . & \dots & -1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ . \\ . \\ . \\ a_n \\ a_{n+1} \\ a_{n+2} \\ a_{n+3} \\ . \\ . \\ . \\ a_{2n+1} \\ a_{2n+2} \\ . \\ . \\ . \\ a_{mn+m+1} \\ a_{mn+m+2} \\ . \\ . \\ . \\ a_{mn+m+n+1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} .$$

The vector $\underline{a}=(a_i : i \in I)$ satisfies the linear system $My=b$ and contains some sensitive cells also. Let us denote the subset of sensitive cells by S. Let there be r sensitive cells each having one internal attacker denoted by k_s ($s= 1 \dots r$), where k denotes the set of attackers in different sensitive cells. Now suppose that by

observing the published pattern, attacker k_s will compute the interval $(\underline{y}_s^{k_s} ... \bar{y}_s^{k_s})$, where $\underline{y}_s^{k_s}$ is the minimum and $\bar{y}_s^{k_s}$ is the maximum value of the interval. The sensitive cell s will be protected against the attacker k_s if the interval computed by the attacker k_s is wide enough. To decide whether the interval computed by the attacker k_s is wide enough or not we need three parameters defined as follows:

Upper protection level: It is a number $UPL_s^{k_s}$ representing a desired lower bound for

$$\bar{y}_s^{k_s} - a_s .$$

Lower protection level: It is a number $LPL_s^{k_s}$ representing a desired lower bound for

$$a_s - \underline{y}_s^{k_s} .$$

Sliding protection level: It is a number $SPL_s^{k_s}$ representing a desired lower bound for

$$\bar{y}_s^{k_s} - \underline{y}_s^{k_s} .$$

The values of these parameters are provided by statistical offices for each sensitive cell and for each attacker k_s. These values can also be defined by using common sense rule (see, Sande, 1984). Protection values are assumed to be unknown to the attacker. Let us assume that the attacker k_s knows two bounds $lb_i^{k_s}$ and $ub_i^{k_s}$ such that $a_i \in (lb_i^{k_s} ... ub_i^{k_s})$ for each cell $i \in I$. Thus the sensitive cells in the published table will be protected if,

$$lb_i^{k_s} \le \underline{y}_s^{k_s} \le a_i - LPL_s^{k_s} \le a_i \le a_i + UPL_s^{k_s} \le \bar{y}_s^{k_s} \le ub_i^{k_s} . \qquad (2.1)$$

This protection level is obtained by satisfying the protection equations which are determined with the help of the attacker's problem. Suppose the attacker is provided with the information that some values of the table are rounded to a common rounding base b. Then the attacker's problem becomes

$$\sum_i M_{ji} y_i = b_j$$

$$x_i - b \le y_i \le x_i + b \qquad (2.2)$$

$$lb_i^{k_s} \le y_i \le ub_i^{k_s} \quad , \quad \forall \quad i \in I$$

where j represents the number of equations ($j = 1,\ldots,m+n+1$) and $(x_i : i \in I)$ is the published pattern. The attacker can compute the value of $\bar{y}_s^{k_s}$ and $\underline{y}_s^{k_s}$ by maximizing $\bar{y}_s^{k_s}$ and minimizing $\underline{y}_s^{k_s}$, respectively, subject to the constraints (2.2).

The published table will be protected if,

$$\text{Maximize } [\bar{y}_s^{k_s} : (2.2) \text{ holds}] \ge u_s l_s^{k_s} \qquad (2.3)$$

$$\text{Minimize } [\underline{y}_s^{k_s} : (2.2) \text{ holds}] \le l_s l_s^{k_s} \qquad (2.4)$$

$$\text{Maximize } [\bar{y}_s^{k_s} : (2.2) \text{ holds}] - \text{Minimize}[\underline{y}_s^{k_s} : (2.2) \text{ holds}] \ge SPL_s^{k_s}, \qquad (2.5)$$

where $u_s l_s^{k_s} = a_s + UPL_s^{k_s}$ and $l_s l_s^{k_s} = a_s - LPL_s^{k_s}$.

