Journal of Official Statistics, Vol. 14, No. 4, 1998, pp. 437–447

Models and Methods for the Microdata Protection Problem

C.A.J. Hurkens and S.R. Tiourine[1]

This article focuses on the mathematical modeling and an algorithmic approach to microdata protection based on a strategy of local suppression and global recoding. Protecting sensitive individual information using such an approach must be balanced against the usefulness of the resulting edited datafile.

We discuss the two subproblems of suppression and recoding of microdata in light of the information loss that occurs. We combine them into an overall disclosure problem formulated as a combinatorial optimization problem. We describe how to decompose the problem into smaller more tractable parts and how to use local search methods to find good approximations of the optimal solution.

Key words: Mathematical modeling; optimization algorithms.

1. Introduction

Statistical disclosure control in microdata is a relatively new problem for statistical offices. The problem arises from contradictory objectives with respect to the public release of microdata files. A microdata file consists of records with information collected from individual respondents, typically by means of a survey. With the release of an anonymized version of such a file, the statistical office aims at providing the most detailed information under the condition that no sensitive information from this file can be attributed with certainty to a particular respondent. Clearly, the dilemma is releasing very detailed information at a high risk of misuse of this information versus providing much less detailed information and guaranteeing the privacy of a respondent. For the definitions used and the background of the problem we refer to the work of Willenborg and De Waal (1996).

Disclosure of information in the microdata file occurs when sensitive information in the file is identified with a certain respondent. Therefore, there are two conditions for a disclosure to occur. First, the file must contain sensitive information. Second, it should be possible to identify this information with a respondent. If one of these conditions is not satisfied, we assume that the file is safe from disclosure.

It is up to a statistical office to judge whether the microdata contains sensitive information. We assume that such information is indeed present and that the issue is to prevent its identification with the individuals from a population. In the absence of directly identifying information like a name or an address, a record in the microdata can be identified by a combination of identifying variables, called a key. There are several scenarios described

[1] Department of Mathematics and Computing Science, Eindhoven University of Technology, Eindhoven, The Netherlands.

Acknowledgments: This work was sponsored by the European ESPRIT project 20462-SDC.

in the literature as to how this identification can occur. The easiest one concerns the population uniques. If someone is unique in a population on a certain key and the corresponding record is present in a microdata file, the respondent can be identified. A more subtle situation occurs when a group of people identical on a certain key also have similar scores of a sensitive variable. In this case, the sensitive information about a member of this group can be disclosed although no personal identification occurs.

In practice, it is assumed that, if a record in the microdata file has a rare score for a low-dimensional key, it is potentially unsafe. This is a natural extension of the concept of uniqueness. Indeed, if someone is unique in the population on a key, then so is the corresponding record in the microdata file, if present. On the other hand, if a group of individuals can be identified on a low-dimensional key, then it is likely that either the members of the group are unique on the higher-dimensional key or that the sensitive information shared by the group is revealed.

Frequency tables are used to compute the set of the rare combinations. More precisely a frequency table is set up for each potentially unsafe key. A cell in the table gives the number of records that contain the corresponding combination of values of a key. If this number is between one and a certain threshold, defined for each table, then the combination is declared to be unsafe. These unsafe combinations have to be screened before the microdata file can be released for public use. We will consider two protective measures for microdata: *local suppressions* and *global recodings*, as suggested by Willenborg and De Waal (1996). Global recoding is an operation defined for all records in the microdata file. If, for example, each record in the file contains a field specifying the municipality of a respondent, then a possible global recoding will be to replace the value of that field by the corresponding province for all records. Local suppression on the contrary is applied to a single record by replacing the values of some of its fields by ''missing.'' The protective effect of these techniques is determined from the updated frequency tables.

There is no consensus in the community of statistical offices as to what the measure of information loss should be (Willenborg 1997). De Waal and Willenborg (1995) proposed to use an entropy function, but the exact implementation of this function has yet to be worked out. We model the information loss by a linear function. Under the assumption of independence, most of the information loss functions known from the literature can be represented in this way.

In the following sections we proceed by formulating the microdata protection problem as an optimization problem. We will have to deal with the size issue, in particular, as the microdata sets themselves may be of enormous magnitude.

