USING NOISE FOR DISCLOSURE LIMITATION OF ESTABLISHMENT TABULAR DATA

Laura Zayatz, Timothy Evans, and John Slanta, Bureau of the Census[1]
Laura Zayatz, Commerce/Census/SRD/3209-4, Washington, DC 20233
laura.zayatz@ccmail.census.gov

ABSTRACT

We propose a new disclosure limitation method for establishment magnitude tabular data in which noise is added to the underlying microdata prior to tabulation. The proposed method has several advantages compared to the standard method of cell suppression: it enables some information to be provided within more cells of the table, it eliminates the need to coordinate cell suppression patterns between tables, and it is a much less complicated and time-consuming procedure than cell suppression. In this paper we outline the proposed procedure for adding noise to the underlying establishment microdata, discuss the advantages and disadvantages of adding noise as compared to cell suppression, and describe the results of using noise with data from one survey.

KEYWORDS: **Confidentiality, Cell Suppression, Magnitude Data**

1. INTRODUCTION

The responding unit in many economic surveys and censuses conducted by statistical agencies is the establishment. Individual establishments' responses are weighted (where appropriate) and estimates of quantities of interest such as value of shipments are generally produced by categorical variables like Standard Industrial Classification (SIC) code and geography. The categorical variables define a table (for example the rows might be SIC code and the columns might be geographical areas). Then the "quantity of interest" is aggregated over all units of analysis in each cell. Such tables are called tables of **magnitude** data. Given the geographic information and other characteristics on which tables are based, in conjunction with common knowledge and publicly available sources, it is generally a reasonable assumption that the set of establishments contributing to a cell in such a table is well known to data users.

Many statistical agencies collect information promising that all responses will be held confidential. Those same agencies attempt to release as much statistically valid and useful data as possible without violating the confidentiality pledge. Techniques used to protect data confidentiality are called "disclosure limitation" procedures (see Federal Committee on Statistical Methodology, 1994 for a review of disclosure limitation methodologies for all types of data and for an annotated bibliography of the literature on disclosure limitation). The disclosure limitation procedures for magnitude data are designed to prevent data users from being able to recover any respondent's reported values using values appearing in the published tables. A statistical agency must ensure that a cell value does not closely approximate data for any one respondent in the cell and, moreover, that one respondent or a coalition of respondents cannot subtract their contribution(s) from the cell value to achieve a "close" estimate of the contribution of another respondent (Cox and Zayatz, 1993).

2. CELL SUPPRESSION

The current widely accepted disclosure limitation technique used for establishment magnitude tabular data is cell suppression. Cells that pose a disclosure risk are typically identified using one of two rules --- the *n-k* rule or the *p%* rule (see Federal Committee on Statistical Methodology, 1994 for a detailed explanation of these rules). All cells that fail the disclosure rule are called sensitive cells. In the context of cell suppression, these cells are also often called primary suppressions.

---

[1]This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review by the Census Bureau than its official publications. This report is released to inform interested parties and to encourage discussion.

Cell suppression limits disclosure by removing from publication (suppressing) all sensitive cells plus sufficiently many additional cells, called complementary suppressions, to ensure that the values of the primary suppressions cannot be narrowly estimated through manipulation of additive relationships between cell values and totals (Cox and Zayatz, 1993). When a cell is suppressed, its total value is removed from the cell and replaced with some type of symbol indicating that the value is withheld to prevent disclosure.

While the concepts behind determining whether a particular cell is a disclosure risk are relatively simple, the process of choosing complementary suppressions to protect these sensitive cells is quite complicated. The methodology by which complementary suppressions are chosen, as well as the accompanying computer software, is very difficult to understand for anyone without a background in linear programming. Because of the structure of the computer programs, often the process must be performed separately for each data product. Among other things, this means that agency staff must keep track, from one data product to the next, of which cells have previously been suppressed (and hence must be suppressed and protected in all subsequent data products) and which cells have previously been published (and hence can't be used as complementary suppressions). Coordinating suppression patterns among tables becomes impracticable in the presence of multiple requests for special tabulations which frequently follow standard publications. The agency must keep an ongoing record of all interrelationships between all cells across all publication tables and special tabulations, a programming nightmare. Many agencies simply do not have the resources to do this.

Another major drawback of cell suppression is that it suppresses much information that is not at risk for disclosure. Any cell that is used as a complementary suppression but that is not itself a primary suppression represents information that could have been published if there were some other way of protecting the sensitive cells. Particularly at fine levels of detail, including most special tabulations, the need for complementary suppressions often results in tables full of suppressed cells.

3. INTRODUCTION OF NOISE TO MICRODATA PRIOR TO TABULATION

3.1 General Description

We propose an alternative method of protecting individual respondents. The method involves adding noise to their data. **This procedure should not be confused with noise procedures aimed at protecting and releasing public use establishment microdata**. **This procedure is for establishment magnitude tabular data.** We propose perturbing each responding establishment's data by a small amount, say 10% (the percent to remain confidential within the statistical agency). Then if a cell contains only one establishment, or if a single establishment dominates the cell, the value in the cell will not be a close approximation to the dominant establishment's value because that value has had noise added to it (in this case, it has been changed by about 10%). By adding noise, we avoid disclosing the dominant establishment's true value.

To each establishment in our sample or census we assign a multiplier, or noise factor. Then all establishments have their values multiplied by their corresponding noise factors before the data are tabulated. Note that because the same multiplier is used with an establishment wherever that establishment is tabulated, values are consistent from one table to another. That is, if the same cell appears on more than one table, it has the same value on all tables.

Note that we add noise to each establishment prior to any tabulations. This is *not* the same as attempting to add noise on a cell-by-cell basis. We rely on a random assignment of the multipliers to control the effects of the noise on different types of cells. The noise should have its greatest impact on sensitive cells, while the effect of the noise on cells that would not be primary suppressions should be minimal. Thus we aim to protect individual establishments without compromising the quality of our non-sensitive estimates.

3.2 The Multipliers

For purposes of illustration, let us assume we want to introduce roughly 10% noise into each establishment’s values. The actual percentage used by a statistical agency would be confidential. To perturb an establishment's data by about 10%, we multiply its data by a number that is close to either 1.1 or 0.9. We could use any of several types of distributions from which to choose our multipliers, and the distributions would remain confidential within the agency.

Whatever distribution we decide to use for generating multipliers near 1.1, it is of paramount importance that we use the same shape distribution, or rather its "mirror image," to generate multipliers near 0.9. In other words, if we consider the two distributions together, the overall distribution of the multipliers should be symmetric about 1. The reason for this condition is discussed in Section 3.3.

Under current practices, the unit of analysis for disclosure limitation is the *company*. That is, we seek to protect respondent data at the company level as well as for individual establishments within the company. Because company-level values must be protected, all noise for a single company should either inflate or deflate that company's true values. In other words, all establishments from the same company should be perturbed in the same direction and hence have approximately the same multiplier. This way, if all of the establishments contributing to a cell belonged to the same company, the resulting cell estimate would be perturbed by, for our example, about 10%. By perturbing all of a company's establishments in the same direction, we ensure that company-level data is protected.

3.3 Assignment of Multipliers and Its Effect on Estimates

We want to assign the multipliers in such a way that we minimize the effect of the noise on those cells that are not at risk for disclosure. In particular, cell values at higher levels of aggregation are not generally sensitive, and we would like these values to contain as little noise as possible. In this section, we will concern ourselves with data from a census. The next section (3.4) extends the methodology to survey data.

We begin by randomly assigning each responding company a *direction* of perturbation. Using our example with 10% as our base for perturbation, this is equivalent to determining if all establishments in that company will have multipliers close to 1.1 or close to 0.9. We then randomly assign a multiplier to each establishment within a company. The multipliers would be generated from that half of the overall distribution of the multipliers that corresponds to the direction of perturbation assigned to that company. An example of potential assignments is as follows:

Example 1:

Company	Establishment	Direction	Multiplier
Company A		1.1	
	Establishment A1		1.12
	Establishment A2		1.09
	Establishment A3		1.10
	Establishment A4		1.11
Company B		0.9	
	Establishment B1		0.89
	Establishment B2		0.93
Company C		1.1	
	Establishment C1		1.08

Intuitively, the expected value of the amount of noise present in any cell value is zero, thanks to the symmetry of the distribution of the multipliers and the random assignment of direction of perturbation and multipliers within the companies. The probability that a company's establishments will be perturbed in a positive direction is equal to the probability that they will be perturbed in a negative direction. The distribution of the multipliers is symmetric about 1. The expected value of any given multiplier is 1, hence the expected value of the *amount* of noise in any given establishment is 0, and the amount of noise in any cell value is simply the sum of the noise in its component establishments. Let Y be the noise free cell value and Y_N be the noise added cell value. Thus for establishments i in cell j, we have:

$Y = \sum_{i \in j} value_i$ and

$Y_N = \sum_{i \in j} (multiplier_i \times value_i)$.

Let $e = Y - Y_N$.

Given that $E(multiplier_i) = 1$, we have

$$
\begin{aligned}
E(Y_N) &= E(\sum_{i \in j} multiplier_i \times value_i) \\
&= \sum_{i \in j} E(multiplier_i \times value_i) \\
&= \sum_{i \in j} (value_i \times E(multiplier_i)) \\
&= \sum_{i \in j} (value_i \times 1) \\
&= \sum_{i \in j} value_i \\
&= Y,
\end{aligned}
$$

and hence $E(e)=0$. Thus the noise procedure does not introduce any consistent bias into the cell values.

For non-sensitive cells, values are not altered a great deal as we will see in Section 4. For these cells, the establishments that are perturbed in the positive direction and those that are perturbed in the negative direction will generally balance each other out. In contrast, a cell that is dominated by a single contributor will most likely contain a large amount of noise. If the largest contributor is very large compared to all others in the cell, it is much less likely that positively-perturbed establishments and negatively-perturbed establishments will cancel each other out when determining the amount of noise present in the cell value. Looked at another way, the more dominant the largest contributor, the more the amount of noise present in the cell value will resemble the amount of noise present in the largest contributor (about 10%). Thus the cells that are at greatest risk for disclosure in general receive the most noise, and the noise in the cell total prevents users from being able to recover an individual respondent's true value from the published value. This is illustrated in the following examples:

Example 2:

Sensitive Cell

	True Establishment Values	Noise-added Establishment Values
	10000	10000 × 1.11 = 11100
	300	300 × 0.89 = 267
	200	200 × 1.12 = 224
Cell Total	10500	11591

Note that in this example of a sensitive cell, the cell value is changed by (11591 - 10500) * 100 / 10500 = 10.39%.

Example 3:

Non-sensitive Cell

	True Establishment Values	Noise-added Establishment Values
	10000	10000 × 1.11 = 11100
	8000	8000 × 0.89 = 7120
	5000	5000 × 1.12 = 5600
Cell Total	23000	23820

In this example of a non-sensitive cell, the cell value is changed by (23820-23000) * 100 / 23000 = 3.57%.

3.4 Adding Noise to Census Data versus Sample Survey Data

When generating tables from census data using the noise approach for disclosure limitation, a responding establishment's contribution to a cell simply becomes establishment value × multiplier. The examples in Section 3.3 show how noise would be applied and cells would be tabulated for census data.

In sample surveys, each respondent's data is generally weighted inversely proportional to the establishment's probability of being included in the sample. For establishments with large weights, the weight itself offers some protection against disclosing the respondent's actual reported values (Willenborg and de Waal, 1998). For sample data, to reflect the protection already provided by the sample weight, noise is applied as follows:

For each establishment in a cell, the establishment's contribution to the cell becomes

$$\text{establishment value} \times \left[\text{multiplier} + (\text{weight} - 1)\right].$$

These noise-added establishment contributions are then added to obtain total cell value. Note that noise is added only to one multiple of each establishment's value, and the remaining (weight - 1) multiples, which conceptually represent the contributions of other unsampled establishments, have no noise added. Also notice that as the weight approaches 1, i.e., as the establishment comes closer and closer to representing only itself, the formula degenerates to the census case, as the following example illustrates.