In order to solve the constraints (2.3)-(2.5), we convert these constraints into linear form, using duality theory in linear programming. Let us consider the dual variables $\alpha_i^{\,1}, \beta_i^{\,1}, \alpha_i^{\,2}, \beta_i^{\,2}$ and γ_j associated with the inequalities $y_i \le ub_i^{k_s}$, $-y_i \le -lb_i^{k_s}$, $y_i \le x_i + b$, $-y_i \le b - x_i$ and $\sum_i M_{ji} y_i = b_j$, respectively. Thus the attacker's problem

$$\text{Maximize } [\bar{y}_s^{k_s} : (2.2) \text{ holds}]$$

is equivalent to

$$\text{Minimize} \sum_j \gamma_j b_j + \sum_i [\alpha_i^{\,1} ub_i^{k_s} + \alpha_i^{\,2} (x_i + b) - \beta_i^{\,1} lb_i^{k_s} - \beta_i^{\,2} (x_i - b)] \qquad (2.6)$$

subject to the constraints

$$\alpha_s^{\,1} + \alpha_s^{\,2} - \beta_s^{\,1} - \beta_s^{\,2} + \sum_j M_{js}\gamma_j = 1, \quad \text{for all } S$$

$$\alpha_i^{\,1} + \alpha_i^{\,2} - \beta_i^{\,1} - \beta_i^{\,2} + \sum_j M_{ji}\gamma_j = 0, \quad \text{for all non-sensitive cells}$$

$$\alpha_i^{\,1} \geq 0$$

$$\alpha_i^{\,2} \geq 0 \qquad (2.7)$$

$$\beta_i^{\,1} \geq 0$$

$$\beta_i^{\,2} \geq 0$$

γ_j is unrestricted in sign.

Now (2.3) can be written in simplified form as,

Maximize $[\bar{y}_s^{\,k_s}$: (2.2) holds$] \geq u_s l_s^{\,k_s}$

$\Rightarrow$ Minimize (2.6) $\geq u_s l_s^{\,k_s}$: all $\alpha_i^{\,1}, \alpha_i^{\,2}, \beta_i^{\,1}, \beta_i^{\,2}, \gamma_j$ satisfying (2.7)

$$\Rightarrow \text{Minimize} \quad \sum_j \gamma_j b_j + \sum_i [\alpha_i^{\,1} ub_i^{\,k_s} + \alpha_i^{\,2}(x_i + b) - \beta_i^{\,1} lb_i^{\,k_s} - \beta_i^{\,2}(x_i - b)] \geq u_s l_s^{\,k_s}$$

$$\Rightarrow \text{Minimize} \quad \sum_j \gamma_j b_j + \sum_i [\alpha_i^{\,1} UB_i^{\,k_s} + \alpha_i^{\,1} x_i + \alpha_i^{\,2} x_i + \alpha_i^{\,2} b - \beta_i^{\,1} x_i + \beta_i^{\,1} LB_i^{\,k_s}$$

$$- \beta_i^{\,2} x_i + \beta_i^{\,2} b] \geq a_s + UPL_s^{\,k_s}$$

$$\Rightarrow \quad \sum_i [\alpha_i^{\,1} UB_i^{\,k_s} + \alpha_i^{\,2}(x_i + b - a_i) + \beta_i^{\,1} LB_i^{\,k_s}$$

$$- \beta_i^{\,2}(x_i - b - a_i)] \geq UPL_s^{\,k_s} \qquad (2.8)$$

where $UB_i^{\,k_s} = ub_i^{\,k_s} - a_i$ and $LB_i^{\,k_s} = a_i - lb_i^{\,k_s}$,

for all $\alpha_i^{\,1}, \alpha_i^{\,2}, \beta_i^{\,1}, \beta_i^{\,2}, \gamma_j$ satisfying (2.7).

