

# Additive noise and multiplicative bias as disclosure limitation techniques for continuous microdata: A simulation study

M. Trottini<sup>a,\*</sup>, S.E. Fienberg<sup>b</sup>, U.E. Makov<sup>c</sup> and M.M. Meyer<sup>d</sup>

<sup>a</sup>*Department of Statistics, University of Valencia, Valencia 46100, Spain*

<sup>b</sup>*Department of Statistics and Center for Automated Learning and Discovery, Carnegie Mellon University, Pittsburgh, PA 15213, USA*

<sup>c</sup>*Department of Statistics, Haifa university, Haifa 31905, Israel*

<sup>d</sup>*Intelligent Results Inc., Bellevue, WA 98004, USA*

Received 15 October 2002

Revised 10 January 2003

Accepted 10 January 2003

**Abstract.** This paper focuses on a combination of two disclosure limitation techniques, *additive noise* and *multiplicative bias*, and studies their efficacy in protecting confidentiality of continuous microdata. A Bayesian intruder model is extensively simulated in order to assess the performance of these disclosure limitation techniques as a function of key parameters like the variability amongst profiles in the original data, the amount of users prior information, the amount of bias and noise introduced in the data. The results of the simulation offer insight into the degree of vulnerability of data on continuous random variables and suggests some guidelines for effective protection measures.

Keywords: Confidentiality, disclosure limitation, identity disclosure, intruder behavior, simulated data

Mathematics Subject Classification: 62F15, 62P25

## 1. Introduction

Most data handled by statistical agencies are collected under a pledge of confidentiality, i.e., an implicit or explicit promise made by the agency to the data providers that it will prohibit or prevent improper uses of the data aimed to disclose confidential information. Depending on whether the data are for continuous or categorical variables, and whether they are for an entire population or simply a sample, researchers have proposed a variety of *disclosure limitation techniques* (or *data masks*) to protect data confidentiality [3,17,18]. The effectiveness of these methods, however, has been only partially explored and statistical agencies are increasingly interested in learning how well these techniques perform in different disclosure scenarios.

---

\*Corresponding author. Tel.: +34 9638 643 62; Fax: +34 9638 647 35; E-mail: mario.trottini@uv.es.

In this paper we address this issue for *continuous microdata* when the disclosure limitation technique is a combination of *additive noise* and *multiplicative bias*. We focus on *identity* disclosure, where an intruder uses the published data to identify individual respondents. For simplicity we consider the case of population data. The model that we present is an extension to and variation on the model proposed by Fienberg et al. [6]. This extension provides an alternative way to introduce bias and noise into the data and allows for the intruder to use the data on several units to enhance his chances of disclosure.

Little has been done in the literature to study the performance of masking for continuous microdata. The most systematic work is by Domingo-Ferrer and Torra [2], who compare several disclosure limitation techniques using actual data from the U.S. Census Bureau. Their approach, however, differ from ours in several ways. That study uses different intruder's attack and data masks and their goals are different from ours. Here we focus on one technique and we study how it performs under a variety of disclosure scenarios; Domingo-Ferrer and Torra [2] fix the disclosure scenario and compare alternative masking methods.

The framework in Fuller [9] is much closer to ours. Although he studies effectiveness of additive noise only and considers disclosure scenarios complementary to ours – where for example the number of variables used for re-identification changes and sample data are released – the intruder's attack, in spirit, is similar to the one that we describe here. Indeed, our approach can be seen as an extension of the intruder's attack in Fuller [9] where the intruder can use data on several units to disclose the identity of individual respondents.

In Sections 2 through 4 we describe in some detail the different components of our disclosure scenario, that is: (i) the original data, (ii) the masking method used, and (iii) the intruder's attack, and we explain the importance of the scenario. In Section 5 we report on a simulation study intended to investigate the effectiveness of additive bias and multiplicative noise as disclosure limitation techniques for continuous microdata, simulating the intruder's attack under a variety of scenarios. In Section 6 we briefly discuss the results of the simulation study. Section 7 contains final comments and outlines ideas of future work.

## 2. The data

We assume that the original data consist of the values for  $s + p + q$  variables in a population  $P$  of size  $N$ . The resulting microdata,  $\mathbf{M}$ , can be represented as a collection of  $N$  records of dimension  $s + p + q$ ,  $\mathbf{M} = \{\mathbf{m}_1, \dots, \mathbf{m}_N\}$ , where “ $\mathbf{m}_i = (\mathbf{I}_i, \mathbf{x}_i, \mathbf{y}_i)$ ” is a comprised of an  $s$ -vector  $\mathbf{I}_i$  of *identifying variables* (such as “name”, “social security number”, etc.), a  $p$ -vector  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$  of *key variables* (i.e. variables that can be used to re-identify respondents linking the information about  $\{x_{ij}\}$  in the released data with the information about the  $\{x_{ij}\}$  contained in a public file available to the users), and a  $q$ -vector  $\mathbf{y}_i = (y_{i1}, \dots, y_{iq})$  of other variables (often referred to as *sensitive variables*). We take the key variables to be continuous. In particular, we assume:

**Assumption A1.** *The key variables in the microdata  $\mathbf{M}$  are generated according to the hierarchical model:*

$$\begin{aligned} x_{ij} | \mu_j, \xi_j^2 &= N(\mu_j; \xi_j^2), \quad i = 1, \dots, N, \quad j = 1, \dots, p; \\ \mu_j &\sim N(\eta; \omega^2); \\ \xi_j^2 &\sim IG(\rho; \lambda); \end{aligned} \tag{1}$$

where  $\eta$ ,  $\omega$ ,  $\rho$ ,  $\lambda$  are fixed constants,  $N(\eta; \omega^2)$  denotes a normal distribution with mean  $\eta$  and variance  $\omega^2$ ,  $IG(\rho; \lambda)$  an inverse gamma distribution with mean  $\lambda/(\rho - 1)$  and  $\mu_j$  and  $\xi_j^2$  are assumed to be independent.

