Journal of Official Statistics
Vol. 9. No. 2, 1993, pp. 469–474

Discussion
Statistical Disclosure Limitation

Leon C.R.J. Willenborg[1]

National statistical offices engaged in the release of statistical information to third parties face a dilemma: they have to protect the privacy of their respondents, while at the same time they should provide users with data that are sufficiently rich statistically. The task of finding the right balance between these two goals is not easy. There are several reasons for this: it is generally difficult to assess the risk of disclosure associated with a particular data set; users are very often inclined to ask for as much and as detailed data as are possibly available to maintain sufficient flexibility for subsequent analyses. Demands on a given source of data often vary from one user to another. Statistical offices have to develop a policy for data dissemination. A major issue in this policy concerns finding a practical solution to this dilemma. Any solution is by necessity multifaceted, among which statistical disclosure control features prominently, but legal, technical and logistic aspects play roles as well.

In the following discussion, I want to outline the disclosure control policy of the Netherlands Central Bureau of Statistics (CBS) with respect to the dissemination of statistical data, both tabular data and microdata. I also want to say a few words about the research carried out at the bureau in this area, and present our current views on disclosure control methodology. From a CBS perspective, I want in particular to comment on the paper by Fuller. I offer my personal opinion about the potential practical relevance of the method he proposes for an agency such as the CBS. I have little to say about the inferential problems facing a statistician working with a masked data file. Little presents a systematic exposition of these problems in a missing data context to which I have nothing to add. By presenting the disclosure control policy of the CBS the reader can compare the position of the CBS relative to those of the American institutions considered in Jabine's paper. I would like to point out that a paper similar to Jabine's is Citteur and Willenborg (1993), which outlines the dissemination policies practised by various national statistical offices, with particular focus on public use files.

The data disseminated by the CBS can be grouped into two broad categories: microdata and tabular data. Because data of either type have their own particularities as far as disclosure protection is concerned, I want to discuss them separately, starting with the microdata. Because a law prohibits the CBS from disseminating business microdata collected under this law, the discussion focuses on the only microdata we

[1] Department of Statistical Methods, Netherlands Central Bureau of Statistics, P.O. Box 959, 2270 AZ Voorburg, The Netherlands.
* The views expressed in this paper are those of the author and do not necessarily reflect the policies of the Netherlands Central Bureau of Statistics.

are allowed to distribute, namely, those containing information about individuals or households. When "microdata" are mentioned below, it is implicitly understood that they pertain to persons or households.

I start with a sketch of the current status of the CBS as a data supplier to government, academia, international organizations, the business community and the general public. The CBS policies governing the release of microdata have evolved in response to the demands for microdata made by the institutions named above. In particular, this sketch serves to describe some of the constraints under which the CBS has to operate when disseminating data, and the disclosure control policy which is still evolving.

Recent years have witnessed an increasing demand in The Netherlands for microdata from the CBS for policy making, academic research and teaching. As a result of the widespread use of computers and sophisticated statistical packages, users are now very well able to carry out their own analyses and produce their own tables. In the past, the CBS bore sole responsibility for such tasks. One may safely speculate that this trend will persevere in the years to come, and that data in electronic form will soon be among the most important products sold by the CBS.

This development creates at least two problems for the CBS. First we have to find efficient ways to meet the external demand for data. Timeliness is an important factor here. This demand is already substantial and it undoubtedly will increase in the future and might even become massive. However, users request not only increasing amounts of microdata, they also want more detailed microdata. It is clear that this is in direct conflict with the second problem the CBS has to solve, namely, to safeguard the privacy of respondents. Highly detailed microdata entail a greater risk that a respondent can be recognized and certain (confidential) information about him/her can be disclosed. Furthermore, this risk is clearly proportional to the degree of proliferation of the microdata.

