Users' Uncertainty, Disclosure Risk and Data Utility

Mario TROTTINI
Universitat de València and Carnegie Mellon University [1]

Abstract: Duncan and Lambert (1986) have shown that a simple intruder's model can be used to characterize several measures of disclosure risk. In this paper we show that a similar model can be developed to define a general class of measures of data utility.

Key words and phrases: data disclosure limitation; disclosure risk; data utility; Bayesian decision theory.

1. INTRODUCTION

A major concern of statistical agencies is to protect the confidential information contained in the data sets that have to be released. Usually this is achieved applying to the original data *disclosure limitation techniques* that produce a modified data set that is finally released. The most common disclosure limitation techniques are very simple, for example they consist of deleting observations, or aggregating observations or adding noise to the original values, others are more sophisticated and make use of simulation methods (see Duncan and Pearson 1991 for a review). The definition of suitable disclosure limitation techniques, for a given problem, is not trivial since protection of confidentiality should be achieved in a way that is compatible with the agency's mission of providing society with good quality data. The modification of the original data using a disclosure limitation technique not only reduces the risk of disclosure, but it also reduces the information content (and therefore the utility for the users) of the original data. The goal of the agency is to find an optimal balance between the needs of the data providers (protection of confidentiality) and the needs of the data users (good quality data).

The major efforts in statistical confidentiality to date have been devoted to assessment of the risks for the data providers. In particular the approach proposed by Duncan and Lambert (1986) provides a decision-theoretic framework to measure the extent of disclosure associated with different disclosure limitation techniques, that contains as special case most of the measures of disclosure currently used by statistical agencies. Very little has been done, however, to measure to what extent the existing disclosure limitation techniques meet the needs of the data users. In this paper we try to fill in this gap. The measures that we propose to quantify the impact of different disclosure techniques on the inferences of the data users are a natural extension of the measures of disclosure proposed by Duncan and Lambert (1986). The approach that we propose is innovative because (i) is very general; (ii) has a theoretical justification. Although excellent, existing solutions are problem specific, do not have theoretical justification, do not allow to compare two generic

[1] *ADDRESS FOR CORRESPONDENCE: Mario Trottini, Departamento de Estadística e Investigación Operativa, C/ Dr. Moliner, 50 46100, Burjassot, València, España, Tel. +34 -963864362, Fax: +34-963864735, e-mail: mario.trottini@uv.es*

disclosure limitation techniques (see for example Mateo-Sanz and Domingo-Ferrer 1998, Baeyens and Defays 1998, and Hurkens and Tiourine 1998). Willenborg and De Waal (2001) propose to assess the impact of a disclosure limitation techniques on the inferences of data users, comparing the entropy of the original data with the entropy of the released data. The method is very general and simple to implement. However it ignores what is the final use of the data and it does not offer a model that describes the society's behavior when data are released. It seems difficult to measure the extent to which the released data are useful for society without specifying what is the final use of the data and how society uses the released data to achieve its target.

In section 2 and 3 we formalize the problem and we introduce the basic assumptions and notation. In Section 4 we review the work of Duncan and Lambert (1986) that inspired the main ideas of this work. Section 5 presents the new measures of data utility. An optimality criterion for the choice of the best form of data release is proposed in section 6 and illustrated with several examples.

2. PROBLEM DEFINITION

A statistical office produces, as a results of its activities, a data set D_0. In order to take into account confidentiality issues and the needs of the users of the data, different forms of data release are considered. The statistical office can release the original data D_0, decide not to release any data, or decide to apply disclosure limitation techniques and release modified data or synthetic data. We denote by $\mathcal{D}$ the set of alternative forms of data release, and by D_i the generic element in $\mathcal{D}$, $i = 0, \ldots, h$. In our disclosure scenario it is assumed that the set of users can be partitioned in two groups. Those who want to use the released data to perform statistical studies or for research purposes, and those who want to use the released data to disclose confidential information about the data provider. We refer to the first group as *society* and to the second group as the *intruder*. More formally, it is assumed that society is interested in a unknown quantity Θ_{SOC} while the intruder is interested in an unknown quantity Θ_{INT} and they try to infer their target values, Θ_{INT} and Θ_{SOC}, based on the information contained in the released data. We assume that the set $\mathcal{D}$ of alternative forms of data release is given and the agency has to decide which data set in $\mathcal{D}$ is best to release. Ideally the statistical agency should provide maximum information about Θ_{SOC} and minimum information about Θ_{INT}. However in general this is not feasible since there is a trade-off between the two types of information and it is not possible to increase one without also increasing the other. Optimality criteria are needed.

3. NOTATIONS AND ASSUMPTIONS

In order for the problem to be meaningful, we assume that both Θ_{INT} and Θ_{SOC} are somehow related with the original data D_0. Different hypotheses about the distribution of D_0 can me made. In this paper we assume the following:

- **ASSUMPTION** A_1: There is not uncertainty about the mechanism generating the data D_0. In particular:

  A_{1a} : Prior to observing D_0, intruder and society agree that D_0 is a realization of a random variable V whose distribution, P, belongs to a parametric family $\mathcal{P}$ with parameter Ψ and parameter space Ω: $V \sim P_\Psi$, $\psi \in \Omega$;

  A_{1b} : The model specified by the intruder and society is correct.

Assumption A_1 is certainly not reflective of real-world settings and we have made it largely for convenience. Elsewhere we plan to relax this assumption to allow disagreement between intruder and society, within intruders, within individuals in the society or both and to allow the "true" model to be outside of the range of models specified by intruder and society.

The conditional distributions of V given Θ_{INT} and the conditional distributions of V given Θ_{SOC} formalize how the data set D_0 is related to intruder's and society's targets, Θ_{INT} and Θ_{SOC}. We also assume that intruder's and society's prior uncertainty about Ψ, Θ_{INT}, and Θ_{SOC}, can be adequately expressed by probability distributions, $\pi(\cdot)$, $\pi_{INT}(\cdot)$, $\pi_{SOC}(\cdot)$, that we will refer to as *prior distributions* for Ψ, Θ_{INT}, and Θ_{SOC} respectively. The conditional distribution of V given Ψ, the prior distribution for Ψ and the particular disclosure limitation technique used, induce, for each form of data release, D_i, conditional distributions of D_i given Θ_{INT} and D_i given Θ_{SOC} that formalize how the revised data set D_i is related to intruder's and society's targets, Θ_{INT} and Θ_{SOC}. We denote by $\pi_{INT}(\cdot|D_i)$ the conditional distribution of Θ_{INT} given D_i and by $\pi_{SOC}(\cdot|D_i)$ the conditional distribution of Θ_{SOC} given D_i. These are the *posterior distributions* for Θ_{INT} and Θ_{SOC} given D_i and express intruder's and society's uncertainty about their targets, after the data D_i have been released.

4. DUNCAN AND LAMBERT MEASURES OF DISCLOSURE

The choice of the best form of data release requires a measure of risk of disclosure and information content. Among the attempts to measure the extent of disclosure, the most successful so far seems the one proposed by Duncan and Lambert (1986). The approach is simple as it is powerful. Let $L_{INT}(\cdot;\cdot)$ be the loss function for the intruder's decision problem "identify Θ_{INT}". $L_{INT}(a;\theta_{INT})$ is the loss that the intruder pays for an estimate a of Θ_{INT} when $\Theta_{INT}=\theta_{INT}$. Denote by $p(\cdot)$ the probability distribution that formalizes the intruder's beliefs about the target Θ_{INT}. Then the intruder's uncertainty about the true value of Θ_{INT}, can be expressed as the expected loss associated with the optimal estimate of Θ_{INT} with respect to the loss function $L_{INT}(\cdot;\cdot)$ and the distribution $p(\cdot)$, i.e. if the intruder's distribution over Θ_{INT} is $p(\cdot)$, then the intruder's uncertainty about the true value of Θ_{INT}, $U_{INT}(p)$, is given by:

$$\begin{aligned} U_{INT}(p) &= inf_a \int L_{INT}(a;\theta_{INT})p(\theta_{INT})d\theta_{INT} \\ &= \int L_{INT}(\hat{a};\theta_{INT})p(\theta_{INT})d\theta_{INT}, \end{aligned}$$

where:

$$\hat{a} = argmin_a \int L_{INT}(a;\theta_{INT})p(\theta_{INT})d\theta_{INT}.$$

Based on the assumption that the bigger is the intruder's uncertainty after the release of the data, the smaller is the extent of disclosure, the different measures of disclosure are defined as decreasing function of intruder's uncertainty associated with the intruder posterior distribution. In particular the authors define three classes of measure of disclosure, (intruder's) *knowledge*, (intruder's) *knowledge gain*, and (intruder's) *relative knowledge gain*. The posterior knowledge measure of disclosure is obtained applying the uncertainty function $U_{INT}(\cdot)$ to the intruder's posterior distribution for Θ_{INT}. This is the right quantity to consider if the final goal is to protect the intruder's knowledge about the target. However, if the intruder already has precise information about the target value before the release of the data, the knowledge gain, i.e. the difference U_{INT}(prior)-U_{INT}(posterior), or the relative knowledge gain, (U_{INT}(prior)-U_{INT}(posterior))/U_{INT}(prior), might be more appropriate measures of disclosure. The appealing feature of the approach proposed by Duncan and Lambert is that it offers a unifying framework for most of the existing measures of the extent of disclosure. Many measures of disclosure currently used by statistical agencies, in fact, can be obtained as a special case of the general framework described above for suitable choices of the intruder's loss function. Suppose, for example, that the released data consists of a table with k categories, possibly cross-classified. The ad-hoc rules of bounding tabular relative frequencies away from zero and away from one, discussed by the Subcommittee on Disclosure Avoidance Techniques

(1978), and used by many statistical agencies, can be obtained as special case of Duncan and Lambert approach assuming that the intruder's target is to disclose to which of the k categories a particular respondent, say A, belongs and assuming that the intruder's loss, when he incorrectly specifies that A belongs to category i, is a decreasing function of the posterior probability that A belong to that category (many others interesting examples are discussed in Duncan and Lambert (1986)). In this paper we will use as measure of disclosure the inverse of the intruder uncertainty. In particular we define the *disclosure risk* when the intruder's distribution over Θ_{INT} is $p(\cdot)$ as:

$$RISK(p) = \frac{1}{U_{INT}(p)}. \tag{1}$$

The choice of (1) as measure of disclosure, is made only for convenience. The intruder's posterior knowledge is a negative quantity and we prefer to work with a measure of disclosure that take positive values. For the purpose of the paper, however, any decreasing function of the intruder's uncertainty would work just fine. We plan to address somewhere else the problem of which decreasing function is best to use.

5. NEW MEASURES OF DATA UTILITY

The measures of disclosures proposed by Duncan and Lambert have a natural extension as measures of information content. In particular, let $L_{SOC}(\cdot;\cdot)$ be the loss function for society's decision problem "identify Θ_{SOC}". $L_{SOC}(a;\theta_{SOC})$ is the loss that the society pays for an estimate a of Θ_{SOC} when $\Theta_{SOC} = \theta_{SOC}$. If society's distribution over Θ_{SOC} is $q(\cdot)$, then society's uncertainty about the true value of Θ_{SOC}, $U_{SOC}(q)$, can be expressed as:

$$U_{SOC}(q) \quad = \quad \int L_{SOC}(\hat{a};\theta_{SOC})q(\theta_{SOC})d\theta_{SOC};$$

where:

$$\hat{a} = argmin_a \int L_{SOC}(a;\theta_{SOC})q(\theta_{SOC})d\theta_{SOC}.$$

Based on the assumption that the bigger is society's uncertainty after the release of the data, the less informative are the data, measures of information content can be defined as decreasing function of society's uncertainty associated with society's posterior distribution for Θ_{SOC}. Following Duncan and Lambert (1986) we can define three classes of measure of information content, *society's knowledge*, *society's knowledge gain*, and *society's relative knowledge gain* whose definition and interpretation mirror the ones given for the corresponding measures of disclosure. In particular we have:

society's posterior knowledge $= -U_{SOC}(\pi_{SOC}(\cdot|D_i))$;

society's knowledge gain $= U_{SOC}(\pi_{SOC}(\cdot)) - U_{SOC}(\pi_{SOC}(\cdot|D_i))$;

society's relative knowledge gain $= \frac{U_{SOC}(\pi_{SOC}(\cdot)) - U_{SOC}(\pi_{SOC}(\cdot|D_i))}{U_{SOC}(\pi_{SOC}(\cdot))}$.