Similarly (2.4) can be written in simplified form as,

$$\sum_i [\alpha_i^{'1} UB_i^{k_s} + \alpha_i^{'2}(x_i + b - a_i) + \beta_i^{'1} LB_i^{k_s} - \beta_i^{'2}(x_i - b - a_i)] \geq LPL_s^{k_s} \tag{2.9}$$

for all $\alpha_i^{'1}, \alpha_i^{'2}, \beta_i^{'1}, \beta_i^{'2}$ and $\gamma_j^{'}$ satisfying the following constraints:

$$\begin{aligned}
&\alpha_s^{'1} + \alpha_s^{'2} - \beta_s^{'1} - \beta_s^{'2} + \sum_j M_{js}\gamma_j^{'} = 1, \quad \text{for all } S \\
&\alpha_i^{'1} + \alpha_i^{'2} - \beta_i^{'1} - \beta_i^{'2} + \sum_j M_{ji}\gamma_j^{'} = 0, \quad \text{for all non-sensitive cells} \\
&\alpha_i^{'1} \geq 0 \\
&\alpha_i^{'2} \geq 0 \\
&\beta_i^{'1} \geq 0 \\
&\beta_i^{'2} \geq 0 \\
&\gamma_j^{'} \text{ is unrestricted in sign.}
\end{aligned} \tag{2.10}$$

Similarly (2.5) reduces to,

$$\begin{aligned}
\sum_i [&(\alpha_i^{1} + \alpha_i^{'1}) UB_i^{k_s} + (\alpha_i^{2} + \alpha_i^{'2})(x_i + b - a_i) + (\beta_i^{1} + \beta_i^{'1}) LB_i^{k_s} \\
&+ (\beta_i^{2} + \beta_i^{'2})(a_i - x_i + b)] \geq SPL_s^{k_s},
\end{aligned} \tag{2.11}$$

for all $\alpha_i^{1}, \alpha_i^{2}, \beta_i^{1}, \beta_i^{2}, \gamma_j$ satisfying (2.7) and $\alpha_i^{'1}, \alpha_i^{'2}, \beta_i^{'1}, \beta_i^{'2}, \gamma_j^{'}$ satisfying (2.10)

The conditions obtained through (2.8), (2.9) and (2.11) ensure upper protection, lower protection and sliding protection, respectively. Solving (2.7) and (2.10), we obtain the values of the dual variables $\alpha_i^{1}, \alpha_i^{2}, \beta_i^{1}, \beta_i^{2}, \alpha_i^{'1}, \alpha_i^{'2}, \beta_i^{'1}$ and $\beta_i^{'2}$.

3 The proposed methodology

Let us assume that there are *r* sensitive cells in the given array. The x-values are assumed to be *1* for the sensitive cells and *0* for the others. Following the notations given in Section 2, we set the values of $UB_i^{k_s}$, $LB_i^{k_s}$ and the protection levels for the sensitive cells provided by statistical offices. After solving (2.7) and (2.10), we

obtain the dual values $\alpha_i^{1}, \alpha_i^{2}, \beta_i^{1}, \beta_i^{2}, \alpha_i^{'1}, \alpha_i^{'2}, \beta_i^{'1}$ and $\beta_i^{'2}$ for the sensitive cells. Putting these values in (2.8), (2.9) and (2.11), we get protection equation for the sensitive cells.