Example 4:

True Establishment Contributions		Noise-added Establishment Contributions		Percent
Value × Weight		Value × [Multiplier + (Weight - 1)]		Difference
10000 × 1	= 10000	10000 × [0.89 + (1 - 1)]	= 8900	11.00%
800 × 5	= 4000	800 × [1.12 + (5 - 1)]	= 4096	2.40%
500 × 7	= 3500	500 × [0.91 + (7 - 1)]	= 3455	1.29%
Cell Total	17500		16451	5.99%

This procedure has the effect of changing the relative contributions for certainty or near-certainty establishments (those having weights close to or equal to 1) by a large amount while changing the relative contributions for establishments with large weights by a small amount. This is desirable because we are more concerned with the disclosure risk of certainty establishments, whose values aren't protected by their weights. It can be shown (Evans, et al, 1998) that the noise procedure does not introduce any consistent bias into the cell values. Note also that we have the following properties:

$COV(\hat{Y},e)=0$ and

$$\begin{aligned}\sigma^2(\hat{Y}_N) &= \sigma^2(\hat{Y}+e)\\ &= \sigma^2(\hat{Y})+\sigma^2(e)+2\,COV(\hat{Y},e)\\ &= \sigma^2(\hat{Y})+\sigma^2(e)+0\\ &= \sigma^2(\hat{Y})+E(e^2)-(E(e))^2\\ &= \sigma^2(\hat{Y})+E(e^2).\end{aligned}$$

Thus the introduction of noise causes the variance of an estimate to increase by an amount equal to the square of the noise term. An unbiased estimator of the variance is:

$$\hat{\sigma}^2(\hat{Y}_N)=\hat{\sigma}^2(\hat{Y})+e^2=\hat{\sigma}^2(\hat{Y})+(\hat{Y}_N-\hat{Y})^2.$$

3.5 Flagging Cells with a Large Amount of Noise

The percentage of noise in a cell is defined as the percent by which the noise-added value for the cell differs from the true noise-free value. All resulting table cells containing a large percentage of noise, say a 7% change in value or more (the chosen percentage would be confidential) are flagged so users would know that the values may not be useful. This set of cells will encompass most sensitive cells, as well as a few non-sensitive cells that received a lot of noise simply through randomness. The description of the flag explains how and why noise was added and lets users know that disclosure limitation has been performed. We also use the same flag on any cells that were identified as sensitive (i.e., failed the disclosure rule) before noise was added but that, because of randomness of multipliers, did not exceed our noise threshold (7% in our example). In this case, users at least *think* the cell contains a lot of noise and hesitate to treat the cell value as reliable.

4. RESULTS WITH ACTUAL SURVEY DATA

To get an idea of how well the noise technique actually works in practice, we tested it with data from the U.S. Bureau of the Census' Research and Development (R&D) Survey, a survey of companies' research and development expenses. In this survey, estimates of R&D expenses are computed for 26 SICs or SIC groupings, and within each SIC the expenses are separated into corporate-sponsored R&D and federally-sponsored R&D.

We randomly assigned responding companies a direction of perturbation. To then generate the establishment multipliers, we experimented with several distributions and chose scaled Beta distributions --- X ~ .1 B(6,2) + 0.8 and X ~ .1 B(2,6) + 1.1. After generating a multiplier for an establishment, we applied the multiplier to the establishment's data items using the formula in Section 3.4 for sample data. We reproduced a table from the R&D publication, which shows R&D expenses broken out by federally-sponsored vs. corporate-sponsored, for the 26 SIC groupings. (Table 4.1 below shows the structure of this R&D table.)

Next we ran 1000 replications of the R&D table, using the Beta distribution to generate multipliers, and computed summary statistics to describe the behavior of the cells over all replications. For each cell in the table, we calculated the ratio of a) the average of the 1000 noise-added values of the cell to b) the true noise-free value. Thus if there is no tendency for the noise to change the value of a cell in any particular direction, the values in the table should be close to 1, i.e., the average noise-added value for any cell should be close to the true cell value. All of the values were indeed close to 1, for both sensitive and non-sensitive cells. The largest and smallest values were, respectively, 1.00326 and 0.99692. It is clear that the symmetry of the distribution of the multipliers and the randomness of the direction of perturbation ensure that the expected value of the noise present in any estimate is zero (i.e., the expected value of the ratio of the noise-added value to the noise-free value is 1). Hence the noise does not introduce any bias into the estimates as was shown in Section 3.3.

Note that while the expected value of the amount of noise in any one *establishment* is also zero (since the symmetry of the distribution of multipliers implies that the expected value of any particular multiplier is 1), in practice the added noise will not be zero because of the bimodality of the distribution; a multiplier can never actually equal 1. In fact, in the degenerate case where an estimate is composed of only one establishment, the estimate is guaranteed to contain at least 10% noise.

To get an idea of how much noise would typically be present in a cell after a *single* application of the noise, we looked at the standard deviation of the 1000 noise-added observations in each cell. We standardized these by dividing by the true cell value. If we consider the true value of the cell estimate $\hat{Y}$ to be "fixed" for purposes of adding noise, then the standard deviation of the noise-added values $\hat{Y}_N$ is simply the standard deviation of the noise itself: writing $\hat{Y}_N = \hat{Y} + e$ and taking $\hat{\sigma}(\hat{Y})$

and Cov($\hat{Y}$,e) to be zero, we have $\hat{\sigma}(\hat{Y}_N|\hat{Y})=\hat{\sigma}(e)$. The value in the table, $\frac{\hat{\sigma}(\hat{Y}_N|\hat{Y})}{\hat{Y}}$, can be thought

of as the coefficient of variation (CV) of the noise-added estimate, given the noise-free estimate, i.e., CV($\hat{Y}_N|\hat{Y}$). Table 4.1 below shows these conditional "CVs" over the 1000 replications. Again, sensitive cells are shaded.

Table 4.1 Conditional "CVs" of Noise-Added Values

standard deviation of the 1000 simulations, divided by the true noise-free estimate

stub #	total R&D	federal	company
1	0.03435	0.04836	0.03426
2	0.03344	0.12550	0.03335
3	0.01512	0.12511	0.01001
4	0.04448	0.05300	0.04414
5	0.06648	(-)	0.06648
6	0.03719	0.12121	0.03959
7	0.03875	0.12505	0.03586
8	0.02008	0.07398	0.01349
9	0.00294	0.10999	0.00747
10	0.01265	(-)	0.01265
11	0.02395	0.12604	0.02399
12	0.03213	0.12507	0.03246
13	0.02200	0.11817	0.00470
14	0.01596	0.01794	0.01589
15	0.04755	0.10916	0.00259
16	0.00937	0.01394	0.00961
17	0.04861	0.10957	0.01592
18	0.03150	0.00757	0.04686
19	0.01954	0.02110	0.01922
20	0.03700	(-)	0.03700
21	0.08972	0.09229	0.08896
22	0.01880	0.11383	0.00473
23	0.00324	0.04377	0.00979
24	0.00367	(-)	0.00367
25	0.04369	0.10209	0.01500
26	0.03716	0.07130	0.03320
TOTAL	0.01912	0.01843	0.01931

Note that the conditional CVs are generally much higher in the sensitive cells than in the non-sensitive ones. The variability of the amount of noise present in sensitive cells is much greater, so a sensitive cell should be much more likely than a non-sensitive cell to contain a large amount of noise after a single application of the noise procedure. This is exactly what we want, since it is the sensitive cells whose values need to be protected.

To confirm this idea, we looked at the amount of noise that was typically present in different types of cell. For each non-zero cell, we computed the *absolute* percent noise present in the cell for each replication. We then computed an overall "percent noise" by averaging these absolute percentages over all 1000 replications. (Note that if we did not use the absolute value of the percentage, the average over all replications would be close to zero and would tell us nothing about the typical behavior of the cell.) Then we looked at the distribution of this "percent noise" variable among cells of various types. Table 4.2 below gives the results.

Table 4.2 Amount of Noise in Non-Zero Cells, By Type of Cell

	amount of noise			
% noise in:	avg	median	max	min
marginal cells (29)	2.88	2.36	8.89	0.24
interior cells (48)	5.18	3.60	12.52	0.21
cells that would have been primary suppressions (11)	11.11	12.08	12.52	5.19
cells that would have been complementary suppressions (12)	2.77	3.24	4.73	0.24
unsuppressed cells (54)	3.27	2.00	11.32	0.21
all (nonempty) cells (77)	4.32	3.31	12.52	0.21

The distinction between marginal cells and interior cells shows that interior cells on average received more noise. This is a desirable result, since interior cells are composed of fewer establishments and are more likely to be sensitive. The noise technique appears to leave marginal estimates with relatively little noise, roughly between 2 and 3 percent.

Cells that would have been primary suppressions receive noticeably more noise than non-sensitive cells. Again, this is what we want, because these are the cells whose values need to be protected. Complementary suppressions are shown separately to illustrate the fact that the noise technique would allow these cells to be published with relatively little noise, thus providing data users with more information than would have been the case with cell suppression.

Because of the element of randomness in assigning multipliers, we don't expect *all* sensitive cells to receive a lot of noise (see Section 3.7), nor do we expect that none of the non-sensitive cells will receive a lot of noise. Table 4.3 below gives the breakdown, by type of cell, of which of the 77 nonzero cells in our test table received a lot of noise (where we define "a lot" as at least 7%) and which didn't receive much.

Table 4.3 Counts of Cells Having Large vs. Small Amounts of Noise

type of cell	$\lvert noise \rvert \geq 7\%$	$\lvert noise \rvert < 7\%$
sensitive (11)	10	1
non-sensitive (66):	7	59
complementaries (12)	0	12
unsuppressed (54)	7	47
marginal (29)	1	28
interior (37)	6	31
total (77)	17	60

This table further illustrates that the noise technique generally leaves non-sensitive cells (including marginal totals) relatively noise-free, while most sensitive cells receive a lot of noise. The few sensitive cells that don't exceed the noise threshold would be flagged as described in Section 3.5, along with both sensitive and non-sensitive cells that do exceed it.

5. CONCLUSIONS

Adding noise to establishment-level data before producing tables has several advantages over the traditional cell suppression techniques. First, it is a far simpler and less time-consuming procedure than cell suppression. Computer programs for adding noise are much easier to write, modify, run, and understand than the programs that currently exist for choosing cell suppression patterns.

Another important advantage of adding noise is that it eliminates the need to coordinate cell suppressions between tables. Under the current cell suppression practices, disclosure analysis involves keeping track, from one data product to another, of all cells that have previously been published and all cells that have previously been suppressed. Keeping track of suppressions is difficult to orchestrate and difficult to understand. However, using noise to protect estimates would make this unnecessary.

Also, with cell suppression, users lose information both for cells which are primary suppressions and for those that are complementary suppressions. With the noise technique, sensitive cells (those that would normally be primary suppressions) would in general contain a lot of noise and be flagged as such. In contrast, non-sensitive cells would end up with little noise, including most of the non-sensitive cells that would have been used as complementary suppressions. Thus for publications which normally contain many complementary suppressions, the noise technique should provide data users with more valuable information.

But what about protection? With cell suppression, although actual values are suppressed, data users can use linear programming techniques to calculate a possible range for each suppressed value. Statistical agencies assign primary and complementary suppressions to ensure that a respondent's value cannot be closely estimated (ranges must meet size requirements). With noise, data users may be able to obtain a point estimate that can be associated with a given respondent, but this estimate would contain "a lot" of noise (statistical agencies would determine the amount they feel comfortable with just as they determine the range size requirements). Some may argue that the added noise does not provide enough protection to values in single-establishment cells. Under the cell suppression approach, if a cell has only a single establishment contributing to it, the cell's value would be suppressed and the cell would simply contain a 'D'.

Using the noise technique, the cell would contain a flag noting that the value in the cell had been severely altered, but the actual value may still be derivable using other cells in the same row or column. The flag may lessen the *appearance* of disclosure, since no value would appear in the cell. However, the respondent may still feel uneasy about the derived number seeming to be an estimate of his actual value, even if the estimate contains a lot of noise and is flagged as being unreliable. The suppression approach may give the appearance of offering more protection.

It is possible that some respondents may resent putting time into preparing good responses if they know the statistical agency is going to add noise to them. We need to emphasize that noise would be added in an unbiased way so as to preserve the statistical properties of the data while having a negligible effect on non-sensitive estimates.

Also there may be concern on the part of some data users as to the quality of the data after noise has been introduced. The users' desire for multiple special tabulations and their desire to see more published cells (at the expense of noise) should be weighed against their desire for true values (at the expense of suppressions).

The results of our test with the R&D Survey indicate that the idea of adding noise as a disclosure limitation strategy warrants further consideration. We have thus far been concerned with the effect of noise on the behavior of the level estimates in our published tables, and in this regard it performs well. Under our scheme for assigning multipliers, the noise does not appear to introduce any bias into the estimates. We have also shown that, in general, sensitive cells end up containing larger amounts of noise than non-sensitive cells; thus noise provides protection where it is most needed. See Evans, Zayatz, and Slanta (1996) for ideas on the use of sorting and raking to better preserve non-sensitive estimates. Looking beyond the behavior of level estimates, further research is required to investigate the effect of noise on trend estimates, longitudinal studies, inter-variable relationships, and other types of analysis that data users typically perform with the published estimates (Evans, Zayatz, and Slanta, 1996; Evans, 1997).