In the rest of the paper we use as measure of information content the inverse of society's uncertainty. In particular we define the *data utility* when society's distribution over Θ_{SOC} is $q(\cdot)$ as:

$$UTILITY(q) = \frac{1}{U_{SOC}(q)}. \tag{2}$$

Again the choice of (2) as measure of information content is made only for convenience. As for the measures of disclosure, any decreasing function of society's uncertainty would work just fine. We do not address in this work which decreasing function is best to use.

6. OPTIMALITY CRITERION

Using the measures of disclosure risk and data utility described in the previous sections we can define an optimality criterion for the choice of the best form of data release. We propose the following:

Optimality Criterion: Consider a given disclosure scenario, i.e. an intruder's target, a society's target, their prior distributions, their loss functions, a set $\mathcal{D}$ of alternative forms of data release. Consider a threshold value t, for the maximum tolerable extent of disclosure. Denote by $\mathcal{D}_S$ the subset of those forms of data release in $\mathcal{D}$ with disclosure risk below t. If $\mathcal{D}_S$ is not empty, release the form of data release in $\mathcal{D}_S$ with maximum data utility. If $\mathcal{D}_S$ is empty do not release any data.

Our optimality criterion is an implementation of the criterion proposed by Duncan and Keller Mc-Nulty (1999). The main difference is that in Duncan and Keller Mc-Nulty the data utility is measured in terms of the MSE of the estimator of society's target. The emphasis of the work of Duncan and Keller Mc-Nulty, however, is on the representation of the trade-off of gain versus risk rather than on the representation of the data utility. The choice of the MSE is made only for convenience.

We now apply the optimality criterion described above to three examples. In the first example it is assumed that the original data set consists of realizations of i.i.d. univariate random variables with normal distribution. The intruder is interested in the value of the first observation in the original sample, society is interested in the mean of the i.i.d. random variables. In the second example we extend the results in the first example to the multivariate case. In the last example we apply the optimality criterion to a simple regression problem. Society's target is the vector of parameters of the linear model, different intruder's targets are considered. In all examples a quadratic loss function is used.

6.1 Example 1: Duncan and Keller-McNulty 1999

Suppose that a statistical office has to decide about the release of a data set, $\underline{x}_n = \{x_1, x_2, \ldots, x_n\}$, $x_i \in \mathcal{R}$, $i = 1, \ldots, n$. Before the data are released neither the intruder nor society know how these n values might be related to each other, or whether any of these values have features that distinguish them from the rest. In particular, prior to observing $\underline{x}_n$, both intruder and society assume that $x_1, x_2, \ldots, x_n$ are realizations of i.i.d. random variables from a normal distribution with unknown mean μ and known variance σ^2, i.e.

$$X_1, X_2, \ldots, X_n \sim_{i.i.d.} N(\mu, \sigma^2), \;\; \sigma^2 \text{ known.}$$

Suppose that:

- society's target, Θ_{SOC}, is to identify μ, $\Theta_{SOC} = \mu$;
- the intruder's target, Θ_{INT}, is to identify X_1, $\Theta_{INT} = X_1$;
- Both society's and intruder's prior uncertainty about μ can be formalized by a normal distribution with known mean θ and known variance τ^2,

$$\mu \sim N(\theta, \tau^2), \;\; \theta, \, \tau^2 \text{ known.}$$

Four forms of data release are considered:

1. release the average of the original sample: $\bar{x}_n = \frac{\sum_{i=1}^n x_i}{n}$;

2. release the average of the subsample of the original data: $\bar{x}_{n-1} = \frac{\sum_{i=2}^{n} x_i}{n-1}$;

3. release the average of the masked data: $\bar{y}_n(l) = \frac{\sum_{i=1}^{n} y_i}{n}$;
   where: $\underline{y}_n = (y_1(l), y_2(l), \ldots, y_n(l))$;

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

4. release the average of the subsample of the masked data: $\bar{y}_{n-1}(l) = \frac{\sum_{i=2}^{n} y_i}{n-1}$.

We want to follow the approach described in section 6 to choose the "best" form of data release among the four considered. A quadratic loss function is used. Note that we make the implicit assumption that if any type of masked data is released, also the value l, for the contamination used, is released.

6.1.1 Quadratic loss function, univariate case

Consider the two decision problems:

1. "identify $\Theta_{SOC} = \mu$",

2. "identify $\Theta_{INT} = X_1$",

with actions spaces $\mathcal{N}_1 = \mathcal{N}_2 = \mathcal{R}$. Assume that both the intruder and society use a quadratic loss function to quantify the loss that they pay for an estimate a of the target values Θ_{INT} and Θ_{SOC}, i.e.

$$L_{INT}(a; x_1) = (a - x_1)^2\,, \quad L_{SOC}(a; \mu) = (a - \mu)^2.$$

Denote by $p(\cdot)$ the probability distribution function that formalizes the intruder's beliefs about X_1. Under quadratic loss function the optimal estimate of X_1 is the posterior mean of X_1 under $p(\cdot)$ and the intruder's uncertainty about X_1 is the variance of X_1 under $p(\cdot)$, i.e

$$\begin{aligned} U_{INT}(p) &= inf_a \int (a - x_1)^2 p(x_1) dx_1 \\ &= \int (x_1 - \hat{X}_1)^2 p(x_1) dx_1 = Var_p(X_1), \end{aligned}$$

where:

$$\hat{X}_1 = argmin_a \int L_{INT}(a; x_1) p(x_1) dx_1 = E_p(X_1),$$

and $E_p(X_1)$ and $Var_p(X_1)$ denote the expected value and the variance of X_1 under $p(\cdot)$.

Similarly if we denote by $q(\cdot)$ the probability distribution function that formalizes society's beliefs about μ, society's optimal estimate of μ is the expected value of μ under $q(\cdot)$ and society's uncertainty about the true value of μ is the variance of μ under $q(\cdot)$, i.e

$$\begin{aligned} U_{SOC}(q) &= inf_a \int (\mu - a)^2 q(\mu) d\mu \\ &= \int (\mu - \hat{\mu})^2 q(\mu) d\mu = Var_q(\mu), \end{aligned}$$

where:

$$\hat{\mu} = argmin_a \int L_{SOC}(a; \mu) q(\mu) d\mu = E_q(\mu),$$

and $E_q(\mu)$ and $Var_q(\mu)$ denote the expected value and the variance of μ under $q(\cdot)$.

The risk of disclosure when the intruder's distribution over X_1 is $p(\cdot)$ is therefore:

$$RISK(p) = \frac{1}{U_{INT}(p)} = \frac{1}{Var_p(X_1)}, \tag{3}$$

and similarly, the data utility when society's distribution over μ is $q(\cdot)$ is:

$$UTILITY(q) = \frac{1}{U_{SOC}(q)} = \frac{1}{Var_q(\mu)}. \tag{4}$$

Table 1 shows intruder's and society's posterior distributions for X_1 and μ for each of the four forms of data release considered .

Data	$\pi(\mu \vert Data)$	$p(x_1 \vert Data)$
No data	$N\left(\theta, \tau^2\right)$	$N\left(\theta, \sigma^2+\tau^2\right)$
$\bar{x}_n$	$N\left(\frac{(\frac{\theta}{\tau^2}+\frac{n\cdot\bar{x}_n}{\sigma^2})}{(\frac{1}{\tau^2}+\frac{n}{\sigma^2})}, (\frac{1}{\tau^2}+\frac{n}{\sigma^2})^{-1}\right)$	$N\left(\bar{x}_n, \frac{n-1}{n}\cdot\sigma^2\right)$
$\bar{x}_{n-1}$	$N\left(\frac{(\frac{\theta}{\tau^2}+\frac{(n-1)\cdot\bar{x}_{n-1}}{\sigma^2})}{(\frac{1}{\tau^2}+\frac{n-1}{\sigma^2})}, (\frac{1}{\tau^2}+\frac{n-1}{\sigma^2})^{-1}\right)$	$N\left(\frac{(\frac{\theta}{\tau^2}+\frac{(n-1)\cdot\bar{x}_{n-1}}{\sigma^2})}{(\frac{1}{\tau^2}+\frac{n-1}{\sigma^2})}, (\frac{1}{\tau^2}+\frac{n-1}{\sigma^2})^{-1}+\sigma^2\right)$
$\bar{y}_n(l))$	$N\left(\frac{(\frac{\theta}{\tau^2}+\frac{n\cdot\bar{y}_n(l)}{\sigma^2+l^2})}{(\frac{1}{\tau^2}+\frac{n}{\sigma^2+l^2})}, (\frac{1}{\tau^2}+\frac{n}{\sigma^2+l^2})^{-1}\right)$	$N\left(\frac{\sigma^2\bar{y}_n(l)}{\sigma^2+l^2}+\frac{l^2}{\sigma^2+l^2}(\frac{\frac{\theta}{\tau^2}+\frac{n\bar{y}_n}{\sigma^2+l^2}}{\frac{1}{\tau^2}+\frac{n}{\sigma^2+l^2}}), \frac{l^4\tau^2}{(\sigma^2+l^2)(\sigma^2+l^2+n\tau^2)}+\frac{(n-1)\sigma^4+n\sigma^2l^2}{n(\sigma^2+l^2)}\right)$
$\bar{y}_{n-1}$	$N\left(\frac{(\frac{\theta}{\tau^2}+\frac{(n-1)\cdot\bar{y}_{n-1}}{\sigma^2+l^2})}{(\frac{1}{\tau^2}+\frac{n-1}{\sigma^2+l^2})}, (\frac{1}{\tau^2}+\frac{n-1}{\sigma^2+l^2})^{-1}\right)$	$N\left(\frac{(\frac{\theta}{\tau^2}+\frac{(n-1)\cdot\bar{y}_{n-1}}{\sigma^2+l^2})}{(\frac{1}{\tau^2}+\frac{n-1}{\sigma^2+l^2})}, (\frac{1}{\tau^2}+\frac{n-1}{\sigma^2+l^2})^{-1}+\sigma^2\right)$

Table 1: Prior and posterior distributions for X_1 and μ

Using (3) and (4) and the distributions in table 1 we can evaluate the disclosure risk and the data utility associated to each of the forms of data release considered and prior to the release of any data. For a fixed threshold t the form of data release with disclosure risk below t and maximum data utility will be selected.

Table 2 shows the data utility for each of the four forms of data release that we are considering and prior to the release of any data.

$Data$	$U_{SOC}(Data)$	$UTILITY(Data)$
No data	τ^2	$\frac{1}{\tau^2}$
$\bar{x}_n$	$(\frac{1}{\tau^2} + \frac{n}{\sigma^2})^{-1}$	$(\frac{1}{\tau^2} + \frac{n}{\sigma^2})$
$\bar{x}_{n-1}$	$(\frac{1}{\tau^2} + \frac{n-1}{\sigma^2})^{-1}$	$(\frac{1}{\tau^2} + \frac{n-1}{\sigma^2})$
$\bar{y}_n(l)$	$(\frac{1}{\tau^2} + \frac{n}{\sigma^2+l^2})^{-1}$	$(\frac{1}{\tau^2} + \frac{n}{\sigma^2+l^2})$
$\bar{y}_{n-1}$	$(\frac{1}{\tau^2} + \frac{n-1}{\sigma^2+l^2})^{-1}$	$(\frac{1}{\tau^2} + \frac{n-1}{\sigma^2+l^2})$

Table 2: Data utility

Similarly table 3 shows the disclosure risk for each of the four forms of data release that we are considering and prior to the release of any data.

