

MODELLING USER UNCERTAINTY FOR DISCLOSURE RISK AND DATA UTILITY

MARIO TROTTINI

*Departamento de Estadística e I. O. Universitat de València, 46100 Burjassot, València, Spain
E-mail: mario.trottini@uv.es*

STEPHEN E. FIENBERG

*Department of Statistics Carnegie Mellon University, Pittsburgh, PA 15213, USA
E-mail: fienberg@stat.cmu.edu*

Received (received date)

Revised (revised date)

In this paper we show how a simple model that captures user uncertainty can be used to define suitable measures of disclosure risk and data utility. The model generalizes previous results of Duncan and Lambert¹. We present several examples to illustrate how the new measures can be used to implement existing optimality criteria for the choice of the best form of data release.

Keywords: Bayesian decision theory; Disclosure limitation; Data release; Information theory; Intruder uncertainty.

1. Introduction

Data Disclosure Limitation denotes a set of techniques aimed to protect confidentiality in the release of statistical data. The goal is to find forms of data release that are both *useful* for the users (researchers, policy makers, public opinion) and *safe*, in the sense that do not violate privacy and confidentiality of respondents represented in the data. This is usually achieved by applying a transformation to the original data, often referred to as a *data mask*, and then releasing the resulting data set^{2,3,4,5,6}.

In this paper we show how a simple Bayesian users' model can be used to assess effectiveness of candidates for data masking. The problem is of great interest in data disclosure limitation since assessment of effectiveness of masking is the basis for any optimality criteria for the choice of the best form of data release^{7,8,9,10,11}.

Current research is primarily heuristic. This has resulted in measures of effectiveness of masking that, although intuitively reasonable, (1) are not very well understood, since the assumptions that characterize these measures are usually unknown, and (2) do not allow comparison of arbitrary masks, since measures defined for a specific data mask sometimes are not well defined for others.

Our model shows a way to overcome these problems and define suitable measures based on solid theoretical ground. In defining effectiveness of data masking we take

the users' perspective. For simplicity we assume only two users: an *intruder* (I) who wants to use the released data to disclose confidential information about the data providers and a *scientist* (S) who wants to use the released data to infer general features of the population represented in the data. The masked data are *useful* (or equivalently they have high *data utility*) if the scientist is satisfied with the quality of the statistical analysis that he can perform using the masked data. The masked data are *safe* (or equivalently have low *disclosure risk*) if the intruder does not find the masked data very useful to make inference about his target. The input to our model is a set of assumptions about the information available to the users before the candidate mask is released, the intruder's and the scientist's target (i.e. what should be protect and what should be released), and the estimation procedure that intruder and scientist will use to estimate their targets once a candidate mask is released. The model's output consists of the users' estimates of their targets, a measure of users' uncertainty about the reliability of their estimates, and a measure of disclosure risk and data utility that depends on these uncertainties. In particular we propose to measure disclosure risk (data utility) as an arbitrary decreasing function of the intruder's (scientist's) uncertainty about the true value of his target. Masking is effective if it results in a data set that has low disclosure risk and high data utility.

The strength of the model is that we are able to characterize measures of disclosure risk and data utility in terms of quantities of straightforward interpretation in Bayesian decision theory such as intruder's and scientist's prior distributions (that reflect the information available to the users before masked data are released) and intruder's and scientist's loss functions (that formalize the estimation procedure used by the users). This provides a general framework to define new and easy understandable measures of effectiveness of masking and, at the same time, it suggests new interpretation of some of the existing measures that we show to be particular cases of the users' model that we present here.

Our entire framework relies on the distinction between the intruder's and the scientist's targets. This distinction is the basis and the premise of Data Disclosure Limitation. The underlying assumption is that researchers, policy makers and public opinion are interested in statistical analysis aimed to discover and investigate general features of the population represented in the data (relationship, dependence, association among variables, modeling of different type of phenomena). In our framework these are the scientist's targets. On the other hand, specific users might want to perform statistical inferences aimed to disclose confidential information about individual respondents represented in the data. These in our framework are the intruder's targets. Legal, ethical and pragmatic considerations force the statistical agency to release the data in a form that the scientist's targets can be achieved and intruder's targets can not. Data disclosure limitation makes sense because, intuitively, it is possible to release information about general features of the population without violating the privacy and confidentiality of individual respondents. Certainly, there are cases where the same entity/person could have interest in general features of the population (thus acting as scientist) as well as

in confidential information about individual respondents (thus acting as intruder). Nevertheless, in these cases, it is also useful to think about the existence of two different users in order to distinguish between targets that the agency wants to achieve (the scientist’s targets) and targets that the agency wants to discourage (the intruder’s targets), the prior information available for these targets and the procedures to be used for their estimation. As illustration consider the release of a microdata with medical, financial and demographic information for individuals in a certain population. In this case an insurance company could play both the role of the intruder and the scientist. As “intruder,” the company might want to use the released data to disclose confidential medical information about specific respondents in the population that are currently applying for a life insurance (in order to deny or accept, on the basis of this information, the policy of these respondents). As “scientist,” the insurance company might want to use the data to estimate the proportion of people in the population over 60 years of age and with income greater than \$4000, since these are the persons eligible for the policy that the company sells and this information is input to the design of company marketing strategies. The statistical agency in this case could release a sample from the population and this could be sufficient to ensure an accurate estimate of the proportion of people over 60 and with income greater than \$4000 and, at the same time, preserve the confidentiality of the respondents in the sample.

