Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

A layered perturbation method for the protection of tabular outputs

SURVEY
METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005

VOLUME 31

NUMBER 1

by Jean-Louis Tambay

Release date: June 22, 2017

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at STATCAN.infostats-infostats.STATCAN@canada.ca

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

• Statistical Information Service	1-800-263-1136
• National telecommunications device for the hearing impaired	1-800-363-7629
• Fax line	1-877-287-4369

Depository Services Program

• Inquiries line	1-800-635-7943
• Fax line	1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard table symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^{s}	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Vol. 43, No. 1, pp. 31-40
Statistics Canada, Catalogue No. 12-001-X

A layered perturbation method for the protection of tabular outputs

Jean-Louis Tambay[1]

Abstract

The protection of data confidentiality in tables of magnitude can become extremely difficult when working in a custom tabulation environment. A relatively simple solution consists of perturbing the underlying microdata beforehand, but the negative impact on the accuracy of aggregates can be too high. A perturbative method is proposed that aims to better balance the needs of data protection and data accuracy in such an environment. The method works by processing the data in each cell in layers, applying higher levels of perturbation for the largest values and little or no perturbation for the smallest ones. The method is primarily aimed at protecting personal data, which tend to be less skewed than business data.

Key Words: Confidentiality; Data perturbation; Tabular outputs.

1 Introduction

Statistical agencies are under pressure to provide more information from their data holdings to external users. Many now enable the creation of custom tables through on-line query systems. But the risks of a disclosure of confidential information increase with the quantity of outputs released. To address this problem agencies can go from one extreme, which is to severely limit the amount of information being released, to another, which is to generate outputs from model-based synthetic microdata. Perturbative methods, which add noise to microdata or aggregate results, lie somewhere in between. This paper proposes a perturbative method for quantitative administrative data, such as personal taxation data, in a custom tabulation environment. Section 2 provides some background information, outlines desirable objectives and reviews standard approaches for the protection of tables of magnitude. Section 3 presents the proposed Layered Perturbation Method (LPM) and provides some of its properties. An empirical evaluation is given in Section 4 and outstanding issues are discussed in Section 5.

2 Background

The proposed strategy aims to protect the confidentiality of tables of magnitude in a semi-controlled custom tabulation environment. It was primarily developed for administrative (census-like) data, notably personal taxation data. At Statistics Canada, such outputs are subject to disclosure control rules including minimum population sizes for identifiable geographic areas, the use of minimum-cell-size and dominance rules to suppress sensitive (confidential) cells, and the application of complementary cell suppression (CCS) to prevent the recuperation of sensitive cell values.

While personal data are inherently safer than business data, they are more readily used in custom tabulations. And with wider access to custom tabulations it becomes increasingly difficult to carry out CCS

[1] Jean-Louis Tambay, Statistics Canada, Ottawa, Canada, K1A 0T6. E-mail: jean-louis.tambay@canada.ca.

effectively. Alternative methods need to be considered. The proposed method consists of applying a perturbative technique, independently, in every non sensitive cell of every table. Only sensitive cells are suppressed, although some may become releasable if perturbed. The method is meant to protect sensitive cells in tables as well as to guard against residual disclosure from multiple tables – especially disclosure by the differencing of nested totals. The focus is on protecting two totals that differ by one unit.

It is assumed that we are in a semi-controlled environment where access is somewhat restricted, or at least not anonymous, so that some monitoring and control of requests is applied. This precaution is needed because offering unrestricted tabulations to anonymous hackers trying to exploit every vulnerability (in particular, through multiple requests involving carefully chosen sets of units) could lead to the approximate disclosure of unit values under certain conditions. The method is developed for census-like data, which are riskier, but it could undoubtedly be adapted to sample data if needed. The strategy is better suited to personal data as they are less subject to dominance than business data, and near-dominant cells get perturbed the most. But with some adaptation users may see to what extent the strategy could meet their needs for other types of data.

If possible, we would like the strategy to address other disclosure issues, such as the protection of ratios and of other types of outputs. Other desirable features are the ability to treat zeroes and negative values, the maintenance of data quality, the preservation of additivity in tables, and operational aspects such as computational simplicity and the use of minimal manual intervention.

In this paper we use a $P-$percent rule to identify sensitive cell totals, meaning that a cell is sensitive if the aggregate contribution from the smallest units, starting with the third-largest, is less than $P\%$ of the value of the largest unit (i.e., if $X - x_1 - x_2 < P\%\ x_1$, where X is the cell total and x_i is the contribution of its i^{th} largest unit). We assume that cells failing a minimum-cell-size rule are also sensitive.

