Working Paper No. 19
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES (EUROSTAT**

**Joint ECE/Eurostat work session on statistical data confidentiality**
(Luxembourg, 7-9 April 2003)

Topic (v): Risk assessment

# ASSESSING DISCLOSURE RISK AND DATA UTILITY: A MULTIPLE OBJECTIVES DECISION PROBLEM

**Invited paper**

Submitted by the University of Alicante (Spain)[1]

[1] Prepared by Mario Trottini (Mario.trottini@ua.es).

***Assessing Disclosure Risk and Data Utility: A Multiple Objectives Decision Problem***

***Mario Trottini***

*mario.trottini@ua.es*

*Dpto. de Estadística e I.O , Universidad de Alicante, Spain*

## 1. Introduction

Statistical agencies and Information Organizations that systematically publish statistical data are increasingly concerned with possible misuses of their data release that might lead to disclosure of confidential information about individual respondents represented in the data. Beside legal and ethical considerations, consequences of such disclosure can seriously compromise the trust of the public opinion and thus quality and availability of statistical data. As a result of these concerns, different transformations of the original data (*data masking)* have been designed aimed to produce new forms of data release that, ideally, should be both *safe,* in the sense that prevent disclosure of confidential information, and *valid* in the sense that inferences of legitimate users based on the transformed data agree with inferences that could be made using the original data.

While the number of *data masking* techniques has increased significantly in the last few years, their efficacy has undergone only little exploration. There is an increasing need to design suitable criteria that allow compare alternative data release and identify which one performs best and when. Usual solutions, discussed in the literature of statistical confidentiality and that find several applications in current practice of statistical agencies (Willenborg and De Waal 1996, 2001, Doyle et al. 2001), consist of two steps:

- Step 1a: define suitable measures, *S* and *V* that can express the extent to which *safety* and *validity* would be achieved if the *masked data* were released;
- Step 2a: define a criterion to compare alternative releases based on the values of *S* and *V*.

The conventional terminology in statistical confidentiality refers to *S* and *V* as *measures of disclosure risk*, and *measures of data utility* respectively, and to the criteria in step 2a, as *optimality criteria*.

Implementation of the above procedures presents two main difficulties:

(i) Although there is general agreement that optimality criteria should be defined on the base of safety and validity associated with the released data, *safety* and *validity* are ambiguous concepts, multivariate in nature.

For a given disclosure limitation problem, in fact, it is possible to imagine several potential intruders, for each intruder several targets and for a given intruder and a given target several alternative intruder's attacks. But few exceptions, like for example the multiple-attacker cell suppression methodology discussed by Salazar (Salazar 2002), existing approaches can not deal with the multidimensionality of S and V (Lambert, 1993, Doyle et al. 2001, Domingo 2002).

(ii) Even assuming that suitable measures *S* and *V* for safety and validity could be defined, still *S* and *V* are usually expressed in different units, they have different meanings and comparison of arbitrary pairs (*s,v*), and (s',v') is not a trivial task.

The problem is even more complex due to the fact that *S* and *V*, from the statistical agency perspective, are random variables. In fact the values that *S* and *V* take when a data mask is released depend on the users' actions that are only partially known to the statistical agency. The agency, for example, has uncertainty about users' prior information, the estimation procedures that they use, etc. Thus each data release induces a distribution over the space of consequences and choosing among alternative data release is equivalent to choose among alternative lotteries for (*S*, *V*), a much more difficult task than just express preferences over pairs (*s*, *v*). Existing optimality criteria are essentially of two types:

- Type I: maximize validity under a constraint of minimum safety, i.e. choose the form of data release that maximize validity among those that have safety above a fixed threshold *t*;
- Type II: build an index or *score* based on S and V.