The paper is organized as follows. Section 2 introduces the basic assumption and notation. In Sections 3 we present the users’ model. In Section 4, we use the model’s output to derive new measures of disclosure risk and data utility. We discuss the relationship of these new measures with the measures of disclosure proposed by Duncan and Lambert¹. Generalizing Duncan and Lambert’s results for disclosure risk, we show that common measures of data utility currently used by statistical agencies or suggested in the literature, such as the measures of data utility based on entropy or the variance inflation measure of data utility⁷, are special cases of our framework. In Section 5 we illustrate how to use the new measures of disclosure risk and data utility to implement the R-U confidentiality map proposed by Duncan, Keller-McNulty, and Stokes⁸ as optimality criterion for the choice of the best form of data release. Section 6 summarizes the main results in the paper.

2. Notation and Assumptions

Our general setting is similar to the one described in Trottini⁹. In part to set the notation, we briefly re-state the basic assumptions. We denote by \tilde{D} the generic masked data and by Θ_I and Θ_S the intruder’s and the scientist’s target. Both Θ_I and Θ_S are unknown to the users and need to be estimated on the basis of the masked data that the agency releases. Following Duncan and Lambert¹ we assume that, prior to the release of any data, the intruder and the scientist can express their beliefs about the true value of Θ_h ($h = I, S$) in terms of *prior probability distributions* (“prior” here means that these probability distributions are based on the information available to the users before \tilde{D} is released). For simplicity we assume

that Θ_h takes values in a finite set Ω_h , $\Omega_h = \{\theta_h^1, \dots, \theta_h^{n_h}\}$ and we denote by $\pi_h(\theta_h)$ the prior probability that user h assigns to the event $\Theta_h = \theta_h$. We assume that in order to quantify disclosure risk and data utility associated with masked data \tilde{D} the agency “simulates” the behaviour of the users when \tilde{D} is released. In particular,

Assumption 1: The masked data are *useful* if their release would result in low uncertainty on the part of the scientist about the true value of Θ_S . The masked data are *safe* if their release would result in high intruder uncertainty about the true value of Θ_I .

The intuition underlying assumption 1 is the same as that which characterizes the measures of disclosure risk proposed by Duncan and Lambert¹. If the intruder’s (scientist’s) uncertainty after the masked data have been release is “high” then the release is of little help for the intruder (scientist) in inferring the true value of his target and thus the intruder (scientist) is discouraged from trying any type of inference and disclosure risk (data utility) is small.

We now describe a simple Bayesian users’ model to assess users uncertainty after the data have been release. Then we derive explicit formulas for disclosure risk and data utility.

3. The Users’ Model

When the agency releases the masked data \tilde{D} , user h updates the probability that $\Theta_h = \theta_h$ in the light of the new information contained in the released data, producing *posterior probabilities* $\pi_h(\theta_h|\tilde{D})$, $\theta_h \in \Omega_h$. If user h knew that $\Theta_h = \theta_h^*$ (i.e. if it was $\pi_h(\theta_h^*|\tilde{D}) = 1$, for some $\theta_h^* \in \Omega_h$) then the estimation problem would be trivial: h would estimate Θ_h by θ_h^* . Even after having seen the data \tilde{D} , however, the user is generally not completely sure about the true value of Θ_h . Thus, he needs to establish an estimation criteria. We assume that both the scientist and the intruder follow an estimation method in two steps (which in Bayesian statistics is known as the *Expected Loss Principle*):

Step 1: User h defines a loss function $L_h(\cdot, \cdot)$ for the decision problem “Estimate Θ_h ”;

Step 2: User h estimates Θ_h by the value \hat{e}_h in Ω_h that produces the smallest average loss, where the average loss associated with a generic estimate e is given by:

$$\sum_{i=1}^{n_h} L_h(e; \theta_h^i) \pi_h(\theta_h^i|\tilde{D}), \quad h = I, S, \quad (1)$$

where $L_h(e, \theta_h)$ represents the loss that user h incurs if he estimates Θ_h by e when the true value of the target is θ_h . Clearly $L_h(e, \theta_h)$ is a non decreasing function of some “distance” between e and θ_h , i.e., the farther the estimate e is from θ_h , the bigger is the loss. Intuitively the value of $e \in \Omega_h$ that produces the smallest loss should be selected. But Θ_h is unknown to user h and therefore he can’t calculate

the exact loss associated with the estimate e . Instead he calculates the average loss associated with e for each possible estimate $e \in \Omega_h$ and then step 2 provides the optimal estimate.

Following De Groot¹², we define the user's h uncertainty about the true value of his target after the masked data have been released as the average loss associated with the estimate \hat{e}_h ,

$$U_h(\tilde{D}) = \sum_{i=1}^{n_h} L_h(\hat{e}_h; \theta_h^i) \pi_h(\theta_h^i | \tilde{D}), \quad h = I, S. \quad (2)$$

The intuition underlying (2) is simple. If after the masked data have been released user h has little uncertainty about the true value of his target than he will be able to find an estimate of the target that is “close” to the true value and the average loss will be small (since for all plausible θ_h , $L_h(\hat{e}_h; \theta_h)$ will be small). On the other hand, the bigger users' h uncertainty about the true value of his target is, i.e., the bigger the number is and the “more different” the θ_h 's that are plausible in the light of the observed data, the more difficult it will be for the user to find a value which is “close” to all these θ_h 's. Then the average loss will be high.

4. Disclosure Risk And Data Utility

Let $f(\cdot)$ and $g(\cdot)$ be two arbitrary real valued strictly decreasing functions. We propose to measure disclosure risk and data utility by

$$\text{Disclosure Risk} = f(U_I); \quad (3)$$

$$\text{Data Utility} = g(U_S). \quad (4)$$

The measures of disclosure risk and data utility in (3) and (4) formalize the intuition underlying assumption 1. High intruder's (scientist's) uncertainty means that the released data are of little help for the intruder (scientist) to estimate his target and as result disclosure risk (data utility) associated with the masked data is small.