We are interested in preserving quality and confidentiality for magnitude data in a custom tabulation environment. Techniques for tables of magnitude such as CCS (Cox and Sande 1979) and Controlled Tabular Adjustment (Cox and Dandekar 2004) do not work very well in such an environment. They require solving optimization problems to find table-specific solutions. Problems start to occur when trying to protect huge, complex and/or related (i.e., linked) tables, such as the inability to reach a solution, or the use of heuristics that may yield inconsistencies in suppression or perturbation patterns that can be exploited by hackers. It is far easier to perturb cell totals directly, e.g., by the application of random noise, but one still needs to look at the microdata to ensure adequate protection while controlling the impact on quality. And without additional measures it can lead to inconsistencies within and between tables that can be exploited by hackers.

Microdata perturbation, where data are perturbed at the microdata level, is better suited for our multi-table environment. Tables are additive and usually without suppression; with consistent results between tables. If custom tables are allowed it may be possible to recover some individual perturbed values directly or by differencing, so the noise level for each unit would need to be high enough to meet target ambiguity levels. As a result, the cumulated noise for specific aggregates can be large. A microdata perturbation method developed and used at the U.S. Census Bureau is the EZS method (Evans, Zayatz and Slanta 1998). EZS multiplies individual values x_i by a weight $w_i = 1 + \varepsilon_i$, where ε_i are i.i.d. random variables with mean

0 and variance σ_ε^2. Two distributions for ε_i of interest are the split triangular distribution (shaped like Figure 2.1) and the split uniform distribution (shaped like Figure 2.2) whose corresponding values of σ_ε^2 are $(3a^2 + 2ab + b^2)/6$ and $(a^2 + ab + b^2)/3$, respectively. The ε_i (or w_i) are permanently attached to their unit i. Applying the same noise to all variables will not affect ratios. If it is necessary to protect ratios different weights w_i should be used for different variables, or unit-specific weights can be used jointly with unit-variable specific weights.

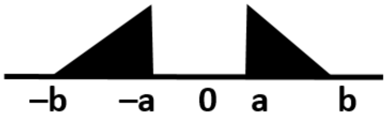

Figure 2.1 Split triangular distribution.

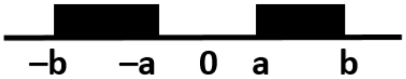

Figure 2.2 Split uniform distribution.

There are ways to attenuate the cumulative impact of microdata perturbation on quality. Massell and Funk (2007) suggest to balance the random noises within cells for a primary table to limit their impact there. Other methods perturb microdata, but not always the same way, allowing some inconsistencies in results. Giessing (2011) proposes to multiply unit values x_i by $w_i = 1 \pm |\varepsilon_i|$, for ε_i i.i.d. $\mathrm{N}(0, \sigma_0^2)$, except in sensitive cells, where the largest value gets multiplied by $w_i = 1 \pm (\mu_0 + |\varepsilon_i|)$. The value μ_0 is chosen to give an appropriate level of protection for sensitive cells, allowing a lower value of σ_0^2 to be used overall. But if σ_0^2 is too low the method may not sufficiently protect against disclosure by differencing. The Australian Bureau of Statistics' Top Contributors Method (TCM), developed for its TableBuilder remote access application, consists of perturbing the largest respondents in each cell in a semi-consistent way, i.e., where parts of their noise is applied consistently (Thompson, Broadfoot and Elazar 2013). The LPM uses some of the same concepts but, as will be explained, protects more against differencing.

Other commonly used strategies such as rounding, (sub-)sampling and swapping units, say between neighbouring areas, are better suited for the protection of frequency tables.

3 The Layered Perturbation Method (LPM)

3.1 Description

The LPM is a perturbative method for totals that focuses on disclosure from differencing. When used in tables of magnitude it allows cell suppression to be restricted to sensitive cells. Three basic ideas underlie the LPM. The first two are similar to the TCM approach.