The use of a threshold value for safety, like in criteria of the type I (the most common case) avoid the problem of different scales for *S* and *V*. However it leads to an "extreme criteria". Suppose, for example, that both *S* and *V* take value in $(0,100)$ and that the threshold for safety is $t=9$. Suppose you have two alternative releases, $R_1$ that produces with certainty $(S=9.1, V=90)$, and $R_2$ that produces with certainty $(S=100, V=89.9999)$. Based on criteria of type I, $R_1$ should be selected. In this case, however, many statistical agencies would choose $R_2$. In fact it is hard to believe that an arbitrary large increase in safety is not worth an infinitesimal decrease in validity. Criteria of type II could be appropriate but should be based on solid theoretical ground. Current implementations are primarily heuristic. This can lead to criteria that do not reflect the actual preferences of statistical agencies. Note that in both criteria of type I and type II the random nature of *S* and *V* is completely neglected thus underestimating the actual agency's uncertainty.
In the terminology used in *multiple objectives decision theory* (Keeney and Raiffa, 1976), step 1a corresponds to define the appropriate set of *attributes* for the decision problem, while step 2a corresponds to assess a suitable *multiattribute utility function*. The best action (form of data release) is the one that maximizes the expected utility. Despite the strong connection between the two problems no tentative has been made so far to apply results and techniques from multiple objectives decision theory to disclosure limitation problems. This paper explores this possibility. The focus is on the elicitation of a multiattribute utility function (step 2a) and its relevance for the definition of suitable optimality criteria as well as for the assessment of safety (validity) when S (V) is multidimensional.
Section 2 describes the general framework of multiple objectives decision theory and its connections with disclosure limitation problems. Assessment of statistical agency preferences for alternative data release through a two-attribute utility function is discussed in Section 3. Section 4 presents the definitions of utility independence and their relevance for data disclosure limitation problems. Sections 5 and 6 illustrates the general theory with two examples. In the first one a multiattribute utility function is used to derive an optimality criterion for alternative data release. In the second, a multiattribute conditional utility function is used to assess safety when S is a two-dimensional vector. Section 7 summarizes the main results in the paper and outlines ideas of future work.

## 2. Multiple Objectives Decision Theory

A multiple objectives decision problem can be summarized as follows. A decision maker D has to choose one action from a set $\mathsf{R}=\{R_1, R_2,\ldots,R_m\}$ of *m* possible actions. In choosing the best action, D has in mind *k* (usually conflicting) objectives $O_1, O_2,\ldots,O_k$. It is assumed that consequences of actions in $\mathsf{R}$ in terms of objectives $O_j$ can be measured by an attribute $X_j$, $X_j \in \mathsf{X}_j$. The attributes $X_j$ can be quantitative, continuous, discrete or categorical. It is also assumed that consequences of actions in $\mathsf{R}$ are not deterministic but random. In particular it is assumed that each action $R_i$ induces a probability measure $F_i$ over the consequence space $X_{.1} \times X_{.2} \times \cdots \times X_{.n}$. This distribution completely defines consequences of action $R_i$ in terms of the $k$ objectives of interest. Thus the decision maker's problem of choosing the best action is equivalent to the problem of choosing the best probability measure or *lottery* $F_i$.

The solution to this problem consists on defining a suitable multiattribute utility function $u(\cdot): X_{.1} \times X_{.2} \times \cdots \times X_{.n} \rightarrow \Re$ such that lottery $F_i$ is preferred to lottery $F_j$ if and only if the expected utility of $u(\cdot)$ under $F_i$ is bigger than the expected utility of $u(\cdot)$ under $F_j$. The optimal action is the action $R_{i^*}$ that maximizes the expected utility,

$$R_{i^*} = \arg\max_i E\{u(\cdot) \mid R_i\} = \arg\max_i \int u(x_1, x_2, ..., x_k) dF_i(x_1, x_2, ..., x_k) .$$

A disclosure limitation problem has a natural formalization as multiple objectives decision problem where the decision maker is the statistical agency, the objectives are "maximize safety" and "maximize validity" and the set of alternative actions is the set of alternative forms of data release. The whole procedure can be summarized in three steps: (Step 1b) definition of suitable set of objectives and attributes; (Step 2b) definition of an appropriate multiattribute utility function; (Step 3b): utility maximization. Because of page limit constraints step 1b is not addressed here and the focus in on the implementation of step 2b. In what follows it is assumed that step 1b has been successfully performed and that a two attributes utility function need to be assessed.

## 3. Eliciting a two attribute utility function

Suppose that attributes S and V for the two objectives ``maximize safety", ``maximize validity" have been identified and are appropriate for the problem ( $S \in \mathsf{S}$, and $V \in \mathsf{V}$ ). The goal is to build an (utility) function $u(\cdot,\cdot): \mathsf{S} \times \mathsf{V} \rightarrow \Re$ which has the property that lottery $F_i$ is preferred to lottery $F_j$ if and only if

$$E\{u(\cdot,\cdot) \mid R_i\} > E\{u(\cdot,\cdot) \mid R_j\} \Leftrightarrow \int u(s,v) dF_i(s,v) > \int u(s,v) dF_j(s,v).$$

