Information loss through global recoding and local suppression

Article · January 1999

CITATIONS
23

READS
253

2 authors:

Ton De Waal
Centraal Bureau voor de Statistiek
135 PUBLICATIONS **1,900** CITATIONS
SEE PROFILE

Leon Willenborg
Centraal Bureau voor de Statistiek
124 PUBLICATIONS **1,797** CITATIONS
SEE PROFILE

Information loss through global recoding and local suppression

Ton de Waal and Leon Willenborg [1]

1. Introduction

Before a microdata set can be disseminated safely by a statistical office it has to be protected against disclosure. More precisely, the risk of disclosure of confidential information should be sufficiently low before a microdata set may be released. Statistical disclosure control (SDC) aims to limit the disclosure risk. This can be done, as is the case at Statistics Netherlands, by checking whether certain combinations of scores occur frequently enough in the population (see e.g. De Waal and Willenborg, 1996b, Willenborg and De Waal, 1996). If such a combination does not occur frequently enough in the population the disclosure risk of the microdata set under consideration is considered too high and appropriate SDC measures should be taken.
A number of SDC measures can be taken to protect a microdata set: recoding, suppression and perturbation, for example. Recoding is collapsing categories of a variable, suppression is the replacement of a value in a record by a missing value, and perturbation is the replacement of one value by another one. Protecting a microdata set by one of these measures leads to a loss of information. Our aim is to retain as much information in the microdata set while making this data set safe. When microdata are interactively protected using μ-ARGUS the data protector employs an intuitive meaning of information loss. If the automatic mode of μ-ARGUS is used – in which the best combination of local suppressions and global recodings has to be found by the package itself – a quantification of information loss should be used that allows the package to make the necessary decisions and trade-offs. To achieve this it is necessary to quantify the information loss due to an SDC measure.
In this article we consider two methods to quantify the information loss in a microdata file due to local recoding, global recoding, or local suppression. One is objective and uses the entropy concept. The other is subjective, and uses weights specified by the data protector. These weights express the predilections of the data protector (who should be enlightened by the users' needs!) for preserving the information of certain variables over others and certain predefined codings for a variable over others.
The basic idea to evaluate the information loss formally is to use the entropy function for some suitably chosen probability distribution. This probability distribution is a stochastic model for the changes attributable to the applied SDC measures. The basis for the probability function is a 'transition probability matrix' for a variable. Such a matrix gives (an estimate of) the probability that an 'old' value in a record is changed to a particular 'new' one as a result of a modification of the microdata. The matrix for a particular variable in a file may be estimated by assuming a model for the changes due to the SDC measures and by estimating the corresponding probabilities. To estimate the probabilities it is necessary to make an assumption about the available information on which these estimates are based. For instance, these probabilities may be estimated by comparing the old and the new files and counting the number of changes that have occurred in the file with respect to the variable(s) under consideration.
With our discussion of information loss in terms of entropy we start by considering information loss caused by a technique that can be seen as more elementary than both global recoding and local suppression, namely local recoding. A subjective method for measuring information loss is described briefly in Section 5. Both these information loss measures are used in μ-ARGUS (cf. Hundepool et al., 1998) in the 'automatic mode' when using a mixture of global recoding and local suppression. The article concludes with a short discussion in Section 6.

2. Information loss due to local recoding

In order to be able to define the information loss due to global recoding and local suppression we first consider a related but simpler action, which we shall refer to as local *recoding*. We define local recoding as recoding a variable for *one* record only, whereas by global recoding we mean that a variable is recoded for *all* records in which one of the recoded categories occurs.
Suppose that a certain combination of scores in the file, e.g. 'age = 17' and 'marital status = widowed' does not occur frequently enough in the population. The records in which this combination occurs have to be protected. This can be achieved, as far as this particular combination is concerned, by recoding the variable 'marital status' in these records. For instance, the value 'marital status = widowed' may be replaced by 'marital status = widowed or divorced', assuming that the combination 'age = 17' and 'marital status = widowed or divorced' occurs frequently enough. In this case there is some uncertainty about the original value of 'marital status' for a user of the microdata set.
Now suppose that we can assign a probability p'_W to the event that the original value of 'marital status' equals 'widowed' given that the new, recoded, value equals 'widowed or divorced', and a probability p'_D that the original value equals 'divorced'. That is,

$$p'_W = \frac{p_W}{p_W + p_D} \tag{2.1}$$

and

$$p'_D = \frac{p_D}{p_W + p_D}\ , \tag{2.2}$$

where P_w is the probability that the original value of 'marital status' in the record under consideration equals 'widowed', and P_d is the probability that the original value of 'marital status' equals 'divorced'.