The noise technique is probably not suitable for all data products; some surveys publish data at such levels of aggregation that disclosure is not an issue. However, for surveys in which cell suppression currently creates problems, the prospects are encouraging. In light of our results and the flexibility and simplicity that the noise technique offers, the addition of noise to the underlying microdata could become a viable alternative to cell suppression for disclosure avoidance with establishment tabular data.

6. REFERENCES

Cox, L.H., and Zayatz, L. (1993), "Setting an Agenda for Research in the Federal Statistical System: Needs for Statistical Disclosure Limitation Procedures," *Proceedings of the Section on Government Statistics*, American Statistical Association, pp. 121-126.

Evans, B. T. (1997), "Effects on Trend Statistics of the Use of Multiplicative Noise for Disclosure Limitation," *Proceedings of the Section on Government Statistics*, American Statistical Association, to appear.

Evans, B. T., Zayatz, L., and Slanta, J. (1996), "Using Noise for Disclosure Limitation of Establishment Tabular Data, " *Proceedings of the Annual Research Conferenc*e, Bureau of the Census, Washington, DC 20233, pp. 65-86.

Evans, B. T., Zayatz, L., and Slanta, J. (1998), "Using Noise for Disclosure Limitation of Establishment Tabular Data, Journal of Official Statistics, Vol. 14, No. 4, pp. 537-551.

Federal Committee on Statistical Methodology (1994). *Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology*, U.S. Office of Management and Budget, Washington, DC.

Willenborg, L. and de Waal, T. (1998), "Statistical Disclosure Control and Sampling Weights," *Journal of Official Statistics*, **13**, No. 3, pp. 417-434.

MODEL BASED DISCLOSURE LIMITATION FOR BUSINESS MICRODATA

Luisa Franconi, Istat, and Julian Stander, University of Plymouth

Luisa Franconi, Istat, Servizio Studi Metodologici,
Via A. Depretis 74/B, 00184 Roma, Italy
franconi@istat.it

ABSTRACT

We outline a methodology for statistical disclosure control for business microdata. For quantitative public variables such as the number of employees our approach releases broader classes instead of the exact value. For other sensitive quantitative variables such as turnover it builds a statistical model that takes account of the geographical area to which each enterprise belongs. We use the Gibbs sampler to draw inferences from our model. We describe how to simulate realisations from the predictive density. We propose limiting disclosure of sensitive quantitative variables such as turnover by releasing predictive intervals or other summaries of the predictive density. We briefly discuss how decision theory could play a role here. We illustrate our approach on data from the Community Innovation Survey. We present interesting estimates of area effects from our model. These estimates suggest a broader categorisation to use when releasing the qualitative public variable geographical area that goes a long way to minimising information loss.

Key Words: Area effects, Community Innovation Survey, Gibbs sampler, Public variables, Predictive density, Spatial model.

1. INTRODUCTION

1.1. Background

Disclosure limitation concerns the methodologies, strategies and techniques employed by National Statistical Institutes to avoid the possibility that sensitive information about individuals or enterprises can be inferred from released data. In this paper we concentrate on disclosure limitation of business microdata. A microdata file comprises a sample from the population of enterprises of interest. It typically contains general information about each enterprise such as economic activity, number of employees and geographical area, and particular information, the nature and form of which depends on the survey. In the case of the Community Innovation Survey this particular information mainly concerns innovation.

The disclosure of sensitive information is usually directly associated with the concept of identification. Identification of a unit or enterprise in the sample being released occurs when it is possible to connect the name of the enterprise to that unit. Although the released microdata file does not contain direct identifiers such as names or fiscal codes, an identification may still be made by using *a priori* information and other data generally available from public registers. The *a priori* information may relate to knowledge about the inclusion of the unit in the sample. In order to provide the best possible representation of the population, business survey designs include the largest and most identifiable enterprises with probability one; see Cox (1995). Moreover, very detailed public registers are available that contain the names of enterprises together with such features as their main economic activities, number of employees and geographical area. The match between public registers and an unprotected sample can often be an easy task, especially when *a priori* information is available, with the result that identification and hence disclosure is accomplished without too much difficulty. In order to make identification a difficult task, some disclosure limitation procedures decrease the information content of all variables that in one way or another may lead to identification. Such procedures are applied to all public variables such as economic activity, number of employees and geographical area, and to all other sensitive quantitative variables such as turnover and export that may lead to indirect identification, as we shall see later.

Disclosure limitation for business microdata has up to now been achieved by masking and microaggregation procedures (Duncan and Pearson 1991; Cox 1994; Defays and Nanopoulos 1992), by data swapping (Fienberg, et al. 1996), or by simulation from relevant distributions (McGuckin and Nguyen 1988; Fienberg 1994). These approaches are not completely satisfactory. In some cases the perturbation imposed on the data to protect the enterprises has to be so large that the information loss is extremely severe. In other cases, such as individual

ranking, the level of perturbation imposed may not be sufficient to protect the data. Finally, some of the simulation processes that have been suggested can be difficult to implement. In this paper we propose a methodology for disclosure limitation of business microdata based on the predictive distribution associated with a particular statistical model for a sensitive quantitative variable. Our method makes use of the geographical structure underlying the data and the characteristics of the variables being surveyed. A full description of the data with which we work is presented in the next section.

1.2. Data Available from the Community Innovation Survey

At the beginning of the 1990s the European Commission and Eurostat began a survey of technological innovation in European manufacturing and services sector enterprises, called the Community Innovation Survey (CIS). The objective of this survey was the production of comparable data harmonised at the European level on all technological activities, not just research and development. The aim was that this data would provide economist and decision-makers with a better understanding of innovation patterns and trends. In 1998 the second CIS, pertaining to the period 1994–96, took place. The results of this survey will be published soon. Economists and the general research community have shown such an interest in CIS microdata that the problem of the release of a 'Microdata for Research' file has arisen.

The definition of technological innovations used in the CIS is taken from the 1997 edition of the Oslo Manual on Innovation written by Organisation for Economic Co-operation and Development and Eurostat. Specifically, technological innovations refer to all new technological products and processes introduced by an enterprise and to all significant technological improvements made by the enterprise to products and processes. The products and processes are required to be new to the enterprise but not necessarily to the market.

The data with which we work come from a representative sample of Italian manufacturing and services sector enterprises with twenty or more employees. These microdata are the result of a single stage stratified random sample, the stratification being with respect to economic activity as based on the two digit NACE rev. 1, the size of the enterprise and geographical area.

From the point of view of disclosure limitation, the variables of the CIS can be divided into two sets. The first contains all the general information about the enterprise such as its main economic activity, number of employees, geographical area, turnover, export, and group membership. The first three of these variables are public, and direct or indirect identification of the enterprise can be made through them. The variables turnover, export and group membership are sensitive, and knowledge of them, together with the public variables, can lead to disclosure, as we will explain in Section 1.3. The second set of variables is related to innovation and contains the remaining sensitive microdata. In particular, for each enterprise in the sample, questions are posed on a range of issues: on the scope of the innovation (new products or process, unsuccessful or not yet completed development projects); on the type of innovation (research and development services, training, government support); on the objectives of innovation (replacement of products, improvement of quality, fulfilment of standards, reduction of labour or energy costs); on the sources of information for innovation (within the enterprise, competitors, clients, universities); on innovation co-operation and on factors hampering innovation. For simplicity we shall work with a variable that indicates whether or not an enterprise is involved in the innovation of products or processes or both.

1.3. The Proposed Approach and Outline of the Paper

Many of the questions on innovation that the enterprises are asked allow an answer that takes the form of a point of view on a subject rather than a precise numerical value. For example, for the question about the objectives of innovation possible replies are 0 for not relevant, and 1, 2 and 3 according to the degree of importance of particular objectives in a given list. As a consequence, most of the answers of interest can be represented by binary or categorical variables, rather than quantitative variables. Categorical variables carry less risk for disclosure limitation than quantitative variables. This is because knowledge of the value of a quantitative variable, even though it is not publicly available, can lead to the identification of an enterprise. For example, information about turnover can lead to the identification of a very large and well-known enterprise in a particular division of the NACE. On the other hand knowledge of a binary variable indicating, for example, whether or not the enterprise has introduced product innovation does not allow for such identification since both small and large enterprises can carry out product innovation.

Protection of the CIS data can therefore be achieved by releasing less detailed information about the public variables and sensitive quantitative variables such as turnover that may lead to identification. In particular, this can be achieved for categorical public variables such as geographical area by combining categories to produce a broader categorisation. The use of two digit NACE already provides sufficient protection without substantial information loss. In addition, for quantitative public variables such as the number of employees, instead of giving the exact value we can release a broader classes such as [20,25), [25,30), and so on. Finally, for sensitive quantitative variables, our proposal to limit disclosure is to release a predictive interval based on a statistical model. We will illustrate this using just turnover. The same idea could be used for releasing other sensitive quantitative variables such as export, and even sensitive binary variables such as group membership, but the model would become more complicated. Once these variables have been protected a careful choice of innovation variables to be included in the microdata file to be released has to be made. These innovation variables will be released without change so as to avoid affecting subsequent analyses.

In this paper we consider data on 870 enterprises, 274 of which belonging to NACE rev. 1 division 18, clothing manufacture, and 596 of which belonging to NACE rev. 1 division 28, metal products manufacture. We omit eleven such enterprises with zero turnovers because the release of data about these requires special consideration. The proposed method takes account of geographical area. In particular each enterprise is assigned to an area based on the NUTS1 territorial classification. This is made more precise in Section 2.1. The variables considered for the model are based on:

- **Turnover**, with associated variable $y = \log(\text{turnover})$ where turnover is measured in millions of Italian lire;
- **Amount of Exports**, with associated variable $x_1 = \log(\text{exports})$ where exports is measured in millions of Italian lire;
- **Number of Employees**, with associated variable $x_2 = \log(\text{number of employees})^5$;
- Whether or not the enterprise is involved in the **innovation of products or processes or both**, with associate variable $x_3 \in \{0,1\}$;
- Whether or not the enterprise is a **member of a group**, with associate variable $x_4 \in \{0,1\}$;
- **NACE**, with associated variable $x_5 \in \{0,1\}$, with 0 corresponding to NACE rev. 1 division 28.

In order to demonstrate our model based disclosure limitation method, we shall concentrate on the variable turnover. In particular, in Section 2 we shall build a model for y in terms of x_1, x_2, x_3, x_4 and x_5 that takes into account the area to which each enterprise belongs. We remark here that the logarithmic transformation was used in the definition of y, x_1 and x_2 to reduce skewness. The fifth power for x_2 was chosen by inspection of residual plots; use of a cubic led to slightly less good residual plots. As we consider data from only two NACE divisions, we need just one indicator variable x_5. In general if we were to have data from D NACE divisions, we would need $D-1$ indicator variables.

Instead of publishing y, we can release an interval based on the predictive distribution associated with our statistical model; this is fully discussed in Section 3. We believe that the approach described in this paper has the potential to contribute to disclosure control methodology. Our aim here is to present our methodology in a straightforward way to arouse interest in it among the official statistics community. In Section 3.3 we sketch a possible way of extending our approach. A by-product of our method is the insight that the model gives into the geographical structure underlying the data. This insight into the area effect suggests a broader categorisation to use when releasing the qualitative public variable geographical area that goes a long way to minimising information loss. This will be discussed in Section 4. Finally, in Section 5 we briefly present our conclusions.

2. THE MODEL

For our model, in addition to data on turnover, exports, employees, whether or not the enterprise is involved in product or process innovation, whether or not the enterprise belongs to a group and whether the associated NACE is 18 or 28, we will also make use of the geographical area to which each enterprise belongs. We begin this section by describing the available spatial data in detail, after which we give a precise formulation of our model.

2.1. Spatial Information

Each enterprise belongs to one of $N = 10$ areas into which Italy has been divided:

Area number (NUTS1)	Area name (abbreviation)	Constituent administrative regions
1	North West (NW)	Val d'Aosta, Piemonte, Liguria
2	Lombardy (LOM)	Lombardia
3	North East (NE)	Trentino Alto Adige, Veneto, Friuli Venezia Giulia
4	Emilia Romagna (ER)	Emilia Romagna
5	Centre (CEN)	Toscana, Umbria, Marche
6	Lazio (LAZ)	Lazio
7	Abruzzo and Molise (ABMO)	Abruzzo, Molise
8	Campania (CAM)	Campania
9	South (SOU)	Calabria, Basilicata, Puglia
10	Sicily and Sardinia (SICSAR)	Sicilia, Sardegna

Note that Sicily (NUTS1 = 10) and Sardinia (NUTS1 = 11) have been combined into one area. One reason for doing this is that only four observations originate from Sardinia.