Now, we round the sensitive cells unbiasedly to base *b*. The rounding base *b* should be chosen in such a way that it is, as far as possible, a factor of the sum of the entries in the sensitive cells. However, if it is not possible to choose a rounding base, which is a factor of sum of the entries in sensitive cells, some other rounding base may be chosen. The advantage of taking the rounding base, a factor of the sum of entries of sensitive cells is that the sum of the rounded values of the sensitive cells will remain unaltered. From these sets of unbiasedly rounded values, we select the set which satisfy the simplified inequalities for upper, lower and sliding protection, i.e., (2.8), (2.9) and (2.11). If more than one set of unbiasedly rounded values satisfy the protection equations, we choose the set which has the minimum distortion between the rounded and the original values, i.e.,

$$\left[\sum_i \left\{ (x_i - a_i)^2 \right\}^{1/2} \right], \tag{3.1}$$

where a_i and x_i represents the original and rounded values, respectively. The sensitive values in the table are then replaced by these unbiasedly rounded values. After replacing the sensitive cell values with the rounded values, the resultant table may not be additive. To make the table additive, some or the entire non-sensitive cell values are then adjusted by as small an amount as possible. This is achieved with the help of the following model:

$$\text{Minimize} \quad z = \sum_{p=1}^{m} \sum_{q=1}^{n} \frac{x_{pq}^{\,2}}{a_{pq}} - 1 \tag{3.2}$$

Subject to the constraints

$$(i) \sum_{q=1}^{n} x_{pq} = X_p - \sum_q S \qquad , \forall\ p = 1 \ldots m$$

$$(ii) \sum_{p=1}^{m} x_{pq} = X_q - \sum_p S \qquad , \forall\ q = 1 \ldots n$$

$$(iii) \sum_p \sum_q x_{pq} = G - \sum_p \sum_q S \tag{3.3}$$

$$(iv)\ lb_{pq}^{\,k_s} \le x_{pq} \le ub_{pq}^{\,k_s} \quad , \quad \text{for all non-sensitive cells,}$$

where x_{pq}*'s* are adjusted non-sensitive integer cell values, X denotes the marginal total of row and column and G (= $a_{..}$) is the grand total. The objective function z is in fact the directed distance D from a_{pq} to x_{pq}, defined as,

$$D = D(a_{pq}, x_{pq}) = E_{a_{pq}} [\frac{x_{pq}}{a_{pq}} - 1]^2 = \sum_{p=1}^{m}\sum_{q=1}^{n} \frac{x_{pq}^2}{a_{pq}} - 1 \qquad (3.4)$$

The distance measure $D(a_{pq}, x_{pq})$ defined in (3.4) is similar to the χ^2-statistic often employed in related problems and is also used by Cassel and Sarndal (1972) and Gabler (1987). Other distance measures are also discussed by Takeuchi, Yanai, and Mukherjee (1983).

The solution obtained through the proposed procedure unbiasedly rounds the sensitive cells to base b while guaranteeing the protection requirements of the cells and also preserves the marginals through (3.2)-(3.3).

4 Empirical results

In what follows, we discuss some empirical examples to illustrate the proposed methodology and demonstrate its superiority by comparing it with the method given by Salazar (2005).

Example 1: Consider the following one-dimensional population of 10 units borrowed from Fellegi (1975).

12 23 34 3 49 23 50 17 8 13

Let the cell values a_4 and a_9 are sensitive. We set the values of $UB_i^{k_s}$ and $LB_i^{k_s}$ as

$$UB_i^{k_s} = a_i \quad \text{and} \quad LB_i^{k_s} = a_i/2.$$

Let the protection level for a_4 provided by statistical office is

$$UPL_4^{k_4} = 2, \quad LPL_4^{k_4} = 1, \quad SPL_4^{k_4} = 5,$$

for a_9, the protection level is

$$UPL_9^{k_9} = 4, \quad LPL_9^{k_9} = 2, \quad SPL_9^{k_9} = 5,$$

and b= 5.

