Survey Methodology, June 1996
Vol. 22, No. 1, pp. 95–103
Statistics Canada

A View on Statistical Disclosure Control for Microdata

A.G. de WAAL and L.C.R.J. WILLENBORG[1]

ABSTRACT

Problems arising from statistical disclosure control, which aims to prevent that information about individual respondents is disclosed by users of data, have come to the fore rapidly in recent years. The main reason for this is the growing demand for detailed data provided by statistical offices caused by the still increasing use of computers. In former days tables with relatively little information were published. Nowadays the users of data demand much more detailed tables and, moreover, microdata to analyze by themselves. Because of this increase in information content statistical disclosure control has become much more difficult. In this paper the authors give their view on the problems which one encounters when trying to protect microdata against disclosure. This view is based on their experience with statistical disclosure control acquired at Statistics Netherlands.

KEY WORDS: Statistical disclosure control; Microdata; Uniqueness.

1. INTRODUCTION

Statistical disclosure control (SDC) is becoming increasingly important as a result of the growing demand for information provided by statistical offices. The information released by these statistical offices can be divided into two major parts: tabular data and microdata. Whereas tables have been released traditionally by statistical offices, microdata sets are released only since fairly recently. In the past the users of data usually did not have the tools to analyze these microdata sets properly themselves. Nowadays every serious researcher is in possession of a powerful personal computer. Analyzing microdata is therefore no longer a privilege of the statistical office. The users of data can and want to analyze these microdata themselves. This creates non-trivial SDC-problems.

A key problem in the theory of SDC for microdata is the determination of the probability that a record in a released microdata set is re-identified. In order to estimate this probability a number of different approaches have been attempted. The aim of these attempts differ considerably. In some publications the aim was to gain a qualitative insight into the probability of re-identification of an unspecified record from a microdata set. In other publications the aim was set much higher, namely to obtain the probability that a specific record is re-identified. These are, of course, extreme cases. The former case is comparatively easy to solve, although still difficult. The latter case is more difficult and may be impossible to solve.

In this paper we give an overview of the problems for which Statistics Netherlands has attempted to provide a solution and problems of which the suggested solution has attracted our attention. We consider the problems and their outline of the solutions, while technical points are skipped. The choice of the problems and the possible solutions we consider is heavily influenced by the experiences of Statistics Netherlands in the field of SDC.

The rest of this paper is organized as follows. Basic concepts are defined in Section 2. Preliminaries on SDC for microdata are the subject of Section 3. Our basic philosophy of SDC for microdata is discussed in Section 4. In Section 5 we describe the ideal situation for microdata: in this case we would have a probability for each record that this specific record can be re-identified. A somewhat less ideal situation is described in Section 6: in this case we have a probability for a data set that an unspecified record can be re-identified. In Section 7 we have to face reality: at the moment we do not have a good disclosure risk model and we have to be satisfied with heuristic arguments. In Section 8 we summarize our conclusions and suggest some possibilities for future research.

2. BASIC CONCEPTS

In this section a number of basic concepts are defined. We will assume that the statistical office wants to release a microdata set containing records of a sample of the population. Each record contains information about an individual entity. Such an entity could be a person, a household or a business enterprise. In the rest of this paper we will usually consider the individual entity to be a person, although this is not essential.

The two most important concepts in the field of SDC are re-identification and disclosure. Re-identification is said to occur if an attacker establishes a one-to-one relationship between a microdata record and a target individual with a sufficient degree of confidence. Following

[1] A.G. de Waal and L.C.R.J. Willenborg, Statistics Netherlands, Division of Research and Development, Department of Statistical Methods, P.O. Box 4000, 2270 JM Voorburg, The Netherlands (E-mail: TWAL@CBS.NL and LWLG@CBS.NL).

Skinner (1992) we distinguish between two kinds of disclosure. Re-identification disclosure occurs if the attacker is able to deduce the value of a sensitive variable for the target individual after this individual has been re-identified. Prediction disclosure (or attribute disclosure) occurs if the microdata enable the attacker to predict the value of a sensitive variable for some target individual with a sufficient degree of confidence. For prediction disclosure it is not necessary that re-identification has taken place. Most research so far has concentrated on re-identification disclosure. In this paper we will use the term disclosure to indicate re-identification disclosure unless stated otherwise.

Now, let us define what is meant by an identifying variable. A variable is called identifying if it can serve, alone or in combination with other variables, to re-identify some respondents by some user of the data. Examples of identifying variables are residence, sex, nationality, age, occupation and education. A subset of the set of identifying variables is the set of direct (or formal) identifiers. Examples of direct identifiers are name, address and public identification numbers. Direct identifiers must have been removed from a microdata set before it is released for else re-identification is very easy. Other identifiers in most cases do not have to be removed from the microdata set. A combination of identifying variables is called a key. The identifying variables that together constitute a key are also called key variables. A key value is a combination of scores on the identifying variables that together constitute the key.

