WP. 47
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Tarragona, Spain, 26-28 October 2011)

Topic (ix): Statistical disclosure limitation for table and analysis servers: how to make outputs of modern data access infrastructures safe

Post-tabular Stochastic Noise to Protect Skewed Business Data

Prepared by Sarah Giessing, Federal Statistical Office, Germany

Post-tabular Stochastic Noise to Protect Skewed Business Data[1]

Sarah Giessing
Statistisches Bundesamt, 65180 Wiesbaden, Germany, Email: Sarah.Giessing@destatis.de

Abstract: Many statistical agencies nowadays operate or envision tools for ad hoc creation and visualization of aggregate tables, ideally with web-access facilities. Users should be able to easily create their own customized tables. However, especially with heavily skewed business data, disclosure control issues usually are a big obstacle in this context, hardly solvable by traditional methods like cell suppression. For population counts data the Australian Bureau of Statistics has implemented a SDL method based on post-tabular noise addition. The US Census Bureau on the other hand has developed a method based on pre-tabular multiplicative noise to protect enterprise tabular data.
This paper proposes a new method based on an idea for post-tabular stochastic noise. The method proved to give encouraging results when tested on tabulations of German business tax statistics.

1 Introduction

Facilities to generate user requested tables in a convenient and flexible way should be part of a modern data production process in a Statistical Office. The problem is how to ensure non-disclosiveness of the tabular outputs, especially in the case of strongly skewed magnitude tabular data. Traditional methods like cell suppression have proven to work well when the aim is to protect a rather fixed set of not too detailed tables. Otherwise, it becomes difficult to manage risks of disclosure by differencing problems. Perturbative SDC methods may offer a way out of the dilemma.

The first question with a perturbation method is whether to perturb the input, i.e. the underlying micro-data, which is what pre-tabular methods do. Or to rather perturb the output, i.e. the tabular aggregates, which is what post-tabular methods do. One of the challenges when designing a post-tabular perturbation method is to ensure between tables consistency. When tables present inconsistent results it may damage the users trust in the data and may also lead to disclosure risk. Disclosure risk can arise for instance when the mean of inconsistently perturbed but logically identical sensitive values is an unbiased estimate of the true value. So by "*consistency*" we mean that different queries should lead to an identically perturbed result whenever they are logically identical, i.e. referring to the aggregate value for the same variable and the same group of respondents. (Fraser and Wooton, 2006) propose a methodology to achieve this by using microdata keys. As for the perturbation, they propose (in the context of frequency counts data) to add noise with zero mean and

[1] This work has been partially supported by the project DwB INFRA-2010-262608 of the EU FP7

constant variance. The method is therefore certainly not suitable for skewed business magnitude data.

In the context of enterprise data sets, (Evans et al., 1998), and (Höhne, 2008) propose several variants of pre-tabular (multiplicative) random noise for masking microdata, i.e. a pre-tabular method.

Combining ideas of those two different concepts, this paper drafts a new post-tabular perturbation method. The paper is organized in 5 sections. Section 2 introduces the methodological concept. One disadvantage of the method is that it is not additive. Section 3 is concerned with this issue. A way out is to extend the method by a rounding strategy which gives users of the data a natural explanation for lacking additivity ("rounding effect"), and also provides a local measure of information loss due to perturbation. Empirical results are presented in section 4. The paper finishes with a summary section and some conclusions.

2 Pre- and post-tabular stochastic noise for masking skewed magnitude data

There is a wide range of literature on micro data masking using random noise. For the sake of simplicity, in this paper we use as starting point a simple variant of multiplicative stochastic noise in the formulation of (Höhne, 2005, c.f. "Zufallsüberlagerungsmodell 3"). Section 2.1 briefly outlines the basic concept of the method. Section 2.2 introduces a new post-tabular variant of the approach, to be used in combination with the micro-data keys concept of the ABS method. In section 2.3 we compare theoretical properties of the two approaches.