In particular, we know the neighbourhood structure of these areas. This information can be summarised by the symmetric matrix W :

$$W = \begin{pmatrix} 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

in which $w_{ik} = 1$, $i, k = 1, \ldots, N$, if areas i and k are neighbours and 0 otherwise.

2.2. Spatial Model Formulation

The data model that we shall adopt takes the following form:

$$y_{ij} \sim N\left(\mu_{ij}, \frac{1}{\eta_y}\right) \qquad (1)$$
$$\mu_{ij} = \mu + \beta_1 x_{1,ij} + \beta_2 x_{2,ij} + \beta_3 x_{3,ij} + \beta_4 x_{4,ij} + \beta_5 x_{5,ij} + u_i + v_i,$$

where $y_{ij} = \log(\text{turnover}_{ij})$ is the response variable for the j^{th} enterprise, $j = 1, \ldots, n_i$, in the i^{th} area, $i = 1, \ldots, N$, and $x_{1,ij} = \log(\text{exports}_{ij})$, $x_{2,ij} = \log(\text{number of employees}_{ij})^5$, $x_{3,ij} = 1$ if the enterprise is involved in product or process innovation or both and 0 otherwise, $x_{4,ij} = 1$ if the enterprise belongs to a group and 0 otherwise, and

$x_{5,ij} = 1$ if the enterprise belongs to NACE rev. 1 division 18 and 0 for division 28. We shall assume that the distribution of the variables u_i, $i = 1,\ldots,N$, is *a priori* independent Gaussian:

$$u_i \sim N\left(0, \frac{1}{\eta_u}\right).$$

As no account is taken here of the available spatial information, the variables u_i, $i = 1,\ldots,N$, are thought of as unstructured random effects. On the other hand, the distribution of the variables v_i, $i = 1,\ldots,N$, does take into account spatial information, the idea being that neighbouring areas should take similar values. Accordingly, these variables will be thought of as structured random effects. This is achieved by adopting a conditional autoregressive scheme defined in terms of the conditional distribution of v_i given $v_1,\ldots,v_{i-1},v_{i+1},\ldots,v_N$:

$$v_i \mid v_1,\ldots,v_{i-1},v_{i+1},\ldots,v_N \sim N\left(\overline{v_{-i}}, \frac{1}{\omega_i \phi}\right),$$

in which $\omega_i = \sum_{k \neq i} \omega_{ik}$ and $\overline{v_{-i}} = \sum_{k \neq i} \omega_{ik} v_k / \omega_i$ for known weights ω_{ik}. This scheme is discussed in Mollié (1996) and Besag, et al. (1991), for example. We follow the standard approach and set the weights $\omega_{ik} = w_{ik}$. Other choices of weights are, however, possible. For example, we could use the distances d_{ik} between the centroids of areas i and k by setting $\omega_{ik} = \exp\left(-\frac{d_{ik}}{\tau}\right)$, where τ is a scale parameter. Diggle, et al. (1998) discuss a joint formulation for $v_1,\ldots,v_N$: $v_1,\ldots,v_N \sim N(0,\Sigma)$, in which the mean 0 is an N dimensional vector and the $N \times N$ covariance matrix Σ is such that $\Sigma_{ik} = \sigma^2 \exp\left\{-\left(\frac{d_{ik}}{\alpha}\right)^{\delta}\right\}$, where σ^2 is a variance, $\alpha > 0$ is a scale parameter and $\delta \in (0,2)$ is a type of smoothing parameter.

We work with precisions η_y, η_u and ϕ instead of variances because of conjugacy considerations. In particular we adopt gamma priors for these parameters:

$$\eta_y \sim \Gamma(a_y, b_y)$$
$$\eta_u \sim \Gamma(a_u, b_u)$$
$$\phi \sim \Gamma(a_\phi, b_\phi)$$

where in general $\eta \sim \Gamma(a,b)$ means that the probability density $f(\eta) \propto \eta^{a-1} \exp(-b\eta)$, with shape parameter $a > 0$ and scale parameter $b > 0$. For such a gamma random variable, we have $E[\eta] = \frac{a}{b}$ and $\mathrm{Var}[\eta] = \frac{a}{b^2}$.

We finish our model specification by giving prior distributions for μ, β_1, β_2, β_3, β_4 and β_5:

$$\mu \sim N\left(0, \frac{1}{\tau_\mu}\right), \quad \tau_\mu > 0$$

$$\beta_m \sim N\left(0, \frac{1}{\tau_{\beta_m}}\right), \quad \tau_{\beta_m} > 0, \quad m = 1,\ldots,5$$

We set the hyper-parameters $a_y = b_y = a_u = b_u = a_\phi = b_\phi = 0.01$ and $\tau_\mu = \tau_{\beta_m} = 10^{-5}$, $m = 1,\ldots,4$. With these choices, our priors have high variances. Space limitation does not permit us to present a full discussion of the effect of different choices of these hyper-parameters. We do, however, believe that the results that we present are relative robust to such choices. A full discussion of these issues in a slightly different context is presented by Bernardinelli, et al. (1995) and Pascutto, et al. (2000).

The above model specification allows us to write down the posterior density for the vector of parameters of interest

$$\theta = \left(\mu, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, u_1, \ldots u_N, v_1, \ldots, v_N, \eta_y, \eta_u, \phi\right)$$

given the data:

$$\begin{aligned} p(\theta \mid \text{data}) \propto & \prod_{i=1,\ldots,N,\, j=1,\ldots,n_i} p(y_{ij} \mid \mu, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, u_1, \ldots, u_N, v_1, \ldots, v_N, \eta_y) \times \\ & \prod_{i=1,\ldots,N} p(u_i \mid \eta_u) \times p(v_1, \ldots, v_N \mid \phi) \times p(\eta_y) \times p(\eta_u) \times p(\phi) \times \\ & p(\mu) \times p(\beta_1) \times p(\beta_2) \times p(\beta_3) \times p(\beta_4) \times p(\beta_5), \end{aligned}$$

in which we use $p(\cdot \mid \cdot)$ and $p(\cdot)$ to represent conditional and prior densities.

3. USING THE PREDICTIVE DENSITY FOR DISCLOSURE CONTROL

In this section we describe our proposal for using the predictive density for disclosure control. We begin by outlining the algorithm for making inferences from the model. We then explain how samples can be drawn from the predictive density and how the released values for turnover are based on this density. This section ends with a brief discussion about possible extensions of this methodology using decision theory.

3.1. Making Inferences from the Model

We choose to make inferences about θ by means of the Gibbs sampler. The Gibbs sampler is an example of a Markov chain Monte Carlo algorithm; for further details see Geman, and Geman (1984) and Gilks, et al. (1996), for example. The main reason for this choice is simplicity of implementation. Also, the `S-PLUS` code that we have developed can be easily modified for other types of data, for example, binary data such as group membership. In a medical application with which one of us has recently been involved, y_{ij} are parasite counts and we assume that $y_{ij} \sim \text{Po}(\mu_{ij})$ or $y_{ij} \sim \text{NegBin}(\mu_{ij}, k)$, where k is the over-dispersion parameter of the negative binomial distribution. Starting from an initial vector of parameters $\theta^{(0)}$, the Gibbs sampler updates each parameter in turn by sampling from the conditional density of that parameter given the other parameters. For the above model these conditional densities turn out to be either Gaussians or gammas and so the required samples can be easily drawn. This process of updating all the parameters in turn is repeated many times to provide a sequence of parameter vectors $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(G)}$. In order to remove the effect of $\theta^{(0)}$, the first B members of this sequence are thrown away; we say that we have used a burn-in of length B. The remaining members $\theta^{(B+1)}, \theta^{(B+2)}, \ldots, \theta^{(G)}$ are considered to have converged to a sequence from the posterior density. Inference is then based upon this sequence. For our model we have always found convergence to be rapid. Accordingly, we take $G = 1000$ and $B = 500$.

3.2. Releasing Values from the Predictive Density

The values that we shall release are based on the predictive density $p(y^{\text{new}} \mid \text{data})$, where y^{new} is a predicted value of the vector $y = (y_{ij}, i = 1, \ldots, N, j = 1, \ldots, n_i)$. Since $p(y^{\text{new}} \mid \text{data}, \theta) = p(y^{\text{new}} \mid \theta)$, it follows that

$$\begin{aligned} p(y^{\text{new}} \mid \text{data}) &= \int_{\theta} p(y^{\text{new}} \mid \text{data}, \theta)\, p(\theta \mid \text{data})\, d\theta \\ &= \int_{\theta} p(y^{\text{new}} \mid \theta)\, p(\theta \mid \text{data})\, d\theta \\ &= E_{\theta}[p(y^{\text{new}} \mid \theta)], \end{aligned}$$

where $p(y^{\text{new}} \mid \theta)$ is the Gaussian density associated with (1) and E_{θ} is the expectation under the posterior. Accordingly, realisations from this predictive density can easily be obtained by simulating a vector from $p(y^{\text{new}} \mid \theta^{(t)})$ for each $t = B+1, \ldots, G$. In this way for each of the original 870 observations we obtain a vector $(y_{ij}^{(B+1)}, \ldots, y_{ij}^{(G)})$ of realisations from the corresponding predictive density. A $(1-\gamma)\%$ predictive interval

can be obtained from this vector by sorting it and taking the $\text{floor}\left\{\frac{\gamma}{2}(G-B)\right\}^{\text{th}}$ and the $\text{ceiling}\left\{(1-\frac{\gamma}{2})(G-B)\right\}^{\text{th}}$ elements, where $\text{floor}(x)$ ($\text{ceiling}(x)$) returns the nearest integer below (above) x. The intervals obtained for each observation with $\gamma = 0.05$ are shown in Figure 1. The observations have been sorted according to the posterior mean of μ_{ij}, which is also shown on the graph.

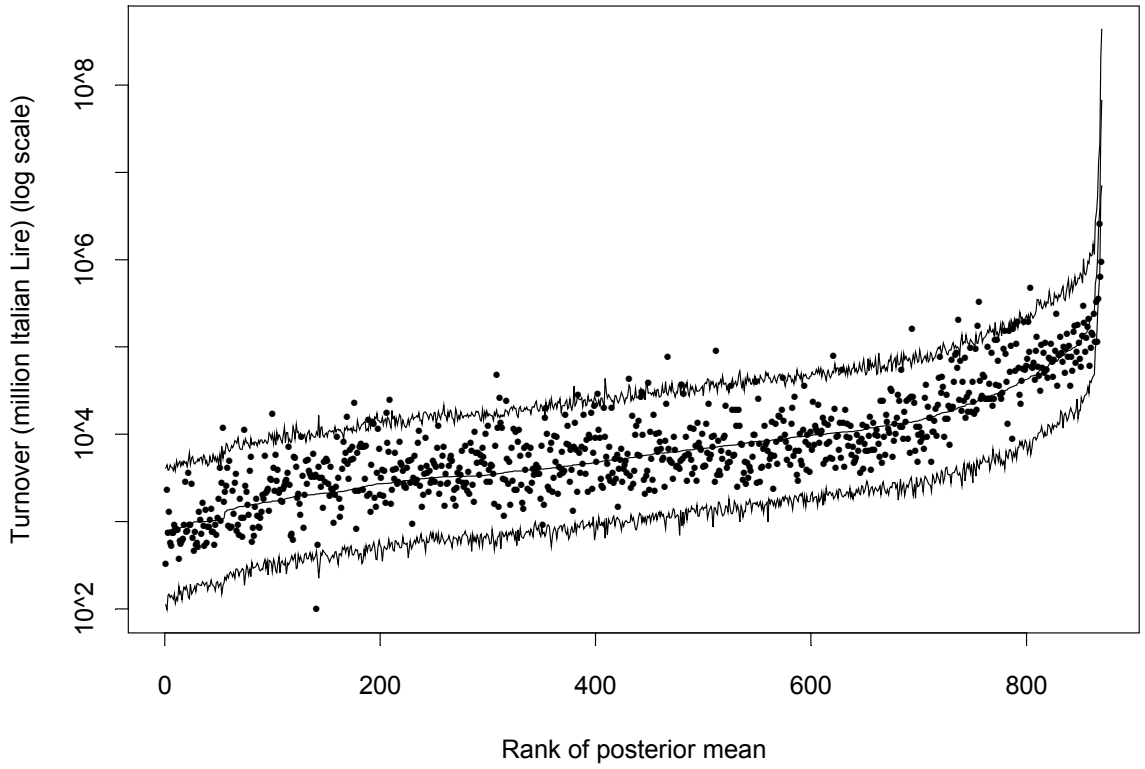


Figure 1: Prediction intervals. For each observed value of turnover the prediction interval obtained from the model in shown; the upper curve joins the upper limits of the intervals, while the lower curve joints the lower limits. The observations have been sorted according to the posterior mean of μ_{ij}, which is shown as the middle curve.

