CASC PROJECT

Computational Aspects of Statistical Confidentiality
15 October 2002

A Methodological Framework for Statistical Disclosure Limitation of Business Microdata

Jim Burridge
Luisa Franconi
Silvia Polettini
Julian Stander

University of Plymouth
School of Mathematics and Statistics
Drake Circus, PL4 8AA
Plymouth UK

ISTAT, MPS/D
Via C. Balbo, 16
00184, Roma
Italy

Deliverable No: 1.1-D4

A Methodological Framework for Statistical Disclosure Limitation of Business Microdata

Jim Burridge[1], Luisa Franconi[2], Silvia Polettini[2], and Julian Stander[1]

[1] University of Plymouth, School of Mathematics and Statistics
Drake Circus
PL4 8AA Plymouth, UK
[2] ISTAT, Servizio della Metodologia di Base per la Produzione Statistica
Via Cesare Balbo, 16
00185 Roma, Italy

Abstract. This document discusses some strategies for statistical disclosure limitation developed under the CASC project for treatment of business microdata. The deliverable mainly results from the composition of different proposals that have been previously or are to be published in the literature by the authors. First, a general framework for microdata protection is sketched, as appears in Polettini, Franconi and Stander (2002); second, a model based method especially designed for the release of business microdata is outlined, following the proposal in Franconi and Stander (2002) and the experiences reported in Polettini *et al.* (2002). A new proposal by Burridge, stemming from the work by Franconi and Stander (2002) is described. Finally, an alternative simulation based protection method, discussed in Polettini and Franconi (2002), is outlined.
Keywords: business microdata, confidentiality, performance assessment, perturbation, protection models, regression models, simulation.

1 Introduction

Dissemination of microdata that allow for reanalysis by different users with different aims is the challenge that NSIs have been facing in the last few years.

The motivating example of the deliverable is the disclosure limitation of microdata from the Community Innovation Survey of manufacturing and services sector enterprises. This will be discussed in detail in Section 3. Business surveys such as the Community Innovation Survey often pose particular problems for disclosure limitation methodology. There are several reasons for this. First, in order to provide the best possible representation of the population, business survey designs include the largest and most identifiable enterprises with probability one; see Cox (1995). Secondly, very detailed public registers are available that contain the names of enterprises together with such features as their main economic activity, geographical area and number of employees. Accordingly, the match between public registers and an unprotected sample can often be an easy task, especially when *a priori* information such as knowledge about the inclusion of an enterprise in the sample is available. In this case identification and hence disclosure is accomplished without too much difficulty.

For the reasons just mentioned disclosure limitation of business microdata requires the use of methods that either perturb the original data or sample from the distribution originating the data themselves. Examples of such methods include masking procedures (Duncan and Pearson, 1991; Cox, 1994), data swapping (e.g. Dalenius and Reiss, 1982), and simulation from relevant distributions (see Fienberg, Makov and Steele, 1998 and references therein). These approaches are not completely satisfactory. In some cases the perturbation imposed on the data to protect the enterprises has to be so large that the errors induced in any subsequent analysis are extremely severe. In other cases, the level of perturbation imposed may not be sufficient to protect the data. Finally, some of the simulation processes that have been suggested can be difficult to implement.

1.1 Plan of the paper

In Part I of this document, as in Polettini *et al.* (2002), we argue that any microdata protection strategy is based on a formal reference model. We use this paradigm to show that different disclosure

limitation methods presented in the literature can be seen in a unified manner, distinguishing themselves according to the degree and number of restrictions imposed on the model.

In Part II we describe the disclosure limitation methodology based on regression models proposed in Franconi and Stander (2002). Section 4.1 outlines the method, which builds regression models for the continuous variables to be protected, and bases the released versions of these variables on the fitted values from these regressions. The assessment of the performance of disclosure limitation methods is discussed in Section 5. In particular, methods of quantifying the level of protection achieved for the file and the error induced by this protection are presented. A popular procedure for disclosure limitation is based on microaggregation; see Defays and Anwar (1998) for example. Section 5.3 briefly illustrates how single axis microaggregation can be applied to data from the Community Innovation Survey.

In Section 5 and 6 we also present the results of the application of this model to the Italian sample of the Community Innovation Survey (CIS), as in Franconi and Stander (2002) and Polettini *et al.* (2002). The experience with CIS data reveals some issues that need to be addressed, such as the use of robust methods, the suitable choice of some parameters, the diagnosis of the protection model, the usefulness of the data and so on.

Section 7 contains some discussion.

Part III describes an alternative approach to data perturbation stemming from the work by Franconi and Stander (2002). This research work has been conducted by Burridge, who suggested that the information related to the assumed model for the data - such as a simple regression model for instance - could be explicitly preserved by using ideas of sufficiency and conditional sampling from mathematical statistics. The resulting method of perturbing data has been termed Information Preserving Statistical Obfuscation (IPSO). Section 8 introduces the basic idea of the model, while Sections 9 to 11 describe application of this idea to multivariate continuous data, mixed continuous and discrete data, and discrete categorical data respectively.