Assumption 1 is made largely for convenience. The hierarchical model (1) enables us to express different features of the original data in terms of few parameters. Thus, for example, we can increase heterogeneity among unit profiles just by increasing the ratio  $\lambda/\rho$ . This formalization turns out to be very useful in the simulation study that we perform in Section 5, where data sets with different characteristics need to be simulated in order to study the performance of the masking.

The strongest component of Assumption 1 is the independence among attributes. In real data attributes in a unit are typically correlated. As we illustrate in Section 4, however, our model can be easily generalized to the case where observations are not normal and the data exhibit non-trivial covariance structure as long as we can write the joint density of  $\mathbf{x}_i$  in closed form.

The key variables can be categorical or continuous with joint distribution  $g$ . Since we focus on identity disclosure the form of  $g$  is irrelevant for our analysis.

Microdata with continuous key variables are not the most common form of data release (for a general discussion see [4,12,18]). They are, however, of particular interest in data disclosure limitation. Researchers, analysts, policy makers, increasingly demand access to microdata where at least a subset of variables is continuous (this is, for example, the case of data on business issues). This demand, however, has been only partially fulfilled by statistical agencies [3,5]. For example in released data on household and individual income level is often too strictly top-coded to allow for full policy analysis [3]. The problem is that the disclosure risk for microdata with continuous key variables is usually too high because each unit in the population is very likely to be “unique” with respect to a combination of the continuous key variables, and thus the probability of re-identification is one for most of the units whose data are released unless some type of masking is applied. Work on data masking for microdata with continuous key variables is therefore, extremely important in order to define suitable forms of data release that can meet the needs of the users while at the same time preserving privacy and confidentiality of respondents represented in the data. In the next section we describe a masking procedure based on a combination of bias and noise.

### 3. The masking method

Consider original microdata  $M$  to be protected. We assume that the data masking method consists of removing the vectors,  $\{\mathbf{I}_i\}$ , of direct identifiers from the original data and then perturbing the values of the key variables by adding bias and noise. In particular, we assume:

**Assumption A2.1** *Bias and noise are introduced to the pure records through the perturbation model:*

$$z_{ij} = x_{ij} \cdot \theta_{ij} + \epsilon_j^2, \quad i = 1, \dots, N; \quad j = 1, \dots, p. \quad (2)$$

**Assumption A2.2** *The bias,  $\theta_{ij}$ 's and the noise  $\epsilon_j^2$ 's are independent, the additive noise varies only with respect to attributes and the distribution of the  $\theta_{ij}$ 's for the same  $j$  is identical for all  $i$ 's. In particular:*

$$\begin{aligned} \theta_{ij} | \phi_j^2 &\sim N(1; \phi_j^2) \quad i = 1, \dots, N, \quad j = 1, \dots, p; \\ \epsilon_j^2 | \sigma_j^2 &\sim N(0; \sigma_j^2). \end{aligned}$$

The released microdata is denoted by  $\mathbf{Z}$ , and the  $i^{\text{th}}$  record in  $\mathbf{Z}$  by  $\mathbf{z}_i$ ,  $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_N\}$ ,  $\mathbf{z}_i = (z_{i1}, \dots, z_{ip}, \mathbf{y}_i)$ ,  $i = 1, \dots, N$ . The larger the values of  $\{\phi_j^2\}$  ( $\{\sigma_j^2\}$ ) the larger is the amount of bias (noise) introduced into the original data. If  $\sigma_j^2 = \phi_j^2 = 0$ , then the agency releases the original data. Our choice of normal distribution for the bias and noise is arbitrary and we could easily consider skewed distributions as well.

We assume that both the model generating the original data in (1) and the perturbation model in (2) are known to the users, but that the values  $\{\sigma_j\}$  and  $\{\phi_j\}$  used to generate the released data  $\mathbf{Z}$  are unknown and must be estimated. In particular, we assume that:

**Assumption A3.** *Users can represent their prior beliefs about  $\sigma_j^2$  and  $\phi_j^2$  in terms of inverse gamma distributions:*

$$\begin{aligned}\sigma_j^2 &\sim IG(\gamma; \delta); \\ \phi_j^2 &\sim IG(\alpha; \beta), \quad j = 1, \dots, p.\end{aligned}$$

The parameters  $\gamma$ ,  $\delta$ ,  $\alpha$ , and  $\beta$  quantify the amount of information that the agency provides to the users about how the masking has been performed. The choice of inverse gamma distributions offers reasonable flexibility to model different degrees of users' prior uncertainty.

The masking method defined in A2.1 and A2.2 is similar to that proposed in Fienberg et al. [6] but it provides a more natural way to incorporate bias and noise into the data. It does not share the nice property of other additive noise methods that preserve correlation structure in the data [10,11,13,14] but it has the advantage of simple interpretation and it can be used to model data with errors where bias and noise are introduced in the data by the respondents when some of the attributes, that are part of the survey, are conceived as a sensitive issue (like income, age etc.). In that context, when  $\theta_{ij}$  is equal to one, an honest reply is given by the respondent  $\mathbf{I}_i$  for the item  $j$  and the exchangeability property of the  $\{\theta_{ij}\}$  implies that individual generates bias which are common in their distribution.

Having defined the structure of the original data and the mechanism behind the masking used by the agency, we now outline the Bayesian model for the intruder's attack.