In order to protect the true value of turnover and hence to reduce the possibility that an enterprise is identified, we propose releasing these intervals instead. Narrower intervals can be obtained by using larger values of γ. We can see from Figure 1 that in a few cases the true value of turnover lies outside its corresponding interval. If this is considered a problem, the interval can be shifted so that it contains the true value of turnover.

Of course, given a predictive interval in which the turnover may lie, one could estimate the true value by the midpoint for example. Such estimates are shown in Figure 2. The fact that many of the points in Figure 2 lie close to the diagonal line without actually being on it means that quite a lot of protection has probably taken place without losing much information.

It may be felt more appropriate to release a point summary of the predictive density instead of an interval. Examples of such point summaries would be the predictive mean, approximated as $\frac{1}{G-B}\sum_{t=B+1}^{G} y_{ij}^{(t)}$ for each data point, and the

predictive median, approximated as median$(y_{ij}^{(B+1)},\ldots,y_{ij}^{(G)})$. We do not present figures analogous to Figure 2 for these summaries, as for this data set they are almost identical.

The next step of this work will be to check that the suggested protection measures are indeed sufficient. This will be done by applying specially designed matching software to the released data.

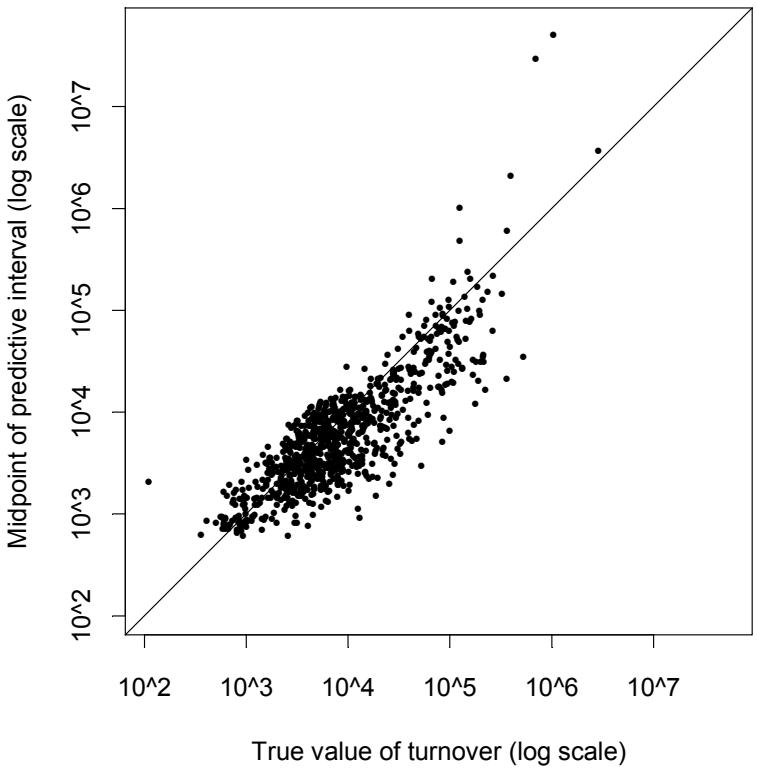


Figure 2: Midpoints of predictive intervals plotted against the true value of turnover.

3.3. A Decision Theoretic Approach

We have seen that it is possible to release a variety of summaries of the predictive density. We feel that a future direction for the proposed approach would be to develop methodology to enable the release of a summary of the predictive density that corresponds to a given loss function. Rue (1997), Frigessi, and Rue (1997), and Rue, and Syversveen (1998) implement such methodology in the context of image analysis using Markov chain Monte Carlo algorithms and simulated annealing (see Geman, and Geman (1984)).

4. THE AREA EFFECT

Our spatial model allows us to produce estimates of the area effect. We display these by presenting $\frac{1}{G-B}\sum_{t=B+1}^{G}\left(u_i^{(t)}+v_i^{(t)}\right)$, an approximation to the posterior means of u_i+v_i, $i=1,\ldots,N$. These area effects are shown in Figure 3. They have been shifted so that their mean is zero. From Figure 3 we see the interesting results that in general the northerly areas have a positive effect on turnover.

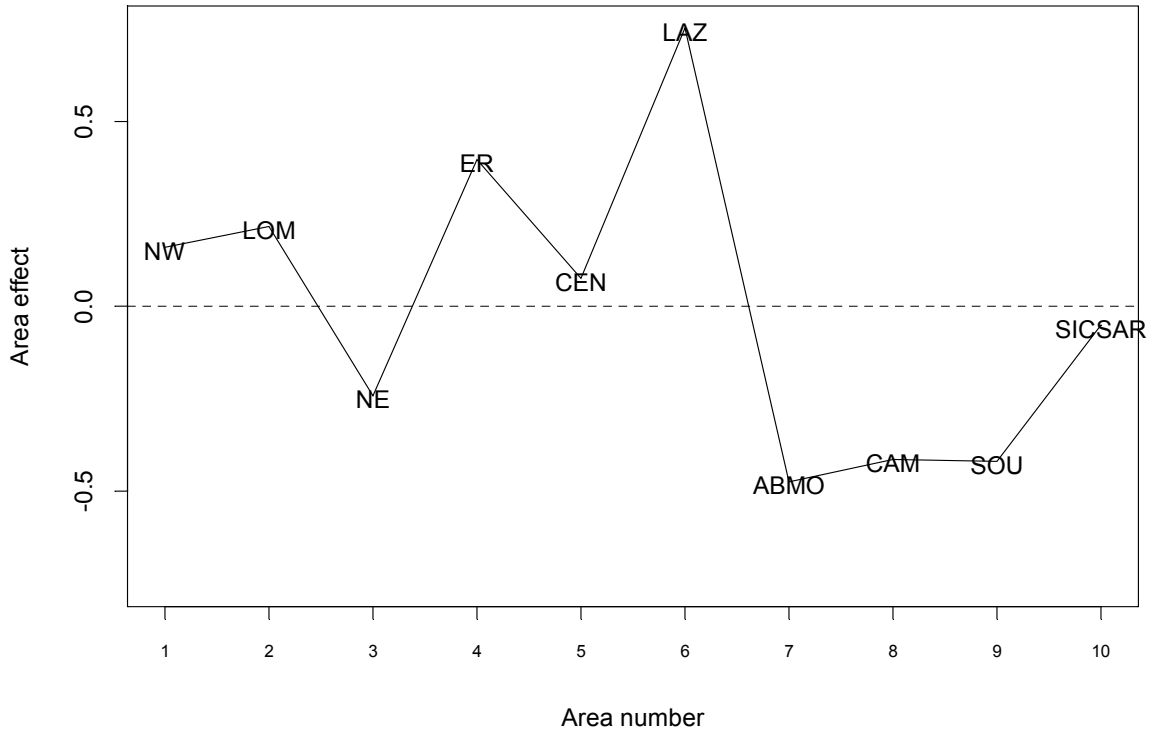


Figure 3: Area effects from our model. In general the northerly areas have a positive effect on turnover.

Mollié (1996) explains how to check which of the unstructured or structured random effects dominate. For this, the posterior mean of $\frac{\eta_u}{\phi}$, approximated by $\frac{1}{G-B}\sum_{t=B+1}^{G}\frac{\eta_u^{(t)}}{\phi^{(t)}}$, is compared with $\overline{W}=\frac{1}{N}\sum_{i=1}^{N}W_i$, where $W_i=\sum_{j\neq i}w_{ij}$. If this posterior mean is smaller (larger) than $\overline{W}$ then the unstructured (structured) random effects dominates. For the above data, the posterior means was very much larger than $\overline{W}$, and so we were able to conclude that the structured random effects dominate the unstructured.

Figure 3 together with the neighbourhood structure of these areas suggest a way to define a broader categorisation to use when releasing the qualitative public variable geographical area. The released variable would have two categories, one comprising Abruzzo and Molise, Campania, the South of Italy, Sicily and Sardinia, and the other comprising the remaining six areas. This categorisation goes a long way to minimising information loss. Our experience is that this categorisation can also be useful for other pairs of NACE divisions.

5. CONCLUSIONS

In this paper we have proposed a model based disclosure limitation method for business microdata. We have illustrated our approach on data arising from the Community Innovation Survey of manufacturing and services sector enterprises. For this data, identification of an enterprise through its NACE classification is rendered difficult by releasing just two digits. For quantitative public variables our approach releases broader classes instead of the exact value. It builds a statistical model for other sensitive quantitative variables such as turnover that takes account of the geographical area to which each enterprise belongs. If such spatial information is not available, groups of enterprises may be defined through cluster analysis or neighbourhood graphs. We discuss how inferences can be

made from the model using the Gibbs sampler and how predictive intervals can be calculated. We illustrate the use of these predictive intervals for disclosure control of the variable turnover for 870 enterprises involved in clothing or metal products manufacture. We briefly discuss other summaries of the predictive density that could be released instead, and ways of extending our approach based on decision theory. A by-product of our spatial model is that estimates of the underlying area effects can be produced. These estimates are presented for the data under consideration and provide useful insights. These insights suggest a broader categorisation to use when releasing the qualitative public variable geographical area that goes a long way to minimising information loss.

ACKNOWLEDGEMENTS

The authors would like to thank Giulio Perani for providing the data, and Rana Moyeed, Cristiana Pascutto and Giovanni Seri for helpful comments.

The views expressed are those of the authors and do not necessarily reflect the policies of Istat or the University of Plymouth.

REFERENCES

Bernardinelli, L., Clayton, D., and C. Montomoli (1995), "Bayesian Estimates of Disease Maps: How Important are Priors?," *Statistics in Medicine*, **14**, pp. 2411–2431.

Besag, J., York, J., and A. Mollié (1991), "Bayesian Image Restoration, with Two Applications in Spatial Statistics (with discussion)," *Annals of the Institute of Statistical Mathematics*, **43**, pp. 1–59.

Cox, L.H. (1994), "Matrix Masking Methods for Disclosure Limitation in Microdata," *Survey Methodology*, **20**, pp. 165–169.

Cox, L. H. (1995), "Protecting Confidentiality in Business Surveys," in B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge, and P. S. Kott (eds.) *Business Survey Methods*, New-York: Wiley, pp. 443-473.

Defays, D., and P. Nanopoulos (1992), "Panels of Enterprises and Confidentiality: The Small Aggregates Method," *Proceedings of Statistics Canada Symposium 92, Design and Analysis of Longitudinal Surveys*, pp. 195–204.

Diggle, P. J., Tawn, J. A., and R. A. Moyeed (1998), "Model-based Geostatistics (with discussion)," *Applied Statistics*, **47**, pp. 299–350.

Duncan, G. T., and R. W. Pearson (1991), "Enhancing Access to Microdata while Protecting Confidentiality: Prospects for the Future," *Statistical Science*, **6**, pp. 219–239.

Fienberg, S. E. (1994), "A Radical Proposal for the Provision of Micro-data Samples and the Preservation of Confidentiality," Technical Report n°. 611, Pittsburg, Pennsylvania: Department of Statistics, Carnegie Mellon University.

Fienberg, S. E., Russel, J. S. and U. Makov (1996), "Statistical Notions of Data Disclosure Avoidance and their Relationship to Traditional Statistical Methodology: Data Swapping and Log-linear Models," unpublished manuscript.

Frigessi, A., and H. Rue (1997), "Bayesian Image Classification with Baddeley's Delta Loss," *Journal of Computational and Graphical Statistics*, **6**, pp. 55–73.

Geman, S., and D. Geman (1984), "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-6**, pp. 721–741.

Gilks, W. R., Richardson, S., and D. J. Spiegelhalter (1996), "Introducing Markov Chain Monte Carlo," in W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (eds.) *Markov Chain Monte Carlo in Practice*, London: Chapman & Hall, pp. 1–19.

McGuckin, R. H., and S. V. Nguyen (1988), "Use of 'Surrogate Files' to Conduct Economic Studies with Longitudinal Microdata," *Proceedings of the Fourth Annual Research Conference, U.S. Bureau of the Census*, **20**, pp. 193–211.

Mollié, A. (1996), "Bayesian Mapping of Disease," in W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (eds.) *Markov Chain Monte Carlo in Practice*, London: Chapman & Hall, pp. 359–379.

Pascutto, C., Wakefield, J. C., Best, N. G., Richardson, S., Bernardinelli, L., Elliott, P., and A. Staines (2000), "Statistical Issues in the Analysis of Disease Mapping Data," *Statistics in Medicine*. In press.

