The Confidentiality Promise: is it safe in our hands?

CCSR Occasional Paper 9

Respondents' informed consent and data release policies

D Holt, Office for National Statistics and University of Southampton
L Willenborg, Research & Development Division, Statistics Netherlands
M Warren, Director General, The Market Research Society

£3.00 Further copies of this paper may be obtained from:

ISBN 1 899005 10 2

CCSR
Faculty of Economic & Social Studies
University of Manchester
Oxford Road
Manchester M13 9PL

Tel: 0161 275 4721; Fax: 0161 275 4722
Email: ccsr@manchester.ac.uk
http://info.mcc.ac.uk/Economics/ccsr

The confidentiality promise: is it safe in our hands?

Foreword

The Social Statistics Section of the Royal Statistical Society and the Social Research Association have decided to hold an annual seminar to commemorate the contribution which Cathie Marsh made to social science. The first seminar, held in November 1994, debated the relative merits of probability versus quota sampling - a debate to which Cathie had made an important contribution.

In the second seminar the focus moved to the topic of confidentiality. Cathie had played a crucial role in work to assess the disclosure risk from the release of samples of anonymised records from the census of population. She chaired an ESRC Working Party on the release of microdata and, with Chris Skinner and other colleagues, produced a paper (Marsh et al, 1991) which has become an important marker in the literature. It was therefore appropriate the second seminar should address this topic from a number of different angles.

The seminar was introduced and chaired by Denise Lievesley, Director of the ESRC Data Archive and a member of the ESRC Working Party on the Samples of Anonymised Records (SARs).

Tim Holt, who had just become Director of the Government Statistical Service, now the Office for National Statistics, gave the first paper, beginning by recalling Cathie's role in making the case for the SARs and then discussing the confidentiality issues which had faced OPCS when taking its decision to release samples of microdata, and, in conclusion, looking at the range of data sources released by government and the challenge faced by the Office for National Statistics in ensuring full and effective use of the data. Leon Willenborg, from Statistics Netherlands, gave the second paper which provided an overview of the statistical methodology being developed to address disclosure risk and the way in which software solutions might be of value, with a particular focus on those being developed within Statistics Netherlands. Finally, Michael Warren, Director General of the Market Research Society, discussed issue of informed consent and the confidentiality promise given to the respondent. He reminded us of the difficulties of ensuring that the consent given by respondents was as fully informed as we might wish to believe and of the ways in which the Market Research Society was preparing guide-lines to assist its members in dealing with confidentiality issues arising from the changing nature of market research and the prospect of restrictive legislation from the European Union.

References

Marsh, C. Skinner, C. Arber, S. Penhale, B. Openshaw, S. Hobcraft, J. Lievesley, D. and Walford, N. (1991), The Case for Samples of Anonymised Records from the 1991 Census, Marsh et al, Journal of the Royal Statistical Society A, 154, Part 2, pp 305-340

Angela Dale
Cathie Marsh Centre for Census and Survey Research, University of Manchester

OFFICIAL STATISTICS AND THE CONFIDENTIALITY PROMISE

D Holt, Office for National Statistics and University of Southampton

1. Introduction

To my mind, there can be no more suitable topic for this second Cathie Marsh Memorial Seminar than the confidentiality of survey data. Cathie was rare among sociologists for her advocacy of the survey method and for her support of rigorous quantitative analysis in conjunction with other sociological methods as a means of understanding society. She believed that such methods could and should affect social policy. Perhaps her greatest contribution was her drive and determination to convince Government that the release of a sample of anonymised records from the 1991 Census of Population could be achieved without unacceptable risk of identification of any particular individual and hence disclosure of census data. In her characteristic way that argument was made, not simply by passionate advocacy, but with the support of others, and perhaps particularly Chris Skinner, by a rigorous analytical assessment of the risks of disclosure. The work, which was instrumental in supporting their case, has been recognised internationally as the most careful and systematic approach yet available for analysing disclosure risks. Thus the topic of confidentiality and the protection of the rights of individuals is entirely appropriate for this seminar.

2. Mission

The new Office for National Statistics will share with the Government Statistical Service a common mission: -

"To provide Parliament, government and the wider community with the statistical information, analysis and advice needed to improve decision making, stimulate research and inform debate."

Underpinning this mission will be two strong policies which are potentially in conflict. The first is a strong commitment, backed by a legal requirement, to preserve the confidentiality of information which is collected from individuals and businesses. The second strong policy is that we must do all we can to serve Government, Parliament and the wider community. We must ensure that our statistics are used.

We are in the middle of an information explosion which is based upon technical innovation. Historically, statistical information would have been made available through paper publication. The statistician would have determined the tabulations which would be produced and would have, in effect, specified the information which it was judged that users would need. Today there is a much

stronger thrust towards other forms of dissemination including CD-ROMs, electronic transfer and of course the Internet. This is coupled with a much stronger demand from users to be able to specify the analyses which they need for their purposes and for a much more interactive process between users and the data. We have moved from a producer oriented statistical system to a user oriented statistical system where the users are the people who determine the form for statistical outputs. As these requirements become more and more sophisticated so there is greater and greater need for additional analysis to be based upon unit level data or on highly disaggregated summary statistics rather than on pre-prepared tabulations. The implications for confidentiality are important and must be treated very carefully.

3. Checking Output

The real threat to confidentiality comes from combinations of variables relating to the same person or household that together make that person or household unique or almost certainly so. Historically, when data were produced in published form by pre-specified, fixed tabulations it was relatively easy to assess the final output and to determine through visual examination whether or not there was any likelihood of disclosure. It was common place to suppress the cell of a table if it was based upon too few observations. The increase in the number of user specified tabulations made this process more onerous and automatic checking procedures were introduced. However the real problem is the risk that two user specified tabulations could be subtracted from each other and that the residual information could apply to identifiable individuals. This so called residual disclosure makes the protection of confidentiality through the output of the analysis extremely difficult indeed. It is, however, the case that some work of this type is still being carried out to see whether or not mechanisms can be introduced which would allow this as a procedure. Work is going on at Statistics Netherlands, in which OPCS is involved, to explore this approach further. In our case the interest is centered on the 2001 Census and it may be that Leon Willenborg will refer to this and I will say no more.

4. Anonymised Data Files

The obvious alternative to checking all output for disclosure is to base statistical analysis on a set of individual data which itself is anonymised so that individuals cannot be identified. The step from this to releasing the individual level data is a small one and data files such as the General Household Survey and the Family Expenditure Survey have been available as unit record files for many years and have been distributed through the ESRC Data Archive. In fact this partnership has been excellent and these large Government surveys have been major sources of secondary analysis for university researchers and have been the data sets which have had the largest number of users through the Data Archive.

The release of a file of individual records from, for example, the General Household Survey, was based upon a judgement. So far as I am aware it was never supported by a rigorous analysis of the risks of disclosure. It was simply felt that if the sampling fraction for the survey was sufficiently small and if there was a very limited geographical identification for any particular record then these two factors together would protect individuals even though the information contained in the record itself may have been extremely detailed. That judgment, underpinned by confidentiality conditions for users, has been fully justified by our experience since there have been no cases of the release of confidential information from these surveys even though they have been used very extensively and have enlarged our understanding of a wide range of social situations.

5. The Census SAR

When the proposal was considered to extend this argument to the release of a sample of anonymised records from the census itself there was, as one would expect, considerable anxiety as to whether this could be done without the risk of disclosure. One could argue that the actual data content of the census is not that sensitive but the census itself is. It catches the imagination of the public when it is collected, has a high political profile and is considered to be special because in principle at least it contains information about every single member of society at a very detailed level of geographical location. It also represents in the public mind 'Officialdom' in a way which no sample survey does. For these reasons the sensitivity to any perceived threat to confidentiality is very great. Apart from the prime responsibility to protect the confidentiality of the census information itself, the risk that any disclosure would discredit future censuses, and therefore have a significant impact upon coverage and the quality of information, and the acceptance by the public of the census process, is real. Despite the recognised benefits of making a sample of anonymised records available it was essential, when the proposal to release a SAR was considered, that the risks of disclosure be fully assessed and that these should be acceptable to Government.