$Data$	$U_{INT}(Data)$	$RISK(Data)$
No data	$\sigma^2 + \tau^2$	$\frac{1}{\sigma^2+\tau^2}$
$\bar{x}_n$	$\frac{n-1}{n} \cdot \sigma^2$	$\frac{n}{(n-1)\cdot\sigma^2}$
$\bar{x}_{n-1}$	$(\frac{1}{\tau^2} + \frac{n-1}{\sigma^2})^{-1} + \sigma^2$	$[(\frac{1}{\tau^2} + \frac{n-1}{\sigma^2})^{-1} + \sigma^2]^{-1}$
$\bar{y}_n(l)$	$\frac{l^4\tau^2}{(\sigma^2+l^2)(\sigma^2+l^2+n\tau^2)} + \frac{(n-1)\sigma^4+n\sigma^2 l^2}{n(\sigma^2+l^2)}$	$[\frac{l^4\tau^2}{(\sigma^2+l^2)(\sigma^2+l^2+n\tau^2)} + \frac{(n-1)\sigma^4+n\sigma^2 l^2}{n(\sigma^2+l^2)}]^{-1}$
$\bar{y}_{n-1}$	$(\frac{1}{\tau^2} + \frac{n-1}{\sigma^2+l^2})^{-1} + \sigma^2$	$[(\frac{1}{\tau^2} + \frac{n-1}{\sigma^2+l^2})^{-1} + \sigma^2]^{-1}$

Table 3: Disclosure risk

In figure 1 we plot the values of "Data Utility" vs. "Disclosure Risk" for the four forms of data release and for the case in which no data are released. We used $n = 10$, $\sigma^2 = 2$, $\tau^2 = 5$.
The red line represents the pairs ("Data utility", "Disclosure risk") when the average of the masked data, $\bar{y}_n(l)$, is released and the contamination parameter l varies in [0,40]. The blue line, instead, represents the pairs ("Data Utility", "Disclosure Risk") when the average of the subsample of the masked data, $\bar{y}_{n-1}(l)$, is released and the contamination parameter l varies in [0,40].

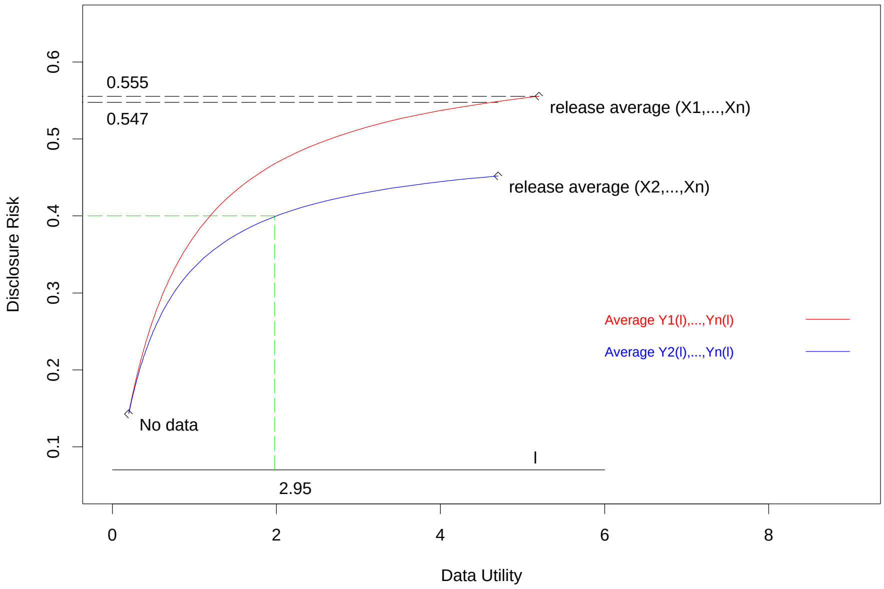

Figure 1: *Original vs. masked data, quadratic loss function,* $n = 10$, $\tau^2 = 5$, $\sigma^2 = 2$

As we expect for large l the release of $\bar{y}_n(l)$ and $\bar{y}_{n-1}(l)$ is equivalent to not release any data, while for l close to zero the release of the averages of the masked data, $\bar{y}_n(l)$ and $\bar{y}_{n-1}(l)$, is equivalent to the release of the averages of the original data $\bar{x}_n$ and $\bar{x}_{n-1}$ respectively. As shown in figure 1 the optimal form of data release will depend on the value of the threshold t. If, for example, $t = 0.4$, the statistical agency should release the average of the masked data using a contamination $l = 2.95$. This would produce a masked data set with disclosure risk 0.399 and data utility equal to 1.98.

Here's the optimal form of data release as a function of the threshold t:

- if $t \leq 0.547$ release the average of the subsample of the masked, data, $\bar{y}_{n-1}(l)$, using the value of the parameter l for which the risk of disclosure is just below t;
- if $0.547 < t \leq 0.555$ release the average of the masked data, $\bar{y}_n(l)$, using the value of the parameter l for which the risk of disclosure is just below t;
- if $t > 0.555$ release the average of the original data $\bar{x}_n$.

Note that while for $t > 0.547$ the decision of the best form of data release depends on the optimality criterion that we are using, for $t \leq 0.547$ the release of the average of $\bar{y}_{n-1}(l)$ dominates the release of the average of $\bar{y}_n(l)$ no matter which optimality criterion we use.

6.1.2 Calibration/Choice of the threshold

Using the notation introduced in the previous section, let $p(x_1|D)$ be the intruder's distribution over X_1 when a data set D is released. We showed in section 6.1.1 that under quadratic loss function the risk of disclosure associated to a particular form of data release, D_i, is the inverse of the variance of X_1 under $p(x_1|D_i)$. We also showed in the previous section that for all forms of

data release considered in this example the posterior distribution of the intruder target, X_1 given the released data, is normal. As a result the risk of disclosure is nothing but that the precision of the posterior mean of X_1 as an estimate of X_1 and the choice of the threshold can be made using heuristic considerations about what is the maximum precision that can be tolerated.
For example we could relate the threshold for the disclosure risk with the length of the 95% credibility interval for X_1. We have:

$$\text{length of the 95\% credibility interval for } X_1 > 4/(\text{threshold})^{1/2}.$$

Figure 2 shows how the length of the 95% credibility interval for X_1 varies as a function of the threshold for the disclosure risk. If we want the length of the 95% credibility interval for X_1 to be at least 6.5, then for the disclosure risk we should use a threshold $t = 0.375$.

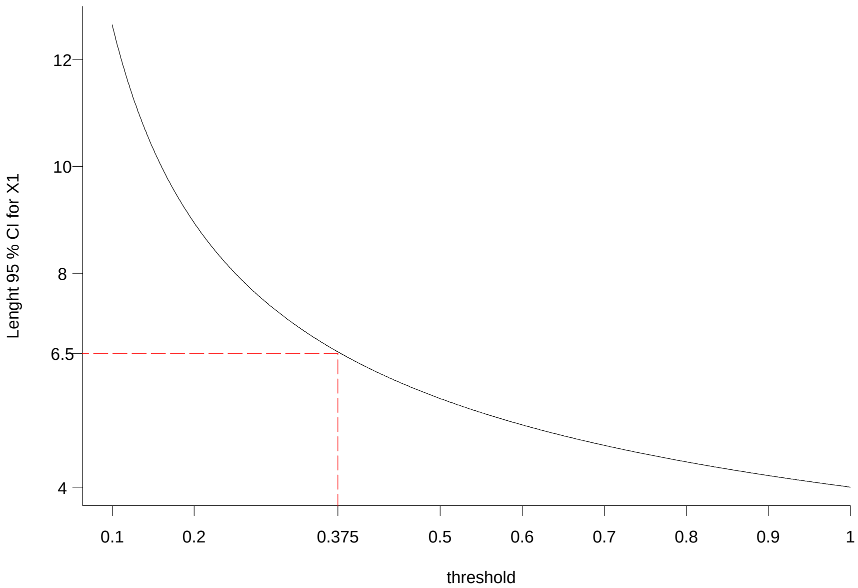

Figure 2: *Threshold vs. length of the* 95% *credibility interval for* X_1

6.2 Example 2: Multivariate case

A statistical office has to decide about the release of a data set, $\underline{x}_n = \{x_1, x_2, \ldots, x_n\}$, $x_i \in R^k$, $k \geq 2$, $i = 1, \ldots, n$. Before the data are released neither the intruder nor society know how these n vectors might be related to each other, or whether any of these vectors have features that distinguish them from the rest. In particular, prior to observing $\underline{x}_n$, both the intruder and society assume that $x_1, x_2, \ldots, x_n$ are realizations of i.i.d. random variables with multivariate normal distribution with unknown vector mean μ and known covariance matrix Σ_1, i.e.

$$X_1, X_2, \ldots, X_n \sim MN(\mu, \Sigma_1), \;\; \Sigma_1 \text{ known.}$$

Suppose that:

- society's target, Θ_{SOC}, is to identify μ, $\Theta_{SOC} = \mu$;

- the intruder's target, Θ_{INT}, is to identify X_1, $\Theta_{INT} = X_1$;
- both society's and 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,

$$\mu \sim MN(\theta, \Sigma_0), \;\; \theta, \, \Sigma_0 \text{ known.}$$

Two forms of data release are considered:

1. release the average of the original sample: $\bar{x}_n = (\frac{\sum_{i=1}^n x_i^1}{n}, \ldots, \frac{\sum_{i=1}^n x_i^k}{n})$, where x_i^j is the j^{th} component of x_i, $i = 1, \ldots, n$, $j = 1, \ldots, k$;
2. release the average of the masked data: $\bar{y}_n(l) = (\frac{\sum_{i=1}^n y_i^1}{n}, \ldots, \frac{\sum_{i=1}^n y_i^k}{n})$, where: $\underline{y}_n = y_1, y_2, \ldots, y_n$,

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

The goal is to choose the "best" form of data release among the two considered. Again a quadratic loss function is used. As in example 1 we make the implicit assumption that if any type of masked data is released, also the value l, for the contamination used, is released.

6.2.1 Quadratic loss function, multivariate case

Consider the two decision problems:

1. "identify $\Theta_{SOC} = \mu$",
2. "identify $\Theta_{INT} = X_1$",

with actions spaces $\mathcal{N}_1 = \mathcal{N}_2 = \mathcal{R}^k$. Assume that both the intruder and society use a quadratic loss function to quantify the loss that they pay for estimate a of their target values Θ_{INT} and Θ_{SOC}, i.e.

$$L_{INT}(a; x_1) = (a - x_1)^T W (a - x_1), \quad L_{SOC}(a; \mu) = (a - \mu)^T M (a - \mu)$$

where:

$$W = \begin{pmatrix} w_1 & 0 & \ldots & 0 \\ 0 & w_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & w_k \end{pmatrix}, \; M = \begin{pmatrix} m_1 & 0 & \ldots & 0 \\ 0 & m_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & m_k \end{pmatrix}, \; w_i > 0, \;\; , m_i > 0, \; i = 1, \ldots, k.$$

The relative magnitude of w_i indicates how important is for the intruder the estimate of the i^{th} component of X_1 compared to the others components. Similarly the relative magnitude of m_i indicates how important is for society the estimate of the i^{th} component of μ compared to the others components (note that in the more general case W and M are $(k \times k)$ positive definite matrices and do not need to be diagonal). Denote by $p(\cdot)$ the probability distribution function that formalizes the complete uncertainty of the intruder about X_1. Under quadratic loss function the optimal estimate of X_1 is the vector of the posterior mean of X_1 under $p(\cdot)$ and the intruder's uncertainty about X_1 is the weighted sum of the variance of the components of X_1 under $p(\cdot)$, with weights given by the diagonal of W,

$$\begin{aligned} U_{INT}(p) &= inf_a \int (a-x_1)^T W(a-x_1)p(x_1)dx_1 \\ &= \int (\hat{X}_1 - x_1)^T W(\hat{X}_1 - x_1)p(x_1)dx_1 = \sum_{i=1}^{k} w_i Var_p(X_1^{(i)}), \end{aligned}$$

where:

$$\hat{X}_1 = argmin_a \int L_{INT}(a; x_1)p(x_1)dx_1 = E_p(X_1),$$

and $E_p(X_1)$ and $Var_p(X_1^{(i)})$ denote the expected value of X_1 and the variance of the i^{th} component of X_1 under $p(\cdot)$.