Not only is the demand for microdata produced by the CBS likely to increase in the future, the budgets for data production are likely to decrease. One solution in meeting this demand may be found in the use of standard microdata sets. On the other hand, the demand for tailor-made data products can be discouraged by high fees and charges for such products. These policies benefit the CBS in the following ways. For each standard data set certain efforts, such as a disclosure control analysis, the production of a code book, etc., need only to be carried out once. This system confers benefits on the data users: requests for standard files can be met quickly and at a modest price. Furthermore, there is no discrimination among data users: what is sauce for the goose is sauce for the gander. Finally, with just one version of a standard microdata file for a particular survey available there is no danger that the safety of each individual microdata set can be compromised by combining various data sets produced from a common parent data set. Of course, the drawback of this policy is that not every wish of a potential user of a microdata file can be honored. In case of conflicting interests those of Mr. Mainstream receive priority over those of Mr. Maverick. Mr. Maverick, however, may be served by on-site access to the original microdata. But his special wishes may cost him dearly, in terms of both financial and logistic inconveniences. This option is a kind of last resort: it can be invoked if a user's specific requests cannot be met by a standard microdata set for external use. Clearly this on-site option can only be used sporadically

because it would put a considerable strain on resources and facilities to be made available by the CBS, such as computing capacity, personal assistance and housing accommodation.

Policy makers and academic researchers require more detailed microdata than the general public. For this reason the CBS offers two types of microdata to external users: microdata under contract (MUCs) and public files (PUFs). MUCs are intended for academic researchers and PUFs for the general public. PUFs may be of particular interest for educational purposes in secondary schools and at universities. Because MUCs have to be rich in detail to be of interest to the intended target group, masking methods alone are not sufficient to guarantee absolute safety of the data. Such safety could only be achieved with considerable modifications of the data, which, however, would seriously affect their quality, and which could render them totally useless (cf. also Paass and Wauschkuhn 1985, p. 17). Therefore a MUC is only delivered to explicitly named statisticians employed at respected research institutions after they have signed a contract. This contract states the purpose of the research and stipulates the conditions under which the MUC may be used: no linking to other files, no copies for third parties, every paper containing results derived from this MUC should be forwarded to CBS prior to its publication for inspection on possible disclosure risks, etc. A rather light form of data masking is applied to produce MUCs. This masking is aimed only at removing characteristics of individuals which could possibly be identified at a glance, or by spontaneous recognition as we usually call it, i.e., on the basis of a rather limited number of identifiers (at most three in our current rules). The removal of these "rare" characteristics is achieved in one of the following ways: by imputing missing values for one or more of these characteristics in the corresponding records, by suitably recoding or deletion of one or more of the variables involved, or (sometimes) even by deletion of the records containing these rare values. In Little's terminology, the imputation of missing values to eliminate rare combinations of characteristics in certain records of the microdata file is by a non-ignorable imputation mechanism, which in principle could complicate subsequent analyses of the data. In practice, however, only relatively few such cases occur in a data file. The number of such cases is limited by choosing a suitable coarsening of variables.

In the case of PUFs one can only rely on masking methods, such as the ones just mentioned, in order to remove every possible risk of disclosure, i.e., by spontaneous recognition by matching (with records in some register) and by response knowledge (the knowledge that a particular respondent has participated in the corresponding survey). See Keller and Bethlehem (1992) for a discussion of these disclosure scenarios. Because PUFs have to be virtually free of disclosure risk, the number and detail of the identifying variables have to be severely restricted. As in the case of MUCs, the same sort of masking procedures are applied to obtain PUFs. Furthermore the number of identifiers allowed to be present in a PUF is at most 15 and no direct regional identifiers may be present, in order to hamper identification.

At the CBS we have formulated some general disclosure protection rules for microdata, one set applying to MUCs and the other to PUFs. These rules have to be applied to the subject matter departments when they want to release a MUC or a PUF (cf. Keller and Willenborg 1992). Only in special cases is it possible to deviate from these general rules.