2. Models for Microdata Protection

We base our modeling approach on the concepts defined by De Waal and Willenborg (1995). Consider a microdata file as a collection R of records r. In the microdata protection problem we are only interested in the part of the file containing identifying information. We denote the index set of identifying fields in a record by F. Let U be a set of unsafe combinations of identifying fields in the microdata. An unsafe combination $u \in U$ has the form $u = (r, S)$, with $r \in R$ and $S \subseteq F$. As we discussed in the introduction, the set U of unsafe combinations is computed from the microdata file using the frequency tables.

We distinguish between two types of critical unsafe combinations in our models: a minimal unsafe combination (minuc) and a maximal unsafe combination (manuc), which are inclusion-wise minimal, respectively maximal, unsafe combinations. An unsafe combination $u = (r, S)$ is minimal if there is no $T \subseteq F$ with $T \subset S$ (we use the sign $\subset$ for strict inclusion) and $(r, T) \in U$. It is maximal if there is no $T \subseteq F$ with $S \subset T$ and $(r, T) \in U$. A minuc is protected by suppressing any of its entries and therefore is useful in the definition of a model for local suppression. Manucs are used to formulate the global recoding problem.

We have to protect the unsafe combinations from U using global recodings and local suppressions and to do so with minimal information loss. We will give the examples of the application of local suppressions and global recodings in Sections 2.1 and 2.2. We will proceed by building our model gradually, starting with simple special cases.

2.1. A pure suppression problem

We begin by formulating the exact local suppression problem, following De Waal and Willenborg (1998). For each record r with at least one unsafe combination we introduce variables

$$x_{rf} = \begin{cases} 1 & \text{if the content of field } f \text{ in record } r \text{ is replaced by ``missing''} \\ 0 & \text{otherwise} \end{cases}$$

where $f \in F$. By setting at least one value of an unsafe combination at ''missing'' the combination becomes untraceable and therefore is considered protected. Hence, the local suppression problem for a microdata file is

$$\begin{array}{lrll} \min & \sum_{r \in R, f \in F} c_{rf} x_{rf} & & \\ \text{s.t.} & \sum_{f \in S} x_{rf} & \geq 1, & \forall (r, S) \in U, \\ & x_{rf} & \in \{0, 1\}, & \forall r \in R, \forall f \in F \end{array} \qquad (1)$$

Under some conditions, we may restrict the constraints to minucs only. This is because if the set U is complete with respect to minucs, or in other words if for each $u \in U$ the set U contains all minimal subsets of u, then the set U is protected by local suppressions if and only if the subset of minucs in U is protected. To prove necessity, suppose that all minucs in U are protected by local suppressions. Let $u = (r, S) \in U$ be any unsafe combination not protected by local suppressions. Consider another combination defined by the not suppressed fields of $u : u' = (r, S') : S' \subset S$, where S' is a subset of not suppressed fields of S. By construction, u' is also an unsafe combination, not necessarily in U, not protected by local suppressions. It is also clear that u' must contain at least one minuc from U, which contradicts the fact that none of the fields of u' are suppressed and therefore the minuc is not protected.

Note that there are variations of the local suppression problem that may connect the problems for the records. For instance, one might want to bound the total number of suppressions of field f from above, by β_f, say. Then the additional restriction $\sum_r x_{rf} \leq \beta_f$ turns the overall suppression problem into one very big problem.

Example: Consider a collection of records and the corresponding list of minucs given in Table 1.

Table 1. Collection of records, with minimal unsafe combinations

record	field 1	field 2
1	10	100
2	11	101
3	19	100
4	19	100
5	10	109
6	10	109

unsafe comb.	in record	minuc	protected by suppression
1	1	10 × 100	x_{11}, x_{12}
2	2	11	x_{21}
3	2	101	x_{22}

An unsafe combination in this example is any unique combination of record field values. The local suppression problem for this data is

$$
\begin{array}{llllll}
\min & c_{11}x_{11} + c_{12}x_{12} + c_{21}x_{21} + c_{22}x_{22} & & & & \\
\text{s.t.} & x_{11} + \quad x_{12} & & & \geq & 1 \\
 & & x_{21} & & \geq & 1 \\
 & & & x_{22} & \geq & 1 \\
 & x_{11}, x_{12}, x_{21}, x_{22} \in \{0, 1\} & & & &
\end{array}
$$

2.2. *A restricted combination of local suppression and recoding*

In the formulation below we require that every unsafe combination is protected by either local suppression or global recoding. This can be done, because, in principle, for each unsafe combination we can list all local suppressions and global recodings that protect it.