#### 4. The intruder's attack

We assume that only one unit  $\mathbf{I}_0$ , (out of the  $N$  in  $P$ ), the so called "target unit", is at the center of the intruder's investigation. Further we assume that the intruder possesses verified information on the value of the key variables for the target unit  $\mathbf{I}_0$  and for  $L$  other units in  $P$  ( $L \leq N - 1$ ). The intruder's goal is to re-identify the record in  $\mathbf{Z}$  that belongs to  $\mathbf{I}_0$ . Clearly by re-identification the intruder gains information on the  $q$  attributes  $(y_{01}, \dots, y_{0q})$  of  $\mathbf{I}_0$ .

We denote the intruder's data by  $\mathbf{E} = (\mathbf{x}_0, \mathbf{X}^{(\mathbf{L})})$  where  $\mathbf{x}_0$  corresponds to the  $p$ -vector of key variables for the target unit and  $\mathbf{X}^{(\mathbf{L})} = (\mathbf{x}_1, \dots, \mathbf{x}_L)$  are the rest of the data in the intruder's possession corresponding to the other units (we can always rearrange the data in such a way that the records in  $\mathbf{X}_L$  corresponds to the first  $L$  records in  $\mathbf{M}$ ). By comparing the information in  $\mathbf{X}^{(\mathbf{L})}$  with the information in the released data  $\mathbf{Z}$ , the intruder "learns" the bias and noise parameters used in the masking and this information facilitates the re-identification of the target unit  $\mathbf{I}_0$ .

We represent the intruder's target as an indicator function  $\tau$  such that  $\tau = i$  iff the record  $\mathbf{z}_i$  in  $\mathbf{Z}$  belongs to  $\mathbf{I}_0$ ,  $i = 1, \dots, N$ . The intruder's decision as to whether re-identification is possible, and if

possible, the decision on a particular match, is based on the posterior distribution of  $\tau$  given  $\mathbf{E}$ , and  $\mathbf{Z}$  that we denote by  $Pr(\tau = i|\mathbf{E}, \mathbf{Z}), i = 1, \dots, n$ . In particular, we assume that the intruder acts according to the following decision rule:

**Assumption A4.** *Let  $t \in [0, 1]$  be a threshold value and let  $\hat{\tau} = \operatorname{argmax}_i Pr(\tau = i|\mathbf{E}, \mathbf{Z})$ . If  $Pr(\tau = \hat{\tau}|\mathbf{E}, \mathbf{Z}) > t$ , then the intruder links  $\mathbf{x}_0$  (i.e.  $\mathbf{I}_0$ ) with  $\mathbf{z}_{\hat{\tau}}$ . Otherwise the intruder does not make any link. The value of  $t$  is fixed by the intruder but unknown to the statistical agency.*

The idea underlying this decision rule is simple. The intruder tries to link  $\mathbf{x}_0$  with the record in  $\mathbf{Z}$  which has the highest posterior probability of belonging to  $\mathbf{x}_0$ , but he makes the link only if he has enough evidence that it is correct. The threshold  $t$  formalizes the amount of evidence that the intruder needs in order to make the link. For example, if  $t = 0$  the intruder always makes the link, while if  $t = 0.9$  than the intruder makes the link only when the posterior probability of  $\tau = \hat{\tau}$  is greater than 0.9. By guessing the value of  $t$  the statistical agency makes an assumption about the intruder's attitude toward risk. A "prudent" intruder will use high values of  $t$  while a "risky" intruder will use values of  $t$  close to 0. Trottini [15] and Trottini and Fienberg [16] have shown that the decision rule A4 corresponds to the intruder using a 0–1 loss function for the estimation problem "estimate  $\tau$ ". Next we discuss the evaluation of  $Pr(\tau = i|\mathbf{E}; \mathbf{Z})$ .

#### 4.1. Evaluation of $Pr(\tau = i|\mathbf{E}; \mathbf{Z})$

If we assume that

$$Pr(\tau = i|\mathbf{X}^{(\mathbf{L})}; \mathbf{Z}) = 1/N \quad \forall i = 1, \dots, N,$$

from Bayes' theorem we get:

$$Pr(\tau = i|\mathbf{E}; \mathbf{Z}) = Pr(\tau = i|\mathbf{x}_0; \mathbf{X}^{(\mathbf{L})}; \mathbf{Z}) \propto f(\mathbf{x}_0|\tau = i; \mathbf{X}^{(\mathbf{L})}; \mathbf{Z}) \cdot Pr(\tau = i|\mathbf{X}^{(\mathbf{L})}; \mathbf{Z}).$$

Thus the evaluation of  $Pr(\tau = i|\mathbf{E}; \mathbf{Z})$  reduces to the evaluation of:

$$f(\mathbf{x}_0|\tau = i; \mathbf{X}^{(\mathbf{L})}; \mathbf{Z}) = \int_{\theta} \int_{\sigma} f(\mathbf{x}_0|\tau = i; \mathbf{X}^{(\mathbf{L})}; \mathbf{Z}; \theta; \sigma^2) \cdot \pi(\theta, \sigma^2|\tau = i; \mathbf{X}^{(\mathbf{L})}; \mathbf{Z}) d\theta d\sigma^2 \quad (3)$$

where  $\theta = (\theta_{i1}, \dots, \theta_{ip})$  and  $\sigma^2 = (\sigma_1^2, \dots, \sigma_p^2)$ . Because of the independence assumptions A1 and A2.2, we can rewrite formula (3) as:

$$f(\mathbf{x}_0|\tau = i; \mathbf{X}^{(\mathbf{L})}; \mathbf{Z}) = \prod_{j=1}^p \int \int f(x_{0j}|\theta_{ij}; z_{ij}; \sigma_j^2) \cdot \pi(\theta_{ij}|\tau = i; \mathbf{X}^{(\mathbf{L})}; \mathbf{Z}) \cdot \pi(\sigma_j^2|\tau = i; \mathbf{X}^{(\mathbf{L})}; \mathbf{Z}) d\theta_{ij} d\sigma_j^2. \quad (4)$$