I shall not give a detailed account of the probabilistic framework which Cathie Marsh, Chris Skinner and others adopted but it did go through in a systematic way all of the conditions that would have to be met if a particular record released onto the sample of anonymised records was to be identified correctly and positively with any member of society. The argument turns to some extent on the amount of information which is available as a set of identification keys which can be used to match an individual from the population with the census records and so reveal the additional census information which is contained in the record. It is perhaps a matter of some amusement to outside observers, but not to census takers, that the lower the level of quality of the individual census variables which are collected then the lower is the risk of being able to correctly match any individual onto the file. In fact errors in data collection and coding help to protect individuals in society. The introduction of random perturbations into the data to protect confidentiality is an extension of the same process.

On the basis of that work it was recognised that we could not provide in one file all of the information which potentially at least an analyst might require. The key issue is the level of detail contained in combinations of variables and this implied that if we were to use a fine geographical identifier then the other information contained in a record would need to be relatively broad. It would be impossible to use fine categories for variables in conjunction with a detailed geographical location without an unacceptable risk of a breach of confidentiality. In fact it would be fair to say that there is no full theoretical framework for determining in practice the fineness of categories which can be used without risk of disclosure. OPCS used a procedure based on the expected frequency of occurrence of each category of a variable at the level of geographical identification chosen but there is much more to be done on this topic. This problem of variable combinations and fineness of categories is even more acute when using household data as opposed to individual data because the combination of variables from different individuals in the same households would more regularly lead to unique sets of information which could be used to identify households in the population.

It is hard to imagine why anyone would want both a very detailed geographical disaggregation of the census data together with a very detailed classification of the census variables such as occupation, industry and education but in any case if that is a requirement it is not one which OPCS felt able to meet because of the very real possibilities of disclosure. The solution adopted, of course, was to release two samples of anonymised records one relating to individuals with a relatively detailed level of geographical location but with a broadening of the categories of other census variables. The second sample of anonymised records related to households and has a very much broader geographical identity but with finer categories for the census variables relating to the members of the household. In this way we have tried to cater for levels of detail for both geography and for other variables but not in the same records.

When the case for the release of a sample of records from the census was made it was felt that, powerful as the theoretical analysis had been, in some ways it had not gone far enough. In effect the argument was based upon the probability of confidentiality being breached for someone chosen at random. What was really required was a set of conditional probabilities given some known information which was to be used as the set of matching keys. This would lead to different probabilities of disclosure for different people in the population depending upon their own personal characteristics. This is an extremely difficult analysis to carry out since it really depends upon both the combination of individual characteristics or household characteristics and the level of knowledge which the would-be discloser may have about that individual household. It also implies that: the greater the information available to the would-be discloser as a set of matching keys, the more the amount of information retained as confidential in the record is at risk.
It was found useful to augment the analysis provided by considering different approaches from which disclosures might occur. The first could occur where some large organisation with an extensive database of individuals could seek to match that file onto the sample of anonymised

records. In effect the attempt would be made as a 'many-to-many' matching operation. The analysis provided by Cathie Marsh and Chris Skinner provided a convincing case that the success rate of such an activity would be so low as to be effectively zero - it would certainly be wholly uneconomic as a commercial proposition.

However the census is an emotive subject and one must consider the possibility that there would be individuals who might attempt to demonstrate that confidentiality could be breached simply for the sake of it rather than for some commercial or other use of the data so disclosed. Such attempts seemed to fall into three categories. The first was the obvious search for a known person. The characteristics of the Prime Minister for example in terms of age, sex, geographical location and other family characteristics are in the public domain and it is entirely possible for someone to take that information and to seek a match on the SAR. This is a 'population-to-SAR' process. There is, of course, no guarantee that any such individual will have been selected in the SAR, but clearly this task would be easier for an individual or household with an abnormal combination of characteristics. Broadening the categories of variables, together with the requirement that the combination of variables should be known to be unique in the population, provides the main safeguard against disclosure.

The second approach is a 'SAR-to-population' process. Someone with an abnormal combination of characteristics is selected from the file and then an attempt is made to identify that person in the population. In this case the individual is known to be included in the SAR but the abnormal combination of characteristics must be in the public domain in some sense if they are to serve as matching keys. The dominant issue now is whether any combination of characteristics matched onto a person is known to be unique in the population and hence provides a true match. In both cases the issue turns on the level of detail contained in the file, its reliability for matching purposes and finally on whether or not a person so identified is known to be unique in the population as a whole. The SAR is itself only a one or two percent sample and there can be no assurance that there are not other individuals with the same combination of characteristics in the population. When this is taken in conjunction with error rates on the census data and other factors the risk of disclosure is low.

The third approach considered was a case where the SAR could be searched within a known geographical area for a number of people who had rather distinctive characteristics and for whom an auxiliary set of information could be available which would allow at least one of the individuals to be identified. Take for example the identification of doctors from the SAR by virtue of their qualifications and occupation. If we assume that it is relatively easy to get hold of a list of doctors in an area, together with perhaps some other personal characteristics, then a 'many-to-many' match could take place between doctors from the SAR with a list of doctors, in the hope that one individual match would be achieved and would be a true linkage. It was an assessment of risks of this kind which led to some further broadening of the detail with which individual data was provided on the SAR.

6. Legal and Contractual Underpinning

It is important to recognise that the final decision to release the sample of an anonymised records did not depend simply on the probabilistic analysis. It also depended upon a stringent set of conditions that would be imposed upon every would-be user of the SAR together with a strong legal backing which would act as both a legal barrier and a deterrent to any would-be discloser. This, together with the professional integrity of those who use the SAR, is an important element in safeguarding confidentiality.

7. The Longitudinal Study

The arrangements for the release of the census data are by no means the most stringent which the Government Statistical Service has to impose in order to protect the confidentiality of individuals. There are data sets which cannot be provided in an anonymised form if they are to remain useful. If we turn to the Longitudinal Study for example which contains individual data of an extremely sensitive kind which is as accurate as possible and has not been subjected to any perturbation or rounding to anonymise it, then it is clear that these data could not simply be put into the public domain at the individual level. The risks to individuals if disclosure were to occur is too great. However the data set has huge potential and was created in order for that potential for longitudinal analyses to be fulfilled. The original purposes have long been exceeded and the Longitudinal Study is now used to understand a wide variety of social and medical situations. The data set is held within OPCS and can only be accessed and used by special arrangements which are provided in cooperation with City University and the ESRC. Of course these arrangements include strict confidentiality requirements being placed upon all analyses. In this way we are working with researchers and allowing them to utilise the full potential of the data that we hold whilst at the same time protecting the confidentiality of the people from whom those data were obtained. The process is cumbersome and expensive but it is a necessary price in order to protect the individuals. Some of the work supported through OPCS requires not only an undertaking to preserve confidentiality but also must be approved by an OPCS ethics committee.

8. Other Data Releases

Apart from the obvious well-known examples, such as the Census SAS and SAR, the GHS, FES, LFS etc., the Government Statistical Service releases a wide range of additional data sets containing individual records or very highly disaggregated data. For example the Department of Transport releases not just the National Travel Survey to the ESRC Data Archive, but also the Road Traffic Accident database. MAFF release Census of Agriculture small area statistics and may do more.

Of particular importance in terms of releasing data are the territorial offices of Scotland, Wales and Northern Ireland who all work closely with their own communities of researchers and analysts to make data available.