The first basic idea is the attachment of pseudo-random hash numbers (PRNs) to units to produce consistent perturbation outcomes when needed. This discourages the use of repeated queries to improve the estimation of unperturbed totals. The EZS method is used to multiply the value of a unit i by a weight $w_i = 1 + \varepsilon_i$, with $\varepsilon_i \sim (0, \sigma_\varepsilon^2)$ as above. To obtain consistent results ε_i are generated from a unit-specific PRN that is uniformly distributed over $[0,1)$. For example, use $h_i / 1000$, where h_i are generated from the Social Insurance Number (e.g., $h_i = Mod(SIN_i \bullet P,\ 1000)$ for P a large prime). Using h_i will always perturb unit i the same way. To perturb unit i the same way only when it appears in the same cell total, generate cell-unit level noise $w_i' = 1 + \varepsilon_i'$ from $h_i' = Mod(h_i + h_{tot},\ 1000) / 1000$, where $h_{tot} = \sum_{i \in cell} h_i$. Primes are used to designate cell-unit specific noises and perturbations. All noise values are derived from h_i or h_i'.

The second idea is the application of perturbation to units in each cell by layers. The largest four units are perturbed in a random but *consistent* manner using perturbation weights w_i generated from h_i. The next largest units, say units 5 to 9, are perturbed in a semi-consistent manner. Their perturbation is a mixture of unit specific weights w_i and unit-cell specific weights w_i'. Smallest units are not perturbed. Their values are protected from differencing by the unit-cell perturbations of units 5 to 9 since adding or removing a unit in a cell, no matter how small, will affect the w_i' for those units. The number of units per layer is flexible, we have found that four and five, respectively gave satisfactory results.

A third set of measures mostly targets the issue of differencing. The direction of noise for even-ranked units is reversed (w_i are set from $(-1)^{i+1}\varepsilon_i$) to increase variances of differences when a top-ranked unit is changed. For units 5 to 9 a random mixture of w_i and w_i' is applied to lessen the risk when a small unit is added or removed. Finally, the noise for the top three units is amplified in nonsensitive cells with greater dominance. This allows lower levels of noise to be used generally, reducing the overall impact of the perturbation on data quality.

A suggested application of the LPM would consist of suppressing all sensitive and small cells (e.g., $n < 10$) and perturbing remaining cells. Because of the protection offered by perturbation, cells that are slightly sensitive may also be publishable. For other cells with cell total $X = \sum_{i \in cell} x_i$, set perturbed value Z as

$$\mathrm{Z} = X + K\varepsilon_1 x_1 - L\varepsilon_2 x_2 + M\varepsilon_3 x_3 - \varepsilon_4 x_4 - \sum_{i=5}^{9} \left\{ (-1)^i \alpha_i \varepsilon_i - (1 - \alpha_i)\varepsilon_i' \right\} x_i .$$

K, L and M are set to increase the noise of Z, when needed (set K, L and $M \geq 1$). The α_i are random variables that are independent of ε_i, e.g., $\alpha_i \sim \mathrm{Uniform}(0,1)$ or $\alpha_i = \mathrm{Mod}(h_i, 8)/7$.

3.2 Some results

Let $\varepsilon_i, \varepsilon_i' \sim (0, \sigma_\varepsilon^2)$, $\alpha_i \sim \mathrm{Uniform}(0,1)$, i.i.d. and let K, L and M be fixed (for now). It follows that:

$$E(Z) = X \text{ and } V(Z) = \left\{ K^2 x_1^2 + L^2 x_2^2 + M^2 x_3^2 + x_4^2 + \tfrac{2}{3} \sum_{i=5}^{9} x_i^2 \right\} \sigma_\varepsilon^2 .$$

Let X_{-1}, X_{-2}, X_{-3} and Z_{-1}, Z_{-2}, Z_{-3} equal X and Z for the cell after removing units 1, 2 and 3, respectively. Keeping subscripts from the original cell (i.e., subscript 2 refers to the unit that was second in X) we have:

$$\begin{aligned}
Z_{-1} &= X_{-1} + K\varepsilon_2 x_2 - L\varepsilon_3 x_3 + M\varepsilon_4 x_4 - \varepsilon_5 x_5 - \textstyle\sum_{i=6}^{10}\left\{(-1)^i\,\alpha_i\varepsilon_i - (1-\alpha_i)\,\varepsilon_i'\right\} x_i,\\
Z_{-2} &= X_{-2} + K\varepsilon_1 x_1 - L\varepsilon_3 x_3 + M\varepsilon_4 x_4 - \varepsilon_5 x_5 - \textstyle\sum_{i=6}^{10}\left\{(-1)^i\,\alpha_i\varepsilon_i - (1-\alpha_i)\,\varepsilon_i'\right\} x_i, \text{ and}\\
Z_{-3} &= X_{-3} + K\varepsilon_1 x_1 - L\varepsilon_2 x_2 + M\varepsilon_4 x_4 - \varepsilon_5 x_5 - \textstyle\sum_{i=6}^{10}\left\{(-1)^i\,\alpha_i\varepsilon_i - (1-\alpha_i)\,\varepsilon_i'\right\} x_i.
\end{aligned}$$