Part IV sketches a simulation approach based on the maximum entropy formalism. This work has been discussed in Polettini and Franconi (2002). Under this approach, the real data are used to estimate a model which preserves some characteristics that have been fixed in advance. The role of simulation methods in data protection is briefly discussed in Section 12; Section 13 outlines the maximum entropy formalism.

Finally Section 14 contains considerations concerning the implementation into Argus of some of the methodologies discussed.

Part I
A Methodological Framework for Data Protection

2 A Unified Framework for Model Based Protection

In this section we express our view about protection methods for data confidentiality. We present a unified framework in which each protection method has its own reference *model*, at least in a broad sense. The extent of specification of such model yields "parametric", "semiparametric", or "nonparametric" strategies. Following this classification, a parametric probability model, such as a normal regression model, or a multivariate distribution for simulation can be specified. Matrix masking (Cox, 1994), covering local suppression, coarsening, microaggregation (Domingo-Ferrer and Mateo-Sanz, 2002), noise injection, perturbation (e.g. Kim, 1986; Fuller, 1993), provides examples of the second and third class of models. Finally, a nonparametric approach (e.g. Dandekar, Cohen and Kirkendall, 2001) can be adopted.

In our view, in order to release protected data the NSIs have basically two options:

1. coarsening, e.g. transforming the data (rows or columns of the data matrix). An extreme version of this consists of artificially introducing missing values (*local suppression*), which includes subsampling as a special case;
2. simulating artificial data set(s) or records.

Coarsening consists of transforming the data by using deterministic or random functions, either invertible or singular. Little (1993) suggests releasing a summary of the data themselves, such as a set of sufficient statistics for the assumed model. An aggregated (marginal) table for categorical data is an example of this. This is also an example of a non invertible transformation -unless the sufficient statistic achieves no reduction of the data. At the extreme of such an approach is reducing the sensitive information carried by the data by artificially introducing missing values. In both cases, Little (1993) discusses post-release analysis of protected data by means of exact or approximate ("pseudo") likelihood inference, heavily relying on the EM algorithm (see Dempster, Laird and Rubin, 1977; Little and Rubin, 1987).

Full imputation, i.e. generation of a set of artificial units, is an alternative option for NSIs. The idea of releasing a synthetic sample seems a way to avoid any confidentiality threat, as the confidentiality of synthetic individuals is not of concern. Rubin (1993) proposes using multiply-imputed data sets for release, and states the superiority of this approach over other methods of data protection. Difficulties in application of these ideas are documented by Kennickell (1998).

The idea of simulation is connected with the principle that the statistical content of the data lies in the likelihood, not in the information provided by the single respondents. Consequently, a model well representing the data could in principle replace the data themselves; alternatively, a simulated data set (or a sample of data sets) drawn from the above mentioned model can represent a more user-friendly release strategy. Indeed, since the proposal of Rubin (1993), several authors have stated the usefulness of disseminating synthetic data sets; see Fienberg, Makov and Steele (1998), and, more recently, Grim, Boček and Pudil (2001) and Dandekar *et al.* (2001; 2002b). In the choice of the generating model, the aim is always to reproduce some key characteristics of the sample.

Each of the alternative options discussed so far can be considered an imputation procedure: a protection model is formalized and released values are generated according to it in substitution of the original ones.

In the sequel, we will consider an observed data matrix X of dimension n by p; the columns of X will correspond to variables and will be denoted by X_l, $l = 1, \ldots, p$. A tilde will denote the corresponding released quantities, so that $\tilde{X}$ will denote the released matrix, $\tilde{X}_l$ the released l-th variable.

We suggest that the basic ingredient of any technique of statistical disclosure control is the *protection model*. As illustrated by the examples provided above, by protection model we mean a relation expressing the protected data in terms of the original data by means of some transformation function. Using this formulation, for the released data $\tilde{X}$ a class of distributions can be specified either directly, or through assumptions about the family of laws governing the matrix X. The degree of specification of the distributional component in the protection model varies from model to model. Some methods make no distributional assumptions, some others specify a parametric class for the probability law of the released data through assumptions on the original data matrix. Moreover, sometimes only a component of the protection model is given a fixed distribution. In some cases, only some characteristics of the distributions, such as conditional means, are specified. In this sense we will distinguish between fully nonparametric, semiparametric and fully parametric models, and from these devise nonparametric, semiparametric and fully parametric methods for data confidentiality. In this view, it is the extent of formalization of the model which makes the strategies inherently different.