In practice, determining whether or not a variable is identifying is a problem that can only be solved by sound judgment. No limitative list of intrinsically identifying variables exists, nor, for that matter, an unambiguous and well-defined set of rules to determine such variables. Selecting a set of identifying variables, and therefore of keys, is generally based on subjective assumptions about the population. Statistics Netherlands applies some criteria, like the visibility of the categories of a variable, to determine whether or not a variable is identifying, but these criteria do not provide a definite answer to this problem for all variables. Whether or not a variable is considered identifying is essentially a matter of judgment. In the remainder of this paper we will assume however that a set of keys has been determined.

The counterparts of identifying variables are the sensitive (or confidential) variables. A variable is called sensitive (or confidential) if some of the values represent characteristics a respondent would not like to be revealed about him. In principle, Statistics Netherlands considers all variables sensitive, but in practice some variables are considered more sensitive than others. Like in the case of identifying variables, determining whether or not a variable is sensitive can be solved only by sound judgment in practice. The variables sexual behavior and criminal past are generally considered sensitive, but for other variables this may depend on, for instance, cultural background. Keller and Bethlehem (1992) give as an example the variable income. In the Netherlands income is considered sensitive, whereas in Sweden it is not. Moreover, there are variables which should be considered both identifying and sensitive. An example of such a variable is ethnic membership. However, in the literature it is usually assumed that the identifying and sensitive variables can be divided into disjoint sets. In the remainder of this paper we will also assume that a set of sensitive variables has been determined which is disjoint from the set of identifying variables.

By using information about the identifying variables a potential attacker can try to disclose information about sensitive variables. Note that this way of disclosure is only possible in case the link between the values of the identifying variables and the values of the sensitive variables has not been perturbed by noise in the data or by a technique like data-swapping.

To end this section, we give a definition of SDC. Statistical disclosure control aims to reduce the risk that sensitive information of individual persons can be disclosed to an acceptable level. What is acceptable depends on the policy of the data releaser. In order to reduce the risk of disclosure an estimate for the risk of disclosure would be very helpful although it is not a necessary requisite (*cf.* Section 7). Some research has been devoted to defining and estimating this risk of disclosure.

3. PRELIMINARIES ON SDC FOR MICRODATA

As a customer of a statistical office, the user of a microdata set should be satisfied with its quality. The user is usually not interested in individual records, but only in statistical results which can be drawn from the total set of records. For instance, he wants to examine tables he has produced himself from the microdata set.

Because a microdata set is meant for statistical analysis it is not necessary that each record in the set is correct. The statistical office has the possibility to perturb records, *e.g.*, by adding noise or by swapping parts of records between different records, in order to reduce the risk of re-identification. By perturbing records the risk of re-identification is reduced because even when a correct re-identification takes place the information which is disclosed may be incorrect. In any case the attacker cannot be sure that the disclosed information is correct. The statistical office 'only' has to guarantee that the statistical quality of, for instance, the tables the user wants to examine is high enough. This may be quite complicated to achieve in practice, however.

Although data perturbation methods may prove to be useful, for the time being Statistics Netherlands does not use them. To protect its microdata sets Statistics Netherlands applies local suppression and global recoding only.

When local suppression is applied some values of variables in some records are set to 'missing', *i.e.*, deleted from the microdata set. When global recoding is applied some variables are given a coarser categorization. In a first step, we try to protect a microdata set by means of global recoding. However, when protecting a microdata set entirely by means of global recodings would result in a considerable information loss, we apply local suppressions as well. In this way we try to avoid that too much information will be lost. It should be clear that local suppressions are only applied parsimoniously.

An advantage of local suppression and global recoding is that these techniques preserve the integrity of the data. A disadvantage of local suppression is that it introduces a bias, because extreme values will be locally suppressed. However, when local suppressions are only applied parsimoniously, this bias will be small.

From the SDC point of view a user of the data should also be looked upon as a potential attacker. Hence, it is useful to consider the ways in which disclosure can take place. An attacker tries to match records from the microdata set with records from an identification file or with individuals from his circle of acquaintances. An identification file is a file containing records with values on direct identifiers and values on some other identifiers of the microdata set. The latter identifiers may be used to match records from the released microdata set with records from the identification file. After matching the direct identifiers in the identification file can be used to determine whose record has been matched, and the sensitive variables in the released microdata set can be used to disclose information about this person. A circle of acquaintances is the set of persons in the population for which the attacker knows the values on a certain key from the microdata set. So, a circle of acquaintances could actually be an identification file, and vice versa. In the rest of this paper we will therefore use the terms 'identification file' and 'circle of acquaintances' interchangeably.

In order for re-identification of a record of an individual to occur the following conditions have to be satisfied:

C_1. The individual is unique on a particular key value K.

C_2. The individual belongs to an identification file or a circle of acquaintances of the attacker.

C_3. The individual is an element of the sample.

C_4. The attacker knows that the record is unique in the population on the key K.

C_5. The attacker comes across the record in the microdata set.

C_6. The attacker recognizes the record of the individual.