We introduce the following variables:

$$y_{fk} = \begin{cases} 1 & \text{if the content of field } f \text{ is recoded according to rule } k \text{ for every record} \\ 0 & \text{otherwise} \end{cases}$$

for $k \in K_f$, where K_f is the set of possible recodings of field f.

A constraint matrix B with columns corresponding to variables y_{fk} and rows corresponding to unsafe combinations u, characterizes the protection of unsafe combinations by global recodings. The column corresponding to a variable y_{fk} has an entry 1 in row $u \in U$ if any only if unsafe cell u is protected by the global recoding k of field f. To construct this column, we effectuate the corresponding global recoding. Consider an unsafe combination $u = (r, S)$. If, in the recoded microdata set, there are enough records r' with the same score in the fields S, the combination u is considered to be protected. Note that we only apply recoding of one specific field f, and leave all other fields unchanged.

Matrix B constructed in this way characterizes protection effect of the global recodings. This characterization is complete under the assumption that the combinations of global recodings have no added effect. However, this is not always the case. Often a combination of two or more global recodings protects unsafe combinations that are not protected by either of the global recodings. We will illustrate this issue in the example below.

We define a constraint matrix A as the incidence matrix of record fields and unsafe combinations. That is, the coefficient in row u and column f of A is 1 if combination u contains field f, and 0 otherwise.

Now our first step is to consider local suppressions in a restricted combination with

Table 2. *Recodings of Fields 1 and 2*

Field 1			Field 2		
orig. value	recoded value	corresp. variable	orig. value	recoded value	corresp. variable
10,11	10–11	y_{11}	100,101	100–101	y_{21}
18,19	18–19	y_{11}	108,109	108–109	y_{21}
10..19	10–19	y_{12}	100..109	100–109	y_{22}
10..19	10..19	y_{13}	100..109	100..109	y_{23}

global recodings. In this approximation we neglect the possible effects of superpositions of the global recodings. We will clarify this issue below. Then the restricted local suppression and global recoding problem is

$$\begin{array}{lrll}
\min & c_x x + c_y y & & \\
\text{s.t.} & Ax + By & \geq 1, & \\
 & \sum_{k \in K_f} y_{fk} & = 1 & \forall f \in F \\
 & x_{rf} & \in \{0,1\} & \forall r \in R, \forall f \in F \\
 & y_{fk} & \in \{0,1\} & \forall k \in K_f, \forall f \in F
\end{array} \tag{2}$$

where c_x and c_y are the cost vectors corresponding to local suppressions and global recodings, respectively. The equality constraints ensure that every field is recorded in one way. By convention we include the recoding ‘‘leaving as it is’’ as one possibility.

Example: (continued). We assume that the fields in the microdata file can be recoded as shown in Table 2. The first column of each table gives the original values of the corresponding fields. The second one gives the recoded values. For example, variable y_{12} corresponds to the replacement of the values of Field 1 from {10,...,19} by the range 10–19. The corresponding local suppression and elementary global recoding problem is given by

$$\begin{array}{llllllllll}
\min c^x_{11}x_{11} + c^x_{21}x_{21} + c^x_{12}x_{12} + c^x_{22}x_{22} + c^y_{11}y_{11} + c^y_{12}y_{12} + c^y_{21}y_{21} + c^y_{22}y_{22} \\
\text{s.t.} & x_{11} + & x_{21} & & & & y_{12} + & & y_{22} & \geq 1 \\
 & & & x_{12} + & & y_{11} + & y_{12} + & & & \geq 1 \\
 & & & & x_{22} + & & & y_{21} + & y_{22} & \geq 1 \\
 & y_{13} + & & & & y_{11} + & y_{12} + & & & = 1 \\
 & y_{23} + & & & & & & y_{21} + & y_{22} & = 1 \\
 & \multicolumn{9}{l}{x_{11}, x_{21}, x_{12}, x_{22}, y_{11}, y_{12}, y_{13}, y_{21}, y_{22}, y_{23} \in \{0,1\}}
\end{array}$$

This model, however, neglects the fact that some manucs which are not secured by elementary recodings may be protected by a combination of them. As an illustration consider unsafe combination 1 of our example, which is protected by the combination of recodings y_{11} and y_{21}, but not by either of them.