After solving (2.7) and (2.10), we get following values for a_4:

$\alpha_4^{\,1} = 0,\ \alpha_4^{\,2} = 1, \beta_4^{\,1} = 0, \beta_4^{\,2} = 0,\ \alpha_4^{\,'1} = 0, \alpha_4^{\,'2} = 0, \beta_4^{\,'1} = 0$ and $\beta_4^{\,'2} = 1$

and for a_9, we get

$\alpha_9^{\,1} = 0,\ \alpha_9^{\,2} = 1, \beta_9^{\,1} = 0, \beta_9^{\,2} = 0,\ \alpha_9^{\,'1} = 0, \alpha_9^{\,'2} = 0, \beta_9^{\,'1} = 0$ and $\beta_9^{\,'2} = 1.$

Putting these values in (2.8), (2.9) and (2.11), we get protection equation for a_4 as

(i) $x_4 + 5 - 3 \geq 2 \Rightarrow x_4 \geq 0$

(ii) $-x_4 + 5 + 3 \geq 1 \Rightarrow x_4 \leq 7$

and for a_9, the protection equations are:

(i) $x_9 + 5 - 8 \geq 4 \Rightarrow x_9 \geq 7$

(ii) $-x_9 + 5 + 8 \geq 2 \Rightarrow x_9 \leq 11.$

Now we unbiasedly round above sensitive cell values and found that only the set (0, 10) of unbiasedly rounded cell values satisfies the protection equation. So we take this set and replace the original sensitive cell values by these unbiasedly rounded values. After substituting these rounded values, we observe that table is not additive. To make the table additive, we apply the model (3.2)-(3.3) and get following values corresponding to the cells of the given table:

12 23 34 0 50 23 50 17 10 13

and $z = 234.4737$.

Solving this example using the Salazar's (2005) procedure, we get following values corresponding to the different cells of the table:

10 25 35 0 50 25 50 15 10 10 and $z = -9$.

The deviation between the rounded and the original values of the table using (3.1) comes out to be 3.74 for the proposed procedure whereas it turns out to be 6.63 for the procedure suggested by Salazar (2005). Thus we see that the deviations is reasonably small for the proposed procedure. Moreover, the proposed procedure rounds the sensitive cells in such a way that the confidential information contained in the sensitive cells is protected against the single internal attacker and the marginal are also not disturbed. To make the table additive only one non-sensitive cell (a_5) has been disturbed and that also by 2.0408% only, while all other non-sensitive cell values are published in their original form. Using the procedure of Salazar (2005), as much as seven non-sensitive cells (a_1, a_2, a_3, a_5, a_6, a_8 and a_{10}) have been disturbed.

Example 2: Consider following example taken from Cox (1995).

20	10	20	10	20	80
10	10	20	5	15	60
40	10	10	20	10	90
5	5	15	10	5	40
75	35	65	45	50	270

Let the values a_1, a_9, a_{16} and a_{22} are sensitive. Let the protection levels for a_1, a_9 and a_{16} provided by the statistical office are:

$UPL_i^{k_i} = 7, \quad LPL_i^{k_i} = 5, \quad SPL_i^{k_i} = 14$ for $i = 1, 9$ and 16

and for a_{22}, the protection levels are:

$UPL_{22}^{k_{22}} = 5, \quad LPL_{22}^{k_{22}} = 2, \quad SPL_{22}^{k_{22}} = 14.$

Now we solve (2.7) and (2.10) to find out the values of the dual variables $\alpha_i^1, \alpha_i^2, \beta_i^1, \beta_i^2, \alpha_i'^1, \alpha_i'^2, \beta_i'^1$ and $\beta_i'^2$ for all the sensitive cells. After solving (2.7) and (2.10), we put these values in (2.8), (2.9) and (2.11) and get only lower protection equation for the cell a_1, given by,

(i) $x_1 \leq 47$.