If the space of possible consequences $\mathsf{S} \times \mathsf{V}$ contains few points (say less than 50) than direct assessment of the utility function is possible using certainty equivalents (see Keeney and Raiffa 1976, pp. 222). Unfortunately in data disclosure limitation problems (as in most of real applications) $\mathsf{S} \times \mathsf{V}$ is too big and direct assessment of utility is not feasible. The idea, in these cases, is to identify relevant features of the decision maker's preferences that allow put strong constraints on the form of the multiattribute utility function. One feature that is of particular importance is *utility independence*. It has been shown that if certain utility independence assumptions hold than the multiattribute utility function must be of a specified form. What makes utility independence operational in data disclosure limitation (as well in several real applications) is that: (i) utility independence assumptions can be ascertained in many real disclosure limitation problems, and they are verifiable in practice; (ii) utility independence allows for great variability in the final form of the multiattribute utility function, that is, it can be used to formalize many different preferences structures for data release, and (iii) under utility independence assessment of multiattribute utility is relatively easy (it is equivalent to assess several lower dimensional (conditional) utility functions and scale them properly). The next section presents the main utility independence definitions and their consequences in terms of the form of the multiattribute utility function.

## 4. Utility independence

In order to introduce the different definitions of utility independence imagine a hypothetical scenario where a decision maker, D, have to choose one of *k* alternative releases and an analyst, A, tries to formalize the preferences structure of D in terms of a multiattribute utility function . Suppose that there are *h* statistical analyses of interest for the users and that, after a careful study of the problem, D comes up with the following measures of safety and validity:

- S= % records that are incorrectly re-identified;
- V= % of the *h* statistical analysis of interest, that have "good" quality compared with the corresponding analysis that could be performed using the original data.

Let $s_0$, $s^*$ ($v_0$ , $v^*$) the least desirable and the most desirable value that S (V) can take. In this example it is $s_0 = v_0=0$ and $s^* = v^*=1$. If S=100 then all records are protected and there is no risk of disclosure. S=0, instead corresponds to the case where all units in the original data are correctly re-identified, and safety takes its minimum. Similarly V=100 means that all the *h* statistical analyses of interest based on the released data are of "good" quality (compared with the corresponding analyses based on the original data). While V=0 means that none of the analyses of interest based on the released data is of "good" quality. The ideal form of data release is the one that yields (S=100, V=100), the worst data release, instead is the one that yields (S=0, V=0).
As a first step, in the preference assessment, A could ask D to choose between two hypothetical forms of data release $R_1$ and $R_2$ that yield the same validity v' but different level of safety. $R_1$, could be, for example, a lottery that produces with equal probability (S=100, V=v') or (S=0, V=v') and $R_2$ a lottery that yields with probability one (S=90, V=v'). Once D has made his decision, A asks D whether in giving the answer he/she was thinking about the actual value of v'. If D's answer is "NO" then it is likely that in the decision maker's preferences *S is utility independent of V.*

**Definition 1**: *We say that S is* <u>*utility independent of V*</u> *if and only if conditional preferences for lotteries on S given V=v do not depend on the particular level v*.

In many real applications, it seems reasonable to assume that S is utility independent of V. This independence assumption essentially reflects the dominant role that safety plays in data disclosure limitation problems. Most of the statistical agencies would probably be comfortable in choosing between alternative data release that yield the same validity v' (but different lotteries for safety) without knowing the actual value of v'. If this is the case the results in the following theorem can be applied to derive the multiattribute utility function.

**Theorem 1**: *If S is utility independent of V, then*

$$u(s,v) = u(s_0,v)\cdot[1-u(s,v_0)]+u(s_1,v)\cdot u(s,v_0)$$

*where* $u(s,v)$ *is normalized by* $u(s_0, v_0)=0$, $u(s_1, v_1)=1$ *and* $u(s_0,v)$, $u(s_1,v)$ *and* $u(s,v_0)$ *are conditional utility functions of V given* $S=s_0$, *of V given* $S=s_1$ *and of S given* $V=v_0$, *respectively.*

Note that under utility independence in order to assess the multiattribute utility function it is sufficient to specify three (lower dimensional) conditional utility functions and scale them properly. An example of such elicitation is presented in section 5.