2.3. *A complete formulation*

We now formulate a model that is complete in the sense that it combines the suppression and recoding problem, and also takes the combined effects of recodings into account. Let

$\mathscr{K} = \{(k_1, k_2, ..., k_{|F|}) \,|K_f \in K_f\}$ denote the set of all possible recodings. Then a vector $\underline{k} \in \mathscr{K}$ denotes a particular recoding of the entire microdata set.

We described in the introduction how the unsafe combinations are determined using frequency tables. It is easy to see that for each unsafe combination there is a unique set of unsafe cells C corresponding to the set U of unsafe combinations. Note that more than one unsafe combination can be mapped to one cell $\alpha \in C$. We therefore define an unsafe cell as a set of unsafe combinations mapped to it.

In addition to 0-1 variables x_{rf} and y_{fk}, we introduce 0-1 variables

$$z_\alpha = \begin{cases} 0 & \text{if unsafe cell } \alpha \text{ is protected by global recodings} \\ 1 & \text{otherwise} \end{cases}$$

for $\alpha \in C$. We introduce 0-1 parameters $P(\alpha, \underline{k})$, where $P(\alpha, \underline{k}) = 1$ if the unsafe cell α is protected by a recoding $k = (k_1, k_2, ..., k_{|F|})$, and $P(\alpha, \underline{k}) = 0$ otherwise. The problem is now

$$\begin{array}{lll} \min & \sum_{r\in R} \sum_{f\in F} c_{rf} x_{rf} + \sum_{f\in F} \sum_{k\in K_f} d_{fk} y_{fk} & \\ \text{s.t.} & z_\alpha + \sum_{\underline{k}\in \mathscr{K}} P(\alpha, \underline{k}) \Pi_f y_{fk_f} \geq 1 & \forall \alpha \in C \\ & \sum_{f\in S} x_{rf} \geq z_\alpha & \forall (r, S) \in \alpha, \quad \forall \alpha \in C \\ & \sum_{k\in K_f} y_{fk} = 1 & \forall f \in F \\ & x_{rf}, y_{fk} \in \{0, 1\} & \forall r \in R, \forall f \in F, \forall k \in K_f \end{array}$$

2.4. Discussion of the model

In principle, the coefficients $P(\alpha, \underline{k})$ of the model are known. The same is true for the cost coefficients c_{rf} and d_{fk}. A major problem is that it is rather hard to define cost coefficients that suitably describe the overall information loss. Especially difficult in this respect is to find a trade-off between the local suppressions and the global recodings.

Another point is that the model does not seem to possess any structure that suggests an efficient solution technique. In particular, the product form of the recoding variables seems hard to work with.

We will use these points to our advantage. We decompose the problem into two, one for the global recoding and one for the local suppression. This is based on the following argument. Each cell in Model (3) has to be protected by either recoding or suppression. For an arbitrary recoding $\underline{k}$ we determine the set $C_{\underline{k}}$ of cells not protected by $\underline{k}$. z is the characteristic vector of this set $C_{\underline{k}} = \{\alpha \in C | z_\alpha = 1\}$. For the cells in $C_{\underline{k}}$ we compute an estimate of the cost of local suppressions. In our approach every cell α gets a weight π_α, which can roughly be interpreted as the cost of local suppressions necessary to protect the cell. Therefore, an overall estimate of the suppression costs amounts to the sum of the weights over the cells in $C_{\underline{k}}$.

We prefer to see the uncertainty in the objective function as an extra degree of freedom. One may think of an iterative setting, where the user may adjust the cost coefficients to reflect his or her preferences about the measure of information loss.

The structure of the model is also justified by the required interface with the existing software. As one of the results of this study, a solver has been developed and incorporated in a decision support system for statistical disclosure control, called ARGUS (Willenborg and Hundepool 1998). It implies extra limitations and certainly extra challenges for our approach. In the decision support system the information about a problem instance is available to us via the coefficients c_{rf}, d_{fk} and $P(\alpha, \underline{k})$.