By assumptions A2.1 and A2.2,  $x_{0j}|\theta_{ij}; z_{ij}; \sigma_j^2 \sim N(z_{ij}/\theta_{ij}; \sigma_j^2/\theta_{ij}^2)$ . In most of the cases the integral in (4) and thus the posterior probability of  $\tau$  can not be evaluated in closed form. If we are able to efficiently simulate from the posterior distributions of  $\theta_{ij}^2$  and  $\sigma_j^2$ , however, then we can obtain a Monte Carlo approximation of (4) using:

$$f(\mathbf{x}_0|\tau = i; \mathbf{X}^{(\mathbf{L})}; \mathbf{Z}) \approx \prod_{j=1}^p \left[ \frac{\sum_{h=1}^k f(x_{0j}|\theta_{ij}^{(h)}; z_{ij}; \sigma_j^{2(h)})}{k} \right], \quad (5)$$

where  $\theta_{ij}^{(h)}$  ( $\sigma_j^{2(h)}$ ) represents the  $h^{th}$  draw from the posterior distributions of  $\theta_{ij}^2$  ( $\sigma_j^2$ ), and  $k$  is the total number of simulations. It follows from (3) that we can extend this calculation to the case where  $\mathbf{X}^{\mathbf{L}}$  shows a non trivial covariance structure as long as the density of  $\mathbf{x}_0$  has a closed form.

We now focus on the evaluation of the posterior distributions for  $\theta_{ij}^2$  and  $\sigma_j^2$ . We evaluate the posterior distributions of  $\theta_{ij}$  and  $\sigma_j^2$  with respect to the combined data sets of  $\mathbf{X}^{(\mathbf{L})}$  and  $\mathbf{Z}$ ; however, only pairs of  $\mathbf{x}$ 's and  $\mathbf{z}$ 's that constitute a genuine match between an intruder record and agency released record can provide information on  $\theta_{ij}$  and  $\sigma_j^2$ . Since no knowledge of such a match is available, we need to consider all pairs  $(x_{mj}, z_{tj})$ ,  $(t \leq N) \cap (t \neq i)$ ,  $m \in \{1, \dots, L\}$ , representing potential links between records. Note that we do not include  $\mathbf{z}_i$  since it is paired with  $\mathbf{x}_0$  which we exclude in the evaluation of the posterior distribution of  $\theta_{ij}$  and  $\sigma_j^2$ . Let  $\nu$  denote the number of such pairs,  $\nu = L - 1, L$ , and  $d(\nu, r)$  the  $r^{th}$  particular collection of such pairs,  $r = 1, \dots, D_\nu$  where

$$D_\nu = \binom{L}{\nu} \binom{N-1}{\nu} \nu!$$

The posterior distribution of  $\theta_{ij}$  is given by:

$$\pi(\theta_{ij} | \tau = i; \mathbf{X}^{(\mathbf{L})}; \mathbf{Z}) = \sum_{\nu} \sum_r \left[ \int \pi(\theta_{ij} | \phi_j^2) \cdot \pi(\phi_j^2 | \mathbf{X}^{(\mathbf{L})}; \mathbf{Z}_{-\mathbf{i}}; d(\nu, r)) d\phi_j^2 \right] p(d(\nu, r)), \quad (6)$$

where  $(x_{mj}, z_{tj})$  is the  $d(\nu, r)$  particular pair in  $D_\nu$ ,  $\mathbf{Z}_{-\mathbf{i}}$  is  $\mathbf{Z}$  excluding  $\mathbf{z}_i$ ,  $\pi(\theta_{ij} | \phi_j^2)$  is the normal distribution defined in A2.2, and the posterior distribution of  $\phi_j^2$  is:

$$\pi(\phi_j^2 | \tau = i; \mathbf{X}^{(\mathbf{L})}; \mathbf{Z}_{-\mathbf{i}}; d(\nu, r)) \propto \prod f(x_{mj} | z_{tj}; \phi_j^2) \cdot \pi(\phi_j^2),$$

where the product is with respect to all pairs  $(x_{mj}, z_{tj})$  defined by  $d(\nu, r)$ ,  $\pi(\phi_j^2)$  is the prior distribution for  $\phi_j^2$  defined in A3, and  $f(x_{mj} | z_{tj}; \phi_j^2)$ ,  $p(d(\nu, r))$  are given, respectively, by

$$f(x_{mj} | z_{tj}; \phi_j^2) = \int_{\theta_{ij}} \int_{\sigma_j^2} f_N(x_{mj} | z_{tj} / \theta_{tj}; \sigma_j^2 / \theta_{tj}^2) \cdot \pi(\theta_{ij} | \phi_j^2) \cdot \pi(\sigma_j^2) d\theta_{tj} d\sigma_j^2,$$

and  $p(d(\nu, r)) = H(\nu; N; n-1; L) / D_\nu$ ,  $r = 1, \dots, D_\nu$ ,  $L-1 \leq \nu \leq L \leq N$ , where  $H(\nu; N; N-1; L)$  is a hypergeometric distribution (providing the probability of obtaining  $\nu$  individuals out of  $L$  whose data are released by the agency, when  $N-1$  of such individuals are sampled from a population of size  $N$ . The  $n^{th}$  individual is the one associated with  $\mathbf{x}_0$ ). Note that for any  $j$  the posterior distribution of  $\theta_{ij}$  is the same for all  $i$ . This is the result of the exchangeability property discussed above. Similarly, we get the posterior distribution for  $\sigma_j^2$  as:

$$\pi(\sigma_j^2 | \tau = i; \mathbf{X}^{(\mathbf{L})}; \mathbf{Z}) = \sum_{\nu} \sum_r \pi(\sigma_j^2 | \tau = i; \mathbf{X}^{(\mathbf{L})}; \mathbf{Z}_{-\mathbf{i}}; d(\nu, r)) p(d(\nu, r)),$$

where

$$\begin{aligned} \pi(\sigma_j^2 | \tau = i; \mathbf{X}^{(\mathbf{L})}; \mathbf{Z}_{-\mathbf{i}}; d(\nu, r)) &\propto \prod \left[ \int_{\theta_{ij}} \int_{\phi_j^2} f_N(x_{mj} | z_{tj} / \theta_{tj}; \sigma_j^2 / \theta_{tj}^2) \cdot \pi(\theta_{ij} | \phi_j^2) \right. \\ &\quad \left. \cdot \pi(\phi_j^2) d\theta_{tj} d\phi_j^2 \right] \cdot \pi(\sigma_j^2), \end{aligned}$$

and the product is with respect to all pairs  $(x_{mj}, z_{tj})$  defined by  $d(\nu, r)$ .

Implementation of this model is computationally demanding due to the permutations between the  $\mathbf{x}$ 's and the  $\mathbf{z}$ 's. In the next section, we provide an implementation in two special cases that are easy to compute and provide useful insight about the performance of the proposed masking method.

## 5. Is the masking method effective?

Now that we have specified all the components of the disclosure scenario, i.e., the original data, the masking used, and the intruders attack, we can return to the question: “Is the masking effective in preventing disclosure?” The performance of the masking method depends on several factors:

- (a) Heterogeneity among profiles (records) in the original data.
- (b) Amount of bias.
- (c) Amount of noise.
- (d) Information provided by the agency to the users about (b) & (c).

In particular we expect disclosure risk to be an increasing function of (a) and (d) – increasing row by row variability in the original data and increasing information amount the mask used should increase the chances of re-identification – and a decreasing function of (b) and (c) – the more the data are perturbed the smaller should be the probability of re-identification.

To study the impact of these factors on the performance of the masking method, we have simulated the intruders attack under a variety of scenarios obtained considering different combinations of (a) through (d). For (d), we considered two extreme cases:

- d.1*: The agency releases the true values of the bias and noise parameters  $\{\phi_j^2\}$  and  $\{\sigma_j^2\}$  used for the perturbation.
- d.2*: The agency does not release the true value of the perturbation parameters and  $L = 0$ , i.e. the intruder knows the value of the key variables only for the target individual.

These two cases make computation straightforward and they provide a range for the performance of the masking as a function of (d). For given amounts of bias and noise introduced in the data and for a given row by row variability of the original data, case *d.1* corresponds to the situation of minimum users uncertainty about the true values of the perturbation parameters. Disclosure risk in this case should be maximum. The second case, (*d.2*), corresponds to the case of maximum users uncertainty about the true value of the perturbation parameters, since the values of the parameters are not told to the users and the users have no data from which to “learn” the true values of the parameters ( $L = 0$ ). Disclosure risk in this case should be minimum. Since computing the performance of the masking for disclosure scenarios “between” *d.1* and *d.2* is extremely difficult due to the permutations between the  $\mathbf{x}$ 's and the  $\mathbf{z}$ 's, these two cases provide a quick and inexpensive way to learn about the performance of the masking method as function of (d).

For both cases *d.1* and *d.2*, we generated two data sets with different row by row variability according to the hierarchical model in (1). We have considered a population with 100 units and only one key variable. For each data set, we use three different amounts of bias and three different amounts of noise to produce the perturbed data. Thus we have  $2 \times 3 \times 3 \times 2 = 36$  possible scenarios corresponding to all possible combinations of choices of (a) through (d). The values of  $\phi_j^2$  we used to generate the bias in the original data were  $\phi_j = 0.01/6$ ,  $\phi_j = 0.2/6$ ,  $\phi_j = 1/6$ , and we chose them to obtain a range

Table 1  
First 5 records of the original and perturbed data in the simulation study

DATA1				DATA2			
Original	minimum bias-noise	medium bias-noise	maximum bias-noise	Original	minimum bias-noise	medium bias-noise	maximum bias-noise
97.492	97.389	100.596	122.380	109.426	108.965	112.792	121.481
94.087	93.912	95.511	106.249	108.178	107.835	109.684	100.163
88.707	88.808	94.861	80.021	84.317	84.233	89.750	90.184
91.528	91.641	90.317	87.651	59.952	60.073	59.081	76.345
92.508	92.647	97.819	93.528	68.824	68.942	74.001	88.247

for  $\theta_{ij}$  equal to  $[0.995, 1.005]$ ,  $[0.9, 1.1]$ , and  $[0.5, 1.5]$  respectively. We defined the noise parameters  $\sigma_j^2$  instead as a fraction of the variance in the original data and we considered  $\sigma_j^2 = c^2 \times \text{variance}(X_j)$  with  $c = 1/100, 1/6$ , and  $1$ .