Rue, H. (1997), "A Loss Function Model for the Restoration of Grey Level Images," *Scandinavian Journal of Statistics*, **24**, pp. 103–114.

Rue, H., and A. R. Syversveen (1998), "Bayesian Object Recognition with Baddeley's Delta Loss," *Advances in Applied Probability*, **30**, pp. 64–84.

STATISTICAL DISCLOSURE CONTROL SOFTWARE

Anco Hundepool, Statistics Netherlands
P.O. Box 4000, 2270 JM Voorburg, The Netherlands
Email: ahnl@krypton.vb.cbs.nl

1. Introduction

In this paper we will give an overview of the developments of the ARGUS-software. In the recent years we have been able to develop and extend ARGUS thanks to the support of the 4th Framework program of the European Union. This work has led to the development of two programs μ-ARGUS for the disclosure control of microdata and τ-ARGUS for tabular data. Although a major step forwards has been made we still plan to extend this software and introduce new techniques. We hope to realise these plans with the support from the 5th framework program.

The first part of this paper will be devoted to the microdata and the second part we will pay attention to the problems of tabular data. Finally we will discuss our plans for the future developments.

2. Statistical Disclosure Control in General

The aim of Statistical Disclosure Control (SDC) is to limit the risk that sensitive information of individual respondents can be disclosed from a published data set. This data set can be either a microdata-set or a table. A microdata-set consists of a set of records containing information on individual respondents. A table contains aggregated information based on individual entities.

Before a microdata set or a table can be published the safety should be checked. This requires the formulation of criteria that can be used to check the data. However this formulation is not an easy question It basically requires that one tries to model the behaviour of an person who, confronted with a data set, might want to identify an individual and disclose certain information of this person. The data releaser should use this disclosure scenario to develop safety criteria from this. These criteria reflect the target population of potential users of a particular data set (e.g. researchers, the general public, etc.) and the legal and organisational measures that accompany the release of these data. Dependent on this, the criteria can be more tight or less.

In order to publish safe data, one should first have to check whether a particular data set is safe according to these criteria or not. If data are not safe according to these criteria they have to be modified in such a way that the resulting data meet these criteria. These modifications, while decreasing the risk of disclosure, also imply that the information content of the data is decreased, because certain variables are coded in a less detailed fashion or values are suppressed or replaced by other values. The idea is that the modifications should be applied in such a way that the resulting information loss is minimised. As a rule achieving this goal is quite complicated and requires the use of specialised software tools. Such tools are μ-ARGUS for microdata and τ-ARGUS for tabular data.

3. μ -ARGUS for microdata

3.1. Background of μ *-ARGUS*

The main concern of the disclosure control of microdata is to avoid that individual records in a dataset can be identified by a user of this dataset. The SDC-measures concentrate on the identifying variables, i.e. the variables whose scores are easily known of an individual, such as sex, place of residence, age, occupation, etc. Not that these variables represent in general much sensitive information, but the scores of these variables are used to identify a record, link a record to an individual in the population. When several values of these identifying variables are

combined a respondent may be identified. An example of such a combination might be: "Place of residence = some smaller town, Sex = female and occupation = dentist". Just a telephone directory might be enough to identify this dentist. Once a record has been linked to an individual by using these variables, the other variables in the dataset are revealed and the disclosure has been made. These other variables contain the sensitive information, like income or whatever is considered sensitive.

The SDC for microdata concentrates on these set of identifying variables. Combinations of these identifying variables (called keys) are inspected to find out which combinations of scores occur rarely in the population. So tables of frequency counts of these variables are calculated and if the score of a certain combination is below a certain value (threshold) this combination is considered unsafe and therefor will be a possible risk of disclosure.

What can be done with these unsafe combinations? The two major tools available are global recoding and local suppression. Global recoding will change/collapse the coding scheme for one or more variables. E.g. the coding scheme for occupation would combine all medical workers (doctors, dentists and other) into one new category. This would be done for the whole dataset. The other option is local suppression i.e. changing one of the codes to missing. So the record would then become a female person in this smaller town with unknown profession.

Both global recoding and local suppression lead to a loss of information, because either less detailed information is provided or some information is not given at all. A balance between global recoding and local suppression has to be found in order to make the information loss due to the application of SDC measures as low as possible.

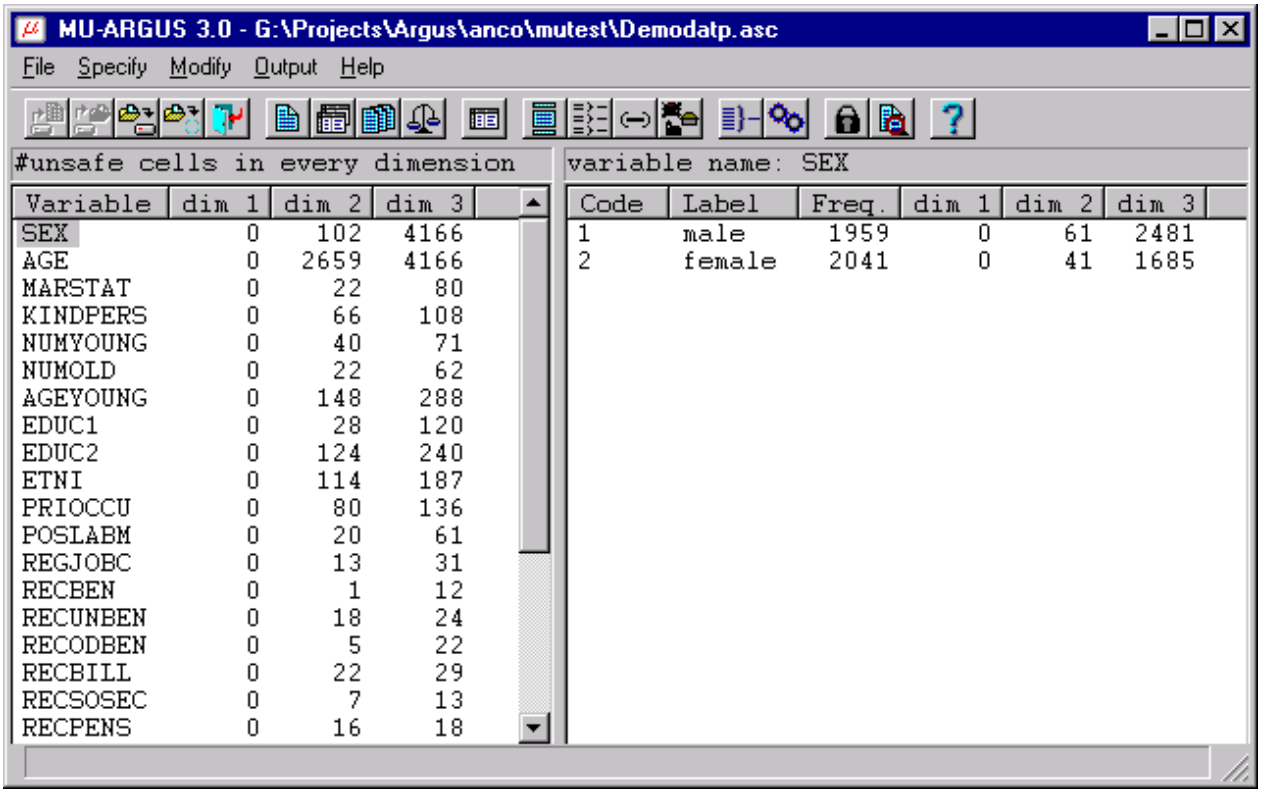
A separate problem is the decision which variables are to be considered identifying and which set of combinations should be inspected. This is a decision to be made by the data-protector. At Statistics Netherlands we have formalised this. Variables are categorised by a degree of identification. As a rule region (place of residence) is considered very identifying, sex and ethnicity more identifying and many other variables are considered identifying. This leads to a set of combinations to be inspected. Each very identifying variable is combined with the class of very and more identifying and with the class of all identifying variables.

3.2 The μ-ARGUS program

The μ -ARGUS program has been build to protect a microdata set based on the above principles. The software will assist the user to find an appropriate set of global recodings. It is an easy task to see the results of the various sets of recodings. Even if large datasets are to be protected, this can be done interactively.

As input μ-ARGUS expects the data in a (commonly used) fixed ASCII format. Irrespective of the data format used to store the data internally it is not problem to export the data in this format. It might very well also be the format in which the data will eventually be made available to the outside users. When the basic metadata (record description etc) has been supplied, the data-protector can select the set of combinations to be checked. Either he uses a predefined set of combinations based on the Dutch approach or he specifies the set of combinations himself. When this set has been selected and the threshold has been chosen, all these combinations are calculated and the sensitive combinations can be identified. This process might take some time for a really large dataset, but in general this goes rather quickly. From now on however the process of disclosure protection μ-ARGUS is at the level of these tables combinations, so it will be very quick and can easily be done interactively.

The user is shown an overview of the unsafe combination per variable:



Variable	dim 1	dim 2	dim 3
SEX	0	102	4166
AGE	0	2659	4166
MARSTAT	0	22	80
KINDPERS	0	66	108
NUMYOUNG	0	40	71
NUMOLD	0	22	62
AGEYOUNG	0	148	288
EDUC1	0	28	120
EDUC2	0	124	240
ETNI	0	114	187
PRIOCCU	0	80	136
POSLABM	0	20	61
REGJOBC	0	13	31
RECBEN	0	1	12
RECUNBEN	0	18	24
RECODBEN	0	5	22
RECBILL	0	22	29
RECSOSEC	0	7	13
RECPENS	0	16	18

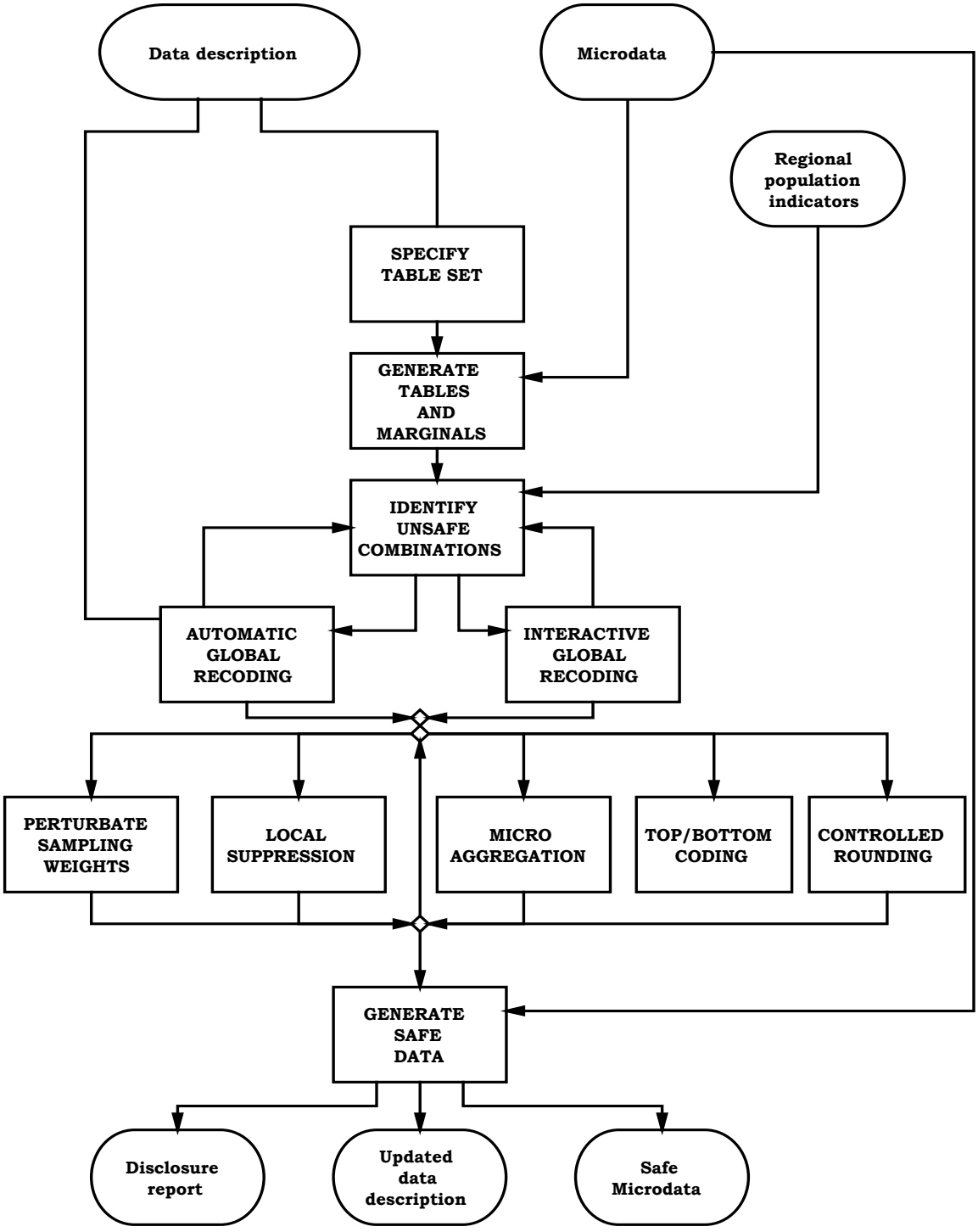
Code	Label	Freq.	dim 1	dim 2	dim 3
1	male	1959	0	61	2481
2	female	2041	0	41	1685

The job he is facing now is to remove these unsafe combinations from this dataset. The tools he has at his disposal are global recoding and local suppression. In general one can say that the global recoding will do the main job. One can assume that the data-protector knows best which global recodings are best suited for his problem. In the daily practice of a statistical office the available recoding schemes are in general already available. It is not useful to invent some peculiar and uncommon recoding scheme to protect some unsafe cell as this would lead to a dataset with a codelist that is not familiar to the end-users of the data.