Therefore, we think that the model we have chosen is a good compromise between the phenomenon we are studying, the data available to us, and the solution techniques we have in mind. We proceed by describing the solution techniques for this model.

3. Solution Approach

3.1. *The relaxed recoding problem*

We first obtain a lower bound on the local suppression problem for each record. Let C_r denote the projection of set C to record $r : C_r = \alpha\{\in C \mid \exists(r, .) \in \alpha\}$. Then the local suppression problem for record r is

$$\begin{array}{llll} \min & \sum_{f\in F} c_{rf}x_{rf} & & \\ \text{s.t.} & \sum_{f\in S} x_{rf} \geq z_\alpha & & \forall\alpha\in C_r \\ & x_{rf} \in \{0,1\} & & \forall f\in F \end{array} \tag{3}$$

The dual to the linear programming relaxation of this problem is

$$\begin{array}{llll} \max & \sum_{\alpha\in C_r} \pi_{\alpha r}z_\alpha & & \\ \text{s.t.} & \sum_{\alpha\in C_r:(r,S)\in\alpha, f\in S} \pi_{\alpha r} \leq c_{rf} & & \forall f\in F \\ & \pi_{\alpha r} \geq 0, & & \forall\alpha\in C_r \end{array} \tag{4}$$

According to the duality theory of the linear programming, any feasible solution $\pi_{\alpha r}$ to (4) provides a lower bound on the value of (3).

In the following formulation, we obtain a lower bound on the complete formulation. Let $\pi_\alpha = \sum_{r\in R} \pi_{\alpha r}$. We define the *relaxed recoding problem* as follows:

$$\begin{array}{lll} \min & \sum_{\alpha\in C} z_\alpha\pi_\alpha + \sum_{f\in F}\sum_{k\in K_f} d_{fk}y_{fk} & \\ \text{s.t.} & & \\ & z_\alpha + \sum_{\underline{k}\in\mathscr{K}} P(\alpha,\underline{k})\Pi_{f\in F}\, y_{fk_f} \geq 1 & \forall\alpha\in C \\ & \sum_{k\in K_f} y_{fk} = 1 & \forall f\in F \\ & y_{fk}\in\{0,1\} & \forall f\in F, \forall k\in K_f \end{array} \tag{5}$$

3.2. *Strategy*

Our strategy in tackling the original problem is as follows.

- Compute weights $\pi_{\alpha r}$ for all-one vectors z_α (it corresponds to protection of the unsafe cells by only local suppressions).
- Find – by whatever method – a recoding that together with the estimated costs of suppression, yields a good solution to the relaxed recoding problem, and thereby gives a good approximate solution to the overall problem.
- Given the recoding found in the previous set, compute or approximate the real optimal solution to the suppression problem. If the value is close to the computed lower bound, then we have a solution and an estimate of its suboptimality, and we stop.
- If we are not satisfied with the solution at hand, we may recompute the weights $\pi_{\alpha r}$

based on the outcome of the last step. That is, we solve (4) with z_α derived from the last solution. We go back and repeat the solution procedure.

We consider two ways of finding good solutions to our relaxed recoding problem. The first is based on Lagrangean relaxation, the second on local search. The methods are described below.

3.3 Lagrange relaxation

The following method for solving the relaxed recoding problem is motivated by the size of the problem. The idea is to solve the problem (6) to optimality, for each frequency table separately. If this leads to a consistent overall solution, we are done. Otherwise, we will try to enforce consistency by imposing Lagrangean type penalties. Note that the relaxed recoding problem for each table is relatively small and could be solved by complete enumeration.

In the following, let T denote the set of frequency tables used to indicate unsafe combinations, and let T_f denote the set of tables that contain field f as one of their dimensions. For convenience we may view a table as a collection of cells. For table t, let I_t denote the set of its coordinates. We introduce new variables for global recodings for each frequency table:

$$y^t_{fk} = \begin{cases} 1 & \text{if the content of field } f \text{ is recoded according to rule } k \text{ for a table } t \\ 0 & \text{otherwise} \end{cases}$$

Our intention is to use these variables to determine the best recoding for each frequency table. In other words, if only one frequency table t was generated then y^t_{fk} will represent the alternative recodings for microdata. Clearly, there can be different recoding chosen based on the different frequency tables. Therefore, these variables can take values inconsistent with each other and our goal is to find an iterative scheme to eliminate such inconsistencies.