There are plans for the future which may come to fruition and will release important sources of data for new analyses:

(i) the Inland Revenue are exploring the possibility of releasing an anonymised data set of personal incomes;

(ii) the Department of Health is investigating issues around the release of a longitudinal data set on 'children looked after'. This would almost certainly have to follow similar release procedures to the longitudinal study;

(iii) the Home Office has, or is planning, to release several data sets of individual records for research purposes subject to the appropriate confidentiality undertakings. These are:

(a) a subset of registered drug addicts which will be supplied to the Centre for Research on Drugs and Health Behaviour. This will be matched with information from the North Thames Drug Reporting Database and will allow estimates of prevalence of drug misuse in the North Thames region;

(b) a data set of drug offenders in Wales dealt with by courts has been supplied to the University of Wales. Together with other information this will permit estimates of prevalence for injecting drug users in Wales;

(c) anonymised samples can be created from the Offenders Index which contains the criminal histories of people convicted of a standard list of offences since 1963 in England and Wales.

9. The Office for National Statistics

Finally, I turn to the merger of CSO and OPCS and the new responsibility to make full and effective use of the very wide range of data available across Government. The challenges are daunting and first priorities will include greater standardisation of definitions and classifications together with better information about the data that exists within the GSS. The intention is to make these data accessible to a much greater extent and to allow them to be brought together. Geography will be a key element of these developments and we hope to be able to provide more data on a geographically consistent basis. It is too early to say how this programme will develop but there will be issues of confidentiality which will need careful consideration if highly disaggregated data are accessed. However, we see this as a partnership endeavour across the whole of the GSS and a partnership

endeavour with other organisations such as the ESRC Data Archive in helping us to make the data accessible. The whole programme will be developed over the coming months and should be a concrete sign of the creation of the new Office for National Statistics.

10. Conclusion

I started by quoting the GSS mission and drawing attention to the tension between our responsibility to protect confidentiality and our role to make our data accessible and get it used. We have, I think, a good record of balancing these two tensions and providing a service to the whole community through intermediaries such as university researchers. We can and should do more and we need to respond to the changing environment for information transmission and processing as it continues to develop. I also began by paying tribute to Cathie Marsh and I hope that my presentation has illustrated the debt which British social science owes to her for her work on confidentiality and disclosure risk.

SOME METHODOLOGICAL ISSUES IN STATISTICAL DISCLOSURE CONTROL[1]

L.C.R.J. Willenborg, A.G. de Waal and W. J. Keller, Statistics Netherlands, Research and Development Division

1. Introduction

In the last decade the demand for detailed information has increased considerably. This is mainly due to the great power of modern PC's, which enables researchers to analyse large data sets by themselves, whereas in former days only national statistical offices were able to analyse such data. The increased demand for detailed information becomes clear from the data that are released by statistical offices. Whereas in the old days relatively small two-way tables were sufficient to satisfy most of the users' demands, nowadays large three- and higher-dimensional tables are not exceptional. Microdata sets, i.e. data on individual respondents, are relatively new products of statistical offices, and contain a wealth of information. However, both the release of large tables and of microdata sets lead to considerable problems when trying to protect the privacy of respondents. Microdata especially sets create interesting challenges in the field of statistical disclosure control (SDC).

In this paper we examine how Statistics Netherlands is dealing with the problems of SDC. The emphasis is on SDC for microdata sets rather than for tables, because SDC for microdata is a relatively new, and still developing, subject. Statistics Netherlands releases two kinds of microdata sets, namely public use files and microdata sets for research. The current SDC-rules and techniques for these microdata sets that are applied at Statistics Netherlands are examined in more detail in Section 2.

Applying these rules and techniques is not a straightforward matter. On the contrary, a number of methodological problems must be solved in order to apply these rules and techniques appropriately. For instance, the population frequency of a combination of values of identifying variables is usually not known, and has to be estimated from a sample. Another important problem is how to apply the SDC-techniques in such a way that the resulting microdata is considered safe, while the information loss due to these measures is minimised. These and other subjects are examined in Section 3.

Although Statistics Netherlands has SDC-rules for its microdata sets, this does not imply that they will remain fixed forever. In the future these rules and techniques are likely to change. This will occur not only because society itself is constantly changing, but also because of some dissatisfaction with the present rules. We can imagine better rules, but these rules, unfortunately, require a methodological understanding that is greater than our present understanding. In Section 4, a number of potential improvements for our rules are examined. All these potential improvements require further theoretical research, however.

[1]The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

Of course, apart from microdata sets Statistics Netherlands also releases tables. Several SDC-problems for tables are discussed in Section 5. The aim of that discussion is not to provide a complete description of the way in which Statistics Netherlands handles SDC-problems for tables, but rather to illustrate a number of similarities and differences between SDC for microdata and for tables. The paper is concluded with a brief discussion in Section 6.

2. SDC For Microdata at Statistics Netherlands

The basic idea of the SDC-rule for microdata applied at Statistics Netherlands is that certain combinations of values of variables should occur frequently enough in the population. We begin our exposition by explaining the underlying philosophy of this basic idea.

2.1 The basic idea

The aim of statistical disclosure control is to limit the risk that sensitive information of individual respondents can be disclosed from a data set. Such a disclosure of sensitive information of an individual respondent can occur after this respondent has been re-identified, i.e. after it has been deduced which record in a given microdata set corresponds to this particular individual. SDC should therefore hamper the re-identification of individual respondents. Re-identification can take place when several values of so-called identifying variables, i.e. variables of which the value can be used alone or in combination with the values of other such variables to re-identify a respondent, are taken into consideration. The values of these identifying variables can be assumed known to friends and acquaintances of a respondent. Examples of identifying variables are 'Place of residence', 'Sex' and 'Occupation'.

An important concept in the theory of re-identification is a *key*, i.e. a combination of identifying variables. The dimension of a key is the number of identifying variables present in this key. Re-identification of a respondent can occur when this respondent is unique in the population with respect to a certain key value, i.e. a combination of values of identifying variables. Hence, uniqueness of respondents in the population with respect to certain key values should be avoided. When a respondent appears to be unique in the population with respect to a particular key value, then disclosure control measures might be called for to protect this respondent against re-identification. In practice, however, it is not a good idea to try to prevent only the occurrence of respondents in the data file who are unique in the population (with respect to a certain key). Several reasons can be given for this. Firstly, there is a practical reason: unicity in the population, in contrast to unicity in the data file, is hard to establish. There is no way to determine whether a person who is unique in the data file (with respect to a certain key) is also unique in the population. Secondly, an intruder may use other keys than those considered by the data protector. For instance, the data protector may consider only keys consisting of at most three variables while the intruder may use a key consisting of four variables. Therefore, it is better to avoid the occurrence of combinations of values in the data file that are rare in the population rather than restricting attention to avoid population-uniques in the data file.

When the frequency of a combination of values of identifying variables is sufficiently high then this combination is considered safe, otherwise it is considered unsafe. If a record contains any unsafe combinations then this record may not be published in its present form, and appropriate SDC-measures should be applied.