The equations to satisfy the upper protection requirement for the cell a_1 could not be obtained. Since the values of the dual variables for all the other sensitive cells come out to be 0, we could not obtain any protection equation for all the other sensitive cells as well. This may be noted that if we cannot form any lower or upper protection equation for a particular sensitive cell, even then the sensitive cell may be protected. In such situations, we will have to check in the auditing phase whether the sensitive cell for which no protection equation could be obtained or only one protection equation (upper or lower) is obtained, is protected or not. Now we unbiasedly round these sensitive cell values taking b=14 and get the following sets of rounded values, which are protected and nearest to the set of original sensitive cell values:

(i) (28, 14, 14, 14)

(ii) (14, 28, 14, 14)

(iii) (14, 14, 28, 14)

After replacing the original sensitive cell values by the above sets of rounded values and applying the model (3.2)-(3.3), we could not obtain the solution for the set (iii). Also the value of the objective function, which minimizes the distance between original and final table comes out to be 213.9713 and 209.7067 for the set (i) and (ii), respectively. Hence we select set (ii) of rounded values and get the following results:

14	12	18	13	23	80
9	8	28	4	11	60
47	10	8	14	11	90
5	5	11	14	5	40
75	35	65	45	50	270

with z = 209.7067. For this problem, we could not obtain any protection equation for the sensitive cells a_9, a_{16} and a_{22}. Moreover, for the sensitive cell a_1, the upper protection equation could not be obtained. Therefore, we verify whether these sensitive cells are protected or not. In auditing phase, we observe that all the sensitive cells are protected.

We also solved this problem by the procedure of Salazar (2005) and obtained the following results:

14	14	28	14	14	84
14	14	14	0	14	56
42	0	14	14	14	84
0	0	14	14	14	42
70	28	70	42	56	266

with z = -68.

Distortions in the final table obtained by the proposed procedure from the original table, using (3.1) is 16.43, whereas it is 28.53 using the Salazar's procedure. Thus we conclude that using the proposed procedure we get smaller distortions for this problem also.

In this problem, although we could not obtain any protection equations for the sensitive cells a_9, a_{16} and a_{22}, but the final table is still protected using the

proposed procedure. To make the table additive, only 12 non-sensitive cells are disturbed using the proposed procedure, whereas in the procedure of Salazar(2005) all the non-sensitive cells are disturbed and marginal are also not preserved.

Example 3: Consider the following two way table:

200	40	50	200	120	610
20	70	60	100	120	370
40	90	250	100	30	510
100	150	30	80	150	510
360	350	390	480	420	2000

Suppose that the cell values a_1, a_4, a_{10}, a_{15}, a_{19}, a_{20} and a_{23} are sensitive. Let the protection levels provided by the statistical office for these sensitive cells are:
For cells a_4

$UPL = 20, \quad LPL = 10, \quad SPL = 15.$

For cells a_{10} and a_{19}

$UPL = 10, \quad LPL = 5, \quad SPL = 15.$

For cells a_{15}

$UPL = 25, \quad LPL = 20, \quad SPL = 15.$

And for cells a_{20} and a_{23}

$UPL = 15, \quad LPL = 7, \quad SPL = 15.$

Now we solve (2.7) and (2.10) to find out the values of the dual variables $\alpha_i^{1}, \alpha_i^{2}, \beta_i^{1}, \beta_i^{2}, \alpha_i^{'1}, \alpha_i^{'2}, \beta_i^{'1}$ and $\beta_i^{'2}$ for all the sensitive cells. After solving (2.7) and (2.10), we substitute these values in (2.8), (2.9) and (2.11) and get the following protection equations:

(i) $x_4 \leq 209$, for the sensitive cell a_4 to satisfy the lower protection and sliding protection requirement and

(ii) $x_{23} \leq 154$, for the sensitive cell a_{23} to satisfy the lower protection and sliding protection requirement.

We could not obtain the equations to satisfy the upper protection requirement for the cell a_4 and a_{23}. Since the values of the dual variables for all the other

sensitive cells come out to be 0, we could not obtain any protection equation for all the other sensitive cells. Now we unbiasedly round these sensitive cell values taking $b = 19$ and get the following sets of rounded values:

(i) (190, 114, 247, 95, 152, 152)

(ii) (190, 95, 247, 114, 152, 152)

Both of these sets are equidistant from the set of original sensitive cell values and satisfy the protection equations for a_4 and a_{23}.