Suppose, now, that the analyst A has ascertained that in D's preferences S is utility independent of V. As a second step in the utility assessment, A could ask D to choose between a data release $R_1$, that yields with equal probability (S=100, V=0) or (S=0, V=100), and a data release $R_2$ that yields with equal probability (S=100, V=100) or (S=0, V=0). If D answers that $R_1$ and $R_2$ are equally preferable, then it is likely that in the decision maker's *preferences S and V are additive independent.*

**Definition 2**: *Attributes S and V are* <u>*additive independent*</u> *if the paired comparison of any two lotteries, defined by two joint probability on (S,V), depends only on the marginal probability distribution.*

If this is the case then the results in the next theorem greatly simplified the assessment of multiattribute utility function. In particular:

**Theorem 2**: *Attributes S and V are additive independent if and only if the utility function is additive.*

$$u(s,v) = u(s_0,v) + u(s,v_0)$$

*where:*

- $u(s,v)$ *is normalized by* $u(s_0,v_0)$*=0, and* $u(s_1,v_1)$*=1 for arbitrary* $s_1$ and $v_1$ *such that* $(s_0,v_1)$ *is preferred to* $(s_0,v_0)$ *and* $(s_1,v_0)$ *is preferred to* $(s_0,v_0)$;
- $u(s_0,v)$, $u(s_1,v)$ *and* $u(s, v_0)$ *are conditional utility functions of V given S=* $s_0$, *S=* $s_1$, *and of S given V=* $v_0$, *respectively.*

Note that additive independence assumes no interaction of the decision maker preferences for different amounts of the two attributes. This assumption is too restrictive in many real applications. It is often the case that desirability of one attribute increases (or decreases) with the level of the other. For data disclosure limitation problems, in particular, it seems reasonable to expect that desirability of safety (validity) increases with the level of validity (safety) and S and V are not utility independent. Note also that theorem 2 provides a necessary and sufficient condition for additive independence. This means that optimality criteria that can be expressed as an additive function of a measure of safety, S, and a measure of validity, V, implicitly assume that S and V are additive independent. The use of such criteria is difficult to justify since they formalize an independence assumption not realistic for many real applications.
The conditional utility functions that appear in theorems 1 and 2 might be either multidimensional or unidimensional. That is the arguments s and v can be scalars or a vectors. If they are unidimensional standard techniques in univariate utility theory (monotonicity, risk aversion, etc.) are appropriate. If they are vectors, hopefully it is possible to use independence properties of the components of s and v to decompose the assessment of the multiattribute conditional utility in simpler assessments of lower dimensional conditional utilities for the components of s and v.

## 5. Example 1

Suppose that a statistical agency needs to choose among *k* alternative forms of data release for a microdata. Suppose also that after an interview with the person/team in charge of confidentiality protection in the statistical agency, the analyst is able to define two subjective indices for the higher level objectives "maximize safety" and "maximize validity":

- S= % records that are protected (that is not correctly re-identified);
- V= % of the statistical analyses of interest, that have "good" quality compared with the corresponding analysis that could be performed using the original data;

Suppose also that it has been ascertained that:

(i) the multiattribute utility is increasing in both s and v;
(ii) data releases that protect less than 80% of records in the original data are very undesirable;
(iii) u(0,100)>u(100,0), that is a release that with certainty protects all records in the microdata but is useless for the users is preferred to a release that with certainty produces re-identification of all records in the original data but allows for good-quality statistical analysis for all the analyses of interest;
(iv) S is utility independent of V.

Assumptions (i)-(iii) seem reasonable for many disclosure limitation problems. Assumption (i) should be obvious. For a given level of safety (validity) release with higher validity (safety) are at least as preferred as release with a lower level of validity (safety). Assumption (ii) formalizes the idea of a threshold value for safety, very common in the practice of statistical agencies. Assumption (iii) also formalizes a common preference structure of statistical agencies: data releases that lead to complete disclosure are always less preferred than data releases with an acceptable level of disclosure, independently of the level of data utility. Assumption (iv) is not obvious, but as commented before should be reasonable in many situations. Under (iv) the multiattribute utility function will take the form:

$$u(s,v) = u(0,v)\cdot\left[1-u(s,0)\right]+u(s_1,v)\cdot u(s,0)\ ,$$

where $u(s,v)$ is normalized by $u(0,0)=0$, and $u(s_1,0)=1$.
The two conditions $u(0,0)=0$, and $u(s_1,0)=1$ specify the origin for the three conditional utilities and the unit measure for $u(0,v)$. Thus what it remains to specify are the unit measures for the two conditional utilities of V given S. This can be done, for example, finding a consequence $(s_2,0)$ that is indifferent to a consequence $(0, v_2)$, and a consequence $(s_3,0)$ that is indifferent to $(s_1, v_3)$. In this example we take: $s_1=80$, $s_2=70$; $v_2=100$; $s_3=100$; and for $v_3$ we distinguish three cases: $v_{3A}=40$; $v_{3N}=83$; $v_{3P}=94$. The condition on $s_2$, $v_2$ is equivalent to say that a data release that protects with certainty 70% of the records in the original data but is useless to the users is equally preferred to a release that disclose with certainty all the records but that guarantees good quality statistical results for all the analyses of interest ($s_2=70$; $v_2=100$). Similarly the condition $v=v_{3A}$ is equivalent to assume that a data release that protects with certainty 100% of the records in the original data but is useless to the users is equally preferred to a release that protect with certainty only 80% of the records but that guarantees good quality statistical results for 40% of the analyses of interest ($s_3=100$; $v_{3A}=40$). Corresponding interpretations hold when $v=v_{3N}$ and $v=v_{3P}$. Figure 2 shows the plots of conditional utility functions $u(s, v_0)$, $u(s_0, v)$, $u(s_1,v)$ that are compatible with the above assumptions together with the points that have been used to scale them properly. For $u(s_0, v)$, a risk averse utility function has been used while for $u(s, v_0)$ and $u(s_1,v)$ three different choices

are considered: *risk averse* (RA), *risk prone* (RP), and *risk neutral* (RN). These choices, as well as the choice of the scaling constants are clearly arbitrary in this example.

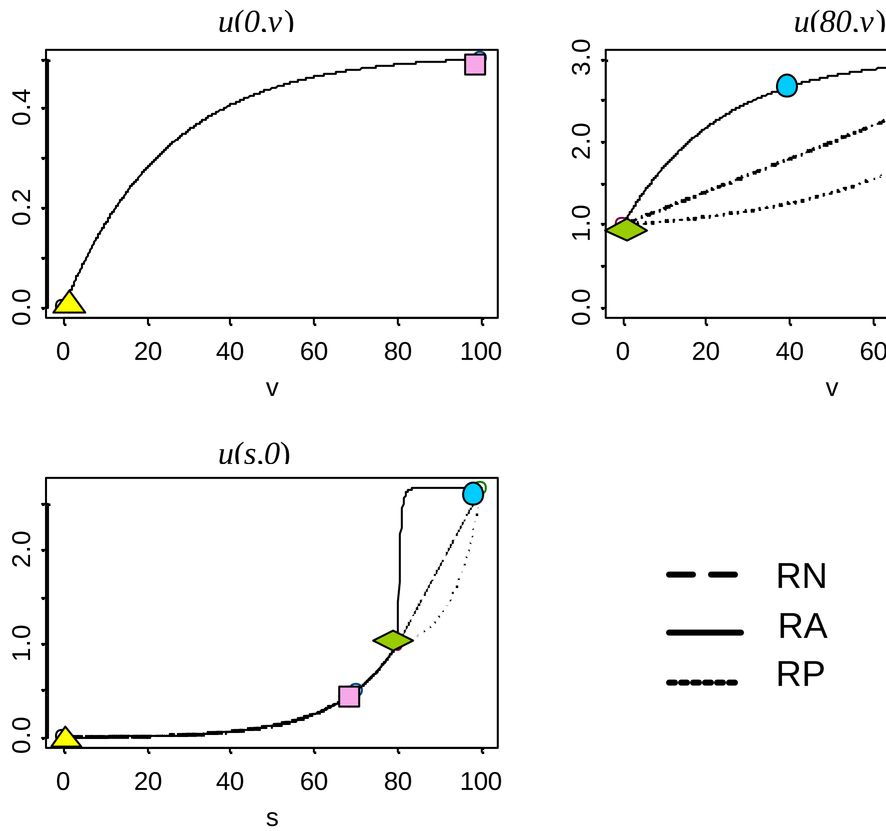


***Figure 1: Conditional utility functions, example1***

In real applications, the definition of scaling constants and conditional utilities should be part of the multiattribute assessment. The statistical agency should be asked to express its preferences for several lotteries on S given V= $v_0$, for V given S= $s_0$, and for S= $s_1$. Simple tests usually allow check certainty equivalents and whether the agency is risk adverse risk prone, risk neutral etc.

Given the alternative expression for u(s, $v_0$), and u($s_1$,v) there are nine possible combinations of the three conditional utilities. The corresponding multiattribute utility functions are shown in figures 3. Note that utility independence of S given V allows for substantial variability in the final form of the multiattribute utility function. Using different choices of the conditional utilities u(s, $v_0$), and u($s_1$,v) is possible to formalize quite different agency's preferences structures over alternative forms of data release. Note also that, under utility independence, multiattribute utility assessment did not require too much work. The problem of elicitating a suitable utility function has been decomposed in the simpler problem of assessing three one-dimensional utilities and identify equivalent pairs ($s_2$,0),(0, $v_2$), and ( $s_3$,0), ($s_1$, $v_3$). A relatively easy task.