From the discussion above, it is clear that the disclosure protection is based on identification disclosure and not on attribute disclosure as considered by Fuller. Furthermore, it is obvious that the masking methods applied by the CBS are fairly simple. The masking also aims to limit (in case of a MUC) or exclude (in case of a PUF) the identification of respondents whose characteristics are stored in the microdata set. In contrast the method proposed by Fuller has the aim to "minimize the information about particular individuals contained in the data set," i.e., by controlling any possibility of attribute disclosure. Controlling the possibilities of identification of respondents hence does not seem the objective of Fuller's method. So the objective of Fuller's approach is quite different from that pursued by the CBS. This also implies that masking methods are applied to different types of variables: the CBS method is aimed at modifying identifiers in a microdata file whereas Fuller's method modifies the non-identifiers. Because a microdata file often contains a large number of non-identifiers and only a limited number of identifiers, Fuller's method would require far greater modification of the file (in terms of the number of variables involved) than the CBS's method. For this reason, I think that Fuller's method would be feasible only when the number of non-identifiers in a file is limited. In my opinion measures to control attribute disclosure risks are, as a rule, not appropriate for disseminating general purpose, standard microdata.

A further comment is that application of Fuller's method requires a clear notion of the statistics that have to remain unchanged or almost unchanged for the analyses intended. This would be the case when a tailor-made microdata file has to be produced (and where the intended analyses should be limited to first and second order moments because higher order moments may be seriously biased). In view of the policy of the CBS to produce general purpose, standard microdata (of PUF- or MUC- type) there is usually no prespecified research goal, so that the application of this method would be "a shot into the dark."

These critical remarks should not conceal my admiration for the ingenuity of Fuller's method. I think his method has value for the intellectual debate on disclosure control and the statistical problems associated with it. I do, however, feel that there is a gap between theory and practice of disclosure control of microdata. At the CBS research in this area is carried out as well and it has so far produced some results which are at best interesting but which have had limited significance in practice (cf. Bethlehem, Keller, and Pannekoek 1990; Keller and Bethlehem 1992; Willenborg, Mokken, and Pannekoek 1990; Mokken, Kooiman, Pannekoek, and Willenborg 1992). In particular it would be of great help if a practically feasible theory would be developed which allows the quantification of the disclosure (or rather, identification) risk of each individual record in a microdata file. Then we would be able to identify the records with highest disclosure risk, and reduce this risk to an acceptable level by applying a data masking procedure. So far we have developed a model to estimate the number of population uniques in a microdata file (with respect to a given set of identifiers) and, based on this, a theory to calculate the risk that at least one respondent in such a file could be identified by a researcher, assuming that he or she has a particular kind of prior knowledge about the population. (Skinner and Holmes (1992) present an alternative model to estimate the number of population uniques. Their model is claimed to outperform the one developed by the CBS.) In all these

approaches we have assumed that there is no measurement error in the data, which is not very realistic. Paass and Wauschkuhn (1985) and Fuller produce individual disclosure risks, and in both cases the presence of measurement error is assumed. These results are worthy of serious consideration.

In view of our current sets of disclosure protection rules for PUFs and MUCs it is necessary to be able to estimate the number of individuals in the population with certain characteristics. Such estimates are used to decide whether these characteristics occur frequently enough, according to specific threshold values used in these rules. (If such a characteristic is found to be below the appropriate threshold level it is considered "dangerous" and a masking procedure is applied to eliminate it from the data, as was explained above.) To estimate the population frequency of such a characteristic, we currently use an interval estimator for a Poisson distribution. This estimation method is difficult to apply for small samples. In order to avoid this problem we are now investigating another estimation technique based on synthetic estimation.

With the introduction of the first set of disclosure protection rules for microdata it was deemed necessary to develop a special computer program to enhance the production of "safe" microdata files. The development of such a program, ARGUS, started two years ago. Due to limited programming resources, only a beta version has been produced so far (see De Jong 1992). Our aim with ARGUS is to create a comprehensive computer program for disclosure control, not only for microdata but for tabular data as well. It should also be a package which is not tailored to our current disclosure protection rules, but which allows the specification of alternative rules as well. This should make ARGUS a useful package for other statistical offices as well.

