WP. 30
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 UNECE/Eurostat work session on statistical data confidentiality
(Geneva, Switzerland, 9-11 November 2005)

Topic (v): Confidentiality aspects of tabular data, frequency tables, etc.

SAFETY RULES IN STATISTICAL DISCLOSURE CONTROL FOR TABULAR DATA

Invited Paper

Submitted by the Office for National Statistics, United Kingdom[1]

[1] Prepared by Giovanni M. Merola.

Safety Rules in Statistical Disclosure Control for Tabular Data

Giovanni M. Merola

Office for National Statistics, 1 Drummond Gate, Methodology Directorate, SW1V 2QQ, London, UK [1].

Abstract. We extend the safety rules used for the Statistical Disclosure Control of magnitude tables to include an intruder who models the ignorance about an unknown confidential quantity with a Uniform distribution. By applying this extension to the generalised p-rule we obtain the safety rules useful also in the presence of groups of respondents. The corresponding disclosure rules for different prior knowledge of the intruder. The different safety rules are then compared to each other by considering some real Structural Business Statistics.

1 Introduction

Statistical Disclosure Control (SDC) consists of a variety of methods used to protect the privacy of respondents when confidential data are published. SDC is mainly applied by National Statistical Institutes (NSIs), but it is also applied by other entities that disseminate confidential data. In order to enforce the confidentiality agreements safely, values that can be estimated "closely" or that can be attributed with "high probability" are considered disclosed, so the values to be published are assessed in terms of *risk of disclosure*. Data-sets are cleared by *safety rule* that sets a level of acceptable risk for for each datum. Once disclosive values have been identified, the whole set of data is then protected with different techniques. Details on SDC theory and methods can be found, for example, in Willenborg and de Waal (2000). Protection of disclosive data unavoidably leads to the suppression or the distortion of some values, so the adoption of an appropriate measure of risk can avoid unnecessary damage to the data while protecting against disclosure (see, for example, Fienberg, 2000; Trottini, 2001; Duncan et al., 2001).

[1]Now at Winton Capital Management, 1-5 St. Mary Abott's Place, London W8 6LS, UK. g.merola@wintoncapital.com. This paper was mostly written while at ISTAT and was partially supported by the European Union project IST-2000-25069 "CASC". We would like to thank Dr. Luisa Franconi of ISTAT for making the data available and for her useful comments. The views expressed are those of the author only.

In this paper we consider the assessment of the risk of disclosure for non-negative values released as sums, which are often released in *magnitude tables*. Magnitude tables are published in large number and they can disclose contributions more easily than other tables. Therefore, SDC for these tables has received great attention in the literature and a computer package mainly devoted to the protection of this type of tables, τ-Argus (Hundepool, 2004), has been developed[2].

In SDC for magnitude tables it is assumed that an intruder with some *prior knowledge* is interested in learning some of the individual responses, in this context called *contributions*, that form a published sum. Cox (1981) defined four different measures of risk for magnitude tables and their properties and mutual relationships are considered, for example, in Willenborg and de Waal (2000); Federal Committee on Statistical Methodology (1994); Cox (2001); Loeve (2001); Merola (2003b). For the SDC of some data-sets it is necessary to consider the existence of groups, that is respondents that are connected and can communicate, must be taken into account when measuring the risk of disclosure. Natural examples of groups of respondents are households and industrial holdings. In this case an intruder may know the contributions of a group and be interested in the total of another group. Ways of including groups of respondents in Cox's rules are proposed in the papers cited above. Merola (2003a,b) extends one of these rules, the p-rule, to groups and shows that all the existing rules can be derived from this generalisation.

In the p-rule it is assumed that an intruder with the knowledge of one of the contributions estimates the largest contribution with its maximum possible value. In this paper, instead, we assume that the same intruder uses the prior knowledge to determine an interval of possible values for the largest contribution and that estimates it by minimizing the expected error. We hypothesise that the ignorance about the unknown quantities is modelled with a Uniform distribution and derive the safety rules for different specification of the prior knowledge. Since the rules so obtained consider safe also contributions with large vales, which are more identifiable than others, we extend the requirements to include large dominating contributions. The rules so obtained are stricter versions of the generalised p-rule.

In the following section we relate the existing safety rules to the identification of the respondents. In Section 3 we recall the generalised p-rule. In Section 3 we derive the new rules and in the following one we give a numerical comparison of the different rules. Finally, in Section 6 we give some final remarks.