Table 1 shows the first five records for the data set with smaller row by row variability (DATA1) and for the other data set (DATA2). For each, we report the original data together with the perturbed data. Because of space constraints, we show only three (of the nine possible) combinations of bias and noise: “min. bias-noise” “med. bias-noise” and “max. bias-noise” which correspond respectively to the cases  $(\phi_j = 0.01/6, c = 1/100)$ ,  $(\phi_j = 0.2/6, c = 1/6)$  and  $(\phi_j = 1/6, c = 1)$ .

As illustrated in Table 1, for a given row by row variability (original data set) the bigger is the amount of bias and noise introduced in the data the more difficult seems the intruder tasks (compare the “original” column with the “min. bias-noise” “med. bias-noise” and “max. bias-noise” columns in both data sets). On the other hand for a given amount of bias and noise if we increase variability in the key variable, re-identification seems to become easier. For example if we compare the “original” column with the “medium bias-noise” column for the two data sets in Table 1, although only a small fraction of the population is represented in the table (5 records out of 100), the comparison suggests that, for a medium amount of bias and noise, re-identification should be quite hard in DATA1 where profiles of the original data are very similar and should become relatively easy in the DATA2 where the original data shows substantial row by row variability. For each of the possible thirty-six scenarios, we have performed a complete simulation study. Each unit has been taken in turn as the “target unit” and we have carried out the calculations in Subsection 4.1, using  $k = 100.000$  simulations. For the prior elicitation in case *d.2* we followed the elicitation scheme discussed in Fienberg et al. [6], who assumed that the intruder believes that the bias parameters  $\theta_{ij}$  used by the statistical agency to perturb the data ranges from 0.75 to 1.25. If we denote by  $\mathbf{z}_{\cdot j}$ , the perturbed value for the  $j^{\text{th}}$  key variable,  $\mathbf{z}_{\cdot j} = (z_{1j}, \dots, z_{Nj})$ , the intruder’s prior for  $\phi_j$  ( $\sigma_j^2$ ) is centered at one-sixth of the range of  $\theta_{ij}$  ( $\mathbf{z}_{\cdot j}$ ), and the coefficient of variation equals 20. The resulting parameters for the intruder’s prior were  $\alpha = \gamma = 402$ ,  $\beta = 33$ , and  $\delta$  equal to one-sixth of the range of the key variable in the released data.

We summarize the results of the simulation study for cases *d.1* and *d.2* in Tables 2 and 3, respectively. There we use “minimum”, “medium”, and “maximum” to denote the different amounts of bias (noise) used in the study, where the notation should be obvious. We have assumed that the value of  $t$  in the intruder decision rule is equal to zero. This corresponds to a worst case scenario where the intruder always makes the link. The first value in each cell represents the number of times the correct record in  $\mathbf{Z}$  was ranked first according with the posterior probability of  $\tau$  (for  $t = 0$  this is the number of correct links made by the intruder). The second value in the cell, the one in parenthesis, gives the average probability of a correct match for the records ranked as first (with standard deviation in square brackets). For example, the  $[1, 1]$  cell for DATA1 in Table 2 says that when data are masked using the minimum amount of bias and noise (among the three considered) the intruder is able to link correctly 30 out of the

Table 2  
Results for case d.1, when the agency release true values of  $\{\phi_j^2\}$  and  $\{\sigma_j^2\}$

DATA1			DATA2				
	minimum noise	maximum noise	minimum noise	medium noise	maximum noise	maximum noise	
minimum bias	30 (0.4439) [0.2860]	12 (0.2663) [0.2777]	2 (0.0273) [0.0077]	minimum bias	71 (0.7687) [0.2466]	12 (0.1823) [0.1773]	2 (0.0223) [0.0016]
medium bias	1 (0.0214) [ - ]	1 (0.0178) [ - ]	0 ( - ) [ - ]	medium bias	10 (0.2764) [0.1760]	9 (0.0880) [0.0375]	2 (0.0203) [0.0042]
maximum bias	1 (0.0152) [0.0003]	0 ( - ) [ - ]	3 (0.0142) [0]	maximum bias	2 (0.0373) [0.0150]	3 (0.0248) [0.0015]	4 (0.0217) [0.0031]

Table 3  
Results for case d.2 when the agency does not release true values of  $\{\phi_j^2\}$  and  $\{\sigma_j^2\}$  and  $L = 0$

DATA1			DATA2				
	minimum noise	medium noise	maximum noise	minimum noise	medium noise	maximum noise	
minimum bias	5 (0.0153) [0.0030]	5 (0.0152) [0.0030]	1 (0.0133) [ - ]	minimum bias	42 (0.0196) [0.0070]	10 (0.0213) [0.0079]	2 (0.0196) [0.0030]
medium bias	3 (0.0131) [0.0020]	0 ( - ) [ - ]	1 (0.0122) [ - ]	medium bias	10 (0.0233) [0.0093]	8 (0.0165) [0.0036]	1 (0.0151) [ - ]
maximum bias	2 (0.0132) [0.0004]	3 (0.0131) [0.0003]	0 ( - ) [ - ]	maximum bias	4 (0.0137) [0.0006]	3 (0.0151) [0.0005]	1 (0.0222) [ - ]

100 records in the released data and that on average the probability of this link being correct is 0.4439 with a standard deviation of 0.2860. It is important to note that the amounts of noise used in DATA1 and DATA2 are not the same since they are defined in terms of the variability of the original data which is greater in DATA2 than in DATA1.