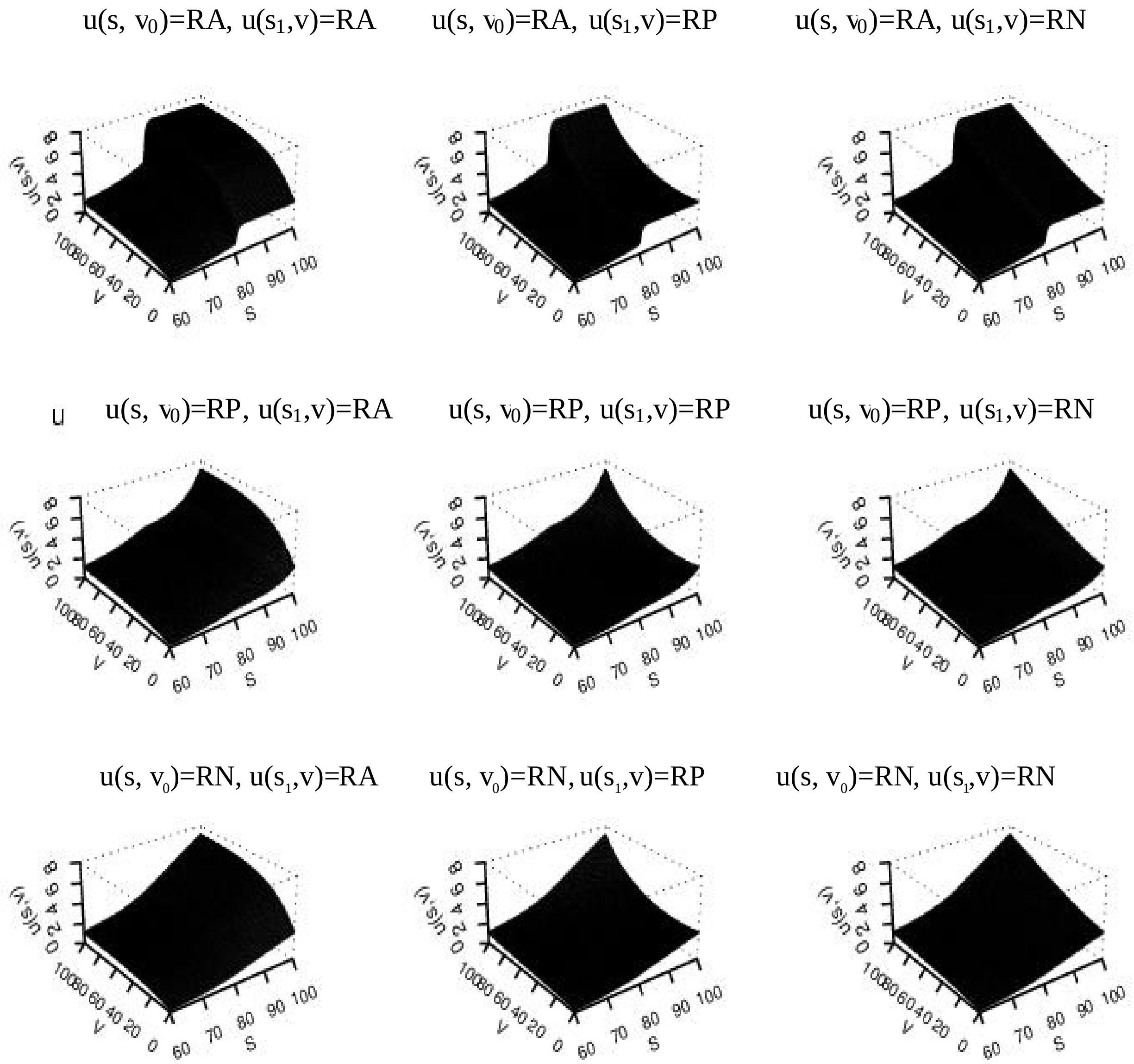


*Figure 2: Multiattribute utility function for different choices of the conditionals* $u(s, v_0)$, $u(s_1,v)$.