[2] τ-Argus was created within the Computational Aspects of Statistical Confidentiality (CASC) project. It can be freely downloaded from the CASC Web Page at *http://neon.vb.cbs.nl/casc/*

2 Disclosure rules and identification of respondents

The identification of a respondent constitutes disclosure by itself when one or more of the categories defining a cell are confidential. For example, if one of the categories is "being infected with HIV". Respondents in a cell can be identified because they are known to have the characteristics defining the cell. The probability of this type of identification depends on the number of respondents in a cell. The safety rule that tackles this risk is the *threshold rule*, by which all the respondents belonging to a cells with less respondents than a given threshold are considered identifiable and the cell is considered disclosive.

Respondents can also be identified because they are known to carry a particularly large contribution. For example, it may be known which respondents have the two largest contributions in a cell. In some cases the shear presence of large contributions may lead to identification; for example, a very high total income for a group of people may give away the presence of a person with a much larger income than the others. Such large contributions are said to *dominate* the others and the rule that tackles this type of identification is called *Dominance rule*. By this rule a cell is considered disclosive if the sum of few of the largest contributions exceeds a certain percentage of the total, regardless of how closely the identifiable contributions can be estimated.

The rule that considers the precision of the estimation of a contribution is the p-rule. In this rule it is assumed that the largest respondent of a cell is identifiable and that the intruder knows the second largest contribution and estimates the largest one by its maximum possible values, that is by subtracting the known contribution from the total. The cells in which this estimate gives a relative error smaller than a given level, typically denoted by p -where from the name,- are considered disclosive.

We would like to stress that the p-rule can be applied only for the protection of the largest contribution as this estimating procedure cannot be extended to other contributions, as sometimes suggested. In fact, if T is the total and $z_1 \geq z_2 \geq z_3$ are the three largest contributions, then z_2 will not be estimated by $T - z_3$, simply because $z_2 \leq T/2$ and $T - z_3 > T/2$. Hence, the p rule properly protects the estimation of the largest contribution. The last of the rules currently used, the so called pq-rule, can be considered a stricter version of the p-rule (*e.g.* Merola, 2003a).

3 The M-rule

In Merola (2003a) we generalise the p-rule to the existence of groups, considering the subtotals of each group as a confidential datum. It could be the case of medical expenses grouped for household, for example. Let T be the published cell total and

$z_1 \geq z_2 \geq \ldots \geq z_n$ be the n ordered contribution, so that $T = \sum_{i=1}^{n} z_i$. We assume that the intruder wants to estimate the sub-total of the m largest contributions, denoted with $t_m = \sum_{i=1}^{m} z_i$, knowing the total of the subsequent l largest ones, denoted by $R_{m,l} = \sum_{i=m+1}^{m+l} z_i$. The largest estimate of t_m is given by

$$\hat{t}_m = T - R_{m,l} = t_m + r_{m+l}, \tag{1}$$

where the remainder $r_m = \sum_{i=m+1}^{n} z_i$ (with $r_0 = 0$) is the estimation error. Like in the p-rule we require that the relative estimation error, denoted by $RE(t_m; l)$, is larger than the level p, with $0 \leq p < 1$, that is:

$$RE(t_m; l) = \frac{|t_m - \hat{t}_m|}{t_m} \tag{2}$$

The generalised p-rule is obtained substituting the estimate (1 into requirement (2:

$$M_p(m; l) : \frac{t_m}{T - R_{l,m}} \leq \frac{1}{1+p}, \tag{3}$$

where the subscript p is used to denote the protection level (but will be omitted when not needed) and l denotes the number of known contributions and will be omitted when equal to zero. The symbol M denotes that the estimate is the maximum possible value. Henceforth we will refer to this rule as the M-rule.

As shown in Merola (2003a), all the existing rules are special cases of the M-rule. The *threshold* rule for m respondents protects against exact disclosure, that is $p = 0$, when the intruder knows l contributions and wants to estimate $m - l$ contributions. The requirement for the $M_0(m - l; l)$ rule is $RE(l; m - l) = (T - R_{l,m-l} - t_m)/z_1 = r_m/t_{m-l} > 0$. It is satisfied if the respondents are more than m or if the reminder r_m is greater than zero. So, this formulation, sensibly, extends the Threshold rule to cells with all zero contributions after the m-th.

The Dominance rule can be obtained by assuming that the intruder does not know any of the contributions, hence by setting $l = 0$. For this case the $M_p(m)$ rule is $t_m/T \leq 1/(1+p)$. As already noted (*e.g.* Cox, 2001), in this way it is possible to express the requirement of the Dominance rule in terms of the minimum relative error of estimation.