Similarly if we denote by $q(\cdot)$ the probability distribution function that formalizes the complete uncertainty of society about μ, society's optimal estimate of μ is the expected value of μ under $q(\cdot)$ and society's uncertainty about the true value of μ is the sum of the variances of the components of μ under $q(\cdot)$ with weights given by the diagonal of M, i.e.

$$\begin{aligned} U_{SOC}(q) &= inf_a \int (\mu - a)^T M(a-\mu)q(\mu)d\mu \\ &= \int (\hat{\mu} - \mu)^T M(\hat{\mu} - \mu)q(\mu)d\mu = \sum_{i=1}^{k} m_i Var_q(\mu^{(i)}), \end{aligned}$$

where:

$$\hat{\mu} = argmin_a \int L_{SOC}(a; \mu)q(\mu)d\mu = E_q(\mu),$$

and $E_q(\mu)$ and $Var_q(\mu^{(i)})$ denote the expected value of μ and the variance of the i^{th} component of μ under $q(\cdot)$.

The risk of disclosure when the intruder's distribution over X_1 is $p(\cdot)$ is therefore:

$$RISK(p) = \frac{1}{U_{INT}(p)} = \frac{1}{\sum_{i=1}^{k} w_i Var_p(X_1^{(i)})}, \tag{5}$$

and similarly, the data utility when society's distribution over μ is $q(\cdot)$ is:

$$UTILITY(q) = \frac{1}{U_{SOC}(q)} = \frac{1}{\sum_{i=1}^{k} m_i Var_q(\mu^{(i)})}. \tag{6}$$

Table 4 and table 5 show the intruder's distribution over X_1 and society's distribution over μ for each of the two forms of data release considered and prior to release of any data. Using (5) and (6) and the distributions in tables 4 and 5 we can evaluate the risk of disclosure and the information content associated to each of the forms of data release considered and prior to the release of any data. For a fixed threshold t we will select the form of data release with disclosure risk below t and maximum information content.

Data	$\pi(\mu \mid Data)$
No data	$MN\left(\theta, \Sigma_0\right)$
$\bar{x}_n$	$MN\left((\Sigma_0^{-1} + n\Sigma_1^{-1})^{-1}(\Sigma_0^{-1}\theta + n\Sigma_1^{-1}\bar{x}_n), (\Sigma_0^{-1} + n\Sigma_1^{-1})^{-1}\right)$
$\bar{y}_n(l)$	$MN\left((\Sigma_0^{-1} + n(\Sigma_1 + \Lambda)^{-1})^{-1}(\Sigma_0^{-1}\theta + n(\Sigma_1 + \Lambda)^{-1}\bar{y}_n(l)), (\Sigma_0^{-1} + n(\Sigma_1 + \Lambda)^{-1})^{-1}\right)$

Table 4: Prior and posterior distributions for μ

Data	$p(x_1 \mid Data)$
No data	$MN\left(\theta, \Sigma_0 + \Sigma_1\right)$
$\bar{x}_n$	$MN\left(\bar{x}_n, \frac{n-1}{n}\Sigma_1\right)$
$\bar{y}_n(l)$	$MN\left(A\theta + \Sigma_1(\Sigma_1 + \Lambda)^{-1}\bar{y}_n(l), \frac{n-1}{n}\Sigma_1 + \frac{1}{n}(\Sigma_1^{-1} + \Lambda^{-1})^{-1} + A(\Sigma_0^{-1} + n(\Sigma_1 + \Lambda)^{-1})^{-1}A^T\right)$ $(A = I - \Sigma_1(\Sigma_1 + \Lambda)^{-1})$

Table 5: Prior and posterior distributions for X_1

6.2.2 Example

Suppose that $k = 2$, $n = 20$ and

$$\Sigma_0 = \begin{pmatrix} 5 & 0.5 \\ 0.5 & 4 \end{pmatrix}, \ \Sigma_1 = \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1.5 \end{pmatrix}, \ W = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}, \ M = \begin{pmatrix} 5 & 0 \\ 0 & 2 \end{pmatrix}.$$

Note that $w_1 < w_2$ while $m_1 > m_2$, i.e. the intruder is more interested in the second component of X_1 while society is more interested in the first component of μ.
In figure Figure 3 we plot the pairs ("Data utility", "Disclosure risk") when the average of the original data is released, when no data are released and when the average of the masked data is released. The magenta line represents the pairs ("Data utility", "Disclosure risk") 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 $(0, 1000)$. The blue line represents the pairs ("Data utility", "Disclosure risk") when $l_1 = 0$ and l_2 varies in $(0, 1000)$. The others colored lines represent the pairs ("Data utility", "Disclosure risk") when the value of l_1 is held constant and l_2 varies in $(0, 1000)$.

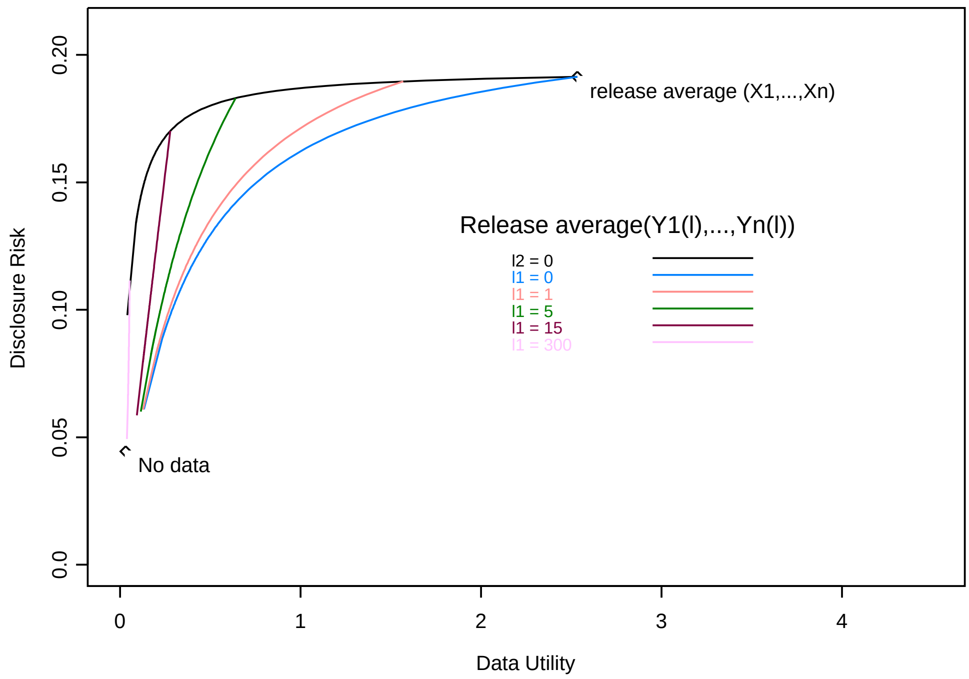

Figure 3: *Release average masked data v.s. average original data, n=20*

As shown in Figure 3, when the threshold t for maximum tolerable disclosure is less than 0.19 the release of the average of the original data is not "safe" and as we expect, the optimal decision would be to release the average of the masked data using among all the pairs (l_1, l_2), for which the risk of disclosure is just below the threshold t, the pair that has the minimum value for l_1, the variance of the contamination of the first component of the original observations. Similarly the worst decision would be to release the average of the masked data using among all the pairs (l_1, l_2), for which the risk of disclosure is just below the threshold t, the pair that has the minimum value for l_2, the variance of the second component of the original observations. For example if the statistical agency believes that a threshold $t = 0.15$ is appropriate for the problem, then the optimal form of data release would be to release masked data with $l_1 = 0$ and $l_2 = 10.95$, that is the statistical agency should add a normal noise with variance 10.95 to the second component of the observations and leave the first component of the observations unchanged. This would produce a masked data set with disclosure risk 0.149 and data utility equal to 0.77. We can repeat the analysis for different choices of the matrix M that defines the society's loss function. In particular we consider:

$$M_1 = \begin{pmatrix} 6 & 0 \\ 0 & 0.5 \end{pmatrix}, \quad M_2 = \begin{pmatrix} 5 & 0 \\ 0 & 8 \end{pmatrix}, \quad M_3 = \begin{pmatrix} 5 & 0 \\ 0 & 15 \end{pmatrix}, \quad M_4 = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}.$$

Table 6 summarizes the relationship between the matrices M_i and the matrix W that defines the intruder's loss function (m_{1i} and m_{2i} denote the elements of the diagonal of M_i, $i = 1, \ldots, 4$).

Matrix	$m_{2i} - m_{1i}$ vs. $w_2 - w_1$	$\frac{m_{2i}}{m_{1i}}$ vs. $\frac{w_2}{w_1}$
M_1	$m_{21} - m_{11} = -5.5 < w_2 - w_1 = 2$	$\frac{m_{21}}{m_{11}} = 0.08 < \frac{w_2}{w_1} = 3$
M_2	$m_{22} - m_{12} = w_2 - w_1 = 2$	$\frac{m_{22}}{m_{12}} = 1.6 < \frac{w_2}{w_1} = 3$
M_3	$m_{23} - m_{13} = 10 > w_2 - w_1 = 2$	$\frac{m_{23}}{m_{13}} = \frac{w_2}{w_1} = 3$
M_4	$m_{24} - m_{14} = w_2 - w_1 = 2$	$\frac{m_{24}}{m_{14}} = \frac{w_2}{w_1} = 3$

Table 6: Diagonal of M_i vs. diagonal of W

In figure 4 we repeat the plot in figure 3 using the different matrices M_i. Only the magenta line and the blue line corresponding to $l_1 = 0$ and $l_2 = 0$ respectively, are plotted. Figure 4 shows

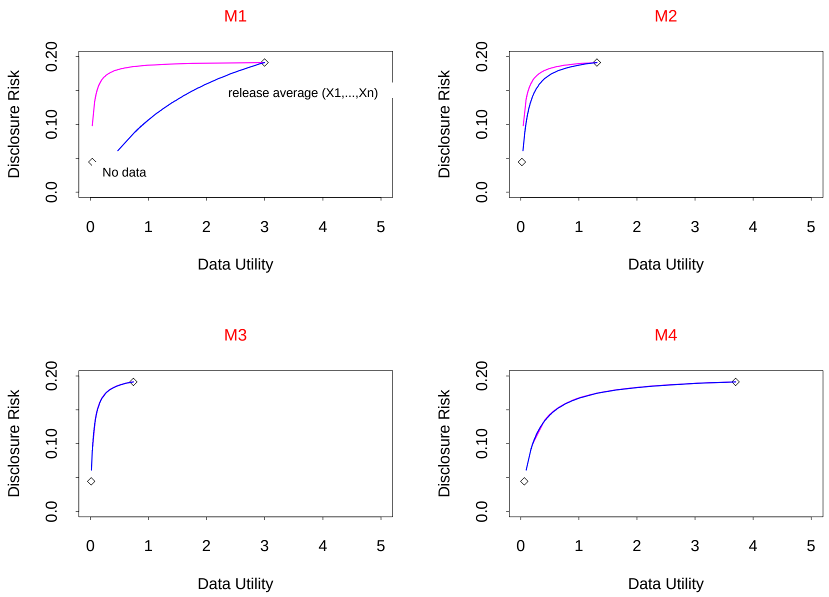

Figure 4: *Disclosure Risk vs. Data utility for different choices of the matrix M that defines society's loss function*

that as far as society's interest for the second component of μ is greater, in relative sense, than the intruder's interest in the second component of X_1, i.e as far as

$$\frac{m_{2i}}{m_{1i}} > \frac{w_2}{w_1}, \ i = 1, \ldots, 4$$

then the optimal strategy is to release the average of the masked data using among all the pairs (l_1, l_2), for which the risk of disclosure is just below the threshold t, the pair that has the minimum value for l_1, the variance of the contamination of the first component of the original observations. Similarly the worst decision would be to release the average of the masked data using among all the pairs (l_1, l_2), for which the risk of disclosure is just below the threshold t, the pair that has the minimum value for l_2, the variance of the second component of the original observations. Note that in this case the choice of the threshold for the risk of disclosure could be made using heuristic considerations similar to those described in section 6.1.2 for the univariate case, but its interpretation is much more complicate than in the univariate case.