After having discussed disclosure control of microdata, I now switch to tabular disclosure. A first remark is that the disclosure problem of tabular data (with marginals), based on an integral observation of a particular, well-defined population, is far easier to formalize than the corresponding problem for microdata. For tabular data a dominance rule is a widely accepted method for identifying potentially sensitive cells. Suppressing these cells yields the primary cell suppressions in the table. Additional, or secondary cells have to be suppressed when additional information is available, e.g., in the form of marginal totals. These secondary cells are suppressed to prevent the value of a suppressed primary cell from being recalculated to sufficient precision. Of course, the problem is to use as few (weighted) secondary cells as possible to meet these goals. A practical problem is to define precisely what should be minimized: the number of cells to be suppressed, the total of the suppressed values, the total number of individual contributions to the suppressed cells, the total of certain sensitivity weights (as employed in the dominance rule for primary suppressions) associated with each suppressed cell, or a (linear) combination of these parameters? In any case we obtain a linear objective function. Note that cell suppression in tables is an attribute disclosure technique, and not an identity disclosure technique. (However, combining rows or columns in a table, which is another masking technique for tables, can be considered a technique of the latter type.)

Secondary cell suppression in tables is a "hard" mixed integer optimization problem for which only heuristics are of practical value. Such algorithms do not solve the original hard problem but rather, they solve a related although easier one. Hence the solutions produced by a heuristic are

not necessarily optimal, but (in most cases) we hope, near optimal ones. Heuristics for secondary cell suppression are treated in, e.g., Kelly (1990), based on a network flow algorithm, and Geurts (1992), employing a branch-and-bound algorithm. Recently Hopfield-like neural networks have successfully been applied to these problems (cf. Wang, Sun, and Golden 1993). In general, secondary cell suppression can be handled well only for low-dimensional tables.

It is clear that there is a difference in methodology for masking microdata on the one hand and tabular data on the other. In the former statistical considerations dominate, whereas in the latter optimization techniques prevail. Disclosure control in practice, however, is not restricted to a study of problems in these academic disciplines, as was remarked in the beginning of this discussion. So the field of disclosure control is strongly interdisciplinary. The papers of Little, Fuller, and Jabine illustrate nicely some of the work that is currently being carried out in this field.

References

Bethlehem, J.G., Keller, W.J., and Pannekoek, J. (1990). Disclosure Control of Microdata. Journal of the American Statistical Association, 85, 38–45.

Citteur, C.A.W. and Willenborg, L.C.R.J. (1993). Public Use Files: Current Practices at National Statistical Bureaus. Journal of Official Statistics, (to appear).

De Jong, W.A.M. (1992). ARGUS: An Integrated System for Data Protection. International Seminar on Statistical Confidentiality, September 8–12, Dublin, Ireland.

Geurts, J. (1992). Heuristics for Cell Suppression in Tables. Report, Netherlands Central Bureau of Statistics, Voorburg, The Netherlands.

Keller, W.J. and Bethlehem, J.G. (1992). Disclosure Protection: Problems and Solutions. Statistica Neerlandica, 46, 5–19.

Keller, W.J. and Willenborg, L.C.R.J. (1992). Microdata Release Policy at the Netherlands CBS. International Seminar on Statistical Confidentiality, September 8–12, Dublin, Ireland.

Kelly, J.P. (1990). Confidentiality Protection in Two- and Three- Dimensional Tables. Ph.D. Dissertation, University of Maryland, College Park, Maryland, USA.

Mokken, R.J., Kooiman, P., Pannekoek, J., and Willenborg, L.C.R.J. (1992). Disclosure Risks for Microdata, Statistica Neerlandica, 46, 49–67.

Paass, G. and Wauschkuhn, U. (1985). Datenzugang, Datenschutz und Anonymisierung. Analysepotential und Identifizierbarkeit von anonymisierten Individualdaten. Oldenbourg Verlag, Munich/Vienna (In German.)

Skinner, C.J. and Holmes, D.J. (1992). Modelling Population Uniques. International Seminar on Statistical Confidentiality, September 8–12, Dublin, Ireland.

Wang, Q., Sun, X., and Golden, B.L. (1993). Neural Networks as Optimizers: A Success Story. Report, College of Business and Management, University of Maryland, College Park, Maryland, USA.

Willenborg, L.C.R.J., Mokken, R.J., and Pannekoek, J. (1990). Microdata and Disclosure Risks. Proceedings of the Sixth Annual Research Conference, Bureau of the Census, Washington, D.C., 167–180.