We can obtain Z_{-i} for other units similarly. If we estimate the dropped units as $\hat{x}_i = Z - Z_{-i}$ it can be shown that, with $G = 2\tfrac{2}{3}x_5^2 + 2\sum_{i=6}^{9} x_i^2 + \tfrac{2}{3}x_{10}^2$,

$$\begin{aligned}
E(\hat{x}_i) &= x_i,\\
V(\hat{x}_1) &= \left\{K^2 x_1^2 + (K+L)^2\,x_2^2 + (L+M)^2\,x_3^2 + (M+1)^2\,x_4^2 + G\right\}\sigma_\varepsilon^2,\\
V(\hat{x}_2) &= \left\{L^2 x_2^2 + (L+M)^2\,x_3^2 + (M+1)^2\,x_4^2 + G\right\}\sigma_\varepsilon^2,\\
V(\hat{x}_3) &= \left\{M^2 x_3^2 + (M+1)^2\,x_4^2 + G\right\}\sigma_\varepsilon^2,\\
V(\hat{x}_4) &= \left\{x_4^2 + G\right\}\sigma_\varepsilon^2,\\
V(\hat{x}_5) &= \left\{\tfrac{2}{3}x_5^2 + 2x_6^2 + 2x_7^2 + 2x_8^2 + 2x_9^2 + \tfrac{2}{3}x_{10}^2\right\}\sigma_\varepsilon^2,\\
V(\hat{x}_6) &= \left\{\tfrac{2}{3}x_5^2 + \tfrac{2}{3}x_6^2 + 2x_7^2 + 2x_8^2 + 2x_9^2 + \tfrac{2}{3}x_{10}^2\right\}\sigma_\varepsilon^2,\\
V(\hat{x}_7) &= \left\{\tfrac{2}{3}x_5^2 + \tfrac{2}{3}x_6^2 + \tfrac{2}{3}x_7^2 + 2x_8^2 + 2x_9^2 + \tfrac{2}{3}x_{10}^2\right\}\sigma_\varepsilon^2,\\
V(\hat{x}_8) &= \left\{\tfrac{2}{3}x_5^2 + \tfrac{2}{3}x_6^2 + \tfrac{2}{3}x_7^2 + \tfrac{2}{3}x_8^2 + 2x_9^2 + \tfrac{2}{3}x_{10}^2\right\}\sigma_\varepsilon^2,\\
V(\hat{x}_9) &= \tfrac{2}{3}\left\{x_5^2 + x_6^2 + x_7^2 + x_8^2 + x_9^2 + x_{10}^2\right\}\sigma_\varepsilon^2, \quad \text{and}\\
V(\hat{x}_i) &= \tfrac{2}{3}\left\{x_5^2 + x_6^2 + x_7^2 + x_8^2 + x_9^2\right\}\sigma_\varepsilon^2, \quad \text{for} \quad i > 9.
\end{aligned}$$

If we assume that $K,\ L$ and M are fixed we can set them based on some requirement for $\mathrm{V}(\hat{x}_i)$. For example, we may want to have $\mathrm{V}(\hat{x}_i) = x_i^2/30$ since, for $z \sim \mathrm{N}(0,1)$, $\Pr(|z| > 0.44) = 0.66$ which for $\hat{x}_i \sim \mathrm{N}(x_i, x_i^2/30)$ gives $\Pr\{|\hat{x}_i - x_i| \geq 8\% x_i\} = 66\%$.

To obtain $\mathrm{V}(\hat{x}_i) = x_i^2/NN$ we can solve (fixed) $K,\ L$ and M in reverse order. This gives

$$M = \frac{\sqrt{(x_3^2 + x_4^2)(x_3^2/NN\sigma_\varepsilon^2 - G) - x_3^2 x_4^2} - x_4^2}{x_3^2 + x_4^2}$$

$$L = \frac{\sqrt{(x_2^2 + x_3^2)(x_2^2/NN\sigma_\varepsilon^2 - G - x_4^2(M+1)^2) - M^2 x_2^2 x_3^2} - Mx_3^2}{x_2^2 + x_3^2}$$

$$K = \frac{\sqrt{(x_1^2 + x_2^2)(x_1^2/NN\sigma_\varepsilon^2 - G - x_3^2(L+M)^2 - x_4^2(M+1)^2) - L^2 x_1^2 x_2^2} - Lx_2^2}{x_1^2 + x_2^2}$$

In practice, L and M are bounded below at 1 and above at some threshold value less than 2, and K is bounded below at 1 and can taper off above the threshold. Also, the target values of K, L and M depend on the situation in each cell. Here, for simplicity of illustration, they were assumed not to change when we removed observations from the cell.