2.1 A simple pre-tabular stochastic noise masking method

The idea of the method in (Höhne, 2005) can be outlined as follows: the disseminator specifies two parameters μ_0 and σ_0^2. Variable value y_i in the ith micro-data record is masked by multiplying it with $(1 \pm (\mu_0 + z_i))$, where z_i is drawn from a $N(0, \sigma_0^2)$ distribution. Parameter σ_0 should be chosen relatively small compared to μ_0. This means we multiply the data alternatively by $(1 + \mu_0)$ or $(1 - \mu_0)$, approximately. In that sense parameter μ_0 determines the strength of the masking. (Höhne, 2005) suggests to select the deviation sense according to the following algorithm: Sort the records according to the values of variable y in descending order. Let the deviation sense be positive, when the total of the perturbed previous values is smaller than the corresponding original total and negative otherwise. Compared to a random selection of deviation senses this scheme helps to better preserve the grand total.

2.2 A simple post-tabular stochastic noise masking method

An advantage when masking table cell values rather than micro-data values is that cell sensitivity can be taken into account. The masking method lined out in the following requires a certain minimum deviation between true and masked cell value for sensitive cells. We can, for example, require that the masked cell value is non-sensitive according to the sensitivity rule employed.

This can be achieved by masking the largest contribution y_1 of a table cell with original value T^{orig} and replacing it by $y_1^{post(T)}$ when computing the perturbed cell value T^{post} : Just as in 2.1 we choose parameters μ_0 and σ_0^2. When a cell c is sensitive, we multiply the largest contribution y_1 by r_c , where r_c := (1 ± (μ_0 + abs(z_c))) (z_c drawn from a N(0, σ_0^2) distribution[2]). When cell c is non-sensitive, y_1 is multiplied by only (1 ± (abs(z_c))). This way, the perturbation of sensitive cell values will be much stronger than that of non-sensitive cell values. When using the p% sensitivity rule, we choose μ_0 := 2*p/100. Then the noisy value T^{post} of a sensitive true value T^{orig} will be non-sensitive, i.e. $|T^{post} - y_1 - y_2| \geq p*y_1$, ($y_1$ and y_2 denoting the two largest contributions to T^{orig})[3]. Finally, we determine the deviation sense according to the deviation sense that would result for this cell when using the pre-tabular method of 2.1. See (Giessing, 2011) for a rational behind that.

In order to achieve between-tables consistency, i.e. to achieve that cells which are logically identical will always have the same masked cell value, we use the random mechanism proposed in (Fraser and Wooton, 2006). For our experiments, we used the SAS random number generator which produces pseudo random numbers distributed uniformly over $[0;2^{31}-1]$. We assign such a random key to each record in the microdata file, the "*microdata keys*". When computing the tables, also the random keys are aggregated. The result is then transformed back into a random number on this interval by applying the modulo function, $\mathrm{mod}_{2^{31}-1}$. If the same group of respondents is aggregated into a cell, the resulting random key will always be the same. Cells which are logically identical thus have identical random keys. This key is then used for drawing the random noise z_c.

2.3 Statistical properties of the two masking methods

In the following, it is assumed that the effect of the scheme for assigning perturbation senses is negligible, i.e. we assume random selection of the deviation sense with

---

[2] Note, abs(z_c) are then distributed according to a normal distribution truncated at zero.

[3] Proof: With denotations from above, and because of our choice of parameter μ_0 it holds:

(1) $|y_1 - y_1^{post(T)}| \geq \mu_u |y_1| \geq 2*(p/100)*|y_1|$. Using reverse triangle inequality we can say $|T^{post} - y_1 - y_2| = |(T^{orig} - y_1 - y_2) - (y_1 - y_1^{post(T)}| \geq | |T^{orig} - y_1 - y_2| - |y_1 - y_1^{post(T)}| | = | |y_1 - y_1^{post(T)}| - |T^{orig} - y_1 - y_2|$. If a cell value is sensitive according to the p%-rule, we have $|T^{orig} - y_1 - y_2| \leq (p/100)*|y_1|$. Hence, because of (1), $|T^{post} - y_1 - y_2| \geq |2\ p/100 - p/100|\ |y_1| = p|\ |y_1|$.