### 6. Example 2

Suppose that M is a microdata that contains a binary variable X that takes value 0 if the individual is HIV positive and zero otherwise. Suppose also that in assessing the disclosure risk the statistical agency has two main concerns, (a) avoid records re-identification, and (ii) avoid an intruder to infer (correctly or incorrectly) that a unit in the data is HIV positive. Safety, in this case, is a two dimensional vector, $S=(S_1, S_2)$ where: $S_1$ is the percentage of records correctly re-identified and $S_2$ is the percentage of units estimated to be HIV positive. Now suppose that there are only two possible intruder's attacks (for example there are two sets of key variables that can be considered) and that they are equally likely. Denoting by $s_{ij}$ the value of $S_i$ when the intruder uses attack j (i, j = 1, 2) each data release can be represented as a lottery that yields with equal probability $(s_{11}, s_{21})$, or $(s_{12}, s_{22})$. Standard approaches to data disclosure limitation in similar cases would try to define an heuristic index representing the disclosure risk associated with alternative data releases. Two intuitive solutions would be averaging over the possible outcomes, or considering the worst case scenario. In this example this would yield the following measures of disclosure risk:

Measure 1= $(s_{11}+s_{21}+s_{21}+s_{22})/4$; Measure 2= $\max(s_{11},s_{21},s_{21},s_{22})$.

Suppose, now, that the statistical agency has in mind two threshold values $t_1$ and $t_2$, ($t_1$, $t_2 \in (0,100)$), such that data releases that either yield a value of $S_1$ bigger than $t_1$ or a value of $S_2$ bigger than $t_2$ are not acceptable. Under this assumption (that seems reasonable in most of real applications) Measure 1 is not an appropriate measure of disclosure. Consider for example the two data releases $R_1$, that yields with equal probability ($S_1$=100, $S_2$=0) or ($S_1$=0, $S_2$=100), and $R_2$, that yields with equal probability ($S_1$=100, $S_2$=100) or ($S_1$=0, $S_2$=0). Under Measure 1 the two releases $R_1$ and $R_2$ are equivalent. For the statistical agency, however $R_2$ should be preferred to $R_1$ since $R_1$ yields with probability one and unacceptable level of disclosure (under $R_1$ with probability one either $S_1 > t_1$ or $S_2 > t_2$) while $R_2$ yields with equal probability an unacceptable level of disclosure ($S_1$=100, $S_2$=100) or zero disclosure ($S_1$=0, $S_2$=0). Measure 2 seems a more realistic measure. However, if $t_1$ and $t_2$ are not the same than Measure 2 is also inappropriate. Suppose, for example, that $t_1$=80 and $t_2$=60 and consider the two releases $R_1$ that yields with probability one ($S_1$=70, $S_2$=0) and $R_2$ that yields with probability one ($S_1$=0, $S_2$=70). Under Measure 2, $R_1$ and $R_2$ are equivalent. The statistical agency, however, would prefer $R_1$ since $R_2$ produces with probability one an unacceptable disclosure level ($S_2$=70>$t_2$=60). If the problem is formalized as a multiple objective decision problem, than empirical assessment of a measure of safety can be avoided. As described in section 5 under utility independence, assessment of preferences will require assessment of conditional utilities of S=($S_1$, $S_2$) given V=v', for specific values of v'. This conditional assessment can be done using utility independence of the components of S. In this example, for instance, it could be assumed that the conditional utility of S given V, takes its minimum, say zero, if either $S_1$ or $S_2$ is bigger than the corresponding threshold value ($t_1$ and $t_2$) and that $S_1$ and $S_2$ are mutually utility independent in the set $(0,t_1)\times(0,t_2)$. The elicitation of the conditional utility, in this case could be done imitating the steps described in section 5. Figure 1 shows an example of the conditional utility of S that can be obtained in such a way. The conditional utility behaves as we expect. If either $S_1$ or $S_2$ takes 'large' values the conditional utility of S, is small, and increases as both components increase.