First, we reformulate the problem in the following way:

$$\begin{aligned} &\min \textstyle\sum_{t\in T}\left(\sum_{\alpha\in t}\pi_\alpha z_\alpha + \sum_{f\in I_t}\sum_{k\in K_f} d_{fk} y^t_{fk}\frac{1}{|T_f|}\right) \\ &\text{s.t.} \\ &\qquad z_\alpha + \textstyle\sum_{\underline{k}\in\mathscr{K}} P(\alpha,\underline{k})\Pi_{f\in F}\, y^t_{fk} \geq 1 && \forall t\in T, \forall\alpha\in t \\ &\qquad y^t_{fk} = \tfrac{1}{|T_f|}\textstyle\sum_{t':f\in I_{t'}} y^{t'}_{fk} && \forall t\in T, \forall f\in I_t \\ &\qquad \textstyle\sum_{k\in K_f} y^t_{fk} = 1 && \forall t\in T, \forall f\in I_t \\ &\qquad y^t_{fk}\in\{0,1\} && \forall t\in T, \forall f\in I_t, \forall k\in K_f \end{aligned} \tag{6}$$

The first equation demands that all table recodings y^t_{fk} are consistent with respect to common fields. In other words, a feasible recoding will have a form $y^t_{fk} = y^t_{fk}$, $\forall t, t'\in T, \forall f\in F, \forall k\in K_f$. We obtain a Lagrangean relaxation of this problem by bringing this equality system into the objective. For arbitrary λ in $(|T|\sum_{f\in F}|K_f|)$–dimensional real space, let $L(\lambda)$ be defined by

$$\begin{aligned} &\sum_{t\in T}\min\left(\textstyle\sum_{\alpha\in t}\pi_\alpha z_\alpha + \sum_{f\in I_t}\sum_{k\in K_f}\left(\frac{d_{fk}y^t_{fk}}{|T_f|} + \lambda_{tfk}\left(y^t_{fk} - \frac{1}{|T_f|}\sum_{t':f\in I_{t'}} y^{t'}_{fk}\right)\right)\right) \\ &\text{s.t.} \\ &\qquad z_\alpha + \textstyle\sum_{\underline{k}\in\mathscr{K}} P(\alpha,\underline{k})\Pi_{f\in F}\, y^t_{fk} \geq 1 && \forall t\in T, \forall\alpha\in t \\ &\qquad \textstyle\sum_{k\in K_f} y^t_{fk} = 1 && \forall t\in T, \forall f\in I_t \\ &\qquad y^t_{fk}\in\{0,1\} && \forall t\in T, \forall f\in I_t, \forall k\in K_f \end{aligned} \tag{7}$$

In principle, the minimization problems in this formulation have exactly the same structure as the original problem, but it is hoped that they are small enough to be solved by complete enumeration. The overall Lagrangean relaxation can be solved using the classical subgradient algorithm (see for example Minoux 1986).

Although this method has not been implemented, it may be of interest. It can always be used as a tool to find lower bounds for the relaxed recoding problem. Moreover there may be possibilities of using it as a means to find approximate solutions.

3.4. A local search approach

The method that has been implemented is based on the principle of local search. Local search algorithms are often chosen to tackle large-scale practical combinatorial optimization problems. These are particularly suitable if there are a lot of feasible solutions. The general idea of local search is to start with an initial solution and iteratively perform small transformations of this solution in an attempt to improve it with respect to a given criterion. The *neighborhood* of a given solution is defined as a set of solutions to which a given one can be transformed in a single iteration. Various search strategies are used to continue the search even when no immediate improvement is found in a neighborhood. An exhaustive treatment of these techniques and their applications is given by Aarts and Lenstra (1997).

We define the neighborhood of solution y for problem (6) as the set of recodings y' such that

$$\exists j \ (y'_{jk} = y_{j,k\pm 1} = 1 \text{ and } \forall i \neq j \ \forall k \ y'_{ik} = y_{ik})$$

In other words, we move from one solution to another by changing the recoding level of one of the fields to an adjacent one.