The p-rule can be obtained straightforwardly by setting $m = l = 1$. The $M_p(1; 1)$ rule requires that $z_1/(T - z_2) \leq 1/(1+p)$. The pq-rule corresponds to the p-rule with protection level equal to p/q, that is $M_{p/q}(1; 1)$. One desirable property of safety rules is sub-additivity, introduced by Cox (1981). By transforming rules in *linear sensitivity measure* he shows that a rule is sub-additive if and only if the corresponding sensitivity measure has nonincreasing coefficients. It can be easily

shown that the generalised p-rule is sub-additive for all values of m and l (Merola, 2003a).

The maximizing estimation procedure assumed in the M-rule, in some cases, may not be realistic. In the next section we derive safety rules under the assumption of a different estimating procedure allowing different prior knowledge to the intruder.

4 The MU-rules

Let us assume that the intruder is interested in estimating t_m and uses the prior knowledge to restrict its possible value within bounds, say $t_m^- \leq t_m \leq t_m^+$. If $F(t_m)$ is the distribution of t_m over this interval, the estimate can be obtained by minimizing the mean squared error (MSE), that is $\int_{t_m^-}^{t_m^+}(t_m - \hat{t}_m)^2 dF(t_m)$. In this paper we assume that the intruder does not know the distribution and that models this ignorance by taking it to be a Uniform over $[t_m^-, t_m^+]$ (for a discussion on modelling ignorance over a finite interval see, for example, Bernardo and Smith, 1994). Then, the estimate that minimises the MSE is

$$\hat{t}_m = \frac{t_m^- + t_m^+}{2}, \tag{4}$$

for a well known property of the mean. Of course, other distributions may lead to different estimates but this estimate would be equally optimal for other symmetric distributions, such as the truncated Normal, for example.

Given estimate (4), the safety rules are derived by requiring that RE is not less than the safety level, p, that is $RE = |\hat{t}_m - t_m|/t_m \geq p$. In Merola (2003b) we show that these conditions are not satisfied by values of t_m within an interval, say $lb \leq t_m \leq ub$. This means that large values of t_m would be considered safe. Since it is plausible to assume that an intruder may know that some values dominate, - that is may have assumptions on the distribution of t_m - s/he could take a maximising estimate. Therefore, we extend the rule to include also this case by dropping the requirement that $t_m < ub$. The resulting rule is in the form $t_m < ub$. We name the resulting rules *MU-rules* as they protect against absolute relative error larger than p for any estimate of t_m that lays between its maximum and the estimate (4).

We now give the MU-rules for different possible prior knowledge of the intruder, without details of the derivation, which can be found in Merola (2003b). We always assume that the intruder has the *basic knowledge* of the cell total, T, and that the number of contributions, n, is larger than m. On top of this we consider four other cases given by the combinations of whether the number of contributions is known and whether one or more contributions are known. Since there is not a generalised solution for all the cases, we will consider the different scenarios separately. The

safety rules will be generically denoted by $MU(m;\cdots)$, where "$\cdots$" are parameters specifying the extra prior knowledge, if present.

$MU(m)$: the MU-dominance

This is the same prior knowledge as for the Dominance rule. With this knowledge t_m can only be bounded by $0 \leq t_m \leq T$. Substituting these values in (4) gives $\hat{t}_m = T/2$. The resulting $MU(m)$ rule is not satisfied when

$$\frac{t_m}{T} \geq \frac{1}{2(1+p)}. \tag{5}$$

Hence, this is a stricter version of the Dominance with half minimum level for the ratio t_m/T.

$MU(m;n)$: the MU-dominance when n is known

This is the prior knowledge of the Dominance with the addition of the number of contributions. As it can be easily verified$nT/m \leq t_m \leq T$, so an intruder knowing T, n and m can bound t_m by $m\bar{T} \leq t_m \leq T$, where $\bar{T} = \sum_{j=1}^{n} z_j/n$ is the average contribution. Substituting these values in (4) gives $\hat{t}_m = T + m\bar{T}/2 = ((n+m)/2n)T$. The resulting $MU(m;n)$ rule is not satisfied when

$$\frac{t_m}{T} \geq \left(\frac{n+m}{2n}\right)\frac{1}{(1+p)}. \tag{6}$$

This rule is slightly less strict than the $MU(m)$ but it is stricter than the Dominance. In this rule, appropriately, the requirement on the concentration of the cell changes with the number of contributions.