probability of ½ for both, positive and negative sense. It is then easy to see (c.f. Giessing, 2011, appendix) that both masking methods will provide unbiased estimates for all cell values. For a first theoretical comparison of information loss caused by either of the two methods, we consider the variance of the noise on the cell level. Denoting by t_{pre} and t_{post} the difference between true and perturbed cell value and by σ^2_{pre} and σ^2_{post} the noise variances, it is straightforward to show (see Giessing, 2011, appendix) that the cell level noise variances are

(2.3.1) $V(t_{pre}) = \sigma^2_{pre} \Sigma_{i=1,\dots,n} (y_i^2)$ in case of the pre-tabular method, and

(2.3.2) $V(t_{post}) = \sigma^2_{post} \; y_1^2$ for the post-tabular method.

(Giessing, 2011, appendix) also proves that the noise variances are

(2.3.3) $\sigma^2_{pre} = (\sigma_0^2 + \mu_0^2)$, and

(2.3.4) $\sigma^2_{post} = (b\,\sigma_0^2 + (l_s \mu_0 + a\,\sigma_0)^2)$. The positive constants a and b are $a=2*(2\pi)^{1/2}\sim 0{,}8$, $b= (2\pi-4)/(2\pi)\sim 0{,}36$, and l_s is a binary variable ($l_s = 1$ for sensitive cells and zero otherwise). Because of $a^2+b=1$ it turns out that the noise variance for non-sensitive cells is $\sigma^2_{post} = \sigma_0^2$. So, for non-sensitive cells, unless μ_0 is chosen to be zero, $V(t_{post}) < V(t_{pre})$. I.e., the cell level noise variance is smaller for the post-tabular method. Note that this is not generally true for non-sensitive cells.

3 How to deal with non-additivity?

If table additivity is really needed, for instance if a user wants to use the output of the tabulation as input to further analysis which eventually requires additiviy, there are ways to restore table additivity. Leaver, V. (2009) points out that restoring additivity can be achieved by iterative methods. Alternatively, a linear programming based method could be considered, like e.g. the Controlled Tabular Adjustement package of (Castro, González, 2009). The algorithm restores additivity to a table, minimizing an overall (L1-norm-) distance to the table provided as input. The distance function implemented is a weighted sum of absolute per-cell-distances. Weights are provided by the user of the software. The user can define for each cell upper and lower bounds on the deviations (‘*a priori bounds’*), and can define a set of cells labelled as ‘*sensitive cells*’. For each sensitive cell, the user defines a ‘*protection interval*’. The adjusted cell value is not allowed to take a value inside the protection interval. Computational complexity of the problem depends strongly on the number of sensitive cells. For empirical experiment, we therefore did not flag any cells as “sensitive”, but gave large weights to them: The post-tabular method always changes the values of sensitive cells sufficiently, if parameters are defined properly (see footnote 3). It should therefore be enough to avoid that the adjustment tends to change those perturbed data back to the original data. Adding suitable *a priori* constraints to the problem ensures that the adjusted values are not too far off from the noisy cell values.

Of course one might consider using the adjustment methodology without previous random perturbation, flagging sensitive cells to have their adjusted values forced out of the respective protection intervals. But unless this yields a fully consistent data base[4], there is then a risk that by averaging (adjusted) cell values over a number of (adjusted) tables a user can recover the original data. With a previous random perturbation, such an approach will only recover the underlying perturbed table, as pointed out in (Leaver, 2009).
Anyway, unless the adjustment yields a fully consistent data basis which is not an option considered in this paper, it creates inconsistency and may further increase the information loss caused by the perturbation. If users do not really need an exactly additive table, and non-additivity is only to be avoided because it might be irritating to users, there might be a better way out by taking up a rounding strategy. The "rounding effect" can serve as a natural explanation for the lack of additivity. Rounding also provides a natural, local measure of information loss caused by the perturbation.
Basically, the idea of the rounding concept suggested here is to publish a kind of confidence interval computed on basis of the perturbed data. Considering the noise variance given by (2.3.2), the size of the confidence interval will not be constant, but will be proportional to the largest contribution to a cell. The same must then hold for any rounding intervals covering the confidence intervals. Therefore, when dealing with strongly skewed data, i.e. data where the largest contributions to cells vary a lot, selecting individual rounding bases for each cell will be the only sensible option.