Tables 2 and 3 illustrate that for both *d.1* and *d.2*, the number of correct links as well as the average probability of the link being correct tend to decrease as a function of the amount of bias and noise introduced into the data and they increase when row by row variability in the original data increases. As we expected, in case *d.1*, for a small amount of bias [noise], we observe a significant decrease (in both probability and number of re-identification) when we increase the noise [bias] perturbation (the cases in the first row and first column of DATA1 and DATA2 in Table 2). Further, increasing bias [noise] seems to be ineffective when the original data are already perturbed with a large amount of noise [bias] (the cases in the last row and last column of DATA1 and DATA2 in Table 2). For example, the number of intruder's correct links when minimum bias is used to contaminate the data and we increase noise from "minimum" to "medium" drops from 30 to 12 in DATA1 and from 71 to 12 in DATA2, while there is no drop at all if "maximum" bias is used and noise is increased (see last row in DATA1 and DATA2 in Table 2).

Similarly, an increase in row by row variability tends to increase significantly the chances of re-identification when small perturbations are introduced into the data and the effect decreases as amount of bias and noise used by the agency increases. For medium amounts of bias and noise introduced into the data, for example, the number of correct re-identification when we move from DATA1 to DATA2

increases from 1 to 9 in case  $d.1$  and from 0 to 8 in case  $d.2$  while there is little or no increase when we used the “maximum” amount of either bias or noise (compare last rows and last columns of DATA1 and DATA2 in Table 2). Note, however, that we need to be careful in interpreting the results since the amounts of noise in DATA1 and in DATA2 are not exactly the same and thus results for the two cases are not quite directly comparable. In accord with our intuition for case  $d.2$ , we observe similar results, although intruder’s performance is less dependent on the amount of bias and noise used by the agency due to the fact the no knowledge of the perturbation used is available to the intruder for this case. This explains for example, why the average probability significantly decreases with the the amount of perturbation in case  $d.1$  and stays almost constant in case  $d.2$  as well as why the number of correct identifications in DATA1 does not vary when “minimum” amount of bias is used and noise increase from “minimum” to “medium” in case  $d.2$  while it dramatically decreases in case  $d.1$ .

Also, as we expected, both the number of correct links and the average probability of the link being correct tend to decrease when we move from case  $d.1$  (which is the case of minimum users’ uncertainty about the true values of the perturbation parameters) to case  $d.2$  (which is the case of maximum users’ uncertainty about the true values of  $\{\phi_j^2\}$  and  $\{\sigma_j^2\}$ ). For the average probability the drop is very high for all of the scenarios we considered except again for those cases where we introduced a large amount of either bias or noise into the data. The drop in the number of correct re-identifications instead is sensible only for the case where we used minimum amounts of bias and noise to perturb the data. If the intruder knows that a small amount of perturbation has been used (case  $d.1$ ) he is likely to correctly re-identify the target individual linking the vector of key variables for the target unit with the records of key variables in the released data. If, instead, the intruder has no information on the perturbation used (case  $d.2$ ) his task is much more difficult since a close match between key variables of the target unit and key variables in the released records could correspond *either* to a real match *or* to a casual erroneous match due to the perturbation. Thus the big discrepancy between cases  $d.1$  and  $d.2$  observed in the intruder’s performance when we introduced “minimum” contamination into the data is not surprising (the number of correct identifications drops from 30 to 5 for DATA1 and from 71 to 42 in DATA2). For all the other scenarios, when at least moderate bias or noise is introduced into the data, the average probability of the links being correct dramatically decreases when we move from case  $d.1$  to case  $d.2$ , while the number of correct re-identifications stays almost constant. This means that if at least moderate perturbation is introduced into the data, not releasing the true values of the perturbation parameters (case  $d.2$ ) does not offer much more protection against the attack of a “risky” intruder than releasing the true values of these parameters (case  $d.1$ ). This result was unexpected, a priori.

## 6. What have we learned?

As we illustrated in Tables 2 and 3, for a fixed threshold for the maximum tolerable risk of disclosure the statistical agency can produce “safe” microdata either by using a suitable amount of perturbation and releasing the parameters used for the masking (case  $d.1$ ) or by fixing the amount of bias and noise and reducing the information provided to the users about the perturbation used (case  $d.2$ ). Although the results are almost equivalent in terms of disclosure risk associated with the released microdata, our simulation study suggests that strategy  $d.1$ , i.e., the release of the perturbation parameters used in the masking method, should be preferred by statistical agencies over the strategy in  $d.2$  where the true values of  $\{\phi_j^2\}$  and  $\{\sigma_j^2\}$  are not released. Not releasing the perturbation parameters, requires the agency to make extra assumptions about the intruder’s behavior. For example, in case  $d.2$  the statistical agency needs to guess the intruder’s prior distribution for  $\{\phi_j^2\}$  and  $\{\sigma_j^2\}$ . These extra assumptions increase

agency uncertainty about the results of the masking method and at the same time reduce the efficacy of the actual masking applied to the data. As we showed in Table 3 the average probability of a correct link is almost invariant to change in the amount of bias and noise used in the perturbation. In addition, as we outlined in Section 4, the task of statistical inference by users of the masked data is extremely difficult when the masking parameters are not released. This difficulty not only prevents the intruder from disclosing confidential information, but it also represents a serious obstacle for legitimate users of the data who want to perform their own statistical analyses of the data. If an agency pursues strategy *d.2*, it is not clear how users can perform standard statistical inferences from the masked data and reach correct conclusions. Under *d.1*, however, statistical analysis are possible. In some special cases, existing statistical tools can be applied directly. Thus, for example, if no bias or very little amount of bias is introduced in the data, statistical inferences for *d.1* can be carried-out using results in Sullivan and Fuller [13]. For the more general case, we would need to derive new statistical tools to deal with *d.1*, a task that goes beyond the goal of this paper. Existing results of [1,8,9,13], suggest a useful starting point for thinking about the problem.