$MU(m;l)$: the MU-p-rule

When $l > 0$ contributions are known, z_{m+1} is the largest contribution known to the intruder. Hence, the value of t_m can be bounded by $mz_{m+1} \leq t_m \leq T - R_{l,m}$. Substituting these values in Equation (4) gives $\hat{t}_m = (T - R_{l,m} + mz_{m+1})/2$. So the the $MU(m;l)$ rule is not satisfied when

$$\frac{t_m}{T - R_{l,m} + mz_{m+1}} \geq \frac{1}{2(1+p)}.$$

This condition is stricter than the corresponding ones for the M-rule. The equivalent of the simple p-rule, $l = m = 1$, the rule reduces to the $MU(1)$ rule, the MU analogous of the Dominance $MU(1)$, because $R_{l,m} = mz_{m+1} = z_2$.

$MU(m;l,n)$: the MU-p-rule when n is known

In this scenario the intruder has maximum knowledge. However when $mz_{m+1} \geq T - R_{l,m} - (n-m-l)z_{m+l}$, the rule reduces to the $MU(m;l)$ because the additional knowledge of n does not improve the bounds on t_m. In the other case, as t_m is estimated by $\hat{t}_m = T - R_{l,m} - ((n-m-l)z_{m+l})/2$ and $MU(m;l,n)$ rule is not satisfied when

$$\frac{t_m}{T - R_{l,m} - (\frac{n-m-l}{2})z_{m+l}} \geq \frac{1}{1+p}.$$

This bound is active for $m = l = 1$.

The resulting MU-rules are shown in Table (1) together with the corresponding *linear sensitivity measures* (LSM), which are a way of representing the safety rules as a linear combinations of the single contributions. From the sensitivity measures it

rule	bound	Sensitivity measure
[Thresh] $M_0(l; m-l)$	$\frac{t_m}{T} < 1$	$-r_m$
[Dom] $M(m)$	$\frac{t_m}{T} < \frac{1}{1+p}$	$pt_m - r_m$
[Gen. p] $M(m;l)$	$\frac{t_m}{T-R_{l,m}} < \frac{1}{1+p}$	$pt_m - r_{m+l}$
$MU(m)$	$\frac{t_m}{T} < \frac{1}{2(1+p)}$	$(1+2p)t_m - r_m$
$MU(m;n)$	$\left(\frac{n}{n+m}\right)\frac{t_m}{T} < \frac{1}{2(1+p)}$	$(n(1+2p)-m)t_m - (n+m)r_m$
$MU(m;l)$	$\frac{t_m}{T-R_{l,m}+mz_{m+1}} < \frac{1}{2(1+p)}$	$(1+2p)t_m - mz_{m+1} - r_{m+l}$
$MU(m;l,n)^*$	$\frac{t_m}{T-R_{l,m}-(\frac{n-m-l}{2})z_{m+l}} < \frac{1}{1+p}$	$pt_m + \frac{(n-m-l)z_{m+l}}{2} - r_{m+l}$

Table 1: Safety bounds and corresponding sensitivity measures for different rules. The asterisk denotes that the bound is active only if conditions are satisfied.

can be seen that the $MU(m)$ and $MU(m;n)$ are subadditive, while the $MU(m;l)$ and $MU(m;l,n)$ rules, that assume some contributions are known, are not. Thus, these last two rules might not be appropriate for the protection of whole tables. However, the cases in which merging two cells safe with respect to these rules results in a disclosive one, can be considered rare. One of the reasons is that it seems unlikely that the contributions known to the intruder will maintain the same rank in the merged cell. This is to say that, in general, the knowledge on specific cells may not be the same as the one on two merged cells and, therefore, in some cases the use of non subadditive rules can be justified.

5 Numerical Comparison of Different Safety Rules

As an example, consider a cell with $n = 12$ contributions (970, 376, 274, 253, 203, 169, 161,121, 86, 62, 21, 10), so that $T = 2706$ and $z_1/T = 0.36$. The estimates

considered in the Dominance and the p-rule give relative error of estimation for z_1 equal to 1.8 and 1.4 respectively. However, RE for the $MU(1)$ and $MU(1;1)$ rules is equal to 0.4 while it is 0.5 for the $MU(1;0,n)$ rule.