Duncan and Lambert' measures of disclosure¹ are special cases of (3) for suitable choices of the function $f(\cdot)$. The major difference between their measures of disclosure and ours in (3) is in the interpretation of the function $f(\cdot)$. Duncan and Lambert suggest the use of different functions $f(\cdot)$ depending on the disclosure scenario. In our framework, however, there is no reason to make this distinction. Given any two real-valued strictly-decreasing functions $f_1(\cdot)$ and $f_2(\cdot)$ with common domain A , $f_1(\cdot) \neq f_2(\cdot)$, there exist a one-to-one transformation z such that $f_1(x) = z(f_2(x))$, $\forall x \in A$. If we consider the two measures of disclosure risk, $Risk_1 = f_1(U_I)$, and $Risk_2 = f_2(U_I)$, any assessment of the risk of disclosure in terms of $Risk_1$ has an equivalent assessment in terms of $Risk_2$ and vice versa. Thus, in our framework formula (3) does not define a class of alternative measures of disclosure risk but rather an equivalence class of measures of disclosure.

An interesting feature of (3) is that many measures of disclosure risk currently used by statistical agencies can be found as special case of (3) for suitable choices

Willenborg and de Waal⁷ propose measuring the data utility associated with suppressing cells in the original table as:

$$\text{Data Utility} = g(\log M(S)), \quad (7)$$

where $g(\cdot)$ is a decreasing function. If we assume a uniform distribution on S (this corresponds to assuming that without additional information each solution in S is equally likely to be the right one), then (7) is a decreasing function of the entropy.

If we further specify the elements of the users' model (scientist's target, loss function, prior information) it is possible to show that (7) belongs to the general class of measures of data utility in (4). In particular, we obtain (7) by assuming:

- (i) The scientist is interested in the cross classification of the k categorical variables in P and he tries to infer the distribution of F from the released table \tilde{F} (it is assumed that both the population size, N , and the m possible cross classifications are known to the users). Note that prior to the release of \tilde{F} , from the scientist's perspective, F is a discrete random variable with support \mathcal{X} , the set of all nonnegative m -vectors with integer entries that add-up to N ,

$$\mathcal{X} = \{(x_1, \dots, x_m) : x_i \text{ non-negative integer and } \sum_{i=1}^m x_i = N\}. \quad (8)$$

Denoting by $M(\mathcal{X})$ the cardinality of \mathcal{X} and by \underline{x}_i the generic element in \mathcal{X} the scientist's target can be represented as:

$$\Theta_S = \{\Theta_1, \dots, \Theta_{M(\mathcal{X})}\} \quad (9)$$

where Θ_i is the probability that F is equal to (the table) \underline{x}_i , $i = 1, \dots, M(\mathcal{X})$.

- (ii) The loss that the scientist incurs for estimating Θ_S by e when $\Theta_S = \theta_S$, is:

$$L_S(e, \theta_S) = D_{KL}(e, \theta_S) + \text{Entropy}(\theta_S) = - \sum_{i=1}^{M(\mathcal{X})} \log(e_i) \cdot \theta_i \quad (10)$$

where e_i and θ_i denote the i^{th} component of e and θ_S ($i = 1, \dots, M(\mathcal{X})$), D_{KL} is the Kullback Leibler divergence between e and θ_S ,

$$D_{KL}(e, \theta_S) = \sum_{i=1}^{M(\mathcal{X})} \log(\theta_i/e_i) \cdot \theta_i, \quad \text{and} \quad \text{Entropy}(\theta_S) = - \sum_{i=1}^{M(\mathcal{X})} \log(\theta_i) \cdot \theta_i. \quad (11)$$

$L_S(e, \theta_S)$ says that the scientist's loss is small when (a) the scientist's estimate agrees with the true distribution θ_S (small Kullback Leibler divergence between e and θ_S), and (b) the true distribution is very informative about the true value of F (small entropy of θ_S). In all the other cases the scientist's loss is large since either scientist's estimate of Θ_S is poor (high values of Kullback Leibler divergence) or, although correct, scientist's estimate of Θ_S contains little information about F (large entropy of $\theta_{(S)}$).

- (iii) The scientist's prior distribution for θ_S is the uniform distribution on the set $\{\pi_1(\cdot), \dots, \pi_{M(\mathcal{X})}(\cdot)\}$ where $\pi_i(\cdot)$ denotes the distribution degenerate at \underline{x}_i , $\underline{x}_i \in \mathcal{X}$.

Under (i)-(iii) the scientist posterior distribution is the uniform distribution on the set of degenerate distributions on \underline{x}_i , $\underline{x}_i \in S_1$, scientist's uncertainty is $\log(M(S_1))$ and the data utility in (4) takes the form

$$\text{Data Utility} = g(U_S) = g(\log(M(S_1))) = g(\log(M(S))) \quad (12)$$

which is exactly the measure of data utility proposed in⁷.

In the example at the beginning of this subsection is $U_S = \log(114)$ and \mathcal{X} is the set of non negative integer vectors (x_1, x_2, x_3, x_4) such that $\sum_{i=1}^4 x_i = 258$.