6.3 Example 3: simple regression

Consider a population P and two quantitative variables X and Y. Suppose that the relationship between X and Y can be adequately described by the linear model:

$$Y_i = \alpha + \beta X_i + \epsilon_i, \quad \epsilon_i \sim N(0, 1),$$

and suppose also that the marginal distribution of X is normal with mean zero and variance 50,

$$X \sim N(0, 50).$$

A statistical office has to decide about the release of the data set $(\underline{x}, \underline{y}) = \{(x_1, y_1), \ldots (x_{10}, y_{10})\}$ consisting of 10 i.i.d. observations of the bivariate random variable (X, Y):

i	1	2	3	4	5	6	7	8	9	10
x_i	4.584	-0.318	0.129	2.037	-2.974	1.441	-12.794	0.320	0.240	7.705
y_i	9.872	-0.234	2.376	4.742	-3.193	5.714	-25.658	3.417	1.745	15.660

Table 7: Original data

We consider two different disclosure scenarios:

(A1) society's target, Θ_{SOC}, is to identify the parameters of the linear model and the intruder's target is to identify the values of (X,Y) for a subject that is not in the sample:

- $\Theta_{SOC} = (\alpha, \beta)$;
- $\Theta_{INT} = (X^*, Y^*)$, $(X^*, Y^*) \notin \bigcup_{i=1}^{10} (X_i, Y_i)$;

(A2) society's target, Θ_{SOC}, is to identify the parameters of the linear model and the intruder's target is to identify the values of Y for a subject that is not in the sample and for which it is known that $X = 5$:

- $\Theta_{SOC} = (\alpha, \beta)$;

- $\Theta_{INT} = Y^*$, $Y^* \notin \bigcup_{i=1}^{10} Y_i$, $X^* = 5$.

For both scenarios we assume that:

- prior to observing the data both society's and intruder's uncertainty about (α, β) can be expressed by a bivariate normal distribution,

$$(\alpha, \beta) \sim MN(\theta_0, \Sigma_0), \;\; \theta_0 = (0,0)^T, \;\; \Sigma_0 = \begin{pmatrix} 50 & 0 \\ 0 & 50 \end{pmatrix};$$

- society and intruder share the information that the marginal distribution of X is normal with mean zero and variance 50, $X \sim N(0, 50)$;

- as a possible alternative to release the original data or not release any data the statistical office consider the possibility of release the masked data $(\underline{w}, \underline{z}) = \{(w_1, z_1), \ldots (w_{10}, z_{10})\}$, obtained adding normal noises to the original data $(\underline{x}, \underline{y})$:

$$w_i(v) = x_i + \epsilon_{x_i}(v), \;\; \epsilon_{x_i}(v) \sim N(0, v^2), \quad z_i(\lambda) = y_i + \epsilon_{y_i}(\lambda), \;\; \epsilon_{y_i}(\lambda) \sim N(0, \lambda^2), \;\; i = 1, \ldots, 10.$$

  In particular $v = 2$, $\lambda = 1.5$ are used and we assume that if the statistical agency decide to release the masked data, also the values of v and λ are released. Table 8 shows the contaminated data.

i	1	2	3	4	5	6	7	8	9	10
w_i	5.079	-4.695	1.885	2.286	-6.624	-0.871	-11.153	2.958	1.795	9.567
z_i	7.412	1.332	1.99	4.669	-1.598	8.347	-27.032	2.468	1.369	15.301

Table 8: Contaminated data, $v = 2$, $\lambda = 1.5$

Figure 5 shows the original data (blue dots) and the contaminated data (red points) together with the true regression line (green line) that we used to generate the original data in table 7 $(Y = 1 + 2X)$.
We want to follow the approach described section 6 to choose the "best" form of data release among the two considered. Again a quadratic loss function is used.

6.3.1 Scenario A_1

Suppose that the society's target is to identify the parameters of the linear model and the intruder's target is to identify the values of (X,Y) for a subject that is not in the sample:

- $\Theta_{SOC} = (\alpha, \beta)$;

- $\Theta_{INT} = (X^*, Y^*)$, $(X^*, Y^*) \notin \bigcup_{i=1}^{10} (X_i, Y_i)$;

and suppose also that both the intruder and society use a quadratic loss function to quantify the loss that they pay for an estimate a of the target values Θ_{INT} and Θ_{SOC}:

$$L_{INT}(a; \theta_{INT}) = (a - \theta_{INT})^T H (a - \theta_{INT}), \;\; L_{SOC}(a; \theta_{SOC}) = (a - \theta_{SOC})^T M (a - \theta_{SOC}), \quad (7)$$

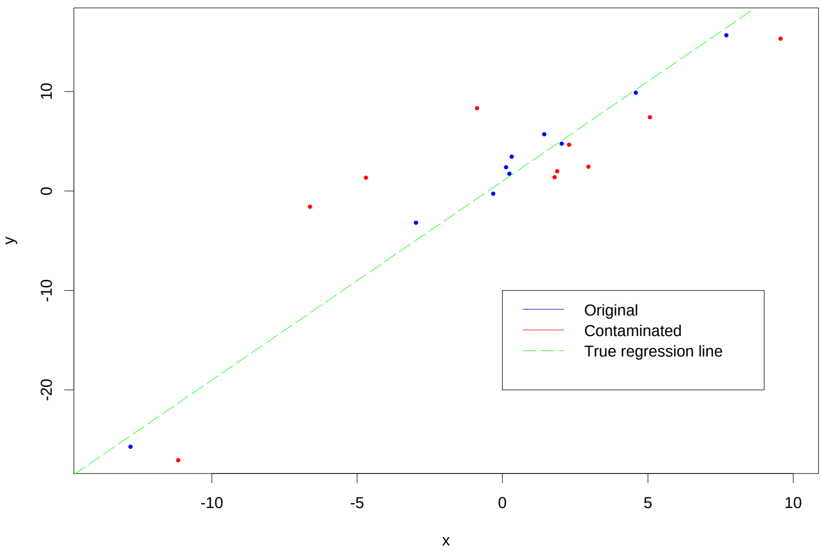

Figure 5: *Original data vs. masked data*

where:

$$H = M = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Under the loss functions in (7) the risk of disclosure when the intruder's distribution over (X^*, Y^*) is $p(\cdot)$ is :

$$RISK(p) = \frac{1}{U_{INT}(p)} = \frac{1}{Var_p(X^*) + Var_p(Y^*)} \tag{8}$$

and similarly, the data utility when society's distribution over (α, β) is $q(\cdot)$ is:

$$UTILITY(q) = \frac{1}{U_{SOC}(q)} = \frac{1}{Var_q(\alpha) + Var_q(\beta)}. \tag{9}$$

Table 9 shows society's prior and posterior distributions for (α, β).

Data	$\pi((\alpha,\beta)\vert Data)$
No data	$MN(\theta_0, \Sigma_0)$
$(\underline{x},\underline{y})$	$MN((\frac{\underline{X}^t\underline{X}}{\sigma^2} + \Sigma_0^{-1})^{-1}(\frac{\underline{X}^t\underline{X}\underline{y}}{\sigma^2} + \Sigma_0^{-1}\theta_0), (\frac{\underline{X}^t\underline{X}}{\sigma^2} + \Sigma_0^{-1})^{-1})$
	$\underline{X}^T = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_{10} \end{pmatrix}.$

Table 9: Prior and posterior distributions of (α, β) given the original data

Note that in this case the posterior distribution of (α, β) given the masked data as well as the intruder's prior and posterior distributions for (X^*, Y^*) are analytically intractable. However we can estimate these distributions using Monte Carlo techniques (for details see the Appendix). Figures 6 ,7, 8, 9 show joint and marginal posterior distributions of (α, β) and (X^*, Y^*) corresponding to each of the two forms of data released and prior to the release of any data.

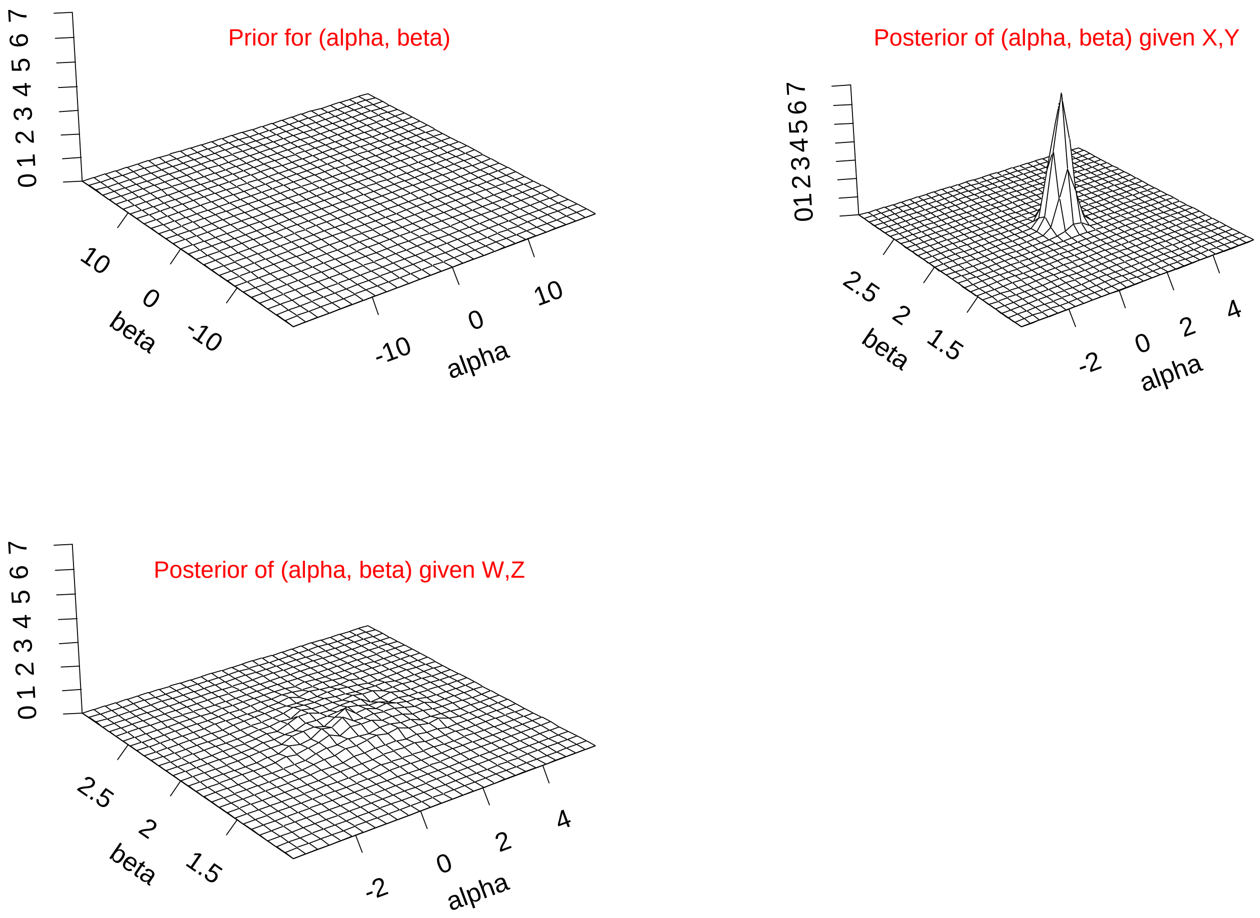

Figure 6: *Joint prior and posterior distributions for* (α, β)