We compared the different rules on the turnover classified by geographical region and NACE with two and three digits from the Italian Structural Business Statistics surveys of enterprises with 20 or more employees for the years 1994 and 1997 (ISTAT, 1997, 2001). First we considered the protection of a single contribution from an intruder with the knowledge of at most one contribution, hence ignoring enterprise groups. The average REs obtained for the different rules are shown in Figure (1); clearly the MU-estimates yield a much lower average RE in all cases.

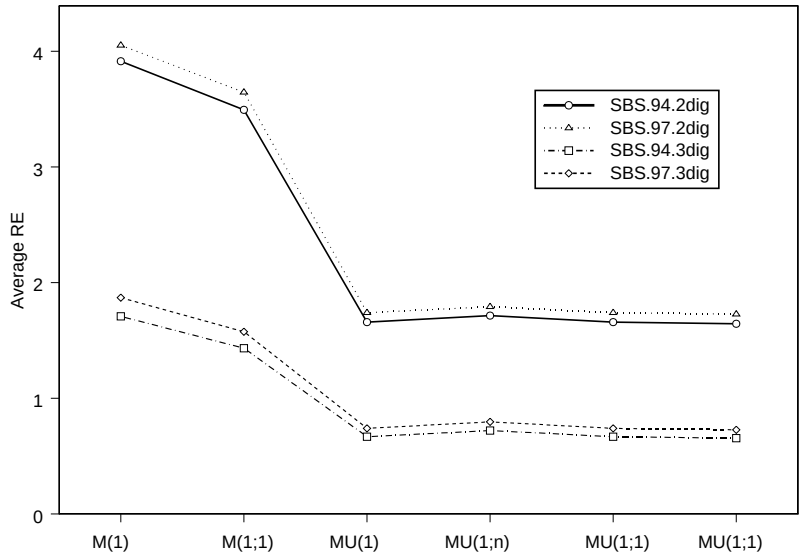

Figure 1: Average relative absolute error committed by one intruder estimating z_1 for different rules.

Table (2) shows the percentage of nonempty cells disclosive at a safety level $p = 0.5$ for the different safety rules. We included also the Dominance $M(2)$ because this is sometimes considered alternative to the p-rule - but it is stricter. The table clearly shows that the number of disclosive cells increases drastically when using the MU-rules. The difference among these is small, though. As expected the tables with finer partition (NACE with 3 digits) present a higher number of disclosive cells.

Rule	NACE 2 dig.		NACE 3 dig.	
	SBS 94	SBS 97	SBS 94	SBS 97
$M_0(1;4)$(Threshold)	14.07	13.52	29.61	29.16
$M(1)$	7.73	7.26	6.68	6.81
$M(1;l=1)$	13.49	14.02	15.48	15.22
$M(2)$	19.26	18.65	21.63	20.68
$MU(1)$	30.33	28.54	32.97	32.05
$MU(1;n)$	25.61	24.28	26.33	25.45
$MU(1;l=1)$	30.33	28.54	32.97	32.05
$MU(1;l=1,n)$	30.33	28.54	32.72	31.94

Table 2: Percentage of unsafe cells for different safety rules requiring that $RE(z_1) > 0.5$.

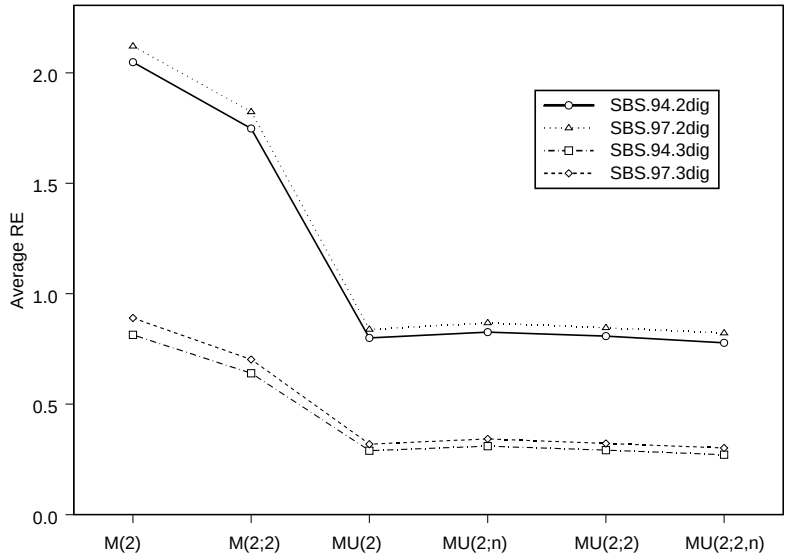

Figure 2: Average relative absolute error committed by one intruder estimating t_2 for different rules.