3.1 Confidence intervals for post-tabular noise

The confidence interval we use to determine a rounding basis is computed considering the following theoretical publication scenario: We assume the formulation of the post-tabular method of sec. 2.2 to be known to the users. Although publishing the exact parameters for the noise might be too risky, we further assume enough information to be made available so that users could imagine (upper) estimates for those parameters. Recalling definitions and denotations of 2.2 and 2.3 the noisy value of a cell c can be written as $T^{post} = T^{orig} - y_1 + r_c\, y_1 = T^{orig} + (r_c - 1)\, y_1$ where $r_c = (1 + d_c\, u_c)$ with binary variable d_c denoting the deviation sense of the noise in cell c and noise u_c drawn from a truncated normal distribution with mean μ_u and variance σ_u^2. Hence $T^{orig} = T^{post} - d_c\, u_c\, y_1$. We can thus compute the upper and lower bound for f.i. a 99% (i.e. 3-σ) confidence interval by $UB(T^{orig}) = T^{post} + (\mu_u + 3\,\sigma_u)\, y_1$ and $LB(T^{orig}) = T^{post} - (\mu_u + 3\,\sigma_u)\, y_1$ [5]. Inserting the formulas for mean and standard

[4] which normally would impose a huge problem, far from easy to define and usually impossible to solve with today's computing capacities

[5] Note that $UB(T^{post}) = T^{orig} + (\mu_0 + 3\,\sigma_0)\, y_1$ and $LB(T^{post}) = T^{orig} (\mu_0 + 3\,\sigma_0)\, y_1$ [5] hold as well, because the bounds for $(r_c - 1)$ and $(1 - r_c)$ are the same (i.e. bounds for $\pm\, d_c\, u_c$).

deviation of a truncated normal distribution, these bounds become
$UB(T^{orig}) = T^{post} + (l_s \mu_0 + a\sigma_0 + 3 b^{½} \sigma_0) y_1$ and $LB(T^{orig}) = T^{post} - (l_s \mu_0 + a\sigma_0 + 3 b^{½} \sigma_0) y_1$.
As the user will often not know, if a cell is sensitive or not, we replace the binary variable l_s by the probability p_c for cell c to be sensitive, assuming the user might estimate this probability. Taking into account also that a user could only guess the true parameters μ_0 and σ_0, we assume a relative error of respective (upper) estimates, denoted by ε_μ and ε_σ. The formulas then turn into

(3.1.1) $UB(T^{orig}) = T^{post} + (p_c \mu_0 (1+\varepsilon_\mu) + (a + 3 b^{½}) (1+\varepsilon_\sigma) \sigma_0) y_1$ and

$LB(T^{orig}) = T^{post} - (p_c \mu_0 (1+\varepsilon_\mu) + (a + 3 b^{½}) (1+\varepsilon_\sigma) \sigma_0) y_1$.

For the empirical experiment of sec. 5 we simulate the user's estimate for p_c by $\min(1; p\, y_1 / (T - y_1 - y_2))$. I.e. for sensitive cells we assume that the user always correctly guesses the cell to be sensitive. For non-sensitive cells the probability is non-zero and thus slightly inflates the confidence interval compared to the one that would be computed, if the non-sensitivity of the cell was actually known.
As for the largest contribution y_1 which will generally be unknown to a user, one might assume that users can guess an upper estimate with a relative deviation of at most $q \geq 0$. For the experiment of sec. 5, we use the worst case assumption $q=0$, i.e. y_1 to be known.