Whenever one of the conditions C_1 to C_6 does not hold, re-identification cannot be accomplished with absolute certainty. If either condition C_1 or C_4 does not hold, then a matching can be made but the attacker cannot be sure that this leads to a correct re-identification.

It is clear from the conditions C_1 to C_6 that a 'good' model for the risk of re-identification should incorporate aspects of both the data set and the user. When a Dutch microdata set is used by someone in, say, China who is essentially unfamiliar with the Dutch population, then the risk of re-identification is negligible. In order to re-identify someone in a microdata set it is necessary to acquire sufficient knowledge about the population. The amount of work that should be done to acquire this knowledge is proportional to the safety of the microdata set.

4. A PHILOSOPHY OF SDC

It seems likely that the attention of a potential attacker is drawn by combinations of identifying variables that are rare in the sample or in the population. Combinations that occur quite often are less likely to trigger his curiosity. If he tries to match records deliberately then he will probably try to do this for key values that occur only a few times. If the user does not try to match records deliberately, but he knows an acquaintance with a rare key value then a record with that particular key value may trigger him to consider the possibility that this record belongs to this acquaintance. Moreover, the probability of a correct match is higher in case the number of persons that score on the matching key value is smaller. Finally, it is also very likely that among the persons that score on a rare key value there are many uniques if the key is augmented with an additional variable. Records that score on such rare combinations of identifying variables are therefore more likely to be re-identified.

In particular key values which occur only once in the population, *i.e.*, uniques in the population, can lead to re-identification. In the past emphasis was placed almost exclusively on uniqueness. It should be noted, however, that uniqueness is neither sufficient nor necessary for re-identification. If a person is unique in the population on certain key variables, but nobody realizes this, then this person may never be re-identified. If on the other hand this person is not unique in the population, but there is only one other person in the population with the same key, then this other person is, in principle, able to re-identify him. Furthermore, suppose a person is not unique, but belongs to a small group of people. Suppose also that the attacker happens to know information about him which is not considered to be identifying by the statistical office, but which is contained in the released microdata set, then it is very well possible that he is unique on the key combined with the new information. So, it is possible that a person is re-identified although he is not unique on the keys of identifying variables in the population. Finally, prediction disclosure may occur. That is, if a person is not unique in the population, but belongs to a group of people with (almost) the same score on a particular sensitive variable, then sensitive information can be disclosed about this individual without actual re-identification. Prediction disclosure is not discussed further in this paper. For more information on prediction disclosure we refer to Skinner (1992), US Department of Commerce (1978), Duncan and Lambert (1986), and Cox (1986).

SDC should concentrate on key values that are rare in the population. A probability that information from a particular respondent, whose data are included in a microdata set, is disclosed should reflect the 'rareness' of the key value of this respondent's record. A probability for the event that information from an arbitrary respondent is disclosed should reflect the 'overall rareness' of the records in the data set. If there are many records in a microdata set of which the key value is rare, then the probability of disclosure for this data set should be high. In the next sections we will examine some attempts to incorporate these ideas within a mathematical framework.

5. RE-IDENTIFICATION RISK PER RECORD

In an ideal world (as far as SDC is concerned) a releaser of microdata would be able to determine a risk of re-identification for each record, *i.e.*, a probability that the respondent of this record can be re-identified. Such a risk per record would enable us to adopt the following strategy. First, order the records according to their risk of re-identification with respect to a single key. Second, select a maximum risk the statistical office is willing to accept. Finally, modify all the records for which the risk of re-identification with respect to the key chosen is too high. Repeat this procedure for each key in case there are more keys.

Unfortunately, we do not live in such an ideal world at the moment. However, steps towards the ideal situation have been made by Paass and Wauschkuhn (1985), and Fuller (1993). In Paass and Wauschkuhn (1985) it is assumed that a potential attacker has both a microdata file, released by a statistical office, and an identification file at his disposal. Between both files there may be many data incompatibilities. These data incompatibilities may be caused by *e.g.*, coding errors, by different definitions of categories or by 'noise' in the data. By assuming a probability distribution for these data incompatibilities and a disclosure scenario Paass and Wauschkuhn develop a sophisticated model to estimate the probability that a specific record from the microdata file is re-identified. The type of distribution of the errors that caused the data incompatibilities was assumed to be known to the attacker. The variance of the errors was assumed unknown to him. A potential attacker had to estimate this variance, on the basis of the (assumed) knowledge of the statistical production process. The model of Paass and Wauschkuhn is essentially based on discriminant analysis and cluster analysis.

Paass and Wauschkuhn distinguish between six different scenarios. Each scenario corresponds to a special kind of attacker. The number of records in the identification file and the information content of the identification file depend on the chosen scenario. An example of such a scenario is the journalist scenario, where a journalist selects records with extreme attribute combinations in order to re-identify respondents with the aim of showing that the statistical office fails to secure the privacy of its respondents.