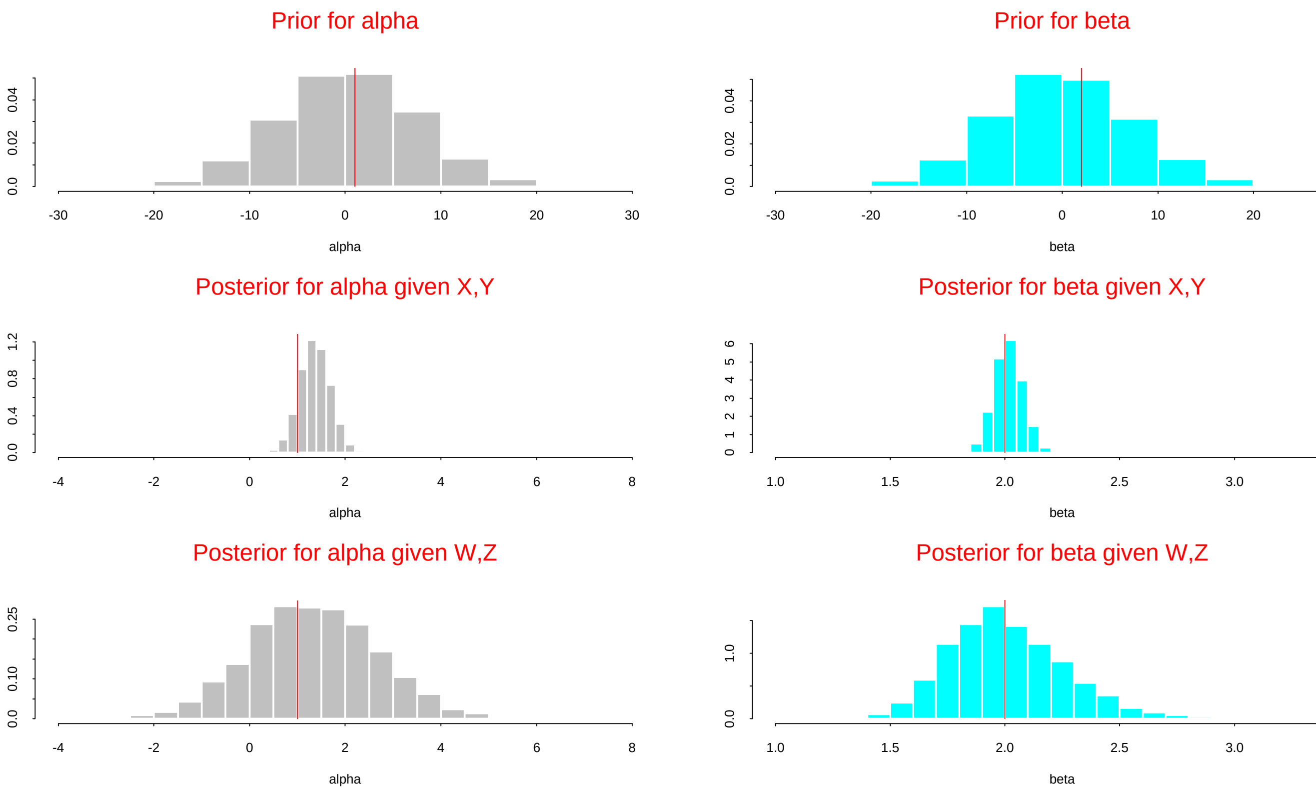

Figure 7: *Marginal prior and posterior distributions for α and β. The red lines denote the true values of the parameters*

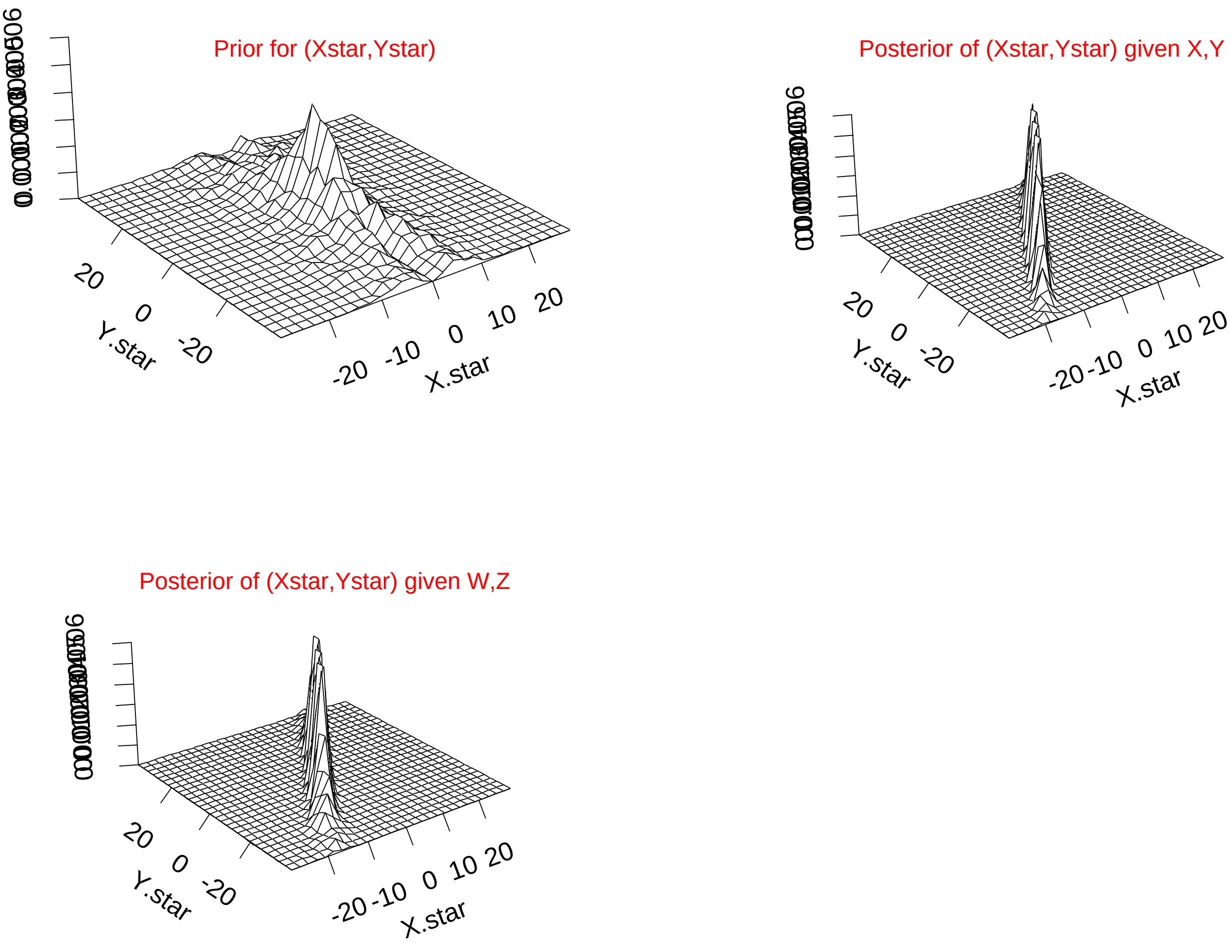

Figure 8: *Joint prior and posterior distributions for* (X^*, Y^*)

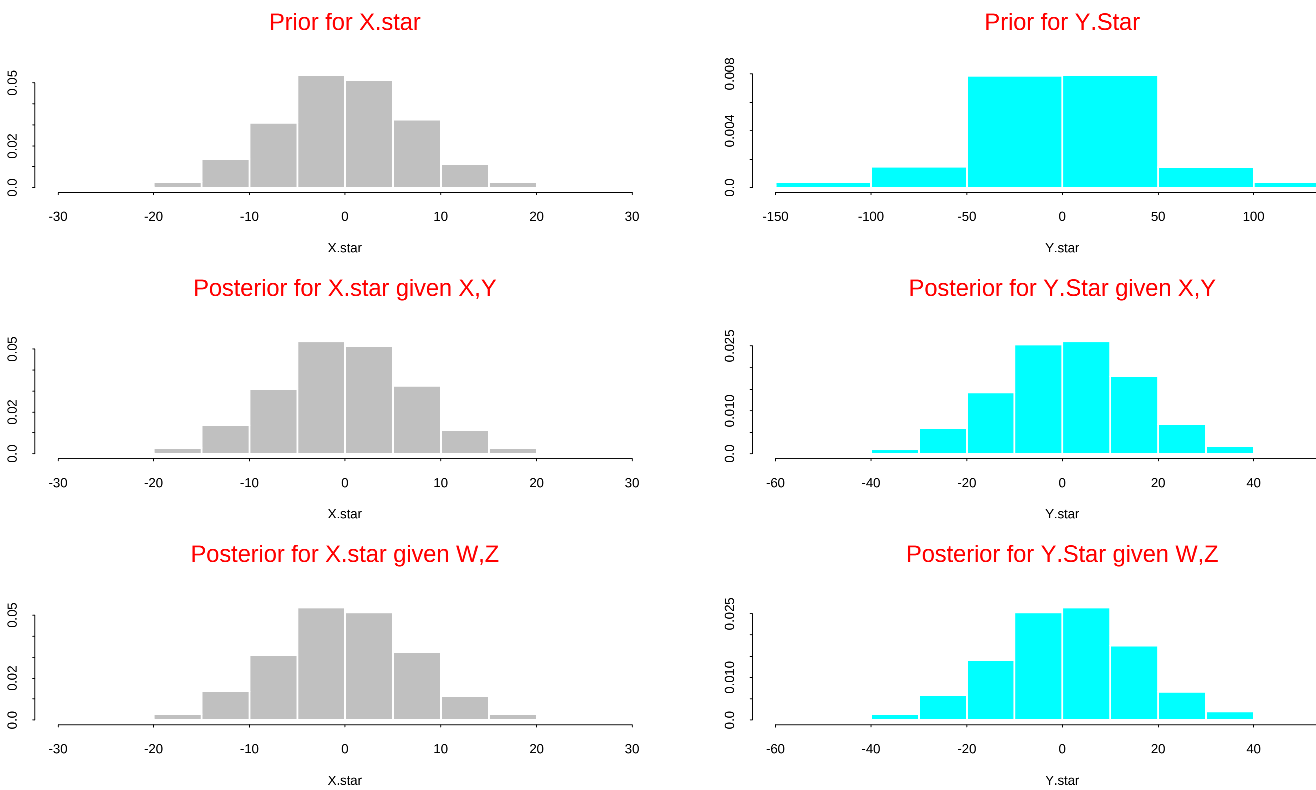

Figure 9: *Marginal prior and posterior distributions for* X^* *and* Y^*

Figure 10 shows the pairs ("Data utility", "Disclosure risk") for each of the two forms of data release and prior to the release of any data.

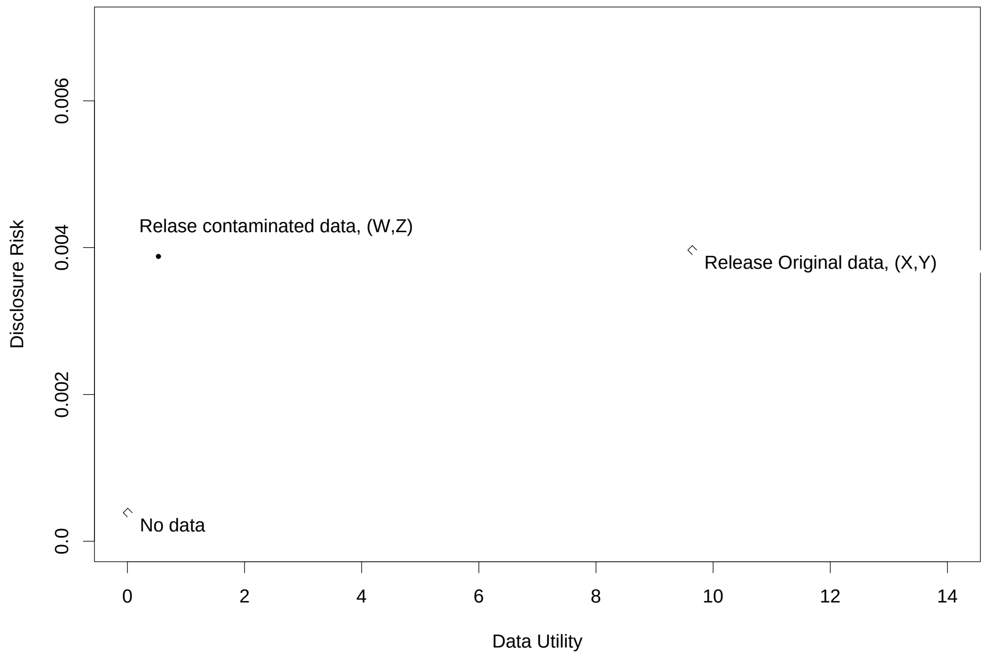

Figure 10: *Data utility vs. disclosure risk, scenario* A_1

The masking in this case does not seem very effective. It reduces to a great extent the data utility and leaves the risk of disclosure almost unchanged. Different values for the contamination parameters probably will give similar results. Figure 11 might help to understand the poor performance of the masking procedure in this scenario. Circles represent unobserved variables, squares what it is observed. The original data carry information about the parameters of the linear model but no information about X^*. The uncertainty about X^* therefore does not depend on the observed data. On the other hand the uncertainty about Y^* depends both on the uncertainty about the parameters of the linear model and the uncertainty about X^*. Prior to observing the data, both the uncertainty about (α, β) and the uncertainty about X^* are very high and they contribute almost in the same way to the uncertainty about Y^* (compare prior distributions for (α, β) and X^* in figures 6, 7 and 9). Note that the variance of Y^* given α, β and x^* is negligible with respect to the prior uncertainty about (α, β) and X^*. After the original data are released, however, the uncertainty about X^* is unchanged but the uncertainty about (α, β) drastically decreases (compare prior and posterior distribution of (α, β) given $(\underline{x}, \underline{y})$ in figure 6). As a result after the release of the original data set the uncertainty about Y^* depends almost exclusively on the uncertainty about X^*.