3.2 From a confidence interval towards a rounding basis

For a naïve user, the concept of a confidence interval might be confusing. A rounded value where the rounding interval covers the confidence interval might be easier to "sell". Another important issue is that official statistics are supposed to report the "truth". We should choose a rounding basis in such a way that the difference between a rounded original cell value and a rounded noisy value becomes very small. Even if users mistake the perturbed rounded data for rounded original data, not too much harm can come from it then. For the kind of magnitude data we consider here, restricting the choice of rounding basis to the powers of 10 (10, 100, 1000, etc) seems to be a natural option.
We discuss now three alternative rules for finding a suitable rounding basis. Starting point of the idea is that when the rounded confidence interval bounds coincide, the true value must be within the corresponding rounding interval. This is the concept of the strictest of the three rules, R1, which rounds the noisy value to the smallest rounding basis where the rounded bounds of the confidence interval deviate only very slightly (e.g. by at most a maximum difference given by parameter *dist*. It is suggested to set *dist* to 1 in general and to 0 for sensitive cells). R2 and R3 are more relaxed. R2 decreases the rounding basis: it is enough, if the rounded true and rounded noisy value deviate by at most *dist*., given that the size of the confidence interval is less than 100 (after rounding). This means normally that the rounded bounds of the confidence interval coincide except for the last two digits on the right-hand side. R3 is even more relaxed. Again the rounded true and rounded noisy value

must not deviate by more than *dist.*, but now no condition is imposed on the bounds of the confidence interval. An illustrative example is: True value T^{orig} = 156 764, noisy value T^{post} = 156 755, confidence interval [155 463;158 047]; parameter *dist.* set to 1. Then for rounding base 100 (=10^2) the distance between the rounded bounds of the confidence interval is 1580-1555=25 and the distance between rounded true and rounded noisy value is 1568-1568=0. Hence, according to rule R2, 10^2 is the appropriate rounding base.
Another – perhaps more theoretically appealing – approach might be the following: Replace in the confidence interval formula (3.1.1) the factor 3 by a variable u_α expressing the α-quantile of the N(0,1) distribution. For a given rounding interval, compute the distances between its bounds and the true cell value. Let *DR* the smaller one. Releasing the rounded value means implicitly that the disseminator confirms that T^{orig} - *DR* (or T^{orig} + *DR*) is a lower (or upper) bound for the true value. Compute then the probability of a (symmetric) confidence interval with width of two times *DR* as follows: equate *DR* with the distance between bounds and midpoint of the confidence interval (3.1.1) and solve this equation for u_α . For more detail see (Giessing, 2011). Looking up the value of the N(0,1)-distribution function at u_α gives the probability for the *2 DR*-width interval. The smaller the probability of this interval, the larger would be the risk connected to release of the rounded value. A sensible approach could be for the disseminator to fix a threshold (like f.i. 90, 95, 97.5 or even 99) for a minimum acceptable probability. If the computed probability is below the threshold, the noisy value must be rounded to a larger basis. With the data of the above illustrative example, we obtain u_α = 1.64 for basis $B = 10^2$. This quantile corresponds to a probabilitiy of 94.95 %.
Regarding the presentation of the rounded data, it would be easy to publish them in scientific notation, like f.i. 1 567.X E+02, using the character '**X**' as a warning that subsequent digits are not necessarily zero. One might imagine that many users would not be comfortable with this format. Filling digits chopped off through the rounding with '**X**' in blocks of three, like 1 567 XXX , might be another option. Note that for sensitive cells the rounding basis will usually turn out to be larger than the cell value. So for sensitive cells, the 'rounded' value would typically look f.i. like this:
X XXX XXX.

4 Disclosure risks

Ideally, the masked value should be at some safe distance from the true value for all sensitive cells. As pointed out in 2.2, for the post-tabular masking method, if parameters are chosen properly, this will always be the case. After restoring additivity this does not generally hold anymore, because some cells are adjusted into proximity of the original value. Also with the pre-tabular noise method of sec. 2.1 a number of masked cells will usually be sensitive according to a concentration rule because positive and negative deviations of the masked individual data tend to

compensate each other. In those cases however, users cannot to be sure that they actually managed to obtain a close estimate of a true value.
A different type of risk arises, when intervals are released for all, or for some cells, like in the case of the rounding method of sec. 3.2. Considering the relations between e.g. inner and marginal cells, users could then try to compute close bounds of sensitive cells by differencing in a certain way between respective interval bounds. Theoretically, they could consider all table relations and the bounds of published intervals and compute - by solving some LP problems - feasibility intervals for sensitive cells (see Hundepool et al., 2009). In the process of fixing parameters and thresholds for the above rounding method, it is therefore important to evaluate the risk of the method by e.g. computing and analyzing those feasibility intervals for a number of test tables.