In this case the entropy H_{ms}, i.e. the information loss due to local recoding of 'marital status', is given by

$$H_{MS} = -p'_W \log(p'_W) - p'_D \log(p'_D). \tag{2.3}$$

In general, when categories C_1, C_2,..., C_n of a variable V in a particular record are combined into a single one, denoted as $C_1+C_2+...+C_n$, then the information loss H_r due to this local recoding is given by

$$H_r = -\sum_{i=1}^{n} p'_i \log(p'_i), \tag{2.4}$$

where p'_i is the conditional probability that the original value of V in the record under consideration is equal to C_i given that the recoded value equals $C_1+C_2+...+C_n$. That is,

$$p'_i = \frac{p_i}{\sum_{i=1}^{n} p_i}\ , \tag{2.5}$$

where p_i is the probability that the original value of V in the record under consideration equals C_i.

So far we have only considered the situation where one variable in one record is recoded. Now we consider the case where several variables are recoded in one record. When variables $V_1, V_2, \ldots, V_m$ are recoded in a particular record k, then the information loss IL_k^r in record k due to these local recodings is measured by

$$IL_k^r = -\sum_{i_1 i_2 \ldots i_m} P(C_{1i_1}, C_{2i_2}, \ldots, C_{mi_m}) \log(P(C_{1i_1}, C_{2i_2}, \ldots, C_{mi_m})), \quad (2.6)$$

where $P(C_{1i_1}, C_{2i_2}, \ldots, C_{mi_m})$ is the simultaneous probability distribution of $V_1, V_2, \ldots, V_m$. When we make the simplifying assumption that the variables $V_1, V_2, \ldots, V_m$ are independent, then the information loss IL_k^r in record k due to the local recodings is measured by the sum of the information losses due to the individual local recodings, i.e.

$$IL_k^r = \sum_{j=1}^{m} IL_{kj}^r, \quad (2.7)$$

where IL_{kj}^r denotes the information loss due to the local recoding of variable V_j, which is given by (2.4).

Lastly, we consider the case where several variables are recoded in several records. When variables are recoded in records 1, 2,..., K then we measure the total information loss IL_{tot}^r due to local recodings in these records by the sum of the information losses in the individual records, i.e.

$$IL_{tot}^r = \sum_{k=1}^{K} IL_k^r, \quad (2.8)$$

where IL_k^r denotes the information loss due to the local recodings applied to record k, which is given by (2.7). In fact, we assume here that local recodings in different records are independent.

In order to be able to calculate the information loss one needs to estimate the probabilities that appear in the entropy expression. We consider a crude model for this. Suppose that some of the old categories $C_1, C_2, \ldots, C_s$ of a variable V are combined in such a way that the new categories are given by $D_1, D_2, \ldots, D_m$ ($m \leq n$), where each D_j is a combination of one or more C_i's. To evaluate the information loss due to this recoding we need to estimate the probability p_{ij}^k that the old, original, category equals C_i given that the new, recoded, category equals D_j in a particular record k. The index k indicates that
depends on (the values in) the record under consideration. We shall assume that D_j is obtained by collapsing $C_1, C_2, \ldots$, and C_s. The probability p_{ij}^k *can then be estimated by*

$$\hat{p}_{ij}^k = \frac{n_i}{\sum_{i=1}^{s} n_i}, \quad (2.9)$$

where n_i denotes the number of times that C_i (i=1,...,s) occurs in the original, unprotected, microdata set. Note that this estimate does not depend on (the values in) record k.

3. Information loss due to global recoding

In practice, local recodings are hardly ever applied, because they lead to a rather odd kind of microdata set in which the categorisation of each variable can differ per record. Local recoding is used as a stepping stone to global recoding. Instead of local recodings one usually applies global recodings. This means that when a variable V is recoded in a particular record, then this recoding is applied to variable V in all records in the microdata set, thus achieving a uniform categorisation, i.e. the categorisation of a variable is the same for each record. For instance, when the categories 'widowed' and 'divorced' of the variable 'marital status' are collapsed into the single category 'widowed or divorced', then this is done for all records in which the value of 'marital status' equals 'widowed' or 'divorced'.