2.2 Public use files and microdata for research

Now that we have examined one of the basic ideas behind the SDC-rules applied at Statistics Netherlands it is time to consider the two kinds of microdata sets and their corresponding rules that are disseminated by Statistics Netherlands. In neither kind of microdata sets, formal identifiers, such as name or address, are published, of course. The first kind of microdata sets are so-called *public use files*. A public use file can be obtained by everybody. The data contained in a public use file should be at least one year old. The keys that have to be examined for a public use file are all combinations of two identifying variables. The (estimated) population frequency of a value of an identifying variable has to be at least d_1, the (estimated) population frequency of a bivariate combination has to be at least d_2, where d_1 and d_2 are fixed threshold values ($d_1 > d_2$). The number of identifying variables is limited, and identifying variables referring directly to a region of residence, work or education, such as 'Place of residence' are not included in a public use file. Very sensitive variables, such as variables on sexual behaviour or criminal activities, may not be included in a public use file. Sampling weights have to be examined before they can be included in a public use file, because there are situations in which these weights can give additional identifying information. For instance, when a certain subpopulation is oversampled then this subpopulation can be recognised by the relatively low weights associated with its members in the sample. Sampling weights may only be published when they do not provide additional information that can be used for disclosure purposes. Re-grouping of households should be prevented, because households are more likely to be unique on a low-dimensional key than the constituting individuals. For this we check whether combinations of so-called household variables, i.e. variables relating directly to a household such as 'Number of persons in the household' and 'Occupation of the head of the household', occur in a sufficient number of households. When this is not the case SDC-measures, such as recoding the household variables, should be taken. Re-grouping of the records by taking into account (by an intruder) that the records in a microdata set are usually sorted in a particular order, e.g. all members of a households are placed consecutively or all respondents from the same region are placed consecutively, should also be prevented. Such a re-grouping of the records could provide additional identifying information. To prevent re-grouping the rules demand that before a public use file is released the order of the records should be randomised. Finally, the rules rarely allow information on region of residence, work or education in a public use file. Only one of these three kinds of regional information may be included in a public use file.

Moreover, it has to be established that the regions that can be distinguished in the microdata set are sufficiently scattered over the country. Where they are not sufficiently scattered, they form compact areas in which it would be relatively easy to trace a particular respondent. Checking whether the regions that can be distinguished in the microdata set are sufficiently scattered over the country

involves all variables that provide any regional information. For example, the variable 'Number of swimming pools in your place of residence' is taken into consideration when performing the checks.

The second kind of microdata sets are so-called *microdata sets 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 identifying variables with much regional detail, such as 'Place of residence', may be included. 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. Detailed regional information may be included. The variables with information on region of residence, work and education should be crossed. The resulting variable is called *the regional variable.*

The keys that have to be examined for a microdata set for research include three-way combinations of the regional variable with variables describing the sex, ethnic group or nationality of a respondent with an ordinary identifying variable. The (estimated) population frequency of these trivariate combinations should be at least d_0, where d_0 is a fixed threshold value. The value of d_0 is less than the threshold value d_2 for bivariate combinations in the case of a public use file. The number of persons in a region that can be distinguished in a microdata set for research should be at least 10,000. Finally, the information that may be given on 'Occupation', 'Employer' and 'Education' of a respondent depends on how much regional information is given in this data set. If much regional information is given then little information may be provided on the other three subjects. For more information on the kinds of microdata sets released by Statistics Netherlands and their rules we refer to Keller and Willenborg (1993).

2.3 Disclosure protection measures

The SDC-rules of a statistical office to determine whether a microdata set is considered safe for release is an important part of its SDC-policy. Another, equally important, part is formed by the techniques that are applied when it turns out that a particular microdata set is considered unsafe for release. Statistics Netherlands advocates two SDC-techniques to protect unsafe combinations in microdata sets, namely global recoding and local suppression. In case of global recoding several categories of a variable are collapsed into a single one. A global recoding is applied to the whole data set, not only to the unsafe part of the set. This is done to obtain an uniform categorisation of each variable. When local suppression is applied one or more values in an unsafe combination are suppressed, i.e. replaced by a missing value. A local suppression is only applied to a particular value. Both global recoding and local suppression lead to a loss of information, either because less detailed information is provided or some information is not given at all. Note that local suppressions may lead to a bias in the data, because the extreme combinations are removed from the data set. This is not considered to be a serious problem by Statistics Netherlands because we try to limit the number of local suppressions as much as possible, i.e. we try to protect most unsafe combinations by global recodings. We only apply local suppression when much information would be lost if certain unsafe combinations were protected by global recodings. An optimal mix of global recodings and local

suppressions has to be found in order to make the information loss due to these SDC -measures as low as possible.

The above two techniques are non-perturbative ones, i.e. they do not modify the values of the variables. There are two main reasons why Statistics Netherlands does not apply any perturbative techniques at the moment. Firstly, it seems hard to ensure the statistical quality of the resulting data when perturbations are applied on a large scale. However, when the number of values that are perturbed is small, i.e. when the perturbations are applied as parsimoniously as the local suppressions, then the statistical quality of the resulting microdata set is bound to be acceptable (Cf. Barnes, 1995). Secondly, a disadvantage of perturbation is that the resulting data may be impossible or very unlikely to occur in real life. For example, when the variable age has been perturbed there may be a record in the resulting data set with the following values 'Age=12 years' and 'Marital status=widowed'. This is highly implausible.

Checking all the SDC-rules is a time-consuming and error prone job. Therefore Statistics Netherlands is developing a general purpose software package for SDC of microdata (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 analyse the data and to carry out suitable protection measures. The structure of the package is such that it will be possible to specify different disclosure control rules. This implies that the package will be suited for other statistical offices too. Moreover, it should be possible to incorporate changes in the rules fairly easily in the package. The strongest feature of the current version of the package is the ability to determine the necessary local suppressions automatically and optimally, i.e. the number of local suppressions is minimised, after the global recodings have been determined interactively.

The further development of ARGUS is a major goal of an ESPRIT-project on SDC. In fact, a package for the SDC of microdata, μ-ARGUS, and a package for the SDC of tabular data, τ-ARGUS will be developed. The participating institutions in this project are 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 Consorzio Padova Ricerche (CPR), and Statistics Netherlands, acting as project co-ordinator.

3. Methodological Problems

As one can conclude from Section 2, rules for SDC of microdata applied at Statistics Netherlands are for a substantial part based on testing whether values on certain key variables occur sufficiently frequently 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 have a value on a certain key, and how to determine appropriate SDC-measures.

The keys that have to be examined at Statistics Netherlands are prescribed by the rules once it has been determined which variables should be considered identifying. For this, Statistics Netherlands applies a number of criteria, such as the visibility of the values of a variable and the tractability of

these values. These criteria do not reach a definite verdict for all variables, however. In many cases deciding whether a variable should be considered identifying is a matter of personal judgement.

When applying one of the threshold rules mentioned in Section 2 to determine whether or not a combination of values of identifying variables occurs sufficiently frequently in the population we are generally posed with the problem that we do not know this population frequency. Often we have only a sample available to us. Estimating the population frequency of a certain combination is especially a problem when one of the values corresponds to a region. In fact, we then have to estimate the population frequency of a combination in a certain region instead of in the entire country.

For large regions it is possible to use an interval estimator to test whether or not a key value occurs sufficiently frequently. 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 this estimator to have a high variance which in turn leads to a lot of records that need to be modified. To estimate the number of times that a key value occurs in a small region we therefore suggest applying a point 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 combining this estimator and the direct estimator. Such a combined estimator has been tested in Pannekoek and De Waal (1995) and the results are encouraging.

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 we at Statistics Netherlands use 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), although we may apply a point estimator in the future. If there is a sufficient number of persons in the population that score on a comparable value, then the extreme value may be published, otherwise the extreme value must be locally suppressed or the corresponding variable should be globally recoded. 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. As has been mentioned in Section 2 Statistics Netherlands currently applies only local suppressions and global recodings to protect microdata sets. Our aim is therefore to determine the optimal mix of these local suppressions and global recodings.

In De Waal and Willenborg (1994a) 0-1 integer programming formulations for determining the optimal local suppressions are presented. These formulations all aim to minimise the information loss while protecting the microdata set. They differ with respect to the way in which a data protector wants to measure this information loss. For example, the data protector can decide to minimise the total number of locally suppressed values, or he can decide to minimise the number of different locally suppressed categories. The data protector can also decide to combine these goals, e.g. minimise the number of different locally suppressed categories given that the total number of locally suppressed values has been minimised. As will be clear the information loss due to local suppressions only can be determined by very simple measures, namely the total number of locally suppressed values or the number of different locally suppressed categories.