The contaminated data contain considerably less information about the parameters of the linear model than the original data (compare the posterior distributions $(\alpha, \beta)|\underline{x}, \underline{y}$ and $(\alpha, \beta)|\underline{w}, \underline{z}$ in figure 6) but again, after the masked data are released, the uncertainty about X^* is unchanged while the uncertainty about (α, β) drastically decreases (compare prior and posterior distributions of α and β given $(\underline{w}, \underline{z})$ in figure 7). As a result after the release of the masked data the uncertainty about Y^* also depends almost exclusively on the uncertainty about X^*. If the masked data were released instead of the original data, we would have almost the same risk of disclosure but much more uncertainty about the parameters of the linear model. In order to significantly reduce the risk of disclosure the values of the contamination parameters, v and λ, should be so big that the resulting contaminated data would be completely useless to estimate the parameters of the linear model and to perform any other type of interesting inference.

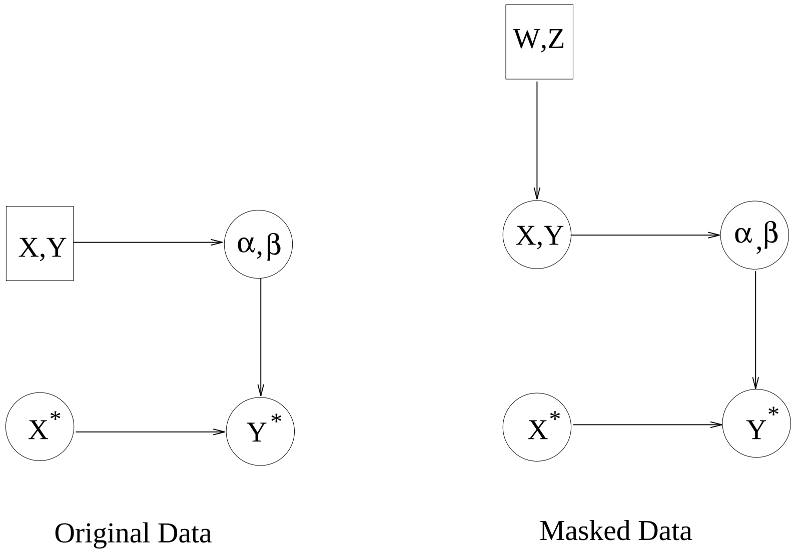

Figure 11: *Updating prior information: masked data vs. original data, scenario* A_1

6.3.1 Scenario A_2

Suppose that society's target is to identify the parameters of the linear model and the intruder's target is to identify the values of Y for a subject that is not in the sample and for which it is known that $X = 5$.

- $\Theta_{SOC} = (\alpha, \beta)$;
- $\Theta_{INT} = Y^*$, $Y^* \notin \bigcup_{i=1}^{10} Y_i$, $X^* = 5$.

Suppose also that both the intruder and society use a quadratic loss function to quantify the loss that they pay for estimates a and b of the target values Θ_{INT} and Θ_{SOC}:

$$L_{INT}(a; \theta_{INT}) = (a - \theta_{INT})^2, \quad L_{SOC}(b; \theta_{SOC}) = (b - \theta_{SOC})^T M (b - \theta_{SOC}), \tag{10}$$

where:

$$M = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Under the loss functions in (10) the risk of disclosure when the intruder's distribution over Y^* is $p(\cdot)$ is :

$$RISK(p) = \frac{1}{U_{INT}(p)} = \frac{1}{Var_p(Y^*)}, \tag{11}$$

and similarly, the data utility when society's distribution over (α, β) is $q(\cdot)$ is:

$$UTILITY(q) = \frac{1}{U_{SOC}(q)} = \frac{1}{Var_q(\alpha) + Var_q(\beta)}. \tag{12}$$

Note that also in this case the intruder's distribution over Y^* prior to release of any data and after the release of the original data or the release of the masked data are analytically intractable and we obtain them through simulations (see the Appendix). Figure 12 shows the posterior distribution of Y^* corresponding to each of the two forms of data released and prior to the release of any data (the prior and posterior distributions for (α, β) are the same than in scenario A_1).

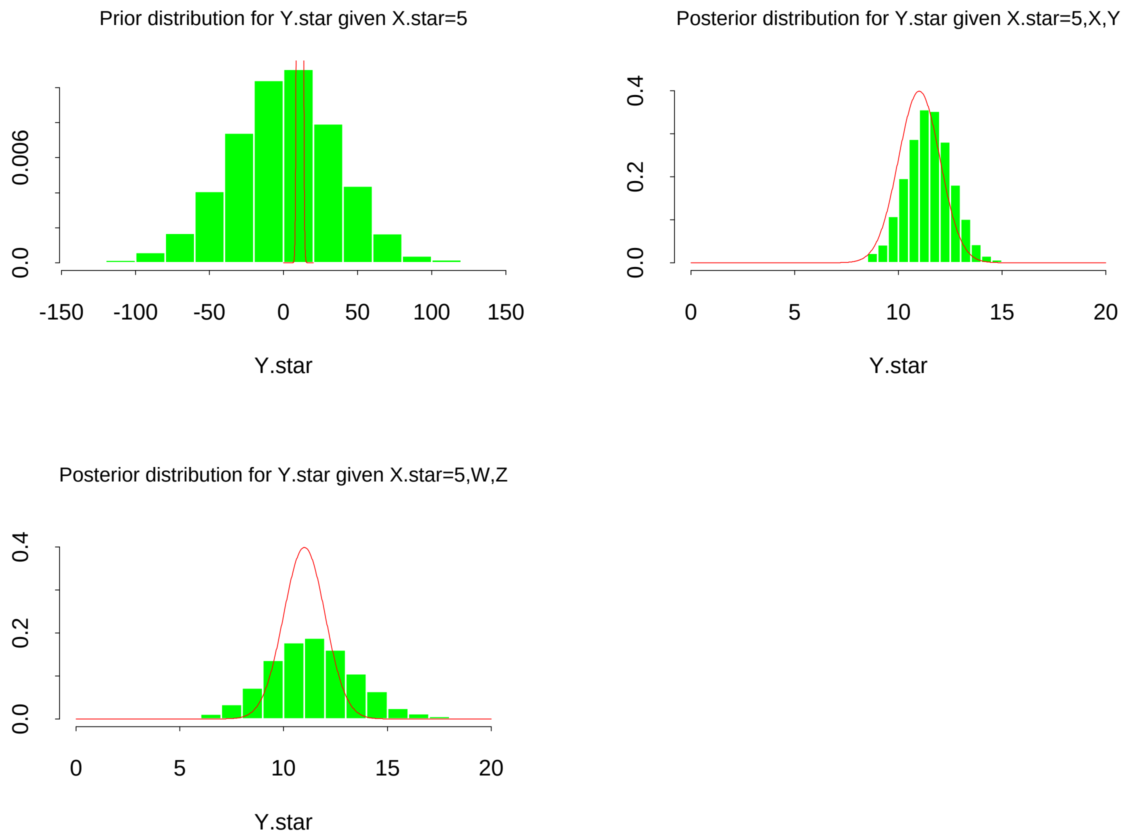

Figure 12: *Prior and posterior distributions of Y^* given $X^* = 5$. The red line denotes the distribution of Y^* given $X^* = 5$, $\alpha = 1$, $\beta = 2$*

Figure 13 shows the pairs ("Data utility", "Disclosure risk") for each of the two forms of data release and prior to the release of any data. The masking in this case seems to be more effective that in scenario A_1. Using the masked data we have both a great reduction of the risk of disclosure and the data utility. It seems however that it is possible to find values of the contamination parameters v and λ that provide a reasonable trade-off between disclosure risk and data utility. Figure 14 explains why. Since in this scenario X^* is known, the uncertainty about Y^* only depends on the uncertainty about the parameters of the linear model. The more we contaminate the original data the more we increase the uncertainty about (α, β) and indirectly the uncertainty about Y^*. Of course what becomes relevant here is the relative value of the contamination with respect to the variance of Y^* given α, β and $X^* = 5$.

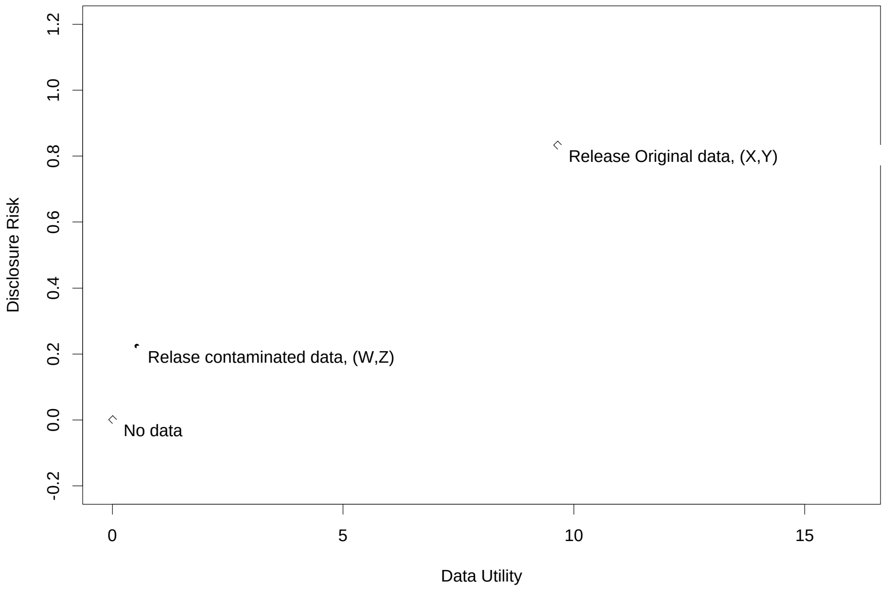

Figure 13: *Data utility vs. disclosure risk, scenario A_2*

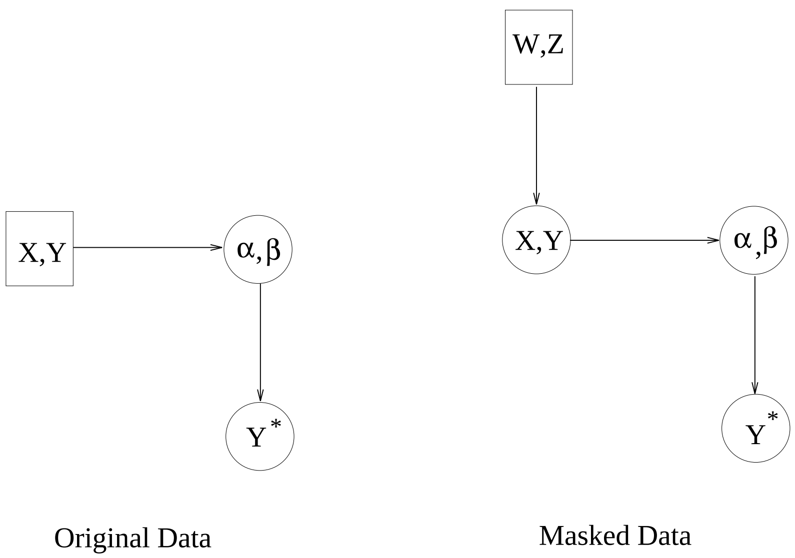

Figure 14: *Updating prior information: masked data vs. original data, scenario A_2*

7. CONCLUSIONS

In this work we presented a general class of measures of data utility that represent a natural extension of the measures of disclosure risk proposed by Duncan and Lambert (1986). In our framework data utility is a decreasing function of society's uncertainty. Two key ideas characterize our approach. The first one is that the problem of measuring the data utility associated with a particular disclosure limitation technique can not be properly addressed without specifying what is the final use of the data set, in which respect its release might be beneficial for society. The second idea is that decision theory offers a natural framework to formalize the problem. Several assumptions characterize our measures of data utility. For example we assume that:

- the statistical agency is able to specify the intruder's and society's targets;
- the statistical agency knows exactly the information available to the intruder and to society prior to the release of any data and it also knows which loss functions they are willing to use;
- the intruder and society are able to formalize the prior uncertainty about their targets through a single probability distribution;
- there is no uncertainty about the mechanism generating the different forms of data release;
- intruder and society share the same information about unknown quantities in the problem (symmetric information).