We also compared the rules assuming the existence of two groups of two respondents in every cell. Hence we applied the rules with $m = 2$ and $l = 2$, together with the threshold rule $M_0(5)$ on the same data. Figure (2) compares the RE obtained with the different estimating procedures and table (3) shows the percentages of nonempty disclosive cells. Again, it is evident how the estimates obtained with the Uniform distribution give much lower average RE than the maximal estimates assumed in the M-rules, and the number of disclosive cells found with the MU-rules is much larger than for the corresponding M-rules.

Rule	NACE 2 dig.		NACE 3 dig.	
	SBS 94	SBS 97	SBS 94	SBS 97
$M_0(2;3)$(Threshold)	22.38	21.78	43.29	42.92
$M(2)$	12.11	11.14	10.14	8.84
$M(2;l=2)$	22.49	20.53	20.92	21.11
$MU(2)$	39.56	37.67	36.64	35.97
$MU(2;n)$	36.10	33.54	32.47	31.59
$MU(2;l=2)$	39.22	36.55	36.29	35.40
$MU(2;l=2,n)$	39.22	36.55	36.29	35.40

Table 3: Percentage of unsafe cells for different safety rules requiring that $RE(t_2) > 0.5$ when it is estimated by a coalition of two intruders.

6 Conclusions

The protection provided by a safety rule depends on the assumptions taken to measure the risk. We show that contributions can be disclosed more precisely than what assumed in the M-rules by adopting a simple ignorance distribution. By clearly specifying the intruder's prior knowledge, safety rules can be adapted to different scenarios. In some situations it is appropriate to expand the safety rules to include groups of respondents. However, taking more stringent hypothesis lead to stricter rules, thus it is important to choose the rules for plausible hypothesis, rather than mechanically apply them, as sometimes may happen. The Uniform distribution used to derive the MU-rules is likely to be inappropriate for many data-sets, other, possibly skewed, distributions could be used and other rules may be derived.

References

Bernardo, J., Smith, A., 1994. Bayesian Theory. Wiley, NY.

Cox, L. H., 1981. Linear sensitivity measures in statistical disclosure control. Journal of Statistical Planning and Inference 5, 153–164.

Cox, L. H., 2001. Disclosure risk for tabular economic data. In: Doyle, P., Lane, J., Theeuwes, J., Zayatz, L. (Eds.), Confidentiality, disclosure and data access: theory and practical application for statistical agencies. Elsevier Science.

Duncan, G., Keller-McNulty, S., Stokes, S., 2001. Disclosure risk vs. data utility: the R-U confidentiality map. technical Report LA-UR-01-6428, Los Alamos National Laboratory.

Federal Committee on Statistical Methodology, 1994. Report on statistical disclosure limitation methodology, working paper 22.Subcommittee on Statistical Limitation Methodology, Washington, DC.

Fienberg, S. E., 2000. Confidentiality and data protection through disclosure limitation: Evolving principles and technical advances. The Philippine Statistician 49, 1–12.

Hundepool, A., 2004. The argus software in the casc project. privacy in statistical databases. Proceedings of the CASC Project International Workshop, June 9-11, 2004, Barcelona, Spain, 323–335.

ISTAT, 1997. Conti economici delle imprese con 20 addetti ed oltre. Anno 1994. Vol. 41 of Collana Informazioni. Istituto Nazionale di Statistica, Roma.

ISTAT, 2001. Conti economici delle imprese. Anno 1997. Vol. 19 of Collana Informazioni. Istituto Nazionale di Statistica, Roma.

Loeve, J. A., 2001. Notes on sensitivity measures and protection levels. research paper no. 0129. Methods and Informatics Department, Statistics Netherlands, Voorburg.Available at the CASC project home page *http://neon.vb.cbs.nl/casc/*.

Merola, G. M., 2003a. Generalized risk measures for tabular data. Proceedings of the 54th Session of the International Statistical Institute.

Merola, G. M., 2003b. Safety rules in statistical disclosure control for tabular data. Contributi Istat 1, istituto Nazionale di Statistica, Roma.

Trottini, M., 2001. A decision-theoretic approach to data disclosure problems. Researchin Official Statistics 4, 7–22.

Willenborg, L., de Waal, T., 2000. Elements of Statistical Disclosure Control. Springer-Verlag, New York.