Measuring the information loss due to a global recoding is rather easy once the information loss due to local recodings has been defined, because a global recoding can be seen as a number of local recodings that have been applied to all records in the microdata set. Suppose that a variable V is recoded by combining some of the old categories $C_1, C_2, \ldots, C_s$ such that the new categories are given by $D_1, D_2, \ldots, D_m$ ($m \leq n$), where each D_j may be a combination of several C_i's. In that case the information loss IL_V^r due to recoding V can be measured, in each record in which the value of V equals D_j, by

$$IL_V^r = -\sum_{i=1}^{n} p_{ij} \log(p_{ij}), \quad (3.1)$$

where p_{ij} denotes the probability that the original category of V in the record under consideration is equal to C_i given that the new category equals D_j. The total information loss in all records due to global recodings can be measured again by (2.8).

4. Information loss due to local suppression

A combination in a microdata set that does not occur frequently enough can also be protected by local suppression, i.e. one or more values in this combination can be deleted. For example, when the combination 'age=17' and 'marital status = widowed' does not occur frequently enough, then this combination can be protected by replacing the value of 'marital status' by a missing value. Local suppression of the value of a variable V is not applied to all records in a microdata set, but to some of the records only.
The information loss due to local suppressions can be measured in different ways. When local suppressions are not applied in combination with global recoding then the situation can be treated relatively easily. The information loss may be expressed as a weighted sum of the numbers of locally suppressed categories. When local suppressions are applied in combination with global recodings then the situation is more difficult. In this case the entropy should again be used to measure the information loss. Both situations, local suppressions not in combination with recodings and local suppressions in combination with recodings, are examined below. It should be noted that the entropy could also be used to measure the information loss when only local suppression is applied. Using the entropy is better from a theoretical point of view, but is more difficult to apply in practice.
De Waal and Willenborg (1994, 1998) considered the problem of finding the minimum number of local suppressions to eliminate a set of unsafe (=rare) combinations in a microdata file. The number of local suppressions was taken as a measure for the information loss due to the suppressions. In these papers this problem was extended to eliminating a set of unsafe combinations while minimising a more general linear target function. For instance, to each variable V_i a weight w_i can be assigned. The information loss in the microdata set due to the local suppressions is then measured by $\sum_i w_i s_i$, where the sum is taken over all variables and s_i equals the number of times that a value of variable V_i is suppressed. Using this linear target function instead of the non-linear entropy has the advantage that the problem of determining the optimal local suppressions is reduced to solving a 0-1 mixed integer programming problem, for which several algorithms are available.
The situation is different in for example Hurkens and Tiourine (1998a,b), where the goal is to find the optimum mix of local suppressions and global recodings to eliminate a given set of unsafe combinations. In this case it was necessary to find a trade-off in information loss due to either type of action. To be able to do this the entropy was introduced.
It is very simple to quantify information loss due to local suppression by means of the entropy once one realises that local suppression is an extreme form of local recoding, namely all categories of a variable are collapsed into a one single category. Information loss for the suppression of a value of variable V, having n categories, in a particular record is

$$H_V^s = -\sum_{i=1}^{n} p_i \log(p_i) \qquad (4.1)$$

where p_i denotes the probability that the original value of variable V in the record under consideration equals C_i.
The total information loss due to local suppressions in record k, H_k^s, is given by formula (2.7) with $H_{V_i}^r$ replaced by $H_{V_i}^s$. The total information loss due to local suppressions in all records, H_{tot}^s, is given by (2.8) with H_k^r replaced by H_k^s. We measure the total information loss due to global recodings and local suppressions by

$$H_{tot}^{r+s} = H_{tot}^r + H_{tot}^s. \qquad (4.2)$$

Suppose that a value C_1, C_2,..., or C_n of variable V is suppressed in a record k. To evaluate the information loss due to this local suppression we need to estimate the probability p_i^k that the original category equals C_i. Again the index k indicates that, in principle, this probability p_i^k depends on the record under consideration. The probability p_i^k can be estimated in the following crude way.

$$\hat{p}_i^k = \frac{m_i}{m_V}, \qquad (4.3)$$

where m_i equals the number of times that C_i has been suppressed and m_V equals the number of times that a value of variable V has been suppressed. Note again that this estimate does not depend on the particular record k.

5. Subjective information loss measure: weights

In μ-ARGUS it is possible to use an automatic mode for protecting a microdata set. In this case the program searches for an optimal mix of global recodes and local suppressions to protect the microdata set. If the program is to perform this task, the data protector should make the necessary preparations, including specifying for each identifying variable a set of possible predefined codings (e.g. for a regional variable codings at the municipality, county, province and area are possible). Then for each variable he should indicate how important each variable is for him, by specifying weights for each variable. Also, for each variable he should indicate how important each of the alternative codings is for him, again by specifying weights. The weight that the system then uses for a particular coding of a particular variable is (proportional to) the product of the variable weight and the coding weight. Furthermore, the user should specify a weight for each identifying variable indicating how important it is that a value of this variable is suppressed.[2] Unfortunately, the data protector is offered no guidance in specifying these weights in the current version of μ-ARGUS, but he has the full power of this weights approach available.