4.2. Variance inflation measure of data utility

Variance inflation is the increase in the variance of an estimate of the scientist's target due to the data mask (the greater the increase in variance the smaller the data utility). In particular, following Willenborg and de Waal⁷ we let

$$\text{variance inflation} = \text{Var}(\hat{\theta}|D_0)/\text{Var}(\hat{\theta}|\tilde{D}) \quad (13)$$

where $\text{Var}(\hat{\theta}_S|D_0)$ and $\text{Var}(\hat{\theta}_S|\tilde{D})$ represent the variance of the estimate of Θ_S when the original data and the masked data, respectively, are released. It can be shown that (13) is a special case of (4) if we assume:

$$L_S(e; \theta_S) = (e - \theta_S)^2; \quad g(U_S) = c/U_S, \quad (14)$$

where $c = \text{Var}(\hat{\theta}_S|D_0)$. Note that $\text{Var}(\hat{\theta}_S|D_0)$ does not depend on the particular form of data release and thus we can consider it to be as a constant.

5. The R-U Confidentiality Map

We now use the measures of disclosure risk and data utility in (3) and (4) to implement the optimality criterion for data release known as the R-U confidentiality map proposed by Duncan, Keller-McNulty, and Stokes⁸. For concreteness we take $f(\cdot)$ and $g(\cdot)$ to be the inverse function;

$$\text{Data Utility} = 1/U_S; \quad \text{Disclosure Risk} = 1/U_I. \quad (15)$$

From the discussion in Section 4, however, it is clear that any other decreasing function of U_I and U_S would work as well and it would produce the same results (since the ordering among alternative forms of data release in terms of disclosure risk and data utility would be exactly the same).

Consider a generic data mask M . We can represent M as a vector with two components, $M = (DLT, \Theta_{DLT})$ where DLT denotes the type of disclosure limitation technique used (additive noise, sampling, microaggregation, etc.) and Θ_{DLT} denotes the parameter vector that characterizes DLT . For example, we can represent the mask that adds Gaussian noise with mean 0 and variance 3 to the original data as $M = (\text{additive noise}, 3)$. Intuitively the best form of data release is the

one that maximizes the data utility among those that have disclosure risk below a fixed threshold t . Duncan, Keller-McNulty, and Stokes⁸ R-U confidentiality map formalizes this idea.

For each data mask DLT we represent in a graph the disclosure risk and the data utility associated with DLT as a function of Θ_{DLT} . We then select the value Θ_{DLT}^{Opt} that maximizes data utility among all those values for which the disclosure risk is below t . If there are k alternative data masks, DLT_1, \dots, DLT_k we evaluate the value of $\Theta_{DLT_i}^{Opt}$ for each data mask DLT_i and then we select as the best form of data release that one produced by the mask $M_{i^*} = (DLT_{i^*}, \Theta_{DLT_{i^*}}^{Opt})$ which maximizes data utility among all the masks $M_i = (DLT_i, \Theta_{DLT_i}^{Opt})$, $i = 1, \dots, k$, i.e.,

$$i^* = \operatorname{argmax}_i \text{Data Utility}[M_i = (DLT_i, \Theta_{DLT_i}^{Opt})], \quad i = 1, \dots, k. \quad (16)$$

We now apply this optimality criterion to two examples. In the first, we assume that the original data set consists of a random sample of a univariate random variable X with normal distribution. The intruder is interested in identifying the value of X for a specific respondent, while the scientist is interested in making inferences about the mean of X . In the second example, we extend the univariate results to the multivariate case. In both examples we assume that the intruder and the scientist use a quadratic loss function and that disclosure risk and data utility are measured as in (15).

5.1. Example 1: univariate normal case

A statistical office collects a random sample of size n of a quantitative variable X in a population of size N where $n \ll N$. We denote the original data by $\underline{x}_n = (x_1, \dots, x_n)$ and we assume the following disclosure scenario:

- The scientist's target, Θ_S , is the mean of X , μ , i.e., $\Theta_S = \mu$;
- The intruder's target, Θ_I , is to disclose the value of X for a particular respondent I_j in the collected sample, $\Theta_I = X_j$;
- Both the intruder and the scientist use a quadratic loss function to quantify the loss that they pay for an estimate e of the target values Θ_I and Θ_S , i.e.

$$L_I(e; x_1) = (e - x_1)^2, \quad L_S(e; \mu) = (e - \mu)^2; \quad (17)$$

- The scientist and the intruder agree that the distribution of X in the population can be very well approximated by a normal distribution with unknown mean μ and known variance σ^2 , i.e., $X \sim N(\mu, \sigma^2)$, σ^2 known;
- The scientist and the intruder formalize their prior beliefs about μ with a normal distribution with known mean θ and known variance ϕ^2 , i.e., $\mu \sim N(\theta, \phi^2)$, θ, ϕ^2 known.

We consider three standard disclosure limitation techniques, *additive noise*, *microaggregation*, and *additive noise - microaggregation* (the data are first perturbed by additive noise and then microaggregation is applied). We make the implicit

assumption that if the statistical agency releases any type of masked data, it also releases the parameter value/s used for the masking.

For this particular example, because the sample mean is a *sufficient statistic* for μ we can show analytically that: (i) the optimal number of groups for microaggregation is one, i.e., microaggregation is most effective when we release instead of the original observations their average; (ii) using additive noise together with microaggregation with the number of groups equals to one (which is equivalent to perturbate the data using additive noise and then release their average) is always better than using additive noise only. We therefore focus only on the two types of masking:

1. $M_1 =$ (microaggregation, $G = 1$), the agency releases: $\bar{x}_n = \frac{\sum_{i=1}^n x_i}{n}$;
2. $M_2 =$ (additive noise + microaggregation, ($r, G = 1$)), the agency releases: $\bar{y}_n(r) = \frac{\sum_{i=1}^n y_i(r)}{n}$, where

$$y_i(r) = x_i + \epsilon, \quad \epsilon \sim N(0, r^2), \quad r > 0; \quad (18)$$

(here r and G denotes the parameters for additive noise and microaggregation respectively).