Using the same noise and changing its direction for even-ranked units means that we take advantage of the correlation between the Z and Z_{-i} to increase the variance of $\hat{x}_i = Z - Z_{-i}$. For example, the contribution to $V(\hat{x}_1)$ from unit 2 is $(K+L)^2\, x_2^2\sigma_\varepsilon^2$. If we had used independent (or unit-cell specific) noises ε_i' instead of ε_i for units 1 to 4 the contribution from unit 2 would have been only $(K^2+L^2)\, x_2^2\sigma_\varepsilon^2$.

3.3 Comparison with the EZS and TCM approaches

With EZS the perturbed cell total is simply $Z = X + \sum_{i\in cell}\varepsilon_i\, x_i$, giving $V(Z) = \sum_{i\in cell} x_i^2\sigma_\varepsilon^2$. For any unit i we have $E(\hat{x}_i) = x_i$ and $V(\hat{x}_i) = x_i^2\sigma_\varepsilon^2$, which is smaller than the equivalent variance with the LPM for the same level of noise σ_ε^2 even when we set $K = L = M = 1$. A possible exception could be unit 5, if subsequent units are relatively quite small. This can be seen by examining $V(\hat{x}_5)$ above.

The TCM applies three multiplicative perturbation factors to the largest, say 4, units in each cell. A magnitude component M_i determines the relative size of the perturbation for the i^{th} ranked unit. The M_i are fixed; typically $M_1 > M_2 > M_3 > M_4$, e.g., $[0.6,\ 0.4,\ 0.3,\ 0.2]$. A permanent random factor $d_i = \pm 1$ fixes the direction of the noise for each unit i. A pseudo-random factor $s_i > 0$ determines unit-cell specific noises. This gives $Z = X + \sum_{i=1}^{4} M_i d_i s_i x_i$. The method can be represented in a form comparable to LPM, with $[M_1,\ M_2,\ M_3,\ M_4] = [K,\ L,\ M,\ 1]$, $d_i = \operatorname{sign}(\varepsilon_i)$ and $s_i = |\varepsilon_i'|$. The way the d_i are fixed is a major difference with the LPM that greatly diminishes the protection offered to $\hat{x}_1$. To illustrate this, consider two adaptions of these methods that yield identical variances for Z:

$$\begin{aligned} Z_{LPM} &= X + K\varepsilon_1 x_1 - L\varepsilon_2 x_2 + M\varepsilon_3 x_3 - \varepsilon_4 x_4, \quad \text{and} \\ Z_{TCM} &= X + Ksign(\varepsilon_1)\,|\varepsilon_1'|\,x_1 + Lsign(\varepsilon_2)\,|\varepsilon_2'|\,x_2 + Msign(\varepsilon_3)\,|\varepsilon_3'|\,x_3 + sign(\varepsilon_4)\,|\varepsilon_4'|\,x_4, \end{aligned}$$

where the same notational conventions as before are used, with fixed $K, L, M > 0$. This yields

$$\begin{aligned} V_{LPM}(\hat{x}_1) &= \left\{K^2x_1^2 + (K+L)^2\, x_2^2 + (L+M)^2\, x_3^2 + (M+1)^2\, x_4^2 + x_5^2\right\}\sigma_\varepsilon^2, \quad \text{and} \\ V_{TCM}(\hat{x}_1) &= K^2x_1^2\sigma_\varepsilon^2 + \left\{(K^2+L^2)\, x_2^2 + (L^2+M^2)\, x_3^2 + (M^2+1)\, x_4^2\right\}\sigma_{|\varepsilon|}^2 + x_5^2\sigma_\varepsilon^2. \end{aligned}$$

Not only are factors such as $(K+L)^2$ larger than (K^2+L^2), but the variance for the noise, σ_ε^2, is often replaced with that of the absolute noise, $\sigma_{|\varepsilon|}^2$, which is much smaller. For the split triangular distribution it goes from $(3a^2 + 2ab + b^2)/6$ to $(b-a)^2/18$. When $b = 2a$ this means dropping from $11a^2/6$ to $a^2/18$.