Paass and Wauschkuhn apply their method to match records from the identification file with records from the microdata file. If the probability that a specific record from the identification file belongs to a specific record from the microdata set is high enough, then these two records are matched. This probability is the probability of re-identification per record, conditional on a particular disclosure scenario.

Müller, Blien, Knoche, Wirth *et al.* (1991) and Blien, Wirth and Müller (1992) applied the method recommended in Paass and Wauschkuhn (1985) to real data. When compared to simple matching, *i.e.*, a record is considered re-identified by an attacker if he succeeds in finding a unique value set in the microdata file which is identical to a value set in the identification file, the method suggested by Paass and Wauschkuhn turned out to be not superior. Apparently, the number of correctly matched records when applying the method by Paass and Wauschkuhn was in disagreement with the probability of re-identification per record.

In the context of masking procedures, *i.e.*, procedures for microdata disclosure limitation by adding noise to the microdata, Fuller (1993) obtained an expression for the probability that a specific record in the released microdata set is the same as a specific target record from an identification file. That is, an expression for the re-identification probability per record is derived. To derive this expression several assumptions are made. It is assumed that the data, the noise and errors in the data are normally distributed. Moreover, it is assumed that the covariance matrices of both the noise and the errors in the data are known to an attacker. Finally, it is assumed that the data have been obtained by simple random sampling. These assumptions allow Fuller (1993) to derive his expression for the re-identification probability by means of probability theoretical considerations. Unfortunately, the approach by Fuller has not been tested on real data yet. Hence, it is hard judge the applicability of this approach. For a comment on the approach by Fuller see Willenborg (1993).

Paass and Wauschkuhn (1985), and Fuller (1993) are mainly interested in the effects of noise that has (unintentionally and intentionally, respectively) been added to the data on the disclosure risk. A weak point of their respective approaches is the, implicit, assumption that the key is a high-dimensional one. Assuming a high-dimensional key implies that (almost) everyone in the population is unique. The probability that a combination or key value occurs more than once in the population is negligible. This makes the computation of the probability of re-identification per record considerably easier. On the other hand, in case of low-dimensional keys it is not unlikely that certain key values occur many times in the population. Therefore, deriving a probability of re-identification per record for low-dimensional keys is much harder than for high-dimensional keys, because for high-dimensional keys the probability of statistical twins in the population is almost zero.

A good model for the re-identification risk per record does not appear to exist at the moment. In Section 6 we therefore consider less ambitious models, namely models for the re-identification risk per file.

6. RE-IDENTIFICATION RISK PER FILE

In a somewhat less ideal world a releaser of microdata would not be able to determine the risk of re-identification for each record, but he would be able to determine the risk that an unspecified record from the microdata set is re-identified. In this case, the statistical office should decide on the maximal risk it is willing to take when releasing a microdata set. If the actual risk is less than the maximal risk, then the microdata set can be released. If the actual risk is higher than the maximal risk, then the microdata set has to be modified. Determining which records have to be modified remains a problem, however.

A basic model to determine the probability that an arbitrary record from a microdata set is re-identified has been proposed by Mokken, Pannekoek and Willenborg (1989) and Mokken, Kooiman, Pannekoek and Willenborg (1992). In Mokken *et al.* (1989) only the case where there is a single researcher, an unstratified population and a single key is considered. It has been extended to include the cases of subpopulations, multiple researchers and multiple keys (*cf.* Willenborg 1990a; Willenborg 1990b; Mokken *et al.* 1992). The model of Mokken *et al.* (1992) takes three probabilities into account. The first probability, f, is equal to the sampling fraction. In other words, f, is the probability that a randomly chosen person from the population has been selected in the sample. The second probability, f_a, is the probability that a specific researcher who has access to the microdata knows the values of a randomly chosen person from the population on a particular key. The third probability, f_u, is the probability that a randomly chosen person from the population is unique in the population on a particular key. Combining these three probabilities, f, f_a and f_u, the probability that a record from a microdata set is re-identified can be evaluated.

For each sample element a number of variables is measured. The values obtained by these measurements (scores) are collected in records, one for each sample element. It is assumed that the variables in the key are either categorical variables or variables for which the measurements fall into a finite number of categories. Together, the records constitute a data set S that will be made available to an researcher R. We recall that whenever we use the term disclosure in fact re-identification disclosure is meant. The model of Mokken *et al.* (1989, 1992) does not take prediction disclosure into account.

In terms of the Paass and Wauschkuhn (1985) set-up f_a and f_u together reflect the *Informationsgehalt der Überschneidungsmerkmale, i.e.*, the information content of the matching values. The various scenarios they consider differ in terms of f_a and f_u. In particular, f_u is influenced by the number of variables and the information content of these variables, *i.e.*, their categorization, an attacker has at his disposal to re-identify a record. The parameter f_a is determined by the number of records that are contained in the information file.

With respect to researcher R and key K there is a circle of acquaintances A. Obviously, A and its size $|A|$ will depend on the particular researcher R as well as on the key K and the variables as registered and coded in the data set.