5 Test results

The methods presented above have been tested on tabular data of the German business tax statistics presenting turnover. Table 1 below presents results regarding the distribution of non-sensitive cells by relative deviation of noisy and true values for three of the test tables: each of them involves a NACE industry classification down to the most detailed (5-digit) level. The first two tables have a size-class dimension. All tables involve geography, either down to the state, the district or the municipality level.
Obviously, the post-tabular noise proposed in sec. 2.2 ("Post") preserves the quality of the data much better than the pre-tabular method of sec. 2.1. ("Pre")[6] . This confirms the theoretical results of section 2.3. We also see that this effect becomes the stronger the more detailed the tables are. Looking at cols. "Rd" we find that the additional information loss observed when the post-tabular noisy data is rounded (according to rule R2 of section 3.3) is not very much. Adjusting the noisy data to restore additivity using the CTA method mentioned in sec. 3 has a stronger effect, c.f. cols. "Adj". Note that the additive solution considered here has been forced to be "consistent" with the rounding, e.g. in the additive solution the cell values of non-sensitive cells are within the bounds of the respective rounding intervals. Comparing the relative deviations it becomes clear that the adjustment further increases the information loss, e.g. the midpoints of the rounding intervals tend to be closer to the true values than the adjusted values.
Generally, the impression is that the methods perform better on the size class tables. This has probably to do with the fact that the contributions to a cell with a size-class grouping tend to be more homogeneous than to cells without size-class grouping.

[6]Note that the same parameters μ_0 and σ_0^2 have been used for both methods

Range of rel. dev. (in %)	NACE5 x State x SizeCl, *124 204 non-sensitive cells*				NACE5 x Distr x SizeCl, *38 256 non-sensitive cells*[7]				NACE5 x Municipality, *4 811 non-sensitive cells*[7]			
	Pre	**Post**	**Rd**	**Adj**	**Pre**	**Post**	**Rd**	**Adj**	**Pre**	**Post**	**Rd**	**Adj**
0-1	*22.8*	*88.6*	*87.2*	*75.5*	*13.9*	*82.8*	*81.1*	*63.8*	*6.5*	*60.1*	*61.6*	*38.7*
1-2	*14.8*	*8.9*	*8.2*	*12.0*	*12.1*	*13.4*	*11.6*	*16.7*	*6.5*	*26.0*	*20.2*	*21.2*
2-3	*10.9*	*1.9*	*2.6*	*5.1*	*10.0*	*2.9*	*4.1*	*7.6*	*6.4*	*8.7*	*8.9*	*12.1*
3-4	*8.5*	*0.4*	*1.0*	*2.7*	*8.7*	*0.7*	*1.6*	*4.2*	*5.5*	*3.4*	*4.8*	*6.8*
4-5	*7.0*	*0.1*	*0.5*	*1.6*	*7.4*	*0.2*	*0.8*	*2.7*	*5.4*	*1.3*	*1.8*	*5.6*
5-6	*5.7*	*0.1*	*0.2*	*1.0*	*6.8*	*0.1*	*0.4*	*1.6*	*4.8*	*0.2*	*1.2*	*3.3*
6-7	*4.9*	*0.0*	*0.1*	*0.6*	*5.9*	*0.0*	*0.2*	*1.0*	*4.8*	*0.2*	*0.8*	*3.0*
7-8	*4.1*	*0.0*	*0.1*	*0.4*	*5.2*	*0.0*	*0.1*	*0.7*	*5.4*	*0.1*	*0.5*	*1.9*
8-9	*3.5*	*0.0*	*0.0*	*0.3*	*4.7*	*0.0*	*0.0*	*0.5*	*6.0*	*0.1*	*0.1*	*1.1*
9-10	*3.1*	*0.0*	*0.0*	*0.2*	*4.2*	*0.0*	*0.0*	*0.4*	*5.2*	*0.0*	*0.2*	*1.4*
≥ 10	*14.7*	*0.0*	*0.0*	*0.5*	*21.2*	*0.0*	*0.1*	*0.8*	*43.6*	*0.0*	*0.2*	*4.8*

Table 1 Distribution of non-sensitive cells by relative deviation of the noise

Regarding disclosure risk that might arise from publishing the rounding intervals, feasibility intervals have been computed for the above three test tables. Indeed, for none of the 47 069, 32 317 and 4 811 sensitive cells in the three tables a feasibility interval was computed where one of the bounds would be too close to the true cell value.