This is not a legitimate comparison of the two methods. We are not using the actual LPM, and method parameters need not be identical. But it shows the impact of the different approaches taken for the d_i.

4 Empirical investigation

We applied the LPM and EZS methods to personal data from a taxation file. Two variables were used: $x = \text{income}\ (\text{if} > 0)$ and $y = x^2$ (to increase skewness). Cells of between 15 and 148 units were generated by combining age groups within postal code, sex and marital status. Different levels of noise (ε_i) from a split triangular distribution were tried. Results presented are those with $\sigma_\varepsilon^2 = 0.006.$ Following the Risk-Utility Framework (Duncan, Keller-McNulty and Stokes 2001) the impacts of the methods on data accuracy and on risk were examined.

Table 4.1 shows the impact of the LPM on the quality of cell totals by cell size range. The LPM was applied 500 times in each cell. For each cell size range the table gives the number of cells, their average coefficient of variation (CV) after perturbation, and the percentage of times that the perturbed total was within 2%, 5%, 8% and 12% of the original cell total. For this study we assumed that cells that failed a $P-$ percent sensitivity rule with $P = 15$ would be suppressed – so they were not included in the results. There were more such cells with variable y (which may resemble business data more). As expected, the impact of the perturbation was higher for smaller cells, and for variable $y.$ All cells perturbed by more than 8% were near-sensitive and would have been suppressed with $P = 20.$

Table 4.1
Impact of layered perturbation method on cell totals

	Variable = Income (x)						Variable = Income^2 (y)				
Cell	**Num.**	**Avg.**	**% times relative distance ≤**				**Num.**	**Avg.**	**% times relative distance ≤**		
size	**cells**	**CV**	**2%**	**5%**	**8%**	**12%**	**cells**	**CV**	**2%**	**5%**	**8%**	**12%**
15 – 18	1,822	2.37	58.5	95.1	99.5	100.0	1,777	4.09	34.5	72.0	92.4	99.6
19 – 25	2,230	2.03	66.2	97.2	99.7	100.0	2,185	3.71	38.1	77.1	94.4	99.7
26 – 40	1,920	1.57	78.2	99.1	99.9	100.0	1,899	3.24	44.2	82.8	96.0	99.8
41 – 148	1,312	1.05	92.1	99.5	99.9	100.0	1,301	2.53	57.1	90.0	97.7	99.9
All	7,284	1.82	72.1	97.6	99.7	100.0	7,162	3.47	42.3	79.7	94.9	99.7

Note: values of 100.0 represent values above 99.95 that were rounded to 100.

Table 4.2 gives the impact of the EZS multiplicative noise, for the same $\sigma_\varepsilon^2,$ on the cell totals. Results for income (x) are fairly similar, while results for y are noticeably better. Similar results were obtained when a value for σ_ε^2 near 0.014 was used (LPM was slightly better with $x,$ EZS slightly better with $y).$

Table 4.2
Impact of EZS multiplicative noise on cell totals

	Variable = Income (x)						Variable = Income^2 (y)				
Cell	**Num.**	**Avg.**	**% times relative distance ≤**				**Num.**	**Avg.**	**% times relative distance ≤**		
size	**cells**	**CV**	**2%**	**5%**	**8%**	**12%**	**cells**	**CV**	**2%**	**5%**	**8%**	**12%**
15 – 18	1,822	2.33	58.7	97.1	100.0	100.0	1,777	3.19	41.2	86.4	99.8	100.0
19 – 25	2,230	2.08	64.5	98.5	100.0	100.0	2,185	2.93	45.2	90.0	99.9	100.0
26 – 40	1,920	1.74	73.9	99.6	100.0	100.0	1,899	2.59	51.4	93.8	99.9	100.0
41 – 148	1,312	1.30	86.9	99.9	99.9	100.0	1,301	2.09	63.4	97.1	100.0	100.0
All	7,824	1.91	69.6	98.7	99.9	100.0	7,162	2.76	49.2	91.4	99.9	100.0

Note: values of 100.0 in the 8% columns represent values above 99.95 that were rounded to 100.