Determining the optimal global recodings, or an optimal mix of global recodings and local suppressions is much more difficult. Measuring the information loss due to global recodings is already a problem. A simple information measure is not available, in contrast to the case of local suppressions only. In De Waal and Willenborg (1995c) this problem of measuring the information loss is solved by using the entropy. Both the information loss due to global recodings and the information loss due to local suppressions can be evaluated by this measure. To evaluate the information measure based on the entropy it is necessary to specify a model for the way in which the users of the microdata will deal with the missing values. For instance, we can assume that for a quantitative variable they will simply replace each missing value by the average value of this variable. In this case the information loss due to local suppression will be rather high. When the users of the data are supposed to be somewhat smarter, we can assume that they use multivariate techniques to impute the missing values. In other words, we could assume that the users of the data explicitly take the logical and statistical dependencies of the data into account. In this case the actual information loss due to local suppression will be less high. Note that if the information loss due to local suppressions is very low, these local suppressions are less effective for protecting data than they appear to be at first sight (also see the remarks on complex microdata in Section 4).

For public use files a number of additional problems have to be solved. For instance, for these files sampling weights may not provide additional identifying information. In De Waal and Willenborg (1995a) it is shown however that in many cases such additional identifying information can be obtained from the sampling weights. There are two ways to prevent this derivation of additional identifying information, namely subsampling and adding noise to the sampling weights. Subsampling, i.e. deleting records from the microdata set, has the advantage that it is easy to apply, but has the disadvantage that it may lead to a considerable loss of information. Adding noise is more difficult to apply. On the one hand it should not be possible to derive additional identifying

information from the sampling weights after noise has been added; on the other hand the statistical quality of the sampling weights should be sufficiently high. The former condition can easily be satisfied by adding much noise to the sampling weights, but then the latter condition will be violated, and vice versa.

4. Towards a Foundation of SDC for Microdata

The SDC-rules and techniques described in the previous sections are based on intuitive reasoning rather than on a formal mathematical model. All rules and techniques reduce the re-identification risk, but it is not possible to evaluate this reduction of the re-identification risk. This is a somewhat undesirable situation. Ideally, we would like to have a model for the re-identification risk per record. When such a model would be available the SDC-rules would only have to prescribe the keys that should be checked and the maximum risk that the statistical office releasing a particular microdata set is willing to take. When the actual re-identification risk of a record is less than this maximum risk then the record may be published without modifications, otherwise the record should be modified.

Several efforts have been made to develop such a re-identification risk per record model. Some of these efforts did not take the 'noise' in the data, e.g. due to measurement errors, into consideration (e.g. Verboon, 1994; Verboon and Willenborg, 1995). Other efforts did take 'noise' in the data into consideration (e.g. Paass and Wauschkuhn, 1985; Fuller, 1993). Unfortunately, these attempts have not yet produced a satisfactory model for the re-identification risk per record.

Somewhat less ambitious is a model for the re-identification risk for an entire microdata set, i.e. the risk that an unspecified record from the microdata set is re-identified. Again the SDC- rules are very simple when such a model is available. Only the keys to be checked and the maximum risk that the statistical office releasing a microdata set is willing to take should be prescribed. If the actual risk for the entire microdata set is higher than the maximum risk then appropriate SDC-measures should be taken. In this case, it is not clear, however, which records should be modified by these measures, because no model for the re-identification risk per record is available.

A model for the re-identification risk per microdata set has been proposed by Mokken et al. (1989, 1992). This model takes three probabilities into consideration. The first probability, 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 set knows the values of a randomly chosen person from the population with respect to a certain key K. The third probability, f_u, is the probability that a randomly chosen person from the population is unique in the population with respect to a certain key K. Under various assumptions, some of which are unrealistic (e.g. that no measurement errors have been made), an expression for the re-identification risk per set, D_R, is derived, namely

$$D_R = 1 - \exp(-Nff_af_u), \qquad (1)$$

where N is the population size.

To evaluate this expression it is necessary to calculate f, f_a and f_u. The sampling fraction f is easy to calculate, of course. The other two probabilities, f_a and f_u, are more difficult to calculate, however. Evaluating f_a seems very hard, because this probability depends on the specific researcher and his knowledge of the population. To estimate f_u a number of models have been proposed in the literature. Models to estimate the number of uniques, and hence the value of f_u, include the Poisson-gamma model (Bethlehem et al., 1989; Mokken et al., 1989; Willenborg et al., 1990; De Jonge, 1990) and the Poisson-lognormal model (Skinner and Holmes, 1992; Hoogland, 1994). Because the results of these and other models are rather unreliable, and because it is very hard to evaluate f_a, the model by Mokken et al. cannot be used in practice to evaluate a re-identification risk per microdata set.

Another approach that is possibly of interest to gain an insight into the re-identification risk per record, although it does not provide a way to actually evaluate such a re-identification risk, is *fingerprinting*. The idea of this approach is that the records that are the most likely ones to be re-identified are those that are often unique on a low-dimensional key. An SDC-rule based on fingerprinting could be the following one: a record is considered unsafe, and hence may not be released unmodified, if it is unique on more than m k-dimensional keys. Such a rule cannot be applied easily, however, because the number of keys that have to be examined becomes astronomically large even for moderate values on k. For example, suppose that there are 50 identifying variables in a microdata set. Suppose furthermore that the values of m and k equal 10 and 6, respectively, i.e. a record is considered unsafe when it is unique on more than 10 keys consisting of at most 6 identifying variables. In this case, an upper bound for the number of combinations of variables that have to checked is about 16 million, equalling the number of ways to select 6 elements from a set of 50 elements. There are several ways to overcome this practical problem. Firstly, one can decide to consider only lower-dimensional keys, say k equals 3 or 4. Secondly, fingerprinting is highly suited for parallel computing. In a distributed computing environment, e.g. a network of PC's, fingerprinting could be done very efficiently. However, further research is needed for fingerprinting to be effectively applied in practice.

So far in this section we have only discussed methods to evaluate the re-identification risk, either per record or per microdata set. There are several other issues that deserve further investigation. An important issue is a practical one, namely determining the local suppressions automatically and optimally in so-called complex microdata, i.e. microdata with logical and statistical dependencies explicitly taken into account. When determining the local suppressions these dependencies should be taken into account. For example, when the value of the variable 'Number of children you have given birth to' in a certain record equals '2', then local suppression of the value of the variable 'Sex' in this record does not offer any protection against disclosure, because it is clear that this value is 'Female'. Such dependencies can be incorporated in the 0-1 integer programming formulations to determine the optimal local suppressions (Cf. De Waal and Willenborg, 1995b). Efficient algorithms to solve the resulting problems (to good approximation) remain to be found.

5. SDC for tables

There are many similarities between SDC of microdata and tables. For instance, when trying to reduce the risk of disclosure one usually starts by modifying the identifying variables. In the case of microdata one collapses the categories of an identifying variable, in the case of tabular data one collapses two columns or rows of the table. After the global modifications have been made local modifications must be made. In the case of microdata values of identifying variables in some records can be changed to 'missing', whilst with tabular data values of sensitive cells can be changed to 'missing'. This example also illustrates a difference between SDC for microdata and SDC for tabular data: with microdata we locally suppress values of identifying variables, whereas with tabular data we suppress values of sensitive data. In this section we examine such similarities and differences between SDC for microdata and tabular data.

First of all note that in the literature on SDC for tables it is generally assumed that the published tables are based on an observation of the entire population. The disclosure problem of tabular data in the situation where only a sample of the population is observed is hardly discussed. In the sequel we also assume that the tables are based on an observation of the entire population.