After replacing the original sensitive cell values by the above sets of rounded values and applying the model (2.12)-(2.13), we observe that the set (i) is nearer to the set of the original sensitive cell values as compared to set (ii). Thus we select set (i) and get the following results:

204	41	52	190	123	610
19	65	58	114	114	370
42	92	247	98	31	510
95	152	33	78	152	510
360	350	390	480	420	2000

and $z = 1056.654$.

Since in this problem we could not obtain the protection equation for some sensitive cells, so in auditing phase we have to check whether these cells are protected or not. In auditing phase, we observe that the sensitive cells a_4 and a_{15} could not satisfy the upper protection requirement, while all other cells are protected.

Solving this problem by the procedure of Salazar (2005), we get the following results:

209	38	57	190	114	608
19	57	57	114	114	361
38	95	247	95	38	513
95	152	38	76	152	513
361	342	399	475	418	1995

with z = -86.

Deviations of the final table obtained by the proposed procedure from the original table using (3.1) is 21.45 and that for the final table obtained by the procedure of Salazar is 34.99. Thus for this example also, the proposed procedure results with smaller loss of information as compared to the procedure of Salazar(2005).

Example 4: We consider the following two-way table borrowed from Fischetti and Salazar (2003):

20	50	10	80
8	19	22	49
17	32	12	61
45	101	44	190

Let the cell value a_7 is sensitive. Let the protection levels provided by the statistical office for a_7 is:

$UPL_7^{k_7} = 7, \quad LPL_7^{k_7} = 5, \quad SPL_7^{k_7} = 5.$

Now we solve (2.7) and (2.10) to find out the values of the dual variables $\alpha_i^1, \alpha_i^2, \beta_i^1, \beta_i^2, \alpha_i^{'1}, \alpha_i^{'2}, \beta_i^{'1}$ and $\beta_i^{'2}$ for the sensitive cell a_7. After solving (2.7) and (2.10) all the values of the above dual variables comes out to be 0, so we cannot form any protection equation for the sensitive cell a_7. After applying rounding procedure with b= 5, we get the rounded value for a_7 as 20. Now we put this value in place of the original sensitive cell value and apply the model (3.2)-(3.3). After applying the model, we get the following results:

20	49	11	80
9	20	20	49
16	32	13	61
45	101	44	190

with z = 172.3245.

In this problem also, we could not obtain the protection equation for the sensitive cell, so in auditing phase we have to check whether the sensitive cell is protected or not. In auditing phase, we observe that the sensitive cell a_7 could not

satisfy the upper, lower and sliding protection requirements and hence we conclude that the cell a_7 is not protected against the single internal attacker.

We also solved this problem by the procedure of Salazar (2005) and get the following results:

20	50	15	85
10	20	20	50
20	30	10	60
50	100	45	195

and z = -9.

Using (3.1), the distortion obtained by the proposed procedure from the original table is 3.16, whereas it is 11.4 for the procedure of Salazar (2005). This result again displays the utility of the proposed procedure.

5 Concluding remarks

In this paper, using the technique of random rounding and quadratic programming, we introduce a new methodology for protecting the confidential information of tabular data with minimum loss of information. The tables obtained through the proposed method consist of unbiasedly rounded values, are additive and have specified level of confidentiality protection. Some numerical examples are also discussed to demonstrate the superiority of the proposed procedure over the existing procedures. One of the limitations of the proposed procedure is that the problem of disclosure control with single internal attacker is only discussed. If there are more than one internal attackers, the formation of the problem may become more complex. Three and more dimensional problems could also not be discussed. Moreover, as in the case of linear programming, there is no guarantee of convergence of a quadratic programming problem. Kuhn and Tucker (1951) have derived some necessary conditions for the optimum solution of a quadratic programming algorithm but no sufficient conditions exist for convergence. Therefore unless the Kuhn-Tucker conditions are satisfied in advance, there is no way of verifying whether a quadratic programming algorithm converges to an absolute (global) or relative (local) optimum. Also, there is no way to predict in advance that the solution of a quadratic programming problem exists or not.