It is assumed that if conditions C_1, C_2 and C_3 of the conditions for re-identification given in Section 3 hold, then conditions C_4, C_5 and C_6 hold too. Condition C_4 is a rather exacting one, but it can be introduced as an assumption for the sake of convenience in formulating a disclosure risk model. Note that it then yields a worst-case situation, in the sense that fallible perception and memory or other sources of ignorance, confusion and uncertainty for a potential discloser are excluded. Taken as an assumption together with C_5 and C_6 the implication is that the occurrence of any unique acquaintance E of R in data set S is equivalent to re-identification by R. It is assumed that re-identification of a record implies disclosure of confidential information. Thus re-identification can be treated as equivalent to disclosure. Implicitly, it is assumed that the link between the identifying variables and the sensitive variables has not been disturbed by a technique such as data-swapping.

Furthermore it is assumed that both the identifying and the confidential information are free of error or noise to researcher R, contrary to *e.g.*, Paass and Wauschkuhn (1985), and Fuller (1993). Clearly, this assumption is unrealistic for most microdata sets.

The disclosure risk D_R for a certain microdata set S with respect to a certain researcher R and a certain key K, is defined to be the probability that the researcher makes at least one disclosure of a record in S on the basis of K. In order to apply a criterion based on the disclosure risk, the value of this quantity for a given data set has to be determined. An expression for this quantity can be derived on the basis of a set of assumptions.

In the model of Mokken *et al.* the following assumptions are made in addition to $C_1 - C_6$:

A_1. The circle of acquaintances A can be considered as a random sample from the population.

A_2. The data set S is a random sample from the population.

Assumption A_1 serves to imply that the probability that a randomly chosen element from the population is an acquaintance of R is $f_a = | A | /N$, where N is the size of the population. As a consequence the expected number of unique elements in A, $| U_a |$, is equal to $f_a | U | = | A | f_u$, where U is the set of unique persons in the population and $| U |$ its size. Obviously assumption A_2 implies that the probability that a specific unique element E is selected in the sample is f. These assumptions allow one to obtain a very simple expression for the disclosure risk D_R in terms of f, f_a and f_u, namely

$$D_R = 1 - \exp(-Nff_af_u). \tag{1}$$

Two of the parameters in the model of Mokken *et al.* (1989, 1992), f_a and f_u, are unknown. The parameter f_a can be 'guestimated', *i.e.*, obtained by inspired guesswork, by assuming different scenarios an attacker may follow. A number of such scenarios has been described in Paass and Wauschkuhn (1985) and Paass (1988). Evaluating f_a seems difficult, however. In order to estimate the other parameter, f_u, a number of models has been proposed in the literature. Models to estimate the number of uniques in the population, and hence f_u, that have been proposed include the Poisson-gamma model (Bethlehem, Keller and Pannekoek 1989; Mokken *et al.* 1989; Willenborg, Mokken and Pannekoek 1990; De Jonge 1990), the negative binomial superpopulation model (Skinner, Marsh, Openshaw and Wymer 1990), the Poisson-lognormal model (Skinner and Holmes 1992; Hoogland 1994), models based on equivalence classes (Greenberg and Zayatz 1992) and models based on modified negative binomial-gamma functions (Crescenzi 1992; Coccia 1992). As we have remarked in Section 4 not only the number of population uniques is important, but the numbers of cells with two, three, *etc.* persons are important as well. The Poisson-gamma model, the Poisson-lognormal model and the negative binomial superpopulation model can be applied to estimate the number of cells with two, three, *etc.* persons as well. It seems that the other models mentioned above can be extended in order to estimate these numbers. A major drawback is that the results are not very reliable in many cases.

From the model by Mokken *et al.* (1989, 1992) it is clear that the statistical office that disseminates the data is able to influence the risk of re-identification. The statistical office basically has two ways to do this. First of all, the size of the data set can be reduced, *i.e.*, the sampling fraction f can be reduced. A reduction of f implies a reduction of the risk. However, lowering f is generally undesirable, because usually f has to be reduced substantially to be effective. This implies that only a small part of the data available can be released. The second way in which the statistical office can influence the re-identification risk is by reducing the number of population uniques, *i.e.*, by reducing f_u. The fraction f_u depends on the information provided by the key variables. The less information the key variables provide the less uniques there are in the population. In order words, f_u can be reduced by collapsing categories (global recoding) and by replacing values by missings (local suppression). Collapsing categories is a global action, because it generally affects many records; replacing values by missings is a local action because it affects only a few individual records. Usually, the loss in information when reducing f_u is considerably less than the loss in information when reducing f. Therefore, a statistical office will usually choose to control the re-identification risk by reducing f_u rather then reducing f. The third possibility of controlling the re-identification risk, *i.e.*, by reducing f_a, is not applied in practice, because f_a is difficult to model.