After some columns and/or rows have been collapsed it is necessary to make some local modifications. A well-known technique to modify tabular data in order to safeguard against disclosure is cell suppression, which can be compared to local suppression in microdata. To apply suppression in tables the cells that contain sensitive information have to be determined. The usual way to determine whether a cell is sensitive is by means of a dominance rule. A dominance rule states that if the values of the data of a certain number of respondents, say 3, constitute more than a certain percentage, say 75%, of the total value of the cell, then this cell is sensitive. The main idea on which this approach is based is that when a cell is dominated by the contributions of a few respondents, then these contributions can be estimated rather accurately. For instance, if there is only one respondent then his contribution can be disclosed exactly. When there are exactly two respondents then each of these respondents can disclose the contribution of the other, and when the value of a cell is dominated by the contributions of two respondents, then each of these respondents is able to estimate the value of the contribution of the other one accurately. In general, if there are k respondents then $k-1$ of them, after pooling their information, can disclose information about the value of the data of the remaining respondent. For small k, say, 2, 3 and 4, this poses a problem.

The sensitive cells in tables can be compared to the unsafe combinations in case of microdata. Like the unsafe combinations in case of microdata, sensitive cells have to be protected by suppressing, recoding or perturbing their values. We first consider suppression of sensitive cells.

The suppression of a cell because the contents of this cell are considered sensitive according to some sensitivity criterion, e.g. a dominance rule, is called *primary suppression*. Primary suppression alone is generally not sufficient to obtain a table which is safe for release. In a table the marginal totals are often given as well as the values of the cells. A cell which has been suppressed can then be computed

by means of the marginal totals. Therefore, other cells have to be suppressed in order to avoid this possibility. This is called *secondary suppression.*

Like local suppression in the case of microdata, secondary suppression in tables should be done in such a way that information loss is minimised. Usually, weights are assigned to the cells in a table. The information loss due to suppression of a cell is then given by the corresponding weight. There are several ways of specifying the weights. For instance, the weight of a cell can be chosen to be equal to the number of respondents in the cell. In this case, one aims at minimising the number of respondents whose data are suppressed in the table. Alternatively, the weight of a cell can be chosen to be equal to the cell value. In this case, one aims to minimise the total value of the data which are suppressed. Selecting 'good' weights involves subjective considerations.

Secondary suppression causes other problems as well. Although it might be impossible to compute the exact values of suppressed cells in a table after secondary suppression, it is still possible to compute the ranges in which the values of the cells lie, when the marginal totals of the table are given and it is, for instance, known that the values of the cells are all non-negative (Cf. Geurts, 1992). If these ranges are small for the sensitive cells, then an attacker is able to obtain good estimates for the values in the suppressed cells. Therefore, secondary suppression must be carried out in such a way that the ranges in which the values of the suppressed cells lie are not too small. Tables with marginal totals can be compared to complex microdata, because in both cases dependencies between (cell) values should be taken into account.

Another well-known technique to protect sensitive cells in a table against disclosure is *rounding*. The most interesting way of rounding is controlled rounding (Cf. Fellegi, 1975; Cox, 1987). The main advantage of controlled rounding compared to conventional rounding and random rounding is that the additivity of the tables is preserved, i.e. after rounding the rows and columns still add up to their rounded marginal totals. A slight disadvantage of controlled rounding is that a cell value is not necessarily rounded to its nearest integer multiple of the rounding base b, but rather to one of its two nearest integer multiples of b. Controlled rounding for two-dimensional tables does not provide serious problems any more. Note that rounding is a 'stylised' way of perturbing the data, i.e. adding noise to the data. Hence, to some extent rounding in tables is comparable to perturbation in a microdata set.

The protection offered by rounding the sensitive cells should be approximately the same as if suppression had been applied. In the case of suppression the range in which the value of a sensitive cell must lie should be sufficiently wide, say the width of this range should be at least p% of the cell value. The range of ambiguity offered by rounding should then also be at least p% of the value of a sensitive cell. From this criterion a value for the rounding base b can be derived for a given value p.

Three- and higher-dimensional tables and linked tables, i.e. tables with common variables obtained from the same base file, pose a lot of theoretical problems (Cf. De Vries, 1993). The theory for these kinds of tables is much more difficult than for ordinary two-dimensional tables. An interesting similarity between microdata and linked tables is the following. Suppose that one wants to protect a

set of linked tables by recoding the variables. Suppose furthermore that one wants to use the same categorisation for each variable in each table where this variable occurs. The aim is to protect the linked tables in such a way that the information loss is minimised. This problem is similar to the so-called global recoding problem for a microdata set, i.e. the problem of applying only global recodings in such a way that the resulting data set is safe while the information loss is minimised. A solution for the global recoding problem for a microdata set implies a solution to the recoding problem for linked tables and vice versa.

For secondary suppression in three- and higher-dimensional tables some heuristic algorithms are available, but much work still has to be done in order to perfect these algorithms. This is another subject that attracts the attention of Statistics Netherlands. Controlled rounding of higher-dimensional tables is a difficult problem. In some cases the problem is impossible to solve (Cf. Cox, 1987). In these cases it is necessary to relax the conditions of controlled rounding. For instance, instead of demanding that each value is rounded to one of its two nearest integer multiples of the rounding base, one could specify a window of values in which the rounded value of a cell should lie. Some heuristics to deal with three-way tables have been developed (Cf. Fagan et al., 1988; Kelly, 1990; Kelly et al., 1990). For four- and higher-dimensional tables satisfactory heuristics are hard to find. Another interesting problem is controlled rounding for linked tables. It is not possible to round each of these tables separately, in case the same marginal totals occur in several tables.

6. Discussion

As a consequence of the increased demand for detailed information Statistics Netherlands has disseminated a considerable number of microdata sets in recent years. The SDC-rules for these microdata sets were, and still are, based on intuitive reasoning and still lack a solid theoretical framework. The main idea of these rules is that the population frequencies of certain combinations of values of identifying variables have to be checked. The population frequency of such a combination should be sufficiently high, otherwise SDC-measures should be taken.

After the laborious process of developing the SDC-rules it was soon realised that applying them in practice is a nontrivial exercise. Instead it turned out that many methodological problems had to be solved. Some of these problems, such as determining the local suppressions automatically and optimally, have been solved (more or less) by now. Others, such as determining the global recodings automatically and optimally still remain to be solved. It was also realised that in order to apply the SDC-rules in practice a software package, ARGUS, should be developed. Without such a package the application of the SDC-rules would be very time-consuming and error-prone.

Although Statistics Netherlands has developed SDC-rules for its microdata sets, this does not imply that it is time to relax. In fact, we are somewhat dissatisfied with our rules. Better rules could be deduced if a good model for the re-identification risk per record would be available. Unfortunately, such models do not seem to be available at the moment.

Fortunately for researchers in the field of SDC, but unfortunately for statistical offices trying to protect their microdata sets, SDC for microdata sets offers many possibilities for future research. In the long run a model for the re-identification risk per record should be developed. Until that time, the present SDC-rules and techniques should be refined. For instance, global recodings should be automated and optimised, local suppressions (and global recodings) in complex microdata should be automated and optimised, and sampling weights should be protected efficiently. Only by continued research efforts the present and future challenges of SDC for microdata can be met adequately.

References

Barnes, G. , 1995, Local perturbation. Report, Statistics Netherlands, Voorburg.

Bethlehem, J.A., W.J. Keller and J. Pannekoek, 1989, Disclosure control of microdata. Journal of the American Statistical Association, Vol. 85, no. 409, 38-45.

Cox, L.H., 1987, A constructive procedure for unbiased controlled rounding. Journal of the American Statistical Association, Vol. 82, 520-524.

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

De Vries, R.E., 1993, Disclosure control of tabular data using subtables. Report, Statistics Netherlands, Voorburg.

De Waal, A.G. and A.J. Pieters, 1995, ARGUS user's guide. Report, Statistics Netherlands, Voorburg.

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

De Waal, A.G. and L.C.R.J. Willenborg, 1994b, Development of ARGUS: past, present and future. Report, Statistics Netherlands, Voorburg.

De Waal, A.G. and L.C.R.J. Willenborg, 1995a, Statistical disclosure control and sampling weights. Report, Statistics Netherlands, Voorburg.