6. Discussion

The operational definition of the entropy-based model that we propose is based on the point of view of a statistical office, not that of the user of the published data. This is a consequence of the fact that to make the probability estimates one needs the original data as well, and these are obviously only at the disposal of a statistical office and not of data users. This is not restrictive in view of the particular application that we have in mind for this measure: as a kind of steering wheel in μ-ARGUS when the system is being used in automatic mode, finding the optimum mix of global recodes and local suppressions. In this mode safety of the resulting microdata is a goal, but no more modifications should be carried out than necessary. The information loss function is defined to steer the data modification process so as to make sure that enough interesting information will be left in the data.

The attractive thing about this entropy-based information loss approach is that it is general and versatile, deriving the information loss for various modification techniques such as global recoding and local suppression from a common principle, thereby making direct comparison in information losses due to different data modification techniques possible. As is shown in De Waal and Willenborg (1996a) - the paper from which the present article is derived - this method can also be used to quantify information loss caused by a perturbative method (such as PRAM, see Gouweleeuw et al., 1998).
The entropy-based measure of information loss is objective, as it is only based on objective information such as variables, domains, and objective probabilities over these domains. There is no option for a data protector to express his preferences for certain variables or certain recodings. Therefore we have discussed a second type of information loss model in the present article, which can actually be called entirely subjective. The model allows the user to express his preferences (over variables, values or codings) in terms of weights. The disadvantage of this approach – which is not shared by the entropy-based approach – is the difficult intercomparability for different data modification techniques. In practice this is likely to be only achievable by a certain degree of experimenting and fine-tuning, through judgement and valuation of the resulting safe microdata. In μ-ARGUS both information loss measures are used in the automatic mode: the entropy-based method at initialisation and the weight method as an option to change the initial weights. It must be admitted, though, that up to now there has been relatively little experience in setting these subjective weights satisfactorily. A lot of experimenting and testing is needed with real data to develop some intuition in applying this method. This is not unlike the problem a data protector faces when defining suitable cell weights in a table that is to be protected by secondary cell suppression.

References

De Waal, A.G. and L.C.R.J. Willenborg. 1994 *Minimizing the Number of Local Suppressions. Statistics Netherlands Report.*

De Waal, A.G. and L.C.R.J. Willenborg. 1996a *SDC Measures and Information Loss for Microdata Sets. Statistics Netherlands Report.*

De Waal, A.G. and L.C.R.J. Willenborg. 1996b. A View on Statistical Disclosure Control for Microdata. In: *Survey Methodology, Vol. 22, No. 1, pp. 95-103.*

De Waal, A.G. and L.C.R.J. Willenborg. 1998. Optimal Local Suppression in Microdata. In: *Journal of Official Statistics, Vol. 14, No. 4, pp. 421-435.*

Gouweleeuw, J, P. Kooiman, L.C.R.J. Willenborg and P.P. de Wolf. 1998. *The Post Randomisation Method for Protecting Microdata. In: Qüestlió, Vol. 22, No. 1, pp. 145-156.*

Hundepool, A.J., L.C.R.J. Willenborg, A. Wessels, L. van Gemerden, S.R. Tiourine and C.A.J. Hurkens. 1998, *μ-ARGUS User's Manual (Version 3.0), Report, Department of Statistical Methods, Statistics Netherlands.*

Hurkens, C.A.J. and S.R. Tiourine, 1998a, *On Solving Huge Set-Cover Models of the Microdata Protection Problem. Paper presented at SDP'98, 25-27 March, Lisbon, Portugal.*

Hurkens, C.A.J. and S. R. Tiourine, 1998b, *Models and Methods for the Microdata Protection Problem. In: Journal of Official Statistics, Vol. 14, No. 4, 437-447.*

Willenborg, L.C.R.J. and A.G.de Waal, 1996, *Statistical Disclosure Control in Practice, Lecture Notes in Statistics, Vol. 111, Springer-Verlag, New York.*

Notes

1) This article is an update of the report 'SDC measures and information loss for microdata sets' by the same authors that appeared in 1996.

2) In principle it is possible to use a more refined weighting system, which allows a user to specify weights for the identifying variables and also weights for each of the categories of a variables, similar to the global recoding case. But this is not implemented in the current version (3.0) of μ-ARGUS.