We next examined the amount of protection offered to the largest units in each cell. For each cell, an estimate $\hat{x}_i$ for unit x_i was obtained by differencing perturbed cell totals with and without the unit. Relative differences $d_i = 100 \lvert \hat{x}_i - x_i \rvert / x_i$ were calculated and incorporated in a score equal to $\sum_{cells} r_i$, where $r_i = 1$ if $d_i < 10$, $r_i = 0$ if $d_i > 15$ and $0 < r_i < 1$ otherwise. Table 4.3 shows the quartiles of d_i and the scores for variables x and y for the largest twelve units in each cell with LPM, and for the largest unit with EZS (EZS offers the same level of protection to all units).

With the LPM the largest three units tend to be protected the most, as expected. Patterns for variables x and y are different. If one looks at the quartiles of d_i for variable x, the level of protection gradually declines until unit 10 and increases afterwards. Since the $\mathrm{V}(\hat{x}_i)$ are the same for $i > 9$ results should keep improving after the 10th largest unit. The scores give a similar story. For variable y the descent is not as regular, with unit 5 being protected the least (unit 10 if one looks at Q1 only). The weaker protection around units 5 and 10 is predicted by the formulas for $\mathrm{V}(\hat{x}_i)$, whose basic form changes around those two units. Unit 10 is vulnerable to repeated targeted attacks the most, where an attack consists of obtaining an estimate $\hat{x}_{10}$ from totals for units 1 to 10, and for units 1 to 9, with some set of smaller units (e.g., obtain $\hat{x}_{10(i)}$ from totals excluding unit i and excluding units i and 10, for $i = 11, 12, 13\ldots$). Averaging the $\hat{x}_{10(i)}$, if there are enough of them, may give good estimates of x_{10}. Such attacks require carefully set up tabulation requests, which a semi-controlled custom tabulation environment could discourage.

Table 4.3
Protection of largest twelve units with LPM and of largest with EZS (quartiles for d_i)

	Variable = Income (x)					Variable = Income² (y)				
	Cells	**Q1**	**Med**	**Q3**	**Score (%)**	**Cells**	**Q1**	**Med**	**Q3**	**Score (%)**
Unit 1	7,962	7.9	15.7	26.6	3,196 (40)	7,823	7.6	14.4	23.2	3,365 (43)
Unit 2	7,962	8.6	17.5	29.3	2,895 (36)	7,782	7.2	15.0	25.2	3,311 (43)
Unit 3	7,962	8.1	16.9	28.7	3,021 (38)	7,782	6.6	14.1	24.2	3,522 (45)
Unit 4	7,962	7.2	15.5	26.2	3,314 (42)	7,799	6.1	13.3	22.5	3,726 (48)
Unit 5	7,962	6.4	13.9	23.8	3,647 (46)	7,808	5.5	11.9	20.5	4,052 (52)
Unit 6	7,962	6.4	13.9	23.3	3,614 (45)	7,811	6.0	12.6	21.6	3,885 (50)
Unit 7	7,962	6.2	13.3	22.4	3,765 (47)	7,814	6.0	12.6	22.2	3,868 (50)
Unit 8	7,962	6.3	13.4	22.3	3,731 (47)	7,818	6.5	13.8	23.7	3,581 (46)
Unit 9	7,962	5.1	11.5	19.9	4,267 (54)	7,818	5.7	13.0	24.2	3,750 (48)
Unit 10	7,962	3.3	10.7	20.9	4,373 (55)	7,818	4.4	13.5	27.4	3,704 (47)
Unit 11	7,962	3.8	11.8	22.4	4,121 (52)	7,818	4.8	15.7	32.1	3,422 (44)
Unit 12	7,962	3.8	12.2	24.7	4,031 (51)	7,820	5.8	17.9	37.9	3,110 (40)
U1/EZS	7,962	6.7	7.5	8.4	7,941 (100)	7,823	6.7	7.5	8.5	7,803 (100)

In contrast, results for EZS show that the level of protection offered to unit 1 (or for any unit for that matter) is fairly constant, and it is generally much poorer than that with the LPM. The score for EZS is almost 100%, a very poor outcome. But EZS was designed to offer protection for totals, not to protect from differencing. If protection from differencing is required then the level of noise would have to be set much higher to protect values at levels comparable to the LPM. But with EZS units around unit 10 would not be more vulnerable to repeated targeted attacks.