## 7. Conclusions

The model developed in this paper is an extension of the one presented in Fienberg et al. [6]. It provides an alternative way to introduce bias and noise into confidential continuous microdata and allows the intruder to use external data on several units to enhance his chances to disclose the identity of a target individual. We have performed a full simulation study to assess effectiveness of the proposed masking method.

Preliminary results show that features of the original data – in particular the row by row variability – as well as the amount of information released to the users about the values of the contamination parameters used for the masking, are important variables for assessing the performance of the masking. While statistical agencies can not control variability in the original data, since this is an intrinsic feature of the data, they can control and decide what to tell users about the masking that has been performed. Our study suggests that releasing the true values of the parameters used for the masking should be preferred to not releasing these values, a common practice in some statistical agencies.

Several assumptions that we have made can be relaxed without altering substantially the analysis that we have presented. In particular, we are currently working on the extension of the simulation study presented here to the case where *sampling* is combined with bias and noise, the original data shows non trivial covariance structures and measures of *data utility* are considered to better assess the trade-off between disclosure risk and needs of the users. Introducing *sampling* as a disclosure limitation technique is important since most microdata currently released are a sample of the original data. Computation in such situations is much more difficult since there is extra uncertainty about whether or not the target individual is in the sample. Considering *data utility* is also important as since performance of the masking method makes no sense if a statistical agency produces “safe” data that are not useful for the users. Previous results of [1,7–10,13,15] should be helpful to define and implement suitable measures of data utility in our model.

## Acknowledgment

Preparation of this paper was supported in part by a Marie Curie Fellowship of the European Community program “Improving the Human Research Potential” under the contract number HPMFCT-2000-00463,

and in part by a U.S. National Science Foundation under Grant EIA-9876619 to the National Institute of Statistical Science. The extension of the model in Fienberg et al. [6] was originally supported by a contract from the U.S. Bureau of the Census. The contents of the paper reflect the authors' personal opinions. Neither the European Commission, the National Science Foundation, nor the U.S. Census Bureau is responsible for any views or results presented. We thank professor M.J. Bayarri for ideas and comments that are reflected in various ways in our work.

## References

- [1] C.A. Clayton and W.K. Poole, *Use of Randomized Response Techniques in Maintaining Confidentiality of Data*, Draft report RTI Project No. 2520-1159, Research Triangle Park, N.C., 1976.
- [2] J. Domingo-Ferrer and V. Torra, A Quantitative Comparison of Disclosure Control Methods for Microdata, in: *Confidentiality, Disclosure, and Data Access. Theory and Applications for Statistical Agencies*, P. Doyle, J. Lane, J.M. Theeuwes and L.V. Zayatz, eds, North-Holland, Amsterdam, 2001, pp. 111-134.
- [3] P. Doyle, J. Lane, J.M. Theeuwes and L.V. Zayatz, eds, *Confidentiality, Disclosure, and Data Access. Theory and Applications for Statistical Agencies*, North-Holland, Amsterdam, 2001.
- [4] G.T. Duncan and D. Lambert, The Risk Disclosure of Microdata, *Journal of Business and Economic Statistics* **7** (1989), 207-217.
- [5] S.E. Fienberg, Conflict Between the Needs for Access Statistical Information and Demands for Confidentiality, *Journal of Official Statistics* **10** (1994), 115-132.
- [6] S.E. Fienberg, E.U. Makov and A.P. Sanil, A Bayesian Approach to Data Disclosure: Optimal Intruder Behavior for Continuous Data, *Journal of Official Statistics* **13** (1997), 75-89.
- [7] J.M. Gouweleeuw, P. Kooiman, L.C.R.J. Willenborg and P.-P. de Wolf, Post Randomisation for Statistical Disclosure Control: Theory and Implementation, *Journal of Official Statistics* **14** (1998), 463-478.
- [8] W.A. Fuller, *Measurement Error Models*, New York, John Wiley, 1987.
- [9] W.A. Fuller, Masking Procedures for Microdata, *Journal of Official Statistics* **9** (1993), 383-406.
- [10] R.J.A. Little, Statistical Analysis of Masked Data, *Journal of Official Statistics* **9** (1993), 407-426.
- [11] J. Kim, A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation, *Proceedings of the Survey Research Method Section*, American Statistical Association, Alexandria, VA, USA, 1986, pp. 370-374.
- [12] G. Paas, Disclosure Risk and Disclosure Avoidance for Microdata, *Journal of Business and Economic Statistics* **6** (1988), 487-500.
- [13] G. Sullivan and W.A. Fuller, *The use of Measurements errors to Avoid Disclosure*, Proceedings of the Survey Research Methods Section, American Statistical Association, Alexandria, VA, USA, 1989, pp. 802-807.
- [14] G. Sullivan and W.A. Fuller, *Construction of Masking Error for Categorical Variables*, Proceedings of the Survey Research Methods Section, American Statistical Association, Alexandria, VA, USA, 1990, pp. 435-439.
- [15] M. Trottini, *A User-Agency Model for Disclosure Limitation Problems*, Technical Report TR04-2002, Departamento de Estadística e I.O., Universitat de València, 2002.
- [16] M. Trottini and S.E. Fienberg, Modelling User Uncertainty for Disclosure Risk and Data Utility, *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems* **10**(5) (2002), 511-528.
- [17] L. Willenborg and T. De Waal, Statistical Disclosure Control in Practice, *Lecture Notes in Statistics* **111** (1996), New York, Springer.
- [18] L. Willenborg and T. De Waal, Elements of Disclosure Control, *Lecture Notes in Statistics* **155** (2001), New York, Springer.