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

De Waal, A.G. and L.C.R.J. Willenborg, 1995c, Optimum global recoding and local suppression. Report, Statistics Netherlands, Voorburg.

Fagan, J., B. Greenberg and B. Hemming, 1988, Controlled rounding of three-dimensional tables, Statistical Research Division Report Series, Bureau of the Census, Washington DC.

Fellegi, I.P., 1975, Controlled random rounding. Survey Methodology, Vol. 1, 123-133.

Fuller, W.A., 1993, Masking procedures for microdata disclosure limitation. Journal of Official Statistics, Vol. 9, no. 2, 383-406.

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 J.A. Bethlehem, 1992, Disclosure protection of microdata: problems and solutions. Statistica Neerlandica, Vol. 46, no. 1, 33-48.

Keller, W.J. and L.C.R.J. Willenborg, 1993, Microdata release policy of the Netherlands CBS. Proceedings of the International Seminar on Statistical Confidentiality, Dublin.

Kelly, J.P., 1990, Confidentiality protection in two- and three-dimensional tables. Ph.D. thesis, University of Maryland, College Park, Maryland.

Kelly, J.P., B.L. Golden and A.A. Assad, 1990, Controlled rounding of tabular data. Operations Research, Vol. 38, 760-772.

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

Mokken, R.J., P. Kooiman, J. Pannekoek and L.C.R.J. Willenborg, 1992, Disclosure risks for microdata. Statistica Neerlandica, Vol. 46, no. 1, 49-67.

Paass, G. and U. Wauschkuhn, 1985, Data access, data protection and anonymisation - analysis potential and identifiability of anonymised individual data (in German). Gesellschaft für Mathematik und Datenverarbeitung, Oldenbourg-Verlag, Munich.

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

Pannekoek, J. and A.G. de Waal, 1995, Synthetic and combined estimators in statistical disclosure control. Report, Statistics Netherlands, Voorburg.

Pieters, A.J. and A.G. De Waal, 1995, A demonstration of ARGUS. Report, Statistics Netherlands, Voorburg.

Skinner, C.J. and D.J. Holmes, 1992, Modelling population uniqueness. Proceedings of the International Seminar on Statistical Confidentiality, Dublin.

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 L.C.R.J. Willenborg, 1995, Comparing two methods for recovering population uniques in a sample. Report, Statistics Netherlands, Voorburg.

Willenborg, L.C.R.J., R.J. Mokken and J. Pannekoek, 1990, Microdata and disclosure risks. Proceedings of the 1990 Annual Research Conference, Bureau of the Census, Washington DC, 167-180.

Willenborg, L.C.R.J., A.G. De Waal, R.E. De Vries and C.A.W. Citteur, forthcoming, Statistical disclosure control in practice. Springer-Verlag, New York.

CONFIDENTIALITY AND INFORMED CONSENT
A VIEW FROM THE COMMERCIAL SECTOR

Michael Warren, Director General, The Market Research Society

1. Introduction

It is common ground that there has tended to be an atmosphere of tension (often constructive tension but not always) between the market and social research worlds. This is less true than it was and I hope that seminars like these can help to bring our two worlds - or our two halves of the same world - closer together.

My viewpoint, not surprisingly, is that of a market research professional. However, by choice I concentrated on the social research end of the spectrum during more than 20 years in the industry, firstly providing or selling research with a major agency, and subsequently buying research, firstly with Consumers' Association and secondly with the Central Office of Information.

It is my assumption that, whether we like it or not, the majority of our informants will get the bulk of their awareness of research from market as opposed to social research. (The opinion polls fall somewhere between the two and also have an impact, but I shall not be considering them in this talk.)

This paper is a series of observations, looking firstly at informed consent, secondly at public attitudes to our work and thirdly at legislative threats. I will leave others to decide whether my observations are accurate and fair and whether, if they are, anything needs to be done.

2. Informed Consent

I have been asked to look in particular at the question of informed consent in the interview process. In other words, to look at the extent to which our informants are fully aware of their rights as interviewees and of the implications of agreeing to be interviewed. It would be easy - too easy - to take the cynical view that informed consent is little more than a theoretical nicety, an intellectual comfort blanket which we can grab and cuddle whenever we feel vulnerable about some of the work we do. The reality may be equally troubling, but it is much less straightforward.

Let us start with a few definitions. My sources are the Code of Conduct of the Market Research Society, the International Code of Marketing and Social Research Practice from ESOMAR (the

European Society for Opinion and Marketing Research), and of course the Social Research Association's own Ethical Guide-lines.

Taking them in that order, the MRS talks of research being founded on "the willing co-operation" of the public and being done "honestly, objectively (and) without unwelcome intrusion". ESOMAR, similarly, talks of respondents' co-operation being "entirely voluntary at all stages".

The SRA's document is a great deal more thorough. It talks of social enquiry being based "as far as practicable" (we must return to this caveat at a later stage) "on the freely given informed consent of subjects". It continues: "In voluntary inquiries, subjects should not be under the impressing that they are required to participate...They should be aware of their entitlement to withdraw at any stage for any reason and to withdraw data just supplied...".

This is followed by a list of a dozen items, ranging from the purpose of the study through to the proposed data storage arrangements, which might be "material to a subject's willingness to participate" and which therefore, in some circumstances, the informant should know, presumably to be able to give genuinely informed consent.

At this point, because my range of knowledge and experience is of course partial, is somewhat out of date and is centred on the commercial end of the social research spectrum, I necessarily fall back on impression. However, I would be surprised if my impressions are significantly in error.

I suspect that in large numbers of studies, the letter and certainly the spirit of these worthy codes are ignored. They are ignored, however, for the very good reason that we cannot risk the quality of our samples, and therefore the validity and value of our research, by offering informants the opportunity or the excuse to opt out.

It is, I think - I'm told - fair to say that commercial pressures, inside and outside government, have pushed research standards down over the last few years. Roger Jowell spoke persuasively on precisely this subject at the MRS Conference 18 months ago. I mention this because my own experience as a research seller, doing a considerable amount of work for government, is now more than 15 years out of date. But I was working for a leading agency and, I think, an agency the right side of the moral line. But even there, and even then, I can remember us doing a project for (I think) the Department of the Environment, where the 'official' style of the covering letter from the Department was thought to be an advantage - we discussed it - because a proportion of the sample would assume that questionnaire completion was compulsory. This is at odds with the SRA code.

At about the same time, a quarter of a million roadside interviews were undertaken with car drivers. The drivers were asked where they lived, where they were going, and which commercial TV station they watched. The essence of the study however, was not asked at all. This was in the days before

seat-belt wearing was compulsory and we were observing whether or not the drivers and front-seat passengers were wearing their seat-belts. The informants knew they were being researched, but didn't know what about. A sin of omission, given the spirit of our codes?

Perhaps the most extreme case - but again, arguably, in a good cause - was the study of learner drivers. The agency's element of the work was to interview people who had been driving for about a year, to find out how they learned to drive and what their experiences had been in the first twelve months. Added into this data set however - and unknown to the informant - was data on precisely how they had performed in their driving test, from government records.

These cases - and there are, I suspect, many others of a similar type - raise serious questions about our approach to research and in particular to the public's awareness and understanding of what we do. But there is another side to this coin.

A few years ago I used to lecture on the MRS questionnaire design course for young researchers, during which I gave a talk on postal questionnaires. In this talk I emphasised the need to avoid what I described, somewhat crudely, as "sod it" points. These are the points at which a complex question, an intimidating grid, or a layout problem, would allow the respondent to say "sod it" and tuck the questionnaire behind the clock, where it would stay until it was thrown out six months later.