Acknowledgement

The author is thankful to the two referees and the editors for their constructive comments and suggestions, which led to considerable improvement in presentation of this work.

References

[1] Carvalho, F.D., Dellaert, N.P., and Osório, M.S. (1994): Statistical disclosure in two-dimensional tables: General tables. *Journal of the American Statistical Association*, **89**, 1547-1557.

[2] Cassel, C.M. and Sarndal, C.E. (1972): A model for studying robustness of estimators and informativeness of labels in sampling with varying probabilities. *Journal of Royal Statistical Society*, Series B, **34**, 279-289.

[3] Causey, B.D., Cox, L.H., and Ernst, L.R. (1985): Applications of transportation theory to statistical problems. *Journal of the American Statistical Association*, **80**, 903-909.

[4] Cox, L.H. (1980): Suppression methodology and statistical disclosure control. *Journal of the American Statistical Association*, **75**, 377-385.

[5] Cox, L.H. (1981): Linear sensitivity measures in statistical disclosure control. *Journal of Statistical Planning and Inference*, **5**, 153-164.

[6] Cox, L.H. and Ernst, L.R. (1982): Controlled rounding. *INFOR*, **20**, 423-432.

[7] Cox, L.H. (1987): A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, **82**, 420-424.

[8] Cox, L.H. (1995): Network models for complementary cell suppression. *Journal of the American Statistical Association*, **90**, 1453-1462.

[9] Duncan, G.T., Elliot, M., and Salazar, J.J. (2011): *Statistical Confidentiality: Principles and Practice* Berlin: Springer.

[10] Fellegi, I.P. (1975): Controlled random rounding. *Survey Methodology*, **1**, 123-135.

[11] Fischetti, M. and Salazar, J.J. (2000): Models and algorithms for optimizing cell suppression in tabular data with linear constraints. *Journal of the American Statistical Association*, **95**, 916-928.

[12] Fischetti, M. and Salazar, J.J. (2003): Partial cell suppression: A new methodology for statistical disclosure control. *Statistics and Computing*, **13**, 13-21.

[13] Gabler, S. (1987): The nearest proportional to size sampling design. *Communications in Statistics-Theory and Methods*, **16**, 1117-1131.

[14] Glover, F., Lawrence, H.C., Kelly, J.P., and Patil, R. (2008): Exact, heuristic and metaheuristic methods for confidentiality protection by controlled tabular adjustment. *International Journal of Operations Research*, **5**, 117-128.

[15] Kuhn, H.W. and Tucker, A.W. (1951): Non-linear programming. *Proceedings of Second Berkely Symposium on Mathematical Statistics and Probability*, 481-492.

[16] Merola, G.M. (2003): Generalized risk measures for tabular data. *Proceedings of the 54th Session of the International Statistical Institute*.

[17] Salazar, J.J. (2005): Controlled rounding and cell perturbation: Statistical disclosure limitation methods for tabular data. *Mathematical Programming*, Ser. B 105, 583-603. 13.

[18] Sande, G. (1984): Automated cell suppression to preserve confidentiality of Business Statistics. *Statistical Journal of the United Nations ECE*, **2**, 33-41.

[19] Takeuchi, K., Yanai, H., and Mukherjee, B.N. (1983): *The Foundations of Multivariate Analysis*. 1st Ed. New Delhi: Wiley Eastern Ltd

[20] Tiwari, N. and Nigam, A. K. (1993): A note on constructive procedure for unbiased controlled rounding. *Statistics & Probability Letters*, **18**, 415-420.

[21] Willenborg, L.C.R.J. and de Waal, T. (2001): Elements of Statistical disclosure control. *Lecture Notes in Statistics*, **155**, Springer.