Although the model by Mokken *et al.* (1989, 1992) provides some insight in how to reduce the disclosure risk it can hardly be used as a basis for the protection of microdata sets. The reason for this is that the two parameters of the model, f_u and f_a, are often difficult to evaluate. Usually there is insufficient data available to estimate f_u and f_a accurately. We conclude that even a model for a re-identification risk for an entire microdata set is difficult to apply in practice. In Section 7 we therefore face reality in which we have no satisfactory model for either the re-identification risk per record or re-identification risk for an entire microdata set.

7. INTUITIVE RE-IDENTIFICATION RISK

In reality we are, unfortunately, forced to base SDC on heuristic arguments rather than on a solid theoretical basis. The SDC rules mentioned in this section all reduce the re-identification risk. It is, however, not possible to evaluate this reduction of the re-identification risk. At Statistics Netherlands, rules for SDC of microdata are based on testing whether scores on certain keys occur frequently enough in the population. A few problems arising here are the determination of the keys that have to be examined, the way to estimate the number of persons in the population that score on a certain key, to make operational the meaning of the phrase 'frequently enough' by determining *e.g.*, (a) threshold value(s), and how to determine appropriate SDC-measures.

Statistics Netherlands distinguishes between two kinds of microdata sets. The first kind is a so-called public use file. A public use file can be obtained by everybody. The keys that have to be examined for a public use file are all combinations of two identifying variables. The number of identifying variables is limited, and certain identifying variables, such as place of residence are not included in a public use file. Moreover, sampling weights have to be examined before they can be included in a public use file, because there are many situations in which weights can give additional information (*cf.* De Waal and Willenborg 1995a).

For instance, when a certain subpopulation is oversampled then this subpopulation can be recognized by the low weights associated with its members in the sample. Weights may only be published when they do not provide additional information that can be used for disclosure purposes. In case sampling weights are not considered suited for publication SDC measures should be taken, such as subsampling the units with a low weight in order to get a subsample in which all units have approximately the same weight. Because the weights are approximately equal assuming that they are exactly equal would introduce only a small error. The second kind of microdata set is a so-called microdata set for research. A microdata set for research can only be obtained by well-respected (statistical) research offices. The information content of a microdata set for research is much higher than that of a public use file. The number of identifying variables is not limited and an identifying variable such as place of residence may be included in a microdata set for research. Because of the high information content of a microdata set for research, researchers have to sign a declaration stating that they will protect any information about an individual respondent that might be disclosed by them. The keys that have to be examined for a microdata set for research consist of three-way combinations of variables describing a region with variables describing the sex, ethnic group or nationality of a respondent with an ordinary identifying variable.

The rules Statistics Netherlands applies for SDC are based on the following idea: a key value, *i.e.*, a combination of scores on the identifying variables that together constitute the key, is considered safe for release if the frequency that this key value occurs in the population is more than a certain threshold value d_0. This value d_0 was chosen after a careful and extensive search considering many different values and comparing the records which have to be modified for each value of d_0. The value that leads to the 'most likely' set of records which have to be modified has been chosen to be the value of d_0. Which records are considered to be the 'most likely' ones to be modified is a matter of personal judgment.

When applying one of the above rules we are generally posed with the problem that we do not know the number of times that a key value occurs in the population. We only have the sample available to us. The population frequency of a key value has to be estimated based upon the sample. For large regions it is possible to use an interval estimator to test whether or not a key value occurs often enough in a region. This interval estimator is based on the assumption that the number of times that a key value occurs in the population is Poisson distributed (*cf.* Pannekoek 1995). However, for relatively small regions the number of respondents is low, which causes the estimator to have a high variance which in turn causes a lot of records to be modified. To estimate the number of times that a key value occurs in a small region we therefore suggest to apply a point estimator. We will now discuss some possibilities for such an estimator.

A simple point estimator for the number of times that a certain key value occurs in a region is the direct point estimator. The fraction of a key value in a region i is estimated by the sample frequency of this key value in region i divided by the number of respondents in region i. The population frequency is then estimated by this estimated fraction multiplied by the number of inhabitants in region i. When the number of respondents in region i is low, which is often the case, the direct estimator is unreliable. Another point estimator is based on the assumption that the persons who score on a certain key value are distributed homogeneously over the population. In this case the fraction of a key value in region i can be estimated by the fraction in the entire sample. The advantage of this, so-called, synthetic, estimator is that the variance is much smaller than the variance of the direct estimator. Unfortunately, the homogeneity assumption is usually not satisfied which causes the estimator to be biased. However, a combined estimator can be constructed with both an acceptable variance and an acceptable bias by using a convex combination of the direct estimator and the synthetic estimator. Such a combined estimator has been tested in Pannekoek and de Waal (1995).

Another practical problem that deserves attention is top-coding of extreme values of continuous (sensitive) variables. These extreme values may lead to re-identification because these values are rare in the population. At the moment Statistics Netherlands uses an interval estimator to test whether there is a sufficient number of individuals in the population who score on a 'comparable' value of the continuous variable (*cf.* Pannekoek 1992). If this is the case, then the extreme value may be published, otherwise the extreme value must be suppressed. In order to apply this method in practice it remains to specify what is meant by 'sufficient' and by 'comparable'.