Under quadratic loss function the intruder's (scientist's) optimal estimate of Θ_I (Θ_S) and the intruder's (scientist's) uncertainty are, respectively, the mean and the variance of the posterior distribution for Θ_I (Θ_S). Thus (15) becomes:

$$\text{Disclosure Risk} = \frac{1}{\text{Var}(X_1)}; \quad \text{Data Utility} = \frac{1}{\text{Var}(\mu)}. \quad (19)$$

where $\text{Var}(X_1)$ ($\text{Var}(\mu)$) represents the variance of the intruder's (scientist's) posterior distribution for X_1 (μ) based on the released masked data.

Table 2 shows the disclosure risk and data utility associated with the masking M_1 and M_2 . For comparison, we also report the pair (*data utility*, *disclosure risk*) for the trivial mask, M_0 , consisting of suppression of all the observations.

Table 2. Disclosure risk and data utility, example 1.

Type of Masking	Disclosure Risk	Data Utility
M_0	$\frac{1}{\sigma^2 + \phi^2}$	$\frac{1}{\phi^2}$
M_1	$\frac{\frac{1}{n}}{(n-1) \cdot \sigma^2}$	$(\frac{1}{\phi^2} + \frac{n}{\sigma^2})$
M_2	$[\frac{r^4 \phi^2}{(\sigma^2 + r^2)(\sigma^2 + r^2 + n\phi^2)} + \frac{(n-1)\sigma^4 + n\sigma^2 r^2}{n(\sigma^2 + r^2)}]^{-1}$	$(\frac{1}{\phi^2} + \frac{n}{\sigma^2 + r^2})$

The disclosure risk and data utility associated with M_0 (i.e., the inverse of users uncertainty before any data are released), represent the baseline that we use to assess to what extent the non trivial masking M_1 and M_2 are effective. In Figure 1, we show the R-U confidentiality map corresponding to the three forms of data release when $n = 10$, $\sigma^2 = 2$, $\phi^2 = 5$. The diamond, the circle, and the dashed line represent, respectively, the pairs (*data utility*, *disclosure risk*) when the trivial mask, M_0 and the non trivial masks, M_1 and M_2 are applied (the contamination parameter

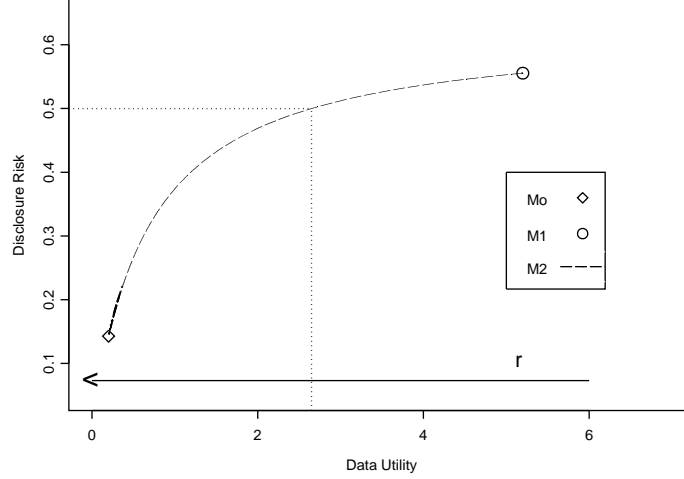


Fig. 1. R-U confidentiality map, univariate case, $n = 10$, $\phi^2 = 5$, $\sigma^2 = 2$.

r for M_2 varies in the interval $[0, 40]$). As we expect for large value of r , the mixed strategy additive noise-microaggregation (M_2) is equivalent to not release any data (M_0), while for r close to zero the mixed strategy additive noise-microaggregation is equivalent to microaggregation with $G = 1$ (M_1). By examining Figure 1, we can see that the optimal choice for our data release depends on the threshold value for the maximum tolerable risk of disclosure. For example, if this threshold value is set equal to 0.5 then the optimal masking (among those considered) is the mixed masking (additive noise + microaggregation, ($r = 2.081, G = 1$)), i.e., the original data should be first perturbed using additive noise with $r = 2.081$ and then microaggregation with $G = 1$ should be applied. This would result in a disclosure risk just below the threshold 0.5 and data utility equals to 2.65.

5.2. Example 2: multivariate case

Consider the same disclosure scenario as in example 1 but suppose the agency collects data on two quantitative variables (the extension to the case of $k > 2$ variables is straightforward). The assumptions that characterize the users' model in example 1 now become:

- $X = (X^{(1)}, X^{(2)}) \sim MN(\mu, \Sigma_1)$, $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$, $\Sigma_1 = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$ known;
- The scientist's target, Θ_S , is to estimate μ , $\Theta_S = \mu$;
- The intruder's target is the value of X for a particular respondent, $\Theta_I = X_j$;
- Both the intruder and the scientist use a quadratic loss function, i.e.

$$L_I(e; x_1) = (e - x_1)^T W (e - x_1), \quad L_S(e; \mu) = (e - \mu)^T M (e - \mu),$$

$$W = \begin{pmatrix} w_1 & 0 \\ 0 & w_2 \end{pmatrix}, \quad Z = \begin{pmatrix} z_1 & 0 \\ 0 & z_2 \end{pmatrix}, \quad w_i > 0, \quad z_i > 0, \quad i = 1, 2; \quad (20)$$

- Both the scientist's and the intruder's prior uncertainty about μ can be formalized by a multivariate normal distribution with known vector mean θ and known covariance matrix Σ_0 , $\mu \sim MN(\theta, \Sigma_0)$, θ, Σ_0 known.