What is true for postal questionnaires is true for face-to-face interviews, and is particularly significant at the point when the interview is being obtained. It is difficult enough for our industry to maintain response rates without giving informants opportunities to opt out. If we think that our work is important (and I for one believe that it contributes substantially to the commercial and social health of our nation), and if we have a duty to make the best possible use of the limited funds that are available, can we really accept the letter and spirit of our various codes? Should we provide potential informants with information that may confuse and intimidate, will certainly take up yet more of their time, and will give them the chance to refuse to co-operate? Is it, to use the SRA's word, 'practicable'?

3. Public Attitudes to Research

I would now like to turn to some thoughts on the public's response to our work. I am going to draw on new data produced by Alan Hedges on behalf of the Research Development Foundation, a body closely linked with the MRS and funded from within the market research industry.

A few caveats should be noted. The data available at this stage, parts of which are detailed below, are from the preliminary and qualitative stage of a longer-term project and it is centred on, though not

limited to, market research. Quantitative validation of these findings will take place in the near future.

The data are not drawn from a representative sample of the public. The RDF's starting point in developing this work was that there are (let us say) some 40%-50% of the population that will agree to be interviewed most of the time. At the other extreme there are (let us say) 10%-15% who will never be interviewed, for reasons of principle, bloody-mindedness or whatever. RDF's current interest is with the remainder - the ones we need to interview if we are to maintain response rates and who, if we lose them, we might never get back.

The initial findings from this work include the following points. I am grateful to the RDF for permission to include this information.

i. Research professionals think of research as a distinct and clearly demarcated subject. It is depressing but necessary to understand that for many, market research is, in many respects, similar to junk mail - unsolicited, mildly intrusive, potentially irritating, time-wasting and unlikely to do the informant much good. (Social research workers may wish to examine this brief list in the light of their own demands on the public.)

ii. Informants' reluctance to co-operate is based mainly on three factors, which are presented here in roughly descending order of importance.

- interviews take too long, and longer than interviewers say they will;

- some questions seem to invade privacy: their relevance is not clear, which leads to suspicion about hidden purposes;

- there seems to be an increasing wariness about getting involved with the many demands made on the public's attention by commercial and other interests.

iii. Faced with research, the potential informant's usual strategy is to avoid it if at all possible, and to cut their losses by minimising involvement if they 'get caught'. (It is worth noting that this avoidance strategy is likely to affect social as well as market research.)

iv. Public sector research sounds more interesting than much market research and is seen as potentially worthwhile, but participants feel they rarely if ever get approached about such matters. The sense that it may be worthwhile can also be dissipated if people are cynical about the possibility that action might be taken about the results.

v. In-home interviews can be more intrusive than on-street interviewing because they invade private space.

vi. Reducing response rates may be a result of, amongst other things, a sense of increasing pressure on time, a growth in the level of research activity and hence frequency of approach, and lengthy and/or poor questionnaires.

vii. In addition - and of particular relevance here - there is a feeling that information on computers is less and less private, and that more and more organisations share and trade in personal data.

viii. Survey co-operation is not a set of technical decisions by an informed public, but a set of responses to stimuli which are largely the output of social relations and forces.

Further details of this study, including of the quantitative main stage, will be available from the RDF and the MRS in the early summer of 1996.

4. Legislative Threats and the Future

Codes of conduct and of ethics are a foundation-stone of our industry. Indeed, the MRS, somewhat to its surprise, discovered a year or two ago that professional standards were, to its members, the most important element of the Society's work. Codes of conduct have also been seen as a way - to use that old rugby union phrase - of getting your retaliation in first. In other words, self-regulation will, with a bit of luck, keep the legislators off our backs.

The problem - an increasing problem - is that commercial pressures and technological developments are changing the nature of market research, and of those areas near and around market research, and, of course, are moving faster than any committee, particularly perhaps an ethics committee, is ever likely to move.

The traditional view of *ad hoc* market research - face-to-face interviews, informant confidentiality, the data being available a month after fieldwork - is now valid for only a small part of the industry. Now there is over-night data, life-style questionnaires, database marketing, and mystery shopping. None of which will go away and all of which are muddying the waters of market research and will, to some extent, impinge on social research as well - the ripples will circle outwards, if only through informant confusion or cynicism. (It is worth noting that mystery shopping is, of course, a half-cousin of participant observation, so - to mix our metaphors - it is not entirely foreign territory.)

The MRS has accepted that many of its members now work not only with confidential survey research as we used to know it, but also with these other forms of data collection. Their employers demand it and, more often than not, the public benefits from it.

In response to these changes, the Market Research Society is preparing guide-lines to assist its members in dealing with different sorts of activity. Guide-lines on the various types of database work are now in place and those on mystery shopping will follow soon. It is important however that we keep these various codes and guide-lines separate.

The Society cannot work in isolation. The Data Protection Act recognises confidential survey research and its aggregate data, and gives it certain exemptions from the provisions of the Act. If confidential survey research were to be watered down however, and individual and non-confidential data were to be allowed in some circumstances, our distinct and special status might come under threat.

There are comparable problems in Europe. This is no place for a discussion of the various threats looming from Brussels - the situation is complex and fluid - but I will mention two particular difficulties.

Firstly, and as many of you will know, there is draft legislation within the EU which would make telephone research impossible except where the potential informant has given their permission to be phoned, in advance. Cold calling would thus not be allowed. The impact of this constraint on the samples that could be obtained, and on the timing and cost of research, would be catastrophic. The legislation is being challenged by professional and trade associations throughout Europe, as well as by ESOMAR and - though perhaps less vociferously - by governments who are themselves users of telephone research. At the time of writing it appears that the immediate problem may have been solved and the legislation altered to remove any reference to research. It would be unwise, however, for us to assume that the threat has gone for all time.

Perhaps more worrying is that there seems to be an increasingly strong anti-research feeling within the EU. The starting point for the challenge to telephone research was not research, but telephone selling - tele-marketing. At some point in the lengthy process of consultation and translation, research was added to the list of things that might be constrained.

This, we assumed, would not be a significant or lengthy problem. After all, we are not salesmen, flogging things to punters over the phone. We are in research, working to improve - to echo a point I made earlier - the commercial and social development of the UK and of Europe. Unfortunately, this view can now been seen as both arrogant and unrealistic.

Increasingly, I gather, research is seen in Brussels as a greater threat than tele-marketing because (firstly) we take up more time than salesmen and (secondly) we ask for a great deal of personal information. So we are, in both senses of the word, more intrusive. And this attitude, clearly, covers social as well as market research.

There is an increasing demand for data but at the same time there seems to be an increasing resistance to the collection of that data. To return to the question I raised earlier in this paper, it is, I think, unreasonable to expect interviewers to do much more than obtain their interviews with a minimum of fuss, and to complete those interviews accurately and thoroughly. So how can we address the problems that we increasingly seem to face?

We should, I think, be aiming to establish research as a central, integral and beneficial part of our society. At the moment it is not seen as any of these things. Our message must be directed at opinion leaders and at the public. We must take every opportunity to emphasise the usefulness of research and to identify and publicise ways in which it has been used, locally or nationally, commercially or 'socially', to change our world. At the same time, the market research industry must be honest: confidential survey research must remain confidential and where data may be used for other purposes, the informants must be aware of it at the time they give the interview.

The Market Research Society will do what it can in this campaign. It will not be easy in the current atmosphere, where resistance to research is increasingly based not on logic or intellect or but more on faith, emotion or fear - on a feeling that asking for people's time or asking them for private or confidential information is, in some way, simply wrong. And as this seminar has shown, there are questions of ethics, and of individual versus group needs, that must be considered.

To put our little local difficulties into a wider context I will finish, if I may, with a quotation from Tom Stoppard's splendid television play 'Professional Foul':

"Ethics were once regarded as a sort of monument, a ghostly Eiffel Tower constructed of Platonic entities like honesty, fairness, and so on, all bolted together and consistent with each other, harmoniously stressed so as to keep the edifice standing up: an ideal against which we measure our behaviour. The tower has long been demolished."