Some important practical problems occur when determining which protection measures should be taken when a microdata set appears to be unsafe. In that case the original data set must be modified in such a way that the information loss due to SDC-measures is as low as possible while the resultant data set is considered safe. In De Waal and Willenborg (1994a) and De Waal and Willenborg (1995b) a model for determining the optimal local suppressions is presented. Determining the optimal global recodings is much more difficult. Comparing the information loss due to global recodings to the information loss to local suppressions is already a problem. In De Waal and Willenborg (1995c) this latter problem is solved by using the entropy.

Currently a general purpose software package for SDC of microdata is being developed at Statistics Netherlands (*cf.* De Jong 1992; De Waal and Willenborg 1994b; Van Gelderen 1995; Pieters and De Waal 1995; De Waal and

Pieters 1995). The package, ARGUS, should enable the statistical office to analyze the data and to carry out suitable protection measures. It will consist of two separate parts: μ-ARGUS for SDC of microdata and τ-ARGUS for SDC of tabular data. The structure of the package is such that it will be possible to specify different disclosure control rules. This implies that ARGUS will be suited for other statistical offices too. Moreover, it will be possible to incorporate changes in the rules fairly easily in the package.

8. CONCLUSIONS

There is one important conclusion one can draw from this paper: SDC still offers a lot of possibilities for future research, despite the considerable amount of research that has been carried out to date. The theory of SDC for microdata has a number of gaps. Among the technical problems that remain to be solved are the following. When we want to release data for small regions we need an acceptable estimator for the number of times that a key value occurs in these regions. Such an estimator is difficult to construct, although the preliminary results obtained at Statistics Netherlands seem encouraging. An important practical problem is the determination of appropriate global recodings and local suppressions. Yet another one is the determination of the number of uniques, or more generally the number of rare frequencies, in the population. Some of the models proposed in Section 6 appear to be acceptable, but can probably be improved upon. An alternative approach is to determine which elements in the sample are unique in the population. In Verboon (1994), and Verboon and Willenborg (1995) this approach is examined. An extension of the model by Mokken *et al.* (1989, 1992) to estimate the risk of re-identification of a file is yet another problem to be solved. This extension should take into account that measurement errors have been made and that population uniqueness is not necessary in order to disclose information. Finally, a model to estimate the re-identification risk per record would be very welcome. In fact, it would yield a sound criterion to judge the safety of a microdata set. This criterion can guide one in producing safe microdata sets by applying SDC-measures such as global recoding and local suppression.

Apart from technical problems there are also some policy problems. Based on the policy that a statistical office wants to pursue the following decisions should be made. The combinations of variables that should be examined should be specified. Suitable threshold values should be selected.

More and better software must be developed in order to deal with time-consuming calculations. For microdata, software must be developed to indicate which records and variables must be modified, and how they should be modified, when applying a particular disclosure rule. At the time of writing an international project on SDC is about to start. The participating institutions in this project are the Eindhoven University of Technology, the University of Manchester, the University of Leeds, the Office of Population Censuses and Surveys (OPCS), the Istituto Nazionale di Statistica (ISTAT), the Consortio Padova Ricerche (CPR), and Statistics Netherlands. One of the major aims of the project is to develop software for the SDC of both microdata (μ-ARGUS) and tabular data (τ-ARGUS).

Finally, some very practical problems remain to be solved. An example of such a problem is the determination of a set of rules for selecting identifying variables. Such a set of rules would be a very valuable asset. Without these rules identifying variables are selected by making subjective choices. Developing such a set of rules is another goal of the above mentioned SDC-project.

REFERENCES

BETHLEHEM, J.A., KELLER, W.J., and PANNEKOEK, J. (1989). Disclosure control of microdata. *Journal of the American Statistical Association*, 85, 38-45.

BLIEN, U., WIRTH, H., and MÜLLER, M. (1992). Disclosure risk for microdata stemming from official statistics. *Statistica Neerlandica*, 46, 69-82.

COCCIA, G. (1992). Disclosure risk in Italian current population surveys. International Seminar on Statistical Confidentiality, Dublin.

COX, L.H. (1986). Comment on Duncan and Lambert (1986). 19-21.

CRESCENZI, F. (1992). Estimating population uniques; methodological proposals and applications on Italian census data. International Seminar on Statistical Confidentiality, Dublin.

De JONG, W.A.M. (1992). ARGUS: An integrated system for data protection. International Seminar on Statistical Confidentiality, Dublin.

De JONGE, G. (1990). The estimation of population unicity from microdata files (in Dutch), Internal note, Statistics Netherlands, Voorburg.

De WAAL, A.G., and PIETERS, A.J. (1995). ARGUS user's guide. Report, Statistics Netherlands, Voorburg.

De WAAL, A.G., and WILLENBORG, L.C.R.J. (1994a). Minimizing the number of local suppressions in a microdata set. Report, Statistics Netherlands, Voorburg.