We start with a random solution, or a solution constructed in some sensible way. We then modify our solution to a neighboring solution. We recompute the solution value, and if the change proves profitable, we effectuate it, otherwise we try another one. A greedy implementation of this strategy will lead to a so-called iterative improvement algorithm. It will find better and better solutions, until it gets stuck in some local minimum.

We have implemented the following search strategies:

- *Iterative Improvement* starts from a randomly generated recoding, or a recoding given by the user. At each step it modifies the current solution to a neighboring solution of lower cost. The method stops when no better neighbor exists.
- *Repeated Iterative Improvement* restarts the iterative improvement procedure from random recodings.
- *Tabu Search* always moves to the best neighbor. In this way the cost of the solutions generated is not necessarily decreasing. To prevent the method from cycling, the reversal of several recently performed moves is disallowed. A stopping criterion is a maximum number of iterations without improvement.
- *Simulated Annealing* modifies a solution to a randomly generated neighbor. Improvements are always accepted. Deteriorations are accepted with a certain probability, decreasing during the run. The method stops when this probability reaches a certain value.

The right choice of a local search algorithm depends on the relative size of the neighborhoods, the probability that a neighboring solution is better, and the effort it takes to evaluate the new solution. The parameter settings of the various approaches are chosen automatically. The efficiency and quality of the various procedures have to be evaluated by performing experiments and judging the solutions within the context of the expected use of the microdata.

4. Conclusions

Statistical disclosure control in microdata gives rise to a constrained decision problem. We work with a mathematical programming formulation of the problem. This approach yields a huge optimization model, the formulation of which requires extensive computations. In this model a set of unsafe combinations has to be protected by application of global recodings and local suppressions at minimum information loss. Each global recoding is characterized by a subset of unsafe combinations protected by it. Local suppressions are used for the unsafe combinations left unprotected by the global recodings.

A practical difficulty of our model lies in the definition of an objective function. Here two different approaches can be used. One is based on the information loss, expressed for example by an entropy function, resulting from global recodings and local suppressions. The other is a subjective assessment by a user as to what the value of a resulting microdata will be for his or her research purposes. These two criteria do not necessarily produce the same result. Moreover, there is no consensus about the exact form of the information loss estimate in the former approach.

This motivated us to construct a cost estimate, where we first calculate the effect of the global recodings and then we estimate the cost of the remaining local suppressions. The estimate is based on the solution to a relaxation of the local suppression problem. In this way we obtain a lower bound on the optimal solution value. We gave an iterative procedure that can be used to strengthen this bound. We proposed a local search approach to obtain a solution to this smaller and computationally less intensive model. Its implementation has been incorporated in a decision support system for statistical disclosure control. This still somewhat cumbersome routine spends about 99% of its run time on cost calculations. Our tests show that the iterative improvement algorithm can be a useful on-line navigation tool for a user to obtain a modified microdata set. Tabu search can perform the same task off-line, and simulated annealing provides a costly alternative.

A Lagrangean relaxation based solution technique as proposed in this article has yet to be implemented and tested.

5. References

Aarts, E.H.L. and Lenstra, J.K. (eds.) (1997). Local Search in Combinatorial Optimization. Wiley, Chichester.

Minoux, M. (1986). Mathematical Programming, Theory and Algorithms. Wiley, Chichester.

De Waal, A.G. and Willenborg, L.C.R.J. (1995). Optimum Global Recoding and Local Suppression. Technical Report, Statistics Netherlands, Voorburg, The Netherlands.

De Waal, A.G. and Willenborg, L.C.R.J. (1998). Optimal Local Suppression in Microdata. Journal of Official Statistics, 14, 421–435.

Willenborg, L.C.R.J. (1997). Personal Communication.

Willenborg, L.C.R.J. and De Waal, A.G. (1996). Statistical Disclosure Control in Practice. Lecture Notes on Statistics 111. Springer, New York.

Willenborg, L.C.R.J. and Hundepool, A. (1998). ARGUS for Statistical Disclosure Control. In Proceedings of the Statistical Data Protection Conference, Lisbon, March 25–27.

Received August 1997
Revised September 1998