The second type of disclosure risk mentioned in sec. 4 arises when users find directly by looking at the noisy data, or indirectly (by taking means etc.) results that are too close to the true value. For the pre-tabular method we observed for example for the three test tables 15, 13 and 4 % of sensitive cells, where the noisy value is too close to the original value. This kind of direct disclosure cannot happen for the post-tabular method. What we found, however, is that when imposing that adjusted values must be within the rounding intervals, adjusted values will quite often be too close to the true values of sensitive cells. Especially for the two size class tables we found about 40 % of sensitive cells that were adjusted into proximity of the original value. Of course users would not know which of the adjusted data would be a close estimate of a true value.

6 Summary and final conclusions

This paper has proposed a post-tabular method to protect skewed business magnitude data by multiplicative stochastic noise in combination with use of micro-data keys. The methodology has been proposed for automatic disclosure control in the context of modern facilities to generate user requested tables in a convenient and flexible way.

[7] The state-level table relates to the whole country, the district-level table to one of the states and the municipality table to only one district.

In order to avoid user irritation because the noisy tables will not be additive and also to make the effect of the noise transparent on the level of the individual table cells, it has been proposed to round the perturbed data. A specific rounding rule determines for each cell an individual rounding basis that depends on the width of a confidence interval around the noisy value.
The method has been proven theoretically and empirically to clearly outperform a corresponding multiplicative noise masking method for micro-data. A theoretical disclosure risk that could be imagined to arise when the rounding intervals cause feasibility intervals for the sensitive cells that are too close has not been confirmed in empirical tests. Hence, the method seems to protect the data properly.
In principle one might imagine to enhance the method by a module to restore exact table additivity. On the other hand this will lead to inconsistency between released tables. It is therefore recommended to rather use rounded data for publication. Rounded data are consistent *and* additive (apart from rounding differences). On special request one might consider to release adjusted tables that add up exactly. Test result show that Controlled Tabular Adjustment can provide additive solutions that are coherent with the rounding intervals.

References

Castro, J., Gonzalez J.A. (2009). *A Package for L1 Controlled Tabular Adjustment*, paper presented at the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality (Bilbao, 2-4 December 2009)

Evans, T., Zayatz, L. and Slanta, J. (1998) *Using Noise for Disclosure Limitation of Establishment Tabular Data*. Journal of Official Statist., 4, 537-551.

Fraser, B., Wooton, J. (2006). A proposed method for confidentialising tabular output to protect against differencing, in *Monographs of Official Statistics. Work session on Statistical Data Confidentiality*, Eurostat-Office for Official Publications of the European Communities, Luxembourg, 2006, pp. 299-302

Giessing, S. (2011), '*Report: Post-tabular Stochastic Noise to Protect Skewed Business Data*', Research Report.

Höhne, J. (2008): *Anonymisierungsverfahren für Paneldaten*. In: Springer, Wirtschafts- und Sozialstatistisches Archiv (2008) Bd. 2, p. 259-275

Hundepool, Anco, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Rainer Lenz, Jane Longhurst, Eric Schulte Nordholt, Giovanni Seri and Peter-Paul De Wolf (2009), *ESSNET handbook on Statistical Disclosure Control,* ESSNET-SDC project, available at http://neon.vb.cbs.nl/casc/handbook.htm

Leaver, V., (2009) '*Implementing a method for automatically protecting user-defined Census tables'*, paper presented at the Joint ECE/Eurostat Worksession on Statistical Confidentiality in Bilbao, December 2009, available at http://www.unece.org/stats/documents/2009.12.confidentiality.htm