So when using µ-ARGUS one can in general expect that the possible recoding schemes are available. However it is still possible in µ-ARGUS to specify a new recoding scheme if the data-protector feels the need to do so.

When he has seen the results of the different global recodings, he makes his final choice. In general there will still be a small remaining set of unsafe combinations. These unsafe combinations will be protected by local suppressions, i.e. a missing value will be imputed for one of the variables in an unsafe combination. The dataset is now safe with respect to the rules stated on the beginning of the process. The safe dataset will be written and can be made available to the end-users of the data. Besides that a report will be written to document the changes in the dataset.

μ-ARGUS overview



4. τ-ARGUS for tabular data

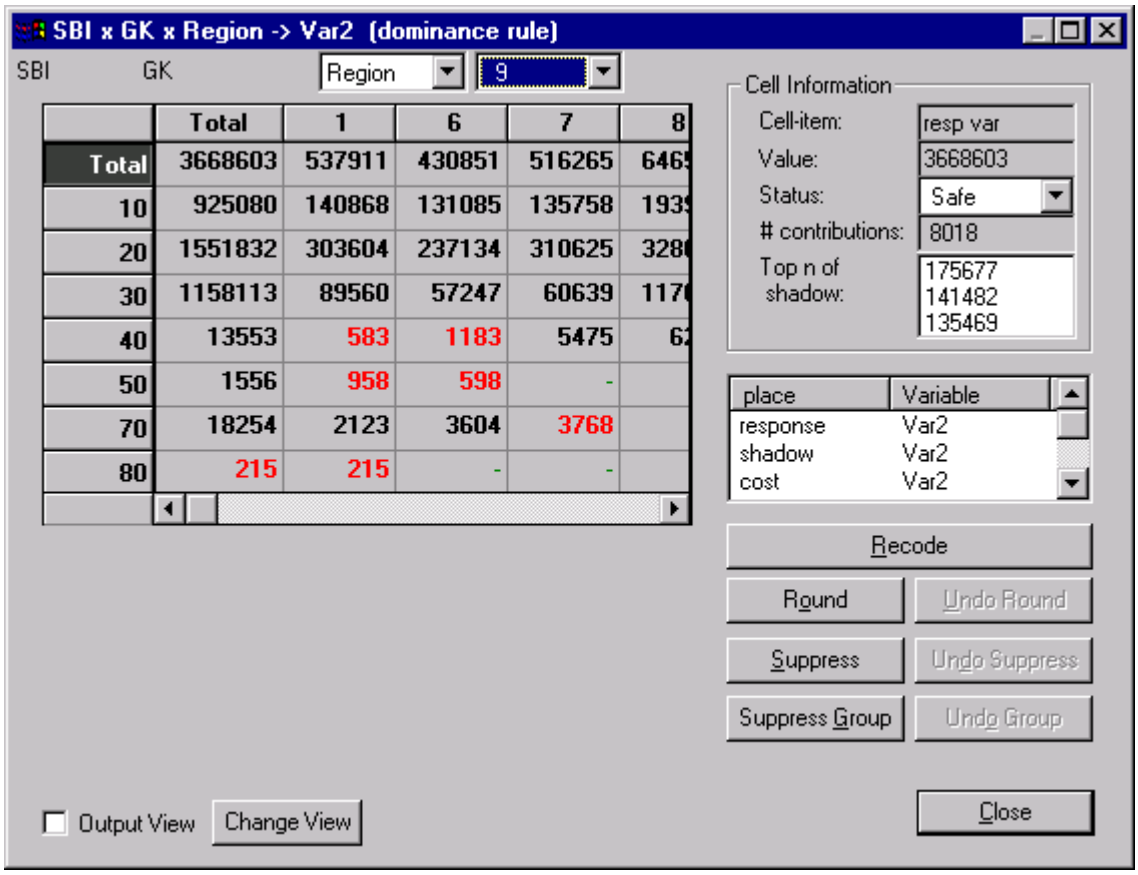
4.1. Background of τ-ARGUS

The other part of Statistical Disclosure Control concentrates on tabular data. Tables are the traditional form of output of the statistical offices. In spite of modern techniques of electronic publishing, tabular data still will be the core of the output. But also the electronic publications concentrate on tabular data. There is a longer history of disclosure control of tabular data. Several rules have been proposed to indicate which cells of a table must be considered unsafe and there for should not be published.

This however is the easy part of disclosure control of tabular data. Suppressing these unsafe cells from a table does not do the job. As tables tend to have marginals or (even worse) all kind of other subtotals, there is much information available in a table to recalculate the suppressed cell using these marginals and sub-totals. Even if the exact recalculation is not possible a very narrow estimate of the interval containing the suppressed cell is also undesirable. So in addition to the (primary) unsafe cells that must be suppressed, additional cells must be suppressed to guarantee that the end-user is not able to re-estimate the cells from the remainder of the table.

A commonly used rule for testing the safety of a cell is the dominance rule. This rule states that a cell of a table is unsafe for publication if a few, n say, major contributors to a cell are responsible, when adding their contributions, for at least a certain percentage p of the total of that cell. A common choice is n=3 and p=70%, but τ-ARGUS allows users to specify other parameter settings. One of the ideas behind this rule is that one large contributor to a cell should not be able to get a narrow estimate for the contribution of its largest competitor from a table published by a statistical office.

The main job of τ-ARGUS however is to find an optimal set of additional (secondary) cells to be suppressed. This tuns out to be a difficult task, requiring state of the art optimisation solutions. This has been implemented in τ-ARGUS.



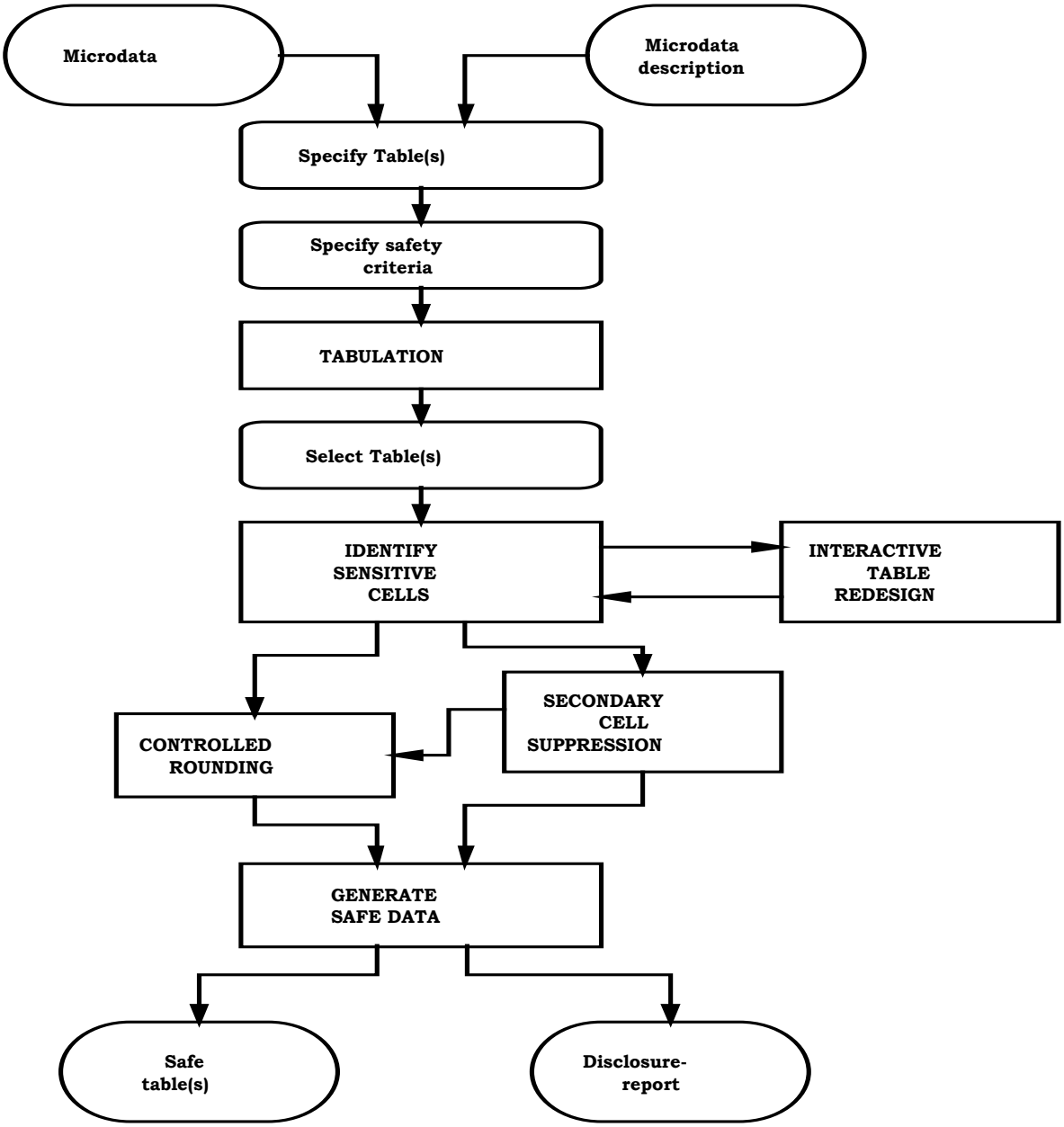
4.2. The τ-ARGUS program

τ-ARGUS is a software solution for the disclosure control of tabular data. It starts with the same ASCII files as its brother μ-ARGUS. It will generate the tables from the raw datafile. We have implemented it this way in order to be able to apply the dominance rule. In order to apply the dominance rule τ-ARGUS needs not only the cell value itself but also the values of the largest contributors to the cell. When τ-ARGUS calculates the table from a micro-datafile it can keep track of the values of the largest contributors. Identifying the primary unsafe cells is then an easy job.

The next step is that τ-ARGUS has facilities to redesign the table. Rows and columns can be collapsed similar to global recoding in μ-ARGUS. This will reduce the number of unsafe cells. Because τ-ARGUS disposes over the top-contributions of each cell it can easily calculate the top-contributions of the collapsed cells and apply the dominance rule to the new (collapsed) cells without going back to the microdata. This has the great advantage that the data-protector can freely play around with various collapsing schemes similar to the various global recoding schemes in μ-ARGUS. This would have been impossible when τ-ARGUS only would have the cell-totals and an indication which cells should be considered unsafe.

When the data-protector is satisfied with the structure of the table he presses the suppress-button and τ-ARGUS will start the search-process to find the optimal set of additional (secondary) cells to be suppressed in order to guarantee the safety of the primary cells with a given safety-interval. This proves to be a mathematical very hard problem. A solution for this has been supplied by JJ Salazar and M. Fischetti (1998).

τ-ARGUS overview



After the suppression process the table with these additional suppressions is now safe and can be published. τ-ARGUS will store the safe table as a text-file or a spreadsheet, which enables further processing of the table in the publication system of the users. To keep track of the actions performed on the table τ-ARGUS will write a log-file describing the actions that have taken place.

5. Future developments

The current versions of ARGUS have been developed with a grant from the 4th Framework program of the EU. A proposal for a further grant form the 5th Framework Program has been submitted. If this proposal will be accepted we can expect that the work on ARGUS will be continued. The key-issues will be the emphasis on business micro data for μ-ARGUS and the hierarchical table for τ-ARGUS.