To investigate the roles of K, L and M we generated random values from a uniform distribution, but created an outlier in each cell by setting the value of x_1 as the highest value that would not make the cell sensitive, i.e., for $P = 15$, set $x_1 = \frac{100}{15}\sum_{i \geq 3} x_i$. The LPM was used with M set to 1, and K and L either calculated as suggested above or set to 1. For our generated data the calculated value of L never left 1. Table 4.4 shows that factor K is useful because when it is set to 1 the level of protection for the outlier is not high enough when $\sigma_\varepsilon^2 = 0.006$.

Table 4.4
Protection of outliers in artificial populations for 1,000 cells (quartiles for *d*₁)

Standard LPM (K ≥ 1)				LPM with *K* = *L* = *M* = 1			
Q1	**Med.**	**Q3**	**Score**	**Q1**	**Med.**	**Q3**	**Score**
11.1	12.6	14.2	472	6.7	7.5	8.6	996

5 Discussion and challenges

We presented a perturbative method for protecting tables of magnitude in a custom tabulation environment. The method is not resource intensive – it is only necessary to keep track of the largest units in each cell and their permanent random number. We have shown that the method is able to protect the largest units from a differencing attack.

Since perturbation is applied to the largest values, and sensitive cells are suppressed, there is less need to use variable-specific noise to protect ratios. Ratios can be calculated using perturbed values (Z). Likewise, means can be calculated using the Z values and perturbed (e.g., rounded) frequencies. Alternatively, if users prefer, means can be calculated by dividing Z by the true frequencies, and totals obtained by multiplying the perturbed means by perturbed frequencies.

Zeroes are not treated, but X (and Z) are suppressed for sensitive and small cells. If a non sensitive cell has less than 5 nonzero values then the addition of another zero-valued unit will not affect Z. So, in that particular situation, users may be able to tell if a unit added to the cell was zero-valued. If unit values x_i can be negative the largest absolute values $|x_i|$ in each cell could be treated (perturbed). Dominance rules would need to be adapted for negative values (e.g., see Tambay and Fillion 2013).

Residual disclosure issues with related outputs such as unperturbed totals and tables of distributions remain. If the Agency released some unperturbed totals, a hacker could try differencing attacks with the unperturbed total as the starting point. It would be preferable to keep unperturbed results to a minimum, e.g., only for official releases. Tables of distribution (e.g., total income by income range) may also present problems of residual disclosure because of the information conveyed by the ranges. One approach would be to severely restrict the ranges that can be used in such tables.

Table additivity is not maintained, and suppressed cells complicate the use of raking to restore additivity. One solution would consist of imputing those cells, raking, then suppressing the imputed cells. We could start by imputing lone suppressions in a row or column based on other cell values (bottom code at 0 if needed) and repeat this if it generated new lone suppressions in a row or column. Other methods can be used to impute values for remaining suppressed cells.

References

Cox, L.H., and Dandekar, R.A. (2004). A new disclosure limitation method for tabular data that preserves data accuracy and ease of use. *Proceedings of the 2002 FCSM Statistical Policy Seminar,* Statistical Policy Working Paper 35, Federal Committee on Statistical Methodology, Washington, DC.

Cox, L.H., and Sande, G. (1979). Techniques for preserving statistical confidentiality. *Proceedings of the 42nd Session of the International Statistical Institute*, Manila, Philippines.

Duncan, G., Keller-McNulty, S. and Stokes, S. (2001). *Disclosure Risk vs. Data Utility: The r-u Confidentiality Map*. Technical Report LA-UR-01-6428, Los Alamos National Laboratory, Statistical Sciences group, Los Alamos, New Mexico.

Evans, T., Zayatz, L. and Slanta, J. (1998). Using noise for disclosure limitation of establishment tabular data. *Journal of Official Statistics,* 14, 537-551.

Giessing, S. (2011). Post-tabular stochastic noise to protect skewed business data. *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality,* Tarragona, Spain, October 26-28, 2011.

Massell, P., and Funk, J. (2007). Recent developments in the use of noise for protecting magnitude data tables: Balancing to improve data quality and rounding that preserves protection. *Proceedings of the Research Conference of the Federal Committee on Statistical Methodology,* Arlington, Virginia.

Tambay, J.-L., and Fillion, J.-M. (2013). Strategies for processing tabular data using the G-Confid cell suppression software. *Proceedings of the Survey Research Methods Section,* American Statistical Association Joint Statistical Meetings, Montreal, August 3-8, 2013.

Thompson, G., Broadfoot, S. and Elazar, D. (2013). Methodology for the automatic confidentialisation of statistical outputs from remote servers at the Australian Bureau of Statistics. *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality,* Ottawa, October 28-30, 2013.