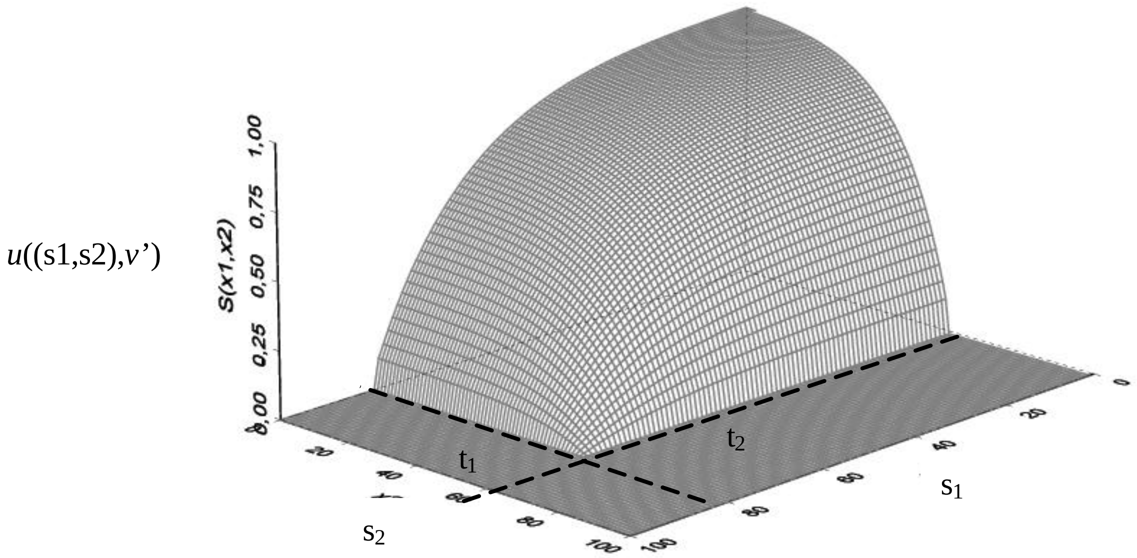


*Figure 3: conditional utility of S, example 2*

**7. Conclusions**

Data disclosure limitation is an important and complex problem. In addition to define new and better transformations of the data there is an increasing need for suitable criteria that allow comparison of alternative forms of data release. In this paper it is argued that the problem has a natural formalization as multiple objectives decision problem. The discussion has focused on elicitation of a multiattribute utility function and its relevance to define a rich class of optimality criteria. The approach presented in the paper has three main advantages: (i) it is based on solid theoretical ground; (ii) it takes into account the multidimensional and the random nature of *safety* and *validity*, and (iii) it allows decompose the multidimensional problem of preferences assessment into several simpler problems of lower dimension. The examples presented in sections 5 and 6 represent a simplified version of what a real problem would look like. The implementation of the whole procedure in real applications would require much more work and interaction both from the part of the decision maker and the analyst that wants to assess the utility. In the examples several important steps have been omitted. Assessment of the conditional utilities, for instance, has been only outlined and no consistency checking for coherence was made. Despite these simplifications, these examples should well illustrate the potential of multiple objectives decision theory to develop a rich class of optimality criteria for the choice of the best form of data release and to assess preferences for safety (validity) when S (V) is multidimensional. Implementation with real world data is currently under study and it will be presented in future works. Because of page limits constraints the problem of the definition of a suitable set of attributes is not addressed in the paper. The reader can refer to Keeney and Raiffa 1976 (chapter 2) for an exhaustive presentation of the problem. It would be desirable to have researchers working on data disclosure limitation, start using and presenting their proposals in terms of the hierarchy of objectives and attributes discussed there. This would beneficiate both understanding and comparability of new and existing methods.

**Acknowledgment**

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 contents of the paper reflect the author's personal opinions. The European Commission is not responsible for any views or results presented. I thank professors M.J. Bayarri and S.E. Fienberg for ideas and comments that are reflected in various ways in this work.

## References

- J. Domingo (editor), 2002, *Inference Control in Statistical Databases*, Lecture Notes in Computer Science, vol. 2316, Berlin: Springer-Verlag, pp. 34-58.
- P. Doyle, J. Lane, J.M. Theeuwes, L.V. Zayatz eds., 2001, *Confidentiality, Disclosure, and Data Access. Theory and Applications for Statistical Agencies*, North-Holland, Amsterdam
- D. Lambert, 1993, Measures of Disclosure Risk and Harm, *Journal of Official Statistics*, 9, pp. 313-333.
- R.L. Keeney and H. Raiffa , 1976, *Decisions with Multiple Objectives*, Wiley, New York.
- J. J. Salazar, 2002, Extending Cell Suppression to Protect Tabular Data Against Several Attackers, in *Inference Control in Statistical Databases*, J. Domingo editor, Lecture Notes in Computer Science, vol. 2316, Berlin: Springer-Verlag, pp. 34-58.
- L. Willenborg and T. De Waal, 1996, *Statistical Disclosure Control in Practice*, Springer-Verlag, New York.
- L. Willenborg and T. De Waal, 2001, *Elements of Disclosure Control*, Lecture Notes in Statistics, Vol 155, Berlin: Springer.