In a realistic scenario most of these assumptions do not hold. The statistical agency has to deal with several users with different targets, prior distributions, loss functions and agency has only imperfect knowledge of the users' models. Incorporate these uncertainties in our framework, however, is not too difficult, at least in principle. "Intruder" and "society" can be used to represent groups of individuals/entities that are typical users of the products released by a statistical agency. Θ_{INT} and Θ_{SOC} can represent multiple targets, and classes of intruder's and society's loss functions, prior distributions and models can be used instead of a single loss function, prior distribution or model. An analysis of robustness then should be performed.

Our agenda of future work includes four main points. First of all we want to complete the examples presented here including other forms of data release and disclosure scenarios. In particular we would like to include, among the possible forms of data release, the released of synthetic data and of a mix of synthetic and masked data. We also would like to consider an alternative disclosure scenario in which the intruder's goal is to discredit the statistical office and any individual represented in the released data can be the target. Secondly we want to apply the theory described in the paper to some real problem. This should give us a better understanding of how feasible is the implementation of the theory and how it performs compared to alternative methods. Another problem that we would like to study more carefully is the problem of the calibration of the threshold for the disclosure risk. The risk of disclosure associated to different forms of data released is measured applying an uncertainty function to the posterior distribution of Θ_{INT} given the particular form of data released considered. The uncertainty function should suggest some "natural" threshold values. In most of the cases however the choice is not so "natural" (like in our second and third example is section 6) and it is not straightforward to decide which threshold value is appropriate for the problem. We should try to understand which features of the posterior distribution for the intruder target better capture the intruder's uncertainty about his target and use these characteristics to decide which loss function (and therefore uncertainty function) and which threshold values are appropriate for the problem.

ACKNOWLEDGEMENT

Preparation of this paper was supported by a Marie Curie Fellowship of the European Community program "Improving The Human Research Potential" under the contract number HPMFCT-2000-00463. The content of the paper reflects the author's personal opinion. The European Commission is not responsible for any views or results expressed.

APPENDIX

Example 3: Let:

$$\Sigma_x = 50 \cdot I_{10},\ \ \Sigma_w = v^2 \cdot I_{10},\ \ \Sigma_z = (\sigma^2 + \lambda^2) \cdot I_{10},\ \lambda = 1.5,\ v = 2,\ \sigma = 1,\ \mu = 0 \cdot 1_{10}$$

where : 1_{10} is a vector of one's of length 10.

- Posterior distribution of (α, β) given the masked data.

$$\pi(\alpha, \beta|\underline{w}, \underline{z}) \propto \pi(\alpha, \beta) \cdot \pi(\underline{w}, \underline{z}|\alpha, \beta),$$

where:

$$\pi(\alpha, \beta) \sim MN(\theta_0, \Sigma_0),\ \ \theta_0 = (0, 0)^T,\ \ \Sigma_0 = \begin{pmatrix} 50 & 0 \\ 0 & 50 \end{pmatrix}; \tag{13}$$

$$\pi(\underline{w}, \underline{z}|\alpha, \beta) = \pi(\underline{w}|\alpha, \beta) \cdot \pi(\underline{z}|\alpha, \beta, \underline{w}) = \pi(\underline{w}) \cdot \pi(\underline{z}|\alpha, \beta, \underline{w}) \propto \pi(\underline{z}|\alpha, \beta, \underline{w})$$

and

$$\pi(\underline{z}|\alpha, \beta, \underline{w}) \quad = \quad \int \pi(\underline{z}, \underline{x}|\alpha, \beta, \underline{w}) d\underline{x} = \int \pi(\underline{z}|\alpha, \beta, \underline{x}, \underline{w}) \cdot \pi(\underline{x}|\alpha, \beta, \underline{w}) d\underline{x} =$$

$$\int \pi(\underline{z}|\alpha, \beta, \underline{x}) \cdot \pi(\underline{x}|\underline{w}) d\underline{x}.$$

It is easy to show that:

$$\pi(\underline{x}|\underline{w}) = MN(\tilde{\mu}, \tilde{\Sigma});$$

with:

$$\tilde{\Sigma} = (\Sigma_x^{-1} + \Sigma_w^{-1})^{-1},$$
$$\tilde{\mu} = \tilde{\Sigma}(\Sigma_x^{-1}\mu + \Sigma_w^{-1}\underline{w}),$$

and

$$\pi(\underline{z}|\alpha, \beta, \underline{w}) = MN(\tilde{\tilde{\mu}}, \tilde{\tilde{\Sigma}}); \tag{14}$$

with:

$$\tilde{\tilde{\Sigma}} = (\Sigma_z^{-1} - \beta^2(\beta^2\Sigma_z^{-1} + \tilde{\Sigma}^{-1})^{-1}\Sigma_z^{-1})^{-1},$$
$$\tilde{\tilde{\mu}} = \tilde{\tilde{\Sigma}}(\alpha 1_n^t \Sigma_z^{-1} + (-\alpha\beta\Sigma_z^{-1}1_n + \tilde{\Sigma}^{-1}\tilde{\mu})^t(\beta^2\Sigma_z^{-1} + \tilde{\Sigma}^{-1})^{-1}\beta\Sigma_z^{-1})^t.$$

Using (13) and (14) we can write the kernel of the posterior distribution of (α, β) given $(\underline{w}, \underline{z})$ and we can use Metropolis algorithm to estimate the posterior distribution of (α, β) given the masked data.

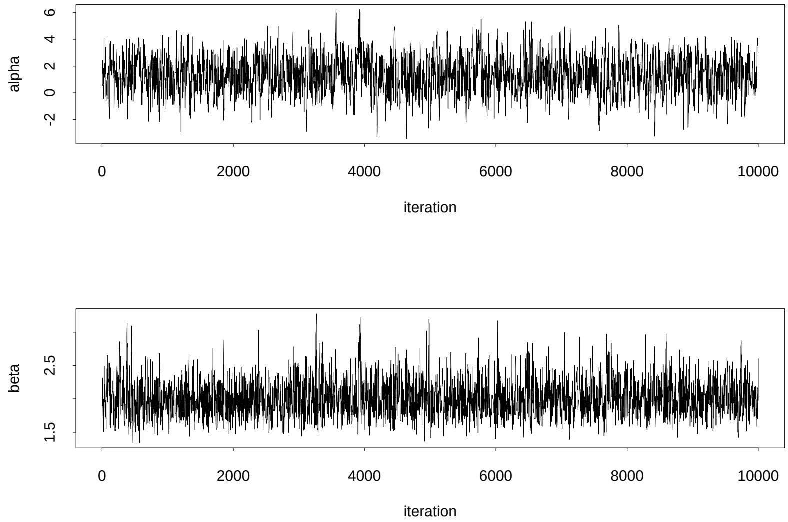

Figure 15: *Time series plots for the posterior distributions of α and β given $(\underline{w}, \underline{z})$, N=10000*

Figure 15 shows the time series plot for the posterior distributions of α and β given $(\underline{w}, \underline{z})$ obtained from Metropolis.

We used N=10000 iterations with no "burn-in". Initial values for the Metropolis were $\alpha = \beta = 2$.

- Prior distribution for (X^*, Y^*).

$$\pi(Y^*, X^*) = \pi(X^*) \cdot \pi(Y^*|X^*)$$

where:

$$\pi(X^*) = N(0, 50);$$

$$\begin{aligned}\pi(Y^*|X^*) &= \int \pi(Y^*, \alpha, \beta|X^*) d\alpha d\beta = \\ &\int \pi(Y^*|X^*, \alpha, \beta)\pi(\alpha, \beta|X^*) d\alpha d\beta = \int \pi(Y^*|X^*, \alpha, \beta)\pi(\alpha, \beta) d\alpha d\beta.\end{aligned}$$

with

$$\pi(Y^*|X^*, \alpha, \beta) = N((1, X^*)(\alpha, \beta)^T, \sigma^2).$$

We obtain the predictive distribution of (X^*, Y^*) through simulation in the following way:

(I) Simulate independently $x^{*(1)}, \ldots, x^{*(N)}$ from $\pi(X^*)$, and $(\alpha^{(1)}, \beta^{(1)}) \ldots (\alpha^{(N)}, \beta^{(N)})$ from $\pi(\alpha, \beta)$;

(II) For each $(x^{*(j)}, \alpha^{(j)}, \beta^{(j)})$ in step (I) simulate $Y^{*(j)}$ from $Y^*|x^{*(j)}, \alpha^{(j)}, \beta^{(j)}$, j=1,...,N

In a similar way we can generate the posterior distributions of (X^*, Y^*) given the original data and the masked data, and the posterior distributions of Y^* given $X^* = 5, \underline{x}, \underline{y}$, and the posterior distribution of Y^* given $X^* = 5, \underline{w}, \underline{z}$.

REFERENCES

Baeyens, Y. and Defays, D. (1998), Estimation of variance loss following from microaggregation by the individual ranking method *Proceedings of "Statistical data Protection SDP '98"*, 101-108. Eurostat, 25-27 March 1998, Lisbon, Portugal.

Disclosure-Avoidance Techniques, S. (1978), *Statistical Working Paper 2, Report on Statistical Disclosure and Disclosure-Avoidance Techniques*, Federal Committee on Statistical Methodology, Office of Federal Policy and Standards, U.S. Department of Commerce, Washington, D.C.

Duncan, G.T. and Lambert, D. (1986), Disclosure-limited Data Dissemination. *Journal of the American Statistical Association, 81*, 10-18.

Duncan, G.T. and Keller-McNulty, S. (2001), Risk of statistical confidentiality disclosure: A comparison of masked and synthetic data release, *Research in Official Statistics*, Vol. 3, (in press).

Duncan, G.T. and Pearson, R.W. (1991), Enhancing Access to Microdata While Protecting Confidentiality. *Statistical Science, 6*, 219-239.

Hurkens C.A.J. and Tiourine S.R.(1998), Models and Methods for the Microdata Protection Problem, *Journal of Official Statistics, 14*, 437-447.

Mateo-Sanz, J.M. and Domingo-Ferrer, J. (1998), A method for data-orientated multivariate microaggregation. *Proceedings of "Statistical data Protection SDP '98"*, 89-99. Eurostat, 25-27 March 1998, Lisbon, Portugal.

Willenborg, L. and de Waal, T. (2001), *Elements of Statistical Disclosure Control*, Lectures Notes in Statistics, Vol. 155. Springer Verlag, New York.