For μ-ARGUS the future developments lie in the field of business data. The methods applied in μ-ARGUS at this moment are more suitable for microdata on individuals than on microdata on businesses. The problem lies in the fact that the data in the business-files is much more skewed distributed than data on individuals. This asks for the development of new techniques. We name here Post-Randomisation (PRAM) and other techniques of noise addition. The advantage of PRAM (see De Wolf 1998) is that the end-user of the data will have the parameters of the randomisation process at his disposal. This will enable him to make better estimates at the level of the population without disclosing individual information. Although much additional research is needed the first results look promising.

For τ-ARGUS the main attention will be paid to the structure of the tables. Up to now τ-ARGUS is only capable of protecting simple 3-dimensional tables. In the daily practice of statistical institutes the tables often have hierarchical structured codelists. This implies that many additional subtotals are present in a table. This gives the intruder much more information to re-estimate the sensitive/suppressed cells. The mathematical problem to solve becomes much harder. New approaches are needed here. On the one hand on can say that it is not always necessary to find the theoretical global optimum of this optimisation problem at high costs. A very good approximation will yield to a very well protected table. So research is needed to find new search-techniques. Very similar to the problem of the hierarchical tables are the linked tables. When e.g. two tables have the same variable as cell item, and at least one common codelist, the some marginals of these tables are identical. So the process of finding the secondary suppressions must be done simultaneously. It cannot be that in one table a marginal is suppressed, which will be published by the other table. These relations between the two tables are very similar to the hierarchical one and we hope that solving the hierarchical problem will at the same time solve the linked problem.

References

- M. FISCHETTI and J.J. SALAZAR (1998). Modelling and Solving the Cell Suppression Problem for Linearly-Constrained Tabular Data. *Proceedings SDP98, Lisbon*.
- L. WILLENBORG and T. DE WAAL (1996). Statistical Disclosure Control in Practice, Springer-Verlag, New York.
- P.-P. DE WOLF, J.M. GOUWELEEUW, P. KOOIMAN and L.C.R.J. WILLENBORG (1998). Reflections on PRAM. *Proceedings SDP98, Lisbon*.

DISCUSSION

Lawrence H. Cox, U.S. Environmental Protection Agency
National Exposure Research Laboratory (MD-75), Research Triangle Park, NC 27711 USA
Cox.Larry@Epa.Gov

The three papers in this session deal with computing, estimates, and computing estimates in the context of statistical disclosure limitation (SDL). With respect to estimation, Hundepool describes a post-randomization method for perturbing frequency tabulations. Zayatz-Evans-Slanta describe a method for perturbing economic microdata prior to tabulation. Both methods release unbiased estimates in lieu of exact tabulations. Franconi-Stander propose Bayesian models for economic microdata, from which predictive intervals are released in place of actual microdata. With respect to computation, Hundepool describes an emerging software environment for SDL developed in the Netherlands. Zayatz-Evans-Slanta offer a method that is computationally simply and flexible within otherwise complex statistical data processing environments. Franconi-Stander make use of sophisticated and computationally intensive methods from modern statistical computing.

SDL is very much about estimation and computation. SDL is aimed at thwarting narrow estimation of confidential data. SDL is typically performed within large and complex data processing environments, placing computational efficiency at a premium, but at the same time theoretical results on SDL methods raise the specter of unmanageable computational complexity. There is also the preeminent issue of balancing disclosure protection with its effects on data quality, completeness and usefulness. As illustrated by Franconi-Stander, this challenging problem, as yet lacking a precise theoretical formulation, is likely to require a computationally intensive Bayesian framework for its examination. Thus, there are three elements: estimation, computation and theory.

The work of Hundepool and colleagues to develop the ARGUS SDL software is an important step towards placing proven SDL tools in the hands of survey practitioners. Two major recommendations of the Subcommittee on Disclosure Limitation Methodology of the U.S. Federal Committee on Statistical Methodology were to "share software and methodology across the (U.S.) government" (Recommendation 3) and "use consistent (SDL) practices" (Recommendation 5) (Federal Committee on Statistical Methodology 1994). The development of the ARGUS software provides a mechanism for realizing these objectives. Its greatest value may be as a hands-on tool to educate survey practitioners about disclosure limitation methodology and its effects, viz., as an *SDL laboratory*.

Regarding post-randomization (PRAM) mentioned by Hundepool and other perturbation methods, I return briefly to theory. Three- and higher-dimensional tables and linked tables present problems quite different from those encountered in two-dimensions. The down-side risk of user-friendly software is that it can be misused, misapplied or get too far ahead of the theory on which it should be based. In a large-scale data processing environment, such as a census or major survey, or in a real-time environment, such as a public-use statistical data base query system, this can lead to unfortunate or embarrassing results for the agency.

As an illustration, consider Table 1, wherein the entries marked "*" can assume any positive value; "0" is zero-restricted. No unbiased perturbation of Table 1 is possible: software based on familiar two-dimensional models such as alternating cycles of cells marked in an alternating +/- fashion will fail to compute.

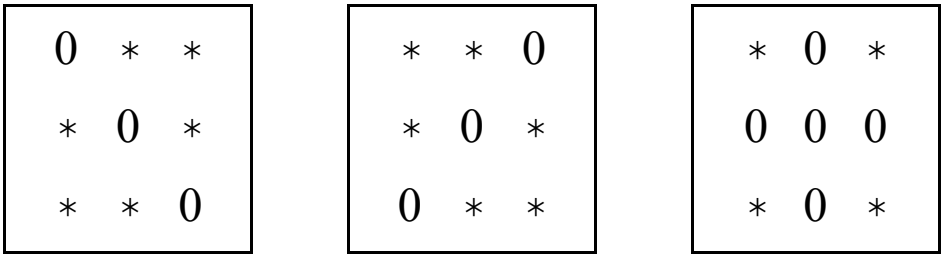


Table 1: 3x3x3 Table With No Alternating Cycle

Conversely, higher-dimensions bring new opportunities for perturbation, as yet not fully understood. Consider Table 2: no alternating cycle exists but a "generalized cycle" of length 17 permits perturbation by **a** units through 16 of the cells and **2a** units through the 17th cell (entry 321, in bold). Tables 1 and 2 are from Cox (2000).

+	−	0
0	−	+
−	**+**	−

−	+	0
+	0	−
0	−	+

0	0	0
−	+	0
+	−	0

Table 2: A Unique Generalized Cycle

With respect to Zayatz-Evans-Slanta, I wonder if the method relies too heavily on unbiasedness, viz., E(e) = 0. This is a useful property when sampling randomly and repeatedly from a large population, or at high levels of aggregation. The authors simulate their method for a survey that produces estimates at the U.S.-totals level for selected SIC, with good results, as might be expected. It would be interesting to see if similar results can be obtained at lower and more diverse levels of aggregation, such as in the Census or Annual Survey of Manufactures. For such data, which exhibit skewed distributions for most economic statistics as well as for the number of respondents per tabulation cell, it may be necessary to do more than simply randomly assign the (up/down) direction of the perturbation.

It is good to see this paper address SDL for sample surveys. Judging at least from the published literature, this important setting seems relatively neglected to its older sibling, censuses. Surveys offer some additional aspects for SDL, mostly favorable ones. As the objective is to hide reported enterprise data in released tabular estimates, sampling weighs and sampling error come into play as they provide some of the required uncertainty almost for free. The problem can be posed in the manner of Franconi-Stander, viz., as incorporating sufficient uncertainty in the estimation of individual respondent data (or verifying that such already exists). Zayatz-Evans-Slanta present and illustrate their method in a clear and concise manner. But, shouldn't it be posed in the reverse direction: rather than incorporating approximately 10% uncertainty into each micro-value, shouldn't a desired level of uncertainty in estimating micro-values be established (say, 10%), the uncertainty already present (due to sampling, weighting) be estimated, and the "residual" uncertainty then incorporated into the pseudo-micro-values to achieve the desired result (10% overall uncertainty)? The method appears relevant to SDL for economic microdata, but the authors do not address this possibility.

The approach of Franconi-Stander is intriguing. They are among the first to explore the use of Bayesian modeling to the release of disclosure-protected microdata and to my knowledge the first to tackle the SDL problem for economic microdata release from this perspective. They illustrate nicely how factors such as regional dependencies can be incorporated into the model. The nagging issue with modeling approaches is of course selection bias: selection of what variables go into the model, selection of link functions, specification of the form of the model and values for its parameters. These latter issues can be addressed through hierarchical modeling, but there remains the problem that the uses to which the data ultimately may be put would have benefitted from the incorporation of different covariates, link functions, etc. It seems to me that this raises the significant issue of how to benefit from the combined advantages of utilizing design-based data to maximum advantage in a model-based world. Another issue is that of large respondents, viz., large businesses. These are important for reliable estimation of economic statistics and for that reason often are included in the sample with certainty. Is it not inevitable for the model to attenuate large values (outliers), and therefore bias estimates? Is it possible to reweight the sample to account for model distortions of this sort? This approach raises many questions about estimation. Regardless, from an SDL perspective, Franconi-Stander have opened an interesting and potentially valuable line of inquiry.

On the theory that discussants are by-and-large frustrated presenters, I cannot resist closing by introducing an additional, related idea–one from a deterministic perspective. In each of these papers, suppression was avoided in favor of releasing either a point or interval estimate of actual data. A deterministic alternative, proposed by Ramesh Dandekar (U.S. Energy Information Administration), is called *controlled estimates*.

Consider a set of economic tabulations. In lieu of suppressing a sensitive value, the value is replaced by a value that is either greater than or equal to its upper protection limit, or less than or equal to its lower protection limit, viz., the actual sensitive value is replaced by an *estimate* that is considered sufficiently broad as to not threaten confidentiality. The problem is then to make adjustments to selected nonsensitive values in order to re-balance the aggregation structure, viz., the estimates are controlled. It is desirable to do so while distorting original values as little as possible. This approach is distinctly different from but is nevertheless analogous to synthetic data SDL methods proposed by Rubin (1993).

I have formulated the controlled estimates problem as an integer linear programming optimization problem. The integer linear program is presented below for an additive two-way table *Table (r, c)* containing (primary) disclosures at table entries (r, c) = *disclosure*, and protection intervals of width *Displacement (disclosure)*. *Direction (disclosure)* is a {0, 1} decision variable indicating whether *Table (disclosure*) is to be replaced by a value below its lower protection limit (0) or above its upper protection limit (1). *Plus (r, c)* indicates a positive quantity to be added to *Table (r, c)*; *Minus (r, c)* is the amount to be subtracted. The integer linear program is as follows.

Minimize: $\sum_{r,\,c} (Plus(r,\,c) + Minus(r,\,c))$

Subject to:
$$Plus(r,\,\cdot) - Minus(r,\,\cdot) = \sum_{c} (Plus(r,\,c) - Minus(r,\,c)) \quad \text{for all r}$$
$$Plus(\cdot,\,c) - Minus(\cdot,\,c) = \sum_{r} (Plus(r,\,c) - Minus(r,\,c)) \quad \text{for all c}$$
$$Table(r,\,c) \geq Minus(r,\,c) \quad \text{for all r, c}$$

For all (r, c) = *(primary) disclosure*:

$$Plus(disclosure) \geq Direction(disclosure) * Displacement(disclosure)$$
$$Minus(disclosure) \geq (1 - Direction(disclosure)) * Displacement(disclosure)$$

Direction binary integer; *Plus*, *Minus* continuous

It is possible to limit the Plus/Minus deviations or impose feasible zero-restrictions using capacity constraints. Subject to appropriate modifications, other cost functions can be accommodated. As this is an integer linear programming formulation, computational approaches such as those of Fischetti and Salazar (1999) are needed for medium sized problems. It is also related to Fischetti-Salazar's *partial suppression*, which, through the introduction of two sets of local variables for each sensitive cell, the authors solve as a linear program, albeit of enormous size.

Disclaimer

The information in this article has been funded wholly or in part by the United States Environmental Protection Agency. It has been subjected to Agency review and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

References

Cox, L.H. (2000), "On Properties of Multi-Dimensional Statistical Tables," unpublished manuscript, submitted.

Federal Committee on Statistical Methodology (1994), *Report on Statistical Disclosure Limitation Methodology*, Statistical Policy Working Paper 22, Washington, DC: U.S. Office of Management and Budget.

Fischetti, M. and J.J. Salazar (1999), "Modeling and Solving the Cell Suppression Problem for Linearly-Constrained Tabular Data," *Statistical Data Protection: Proceedings of the Conference, Lisbon, 25 to 27 March, 1998*,. Luxembourg: EUROSTAT, 401-409.

Rubin, D. (1993), "Discussion: Statistical Disclosure Limitation," *Journal of Official Statistics*, **9**, 460-468.