Note that the relative magnitude of w_i (z_i) indicates the importance to the intruder (scientist) of the estimate of the i^{th} component of X_j (μ) compared with the other component.

As in example 1 (and for the same reasons) we restrict here our attention to only three masking techniques, the trivial mask, M_0 (suppressing all of the observations), *microaggregation* with the number of groups equals to one (M_1), and a mixed strategy *additive noise-microaggregation* (M_2). In this case we have:

- M_1 = (microaggregation, $G = 1$), the agency releases: $\bar{x}_n = (\frac{\sum_{i=1}^n x_i^1}{n}, \frac{\sum_{i=1}^n x_i^2}{n})$, where x_i^j is the j^{th} component of x_i , $i = 1, \dots, n$, $j = 1, 2$;
- M_2 = (additive noise + microaggregation, ($r, G = 1$)), the agency releases: $\bar{y}_n(l) = (\frac{\sum_{i=1}^n y_i^1}{n}, \frac{\sum_{i=1}^n y_i^2}{n})$,

$$y_i(l) = x_i + \epsilon, \quad \epsilon \sim MN(0, \Lambda), \quad \Lambda = \begin{pmatrix} l_1 & 0 \\ 0 & l_2 \end{pmatrix}, \quad l_i \geq 0. \quad (21)$$

The agency's goal is to choose which mask is the best. As in example 1 we make the implicit assumption that if the agency chooses to release any type of masked data, it also releases the parameters used for masking. Under quadratic loss function, the intruder's (scientist's) optimal estimate of X_j (μ) is the vector of the posterior mean of X_j (μ) and the intruder's (scientist's) uncertainty about the true value of X_j (μ) is the weighted sum of the variance of the components of X_j (μ). Thus (15) becomes:

$$\text{Discosure Risk} = \frac{1}{\sum_{i=1}^2 w_i \text{Var}_p(X_j^{(i)})}; \quad \text{Data Utility} = \frac{1}{\sum_{i=1}^2 z_i \text{Var}_q(\mu_i)}. \quad (22)$$

For illustration we assume:

$$n = 20, \quad \Sigma_0 = \begin{pmatrix} 5 & 0.5 \\ 0.5 & 4 \end{pmatrix}, \quad \Sigma_1 = \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1.5 \end{pmatrix}, \quad W = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}, \quad (23)$$

and for the matrix Z we consider the following choices:

$$\begin{aligned} Z_1 &= \begin{pmatrix} 6 & 0 \\ 0 & 0.5 \end{pmatrix}, \quad Z_2 = \begin{pmatrix} 3 & 0 \\ 0 & 0.25 \end{pmatrix}, \quad Z_3 = \begin{pmatrix} 3 & 0 \\ 0 & 1.5 \end{pmatrix}, \quad Z_4 = \begin{pmatrix} 1.5 & 0 \\ 0 & 0.75 \end{pmatrix}, \\ Z_5 &= \begin{pmatrix} 5 & 0 \\ 0 & 15 \end{pmatrix}, \quad Z_6 = \begin{pmatrix} 2.5 & 0 \\ 0 & 7.5 \end{pmatrix}, \quad Z_7 = \begin{pmatrix} 2 & 0 \\ 0 & 9 \end{pmatrix}, \quad Z_8 = \begin{pmatrix} 1 & 0 \\ 0 & 4.5 \end{pmatrix}, \\ Z_9 &= \begin{pmatrix} 0.5 & 0 \\ 0 & 6 \end{pmatrix}, \quad Z_{10} = \begin{pmatrix} 0.25 & 0 \\ 0 & 3 \end{pmatrix}. \end{aligned} \quad (24)$$

Table 3 summarizes the relationship between the matrices Z_i and the matrix W that defines the intruder's loss function (z_{1i} and z_{2i} denote the elements of the diagonal of Z_i , $i = 1, \dots, 10$).

Table 3. Diagonal of Z_i vs. diagonal of W

Matrix	$\frac{z_{2i}}{z_{1i}}$ vs. $\frac{w_2}{w_1}$	Matrix	$\frac{z_{2i}}{z_{1i}}$ vs. $\frac{w_2}{w_1}$
Z_1 and Z_2	$\frac{z_{21}}{z_{11}} = \frac{z_{22}}{z_{12}} = 0.08 < \frac{w_2}{w_1} = 3$	Z_7 and Z_8	$\frac{z_{27}}{z_{17}} = \frac{z_{28}}{z_{18}} = 4.5 > \frac{w_2}{w_1} = 3$
Z_3 and Z_4	$\frac{z_{23}}{z_{13}} = \frac{z_{24}}{z_{14}} = 0.5 < \frac{w_2}{w_1} = 3$	Z_9 and Z_{10}	$\frac{z_{29}}{z_{19}} = \frac{z_{210}}{z_{110}} = 12 > \frac{w_2}{w_1} = 3$
Z_5 and Z_6	$\frac{z_{25}}{z_{15}} = \frac{z_{26}}{z_{16}} = \frac{w_2}{w_1} = 3$		

In Figure 2 we show the R-U confidentiality map corresponding to the different matrices Z_i . The solid line represents the pairs (*data utility*, *disclosure risk*) corresponding to M_2 when the variance of the contamination of the second component of the original observations, l_2 , is zero and the variance of the contamination of the first component of the original observations, l_1 , varies in $(0, 100)$. The dotted line represents the pairs (*data utility*, *disclosure risk*) when $l_1 = 0$ and l_2 varies in $(0, 100)$. It can be shown that all other possible combinations of l_1 and l_2 produce a curve that lays between the two that we have plotted. As in example 1, the diamond and the circle represent respectively the masking M_0 and M_1 (for space constraints we report the full legend only on the first plot). Figure 2 and Table 3 suggest the following optimality criterion for this example:

- If $\frac{z_{2i}}{z_{1i}} < \frac{w_2}{w_1}$ (as for Z_1 - Z_4), then the statistical agency should use M_2 with (l_1, l_2) such that:

$$l_1 = 0, \quad \text{and} \quad l_2 = \operatorname{argmax}_l \{l : \text{Disclosure Risk}[\bar{y}_n(0, l)] < t\}. \quad (25)$$

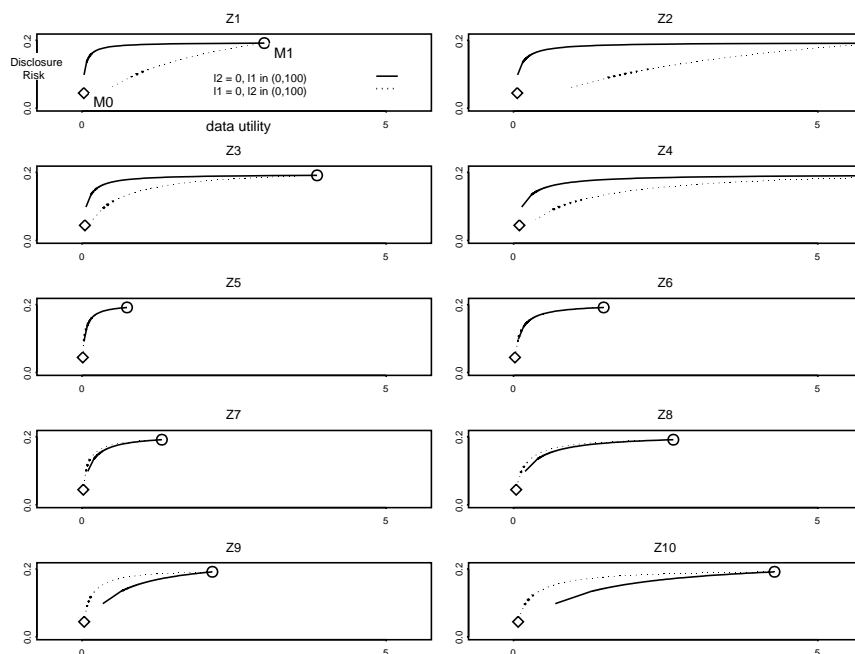
- If $\frac{z_{2i}}{z_{1i}} > \frac{w_2}{w_1}$ (as for Z_7 - Z_{10}), then the statistical agency should use M_2 with (l_1, l_2) such that:

$$l_2 = 0, \quad \text{and} \quad l_1 = \operatorname{argmax}_l \{l : \text{Disclosure Risk}[\bar{y}_n(l, 0)] < t\}. \quad (26)$$

The interpretation underlying this optimality criterion is simple. If the scientist's interest for the first component of μ is greater, in relative sense, than the intruder's interest for the first component of X_j , i.e., if

$$\frac{z_{2i}}{z_{1i}} < \frac{w_2}{w_1}, \quad (27)$$

then the optimal strategy is to leave the first component of the original data unchanged (i.e., choose $l_1 = 0$) and to contaminate the second component as much as necessary to obtain a disclosure risk smaller than the threshold for the maximum

Fig. 2. Disclosure Risk vs. Data utility for different choices of the matrix Z

tolerable disclosure. Similarly, if the scientist's interest for the second component of μ is greater, in relative sense, than the intruder's interest in the second component of X_j , then the optimal strategy is to leave the second component of the original data unchanged (this means $l_2 = 0$) and contaminate the first component as much as necessary to obtain a disclosure risk smaller than the threshold for the maximum tolerable disclosure. Note that from (22) it follows that we can make the disclosure risk as small as we want by increasing either l_1 or l_2 .

We can interpret the difference

$$\Delta = \left| \frac{z_{2i}}{z_{1i}} - \frac{w_2}{w_1} \right| \quad (28)$$

as a measure of the discrepancy between the intruder's and the scientist's targets. In accord with intuition, the smaller this difference is the smaller is the set of strategies available to the agency (as the area between the solid line and the dashed line in Figure 2 increases as a function of Δ). The smaller Δ is, in fact, the stronger is the trade-off between disclosure risk and data utility (since the intruder's and the scientist's targets tend to coincide). For example, if both the intruder and the scientist, are interested in the first component of X_j and μ respectively, then in order to reduce the disclosure risk the agency is forced to contaminate the first component of the original observations. This also implies, however, a drastic drop in data utility. Note also that, as we expect, for a given value of Δ the data utility increases as the entries in the diagonal of scientist's loss function decrease (compare

the different rows in Figure 2).

The two examples that we presented show how the users' model can be used to implement existing optimality criteria. In both examples disclosure risk (data utility) is defined as the inverse of the variance of the posterior distribution for X_j (μ). The bigger is the variance the smaller is the uncertainty. The interpretation of these measures is very intuitive. The strength of our users' model, however, is that we are able to characterize these measures in terms of specific assumptions concerning the users' targets, the information available to the users before the release of the data (users' priors) and the estimation procedures that they use (user's loss functions). This characterization provide an objective criterion to evaluate to what extent the use of these measures is appropriate. If we believe that the set of assumptions in 5.1 and 5.2 (quadratic loss function, formalization of users' prior beliefs through normal distribution, scientist's target is μ , intruder's target is X_j , etc) are realistic, then the measures of disclosure and data utility that we have used are appropriate for the problem. Otherwise, we should modify the model's input accordingly and the model will automatically provide new and more realistic measures.