De WAAL, A.G., and WILLENBORG, L.C.R.J. (1994b). Development of ARGUS: past, present, future. Report, Statistics Netherlands, Voorburg.

De WAAL, A.G., and WILLENBORG, L.C.R.J. (1995a). Statistical disclosure control and sampling weights. Report, Statistics Netherlands, Voorburg.

De WAAL, A.G., and WILLENBORG, L.C.R.J. (1995b). Local suppression in statistical disclosure control and data editing. Report, Statistics Netherlands, Voorburg.

De WAAL, A.G., and WILLENBORG, L.C.R.J. (1995c). Optimal global recoding and local suppression. Report, Statistics Netherlands, Voorburg.

DUNCAN, G.T., and LAMBERT, D. (1986). Disclosure-limited data dissemination. *Journal of the American Statistical Association*, 81, 10-28.

FULLER, W.A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics*, 9, 383-406.

GREENBERG, B.V., and ZAYATZ, L.V. (1992). Strategies for measuring risk in public use microdata files. *Statistica Neerlandica*, 46, 33-48.

HOOGLAND, J. (1994). Protecting microdata sets against statistical disclosure by means of compound Poisson distributions (in Dutch). Report. Statistics Netherlands, Voorburg.

KELLER, W.J., and BETHLEHEM, J.A. (1992). Disclosure protection of microdata: problems and solutions. *Statistica Neerlandica*, 46, 5-19.

MOKKEN, R.J., PANNEKOEK, J., and WILLENBORG, L.C.R.J. (1989). Microdata and disclosure risks, CBS Select 5, Statistical Essays, Staatsuitgeverij (The Hague), 181-200.

MOKKEN, R.J., KOOIMAN, P., PANNEKOEK, J., and WILLENBORG, L.C.R.J. (1992). Disclosure risks for microdata. *Statistica Neerlandica*, 46, 49-67.

MÜLLER, W., BLIEN, U., KNOCHE, P., WIRTH, H. *et al.* (1991). *The Factual Anonymity of Microdata* (in German). Stuttgart: Metzler-Poeschel Verlag.

PAASS G., and WAUSCHKUHN, U. (1985). Data access, data protection and anonymization – analysis potential and identifiability of anonymized individual data (in German). Gesellschaft für Mathematik und Datenverarbeitung, Oldenbourg-Verlag, Munich.

PAASS, G. (1988). Disclosure risk and disclosure avoidance for microdata. *Journal of Business and Economic Studies*, 6, 487-500.

PANNEKOEK, J. (1992). Disclosure control of extreme values of continuous identifiers (in Dutch). Report, Statistics Netherlands, Voorburg.

PANNNEKOEK, J. (1995). Statistical methods for some simple disclosure limitation rules. Report, Statistics Netherlands, Voorburg.

PANNEKOEK, J., and de WAAL, A.G. (1995). Synthetic and combined estimators in statistical disclosure control. Report, Statistics Netherlands, Voorburg.

PIETERS, A.J., and De WAAL, A.G. (1995). A demonstration of ARGUS. Report, Statistics Netherlands, Voorburg.

SKINNER, S., MARSH, C., OPENSHAW, S., and WYMER, C. (1990). Disclosure avoidance for census microdata in Great Britain. *Proceedings of the 1990 Annual Research Conference*, U.S. Bureau of the Census, Washington, DC, 131-143.

SKINNER, C.J. (1992). On identification disclosure and prediction disclosure for microdata. *Statistica Neerlandica*, 46, 21-32.

SKINNER, C.J., and HOLMES, D.J. (1992). Modelling population uniqueness. International Seminar on Statistical Confidentiality, Dublin.

US DEPARTMENT OF COMMERCE (1978). Report on statistical disclosure and disclosure avoidance techniques. Statistical Policy Working Paper 2, Washington DC.

Van GELDEREN, R. (1995). ARGUS: Statistical disclosure control of survey data. Report, Statistics Netherlands, Voorburg.

VERBOON, P. (1994). Some ideas for a masking measure for statistical disclosure control. Report, Statistics Netherlands, Voorburg.

VERBOON, P., and WILLENBORG, L.C.R.J. (1995). Comparing two methods for recovering population uniques in a sample. Report. Statistics Netherlands, Voorburg.

WILLENBORG, L.C.R.J. (1990a). Remarks on disclosure control of microdata. Report, Statistics Netherlands, Voorburg.

WILLENBORG, L.C.R.J. (1990b). Disclosure risks for microdata sets: stratified populations and multiple investigators. Report, Statistics Netherlands, Voorburg.

WILLENBORG, L.C.R.J. (1993). Discussion statistical disclosure limitation. *Journal of Official Statistics*, 9, 469-474.

WILLENBORG, L.C.R.J., MOKKEN, R.J., and PANNEKOEK, J. (1990). Microdata and disclosure risks. *Proceedings of the 1990 Annual Research Conference*, U.S. Bureau of the Census, Washington DC, 167-180.