6. Conclusions

We have presented two new measures of disclosure risk and data utility and we showed how they can be used to implement the R-U confidentiality map proposed by Duncan, Keller-McNulty, and Stokes⁸. We argued that if a statistical agency has good knowledge of how the data will be used (the users' targets), the external information available to the users (the user's prior distributions), and the estimation procedure that will be used (intruder's and scientist's loss function), then any decreasing function of the intruder's (scientist's) uncertainty is a suitable measure of disclosure risk (data utility). We have shown that the choice of the decreasing function is irrelevant since both the optimality criterion and the ordering of alternative disclosure limitation techniques in terms of disclosure risk and data utility are the same whatever decreasing function is used. In Section 4, we illustrated how several measures of data utility currently used by statistical agencies and discussed in the literature are special cases of our framework. In this sense our approach generalizes the result of Duncan and Lambert¹ for disclosure risk.

We can generalize the users model proposed here in several ways. First, we can relax some of the assumptions underlying the model. In a realistic scenario a statistical agency has to deal with several users and it has only imperfect knowledge of users' targets, prior distributions, loss functions. Incorporating these uncertainties into the model is simple, at least in principle. Although we spoke throughout the paper of a single *intruder* and a single *scientist*, these labels can be used to represent groups of individuals/entities that are typical users of the products released by a statistical agency. Θ_I and Θ_S can represent multiple targets (as in example 2), and classes of intruder and scientist loss functions, prior distributions and models can be used instead of a single loss function, prior distribution or model. A more important generalization has to do with the definition of disclosure risk and data utility. Other authors and statistical agencies either implicitly or explicitly use quite different definitions of disclosure risk and data utility^{13,14,15,16,17}. Clearly, it would be helpful to have a unifying theoretical framework able to describe and compare

these different approaches.

Acknowledgements

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 the U.S. National Science Foundation under Grant EIA-9876619 to the National Institute of Statistical Sciences. The contents of the paper reflects the authors’ personal opinion. Neither the European Commission nor the National Science Foundation is responsible for any views or results presented. We thank Susie Bayarri for ideas and comments that are reflected in various ways in our work.

References

1. G.T. Duncan and D. Lambert, “Disclosure-limited Data Dissemination”, *Journal of the American Statistical Association*, **81** (1986) 10–18.
2. G.T. Duncan and R.W. Pearson, “Enhancing Access to Microdata While Protecting Confidentiality”, *Statistical Science*, **6** (1991) 219–239.
3. J.M. Mateo-Sanz and J. Domingo-Ferrer, “A method for data-orientated multivariate microaggregation”, *Statistical Data Protection*, Luxembourg: Office for Official Publication of the European Communities, pp. 89–99.
4. S.E. Fienberg, E.U. Makov and R.J. Steele, “Disclosure limitation using perturbation and related methods for categorical data (with discussion)”, *Journal of Official Statistics*, **14** (1998) 485–512.
5. M. Fischetti and J.J. Salazar-González, “Models and algorithms for cell suppression in tabular data with linear constraints”, *Journal of the American Statistical Association*, **95** (2000) 916–928.
6. P. Doyle, J. Lane, J. Theeuwes and L. Zayatz, *Confidentiality, Disclosure, and Data Access. Theory and Practical Applications for Statistical Agencies*, North-Holland, Amsterdam, 2001).
7. L. Willenborg and T. de Waal, *Elements of Statistical Disclosure Control*, Lectures Notes in Statistics, **155** (Springer Verlag, New York, 2001).
8. G.T. Duncan, S. Keller-McNulty and S.L. Stokes, “Disclosure risk vs. data utility: the R-U confidentiality map”, Technical Report LA-UR-01-6428., Statistical Sciences Group, Los Alamos, N.M.:Los Alamos National Laboratory, 2001.
9. M. Trottini, “A decision-theoretic approach to data disclosure problems”, *Research in Official Statistics*, **4**, 1 (2001) 7–22.
10. G.T. Duncan, S.E. Fienberg, R. Krishnan, R. Padman and S.F. Roehrig, “Disclosure limitation methods and information loss for tabular data”, *Confidentiality, Disclosure, and Data Access. Theory and Practical Applications for Statistical Agencies*, eds. P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (North-Holland, Amsterdam, 2001) pp. 135–166.
11. J. Domingo-Ferrer and V. Torra, “A quantitative comparison of disclosure control methods for microdata”, *Confidentiality, Disclosure, and Data Access. Theory and Practical Applications for Statistical Agencies*, eds. P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (North-Holland, Amsterdam, 2001a) pp. 111–134.
12. M.H. DeGroot, “Uncertainty, Information and Sequential experiments”, *Annals of Mathematical Statistics*, **3** (1962) 404–419.
13. U. Blien, H. Wirth and M. Muller, “Disclosure Risk for Microdata Stemming from

- Official Statistics”, *Statistica Neerlandica*, **46** (1992) 69–82.
14. D. Lambert, “Measures of disclosure risk and harm”, *Journal of Official Statistics*, **9** (1993) 313–333.
 15. C.J. Skinner and M.J. and Elliot, “A measure of disclosure risk for microdata”, *Journal of the Royal Statistical Society*, Series B, forthcoming.
 16. J. Domingo-Ferrer and V. Torra, “Disclosure control methods and information loss for microdata”, *Confidentiality, Disclosure, and Data Access. Theory and Practical Applications for Statistical Agencies*, eds. P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (North-Holland, Amsterdam, 2001b) pp. 91–110.
 17. J.M. Abowd and S.D. Woodcock, “Disclosure limitation in longitudinal linked data”, *Confidentiality, Disclosure, and Data Access. Theory and Practical Applications for Statistical Agencies*, eds. P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (North-Holland, Amsterdam, 2001) pp. 215–277.