2.1 Nonparametric Protection Methods

Suppose that the distribution of $\tilde{X}$ is left completely unspecified. Suppose further that the protection model for the released matrix $\tilde{X}$ has the form of a matrix mask, $\tilde{X} = XB$. The last expression is a compact notation encompassing several different imputation procedures, as discussed in Little (1993) and formalized in Cox (1994). As the latter author has shown, this protection model may provide locally suppressed, microaggregated or swapped data depending on the choice of the *attribute transforming mask* B.

Use of an additive component in the mask accounts for other types of transformations, such as topcoding. In this case, the model takes the more general form $\tilde{X} = XB + C$.

Exclusion of selected units is accomplished by using a different matrix mask, acting on the rows: $\tilde{X} = AX$, A being termed a *record transforming mask*.

Finally, exclusion of selected units followed by deletion of some pre-specified attributes is accomplished by the more general matrix mask $\tilde{X} = AXB$; actually Cox (1994) uses the more general notation $X = AXB + C$.

For protection by simulating artificial data sets, the use of procedures such as the bootstrap, or modified versions of it, give rise to nonparametric disclosure limitation strategies. An example of this is the proposal in Dandekar *et al.* (2001; 2002b) based on Latin Hypercube Sampling. The work by Fienberg *et al.* (1998) discusses analogous strategies that we would classify as nonparametric protection methods.

2.2 Semiparametric Protection Methods

In the previous paragraph, the model contains nothing but the empirical distribution of the data, plus known constants. A semiparametric structure is introduced through assumptions about the masking matrices A, B and C and/or the observed matrix X.

In particular, let us introduce a random matrix C having a known distribution. Then the masked matrix $\tilde{X}$ obtained by adding to AXB a realization of C represents a perturbation of the original data. Of course, C could be masked by introducing a column-selecting matrix D to be used whenever a variable need not be noise injected. In the context of database protection, Duncan and Mukherjee (2000) analyze the extent to which the perturbation method can be applied if valid (e.g. precise) inferences about parameters of interest are to be drawn. In particular, they discuss bounds on the variance of the noise distribution.

A particular case of semiparametric masking is the model discussed in Little (1993), which replaces the observed data by the sample mean plus random noise, obtained by setting $A = \mathbf{1}_{n\times p}$. The model just discussed in general prescribes for the data to be released a convolution of the distribution of the data, possibly after suitable transformation, with the noise distribution.

For a thorough, up-to-date discussion of noise injection, refer to the paper by Brand (2002).

We also define semiparametric the imputation model based on least squares regression; it is a semiparametric version of the naive strategy based on the release of sample mean. The model prescribes a relation between a variable X_l to be protected and a set of covariates extracted from

the observed matrix X, that we will denote by $X_{K\setminus l}$, $K \subseteq \{1, 2 \ldots, p\}$, without further assumptions on the error distribution. For a similar argument, see Little (1993).

2.3 Parametric Models

A step further is represented by the specification of a class of distributions for the released data. If the variables are continuous, very often the multivariate normal distribution is used, possibly after a normalizing transformation.

One option in model based protection is disseminating the fitted values of a normal regression model for one or more variables X_l in X; this is the main idea in Franconi and Stander (2002). A slight variation (see Little, 1993) of the regression method consists of releasing the predicted value plus random noise, taken from the estimated error distribution. This aims to compensate for the reduction of variability of the fitted values compared to that of the original data.

Of course the protection strategy based on regression models may be confined to some of the units of the data matrix; in the previous notation, this may be represented symbolically as $\tilde{X}_l = A(X_{K\setminus l}\hat{\beta} + \eta),\ \eta \sim N(0, \hat{\sigma}^2_{X_l|X_{K\setminus l}})$.

Another example of parametric disclosure protection is the release of prediction intervals for the variables to be protected, based on distributional assumptions, with or without a regression model for the variables to be protected (for an analogous strategy, see Franconi and Stander, 2000).

Finally, simulation of artificial data sets can be based on a fully specified model; the mixture model adopted in Grim *et al.* (2001) and estimated by likelihood methods with the aid of the EM algorithm provides an example of parametric protection.

For categorical variables, the strategy of releasing synthetic data sets drawn from a nonsaturated loglinear model "capturing the essential features of the data", as proposed in Fienberg *et al.* (1998), is another example of parametric procedure.

Several proposals in the literature are present which take advantage of a Bayesian formulation: among the others, Franconi and Stander (2000) develop a Bayesian hierarchical model with spatial structure, making use of the MCMC output to release predictive intervals instead of single values.

For a review of the Bayesian approach to data disclosure, see Cox (2000).

Part II
A Regression Model Approach for the Protection of Business Microdata, with an Application

3 The Microdata

At the beginning of the 1990s the European Commission and Eurostat began a survey of technological innovation in European manufacturing and services sector enterprises, called the Community Innovation Survey. The objective of this survey was the production of comparable data harmonised at the European level on all technological activities. Economists and the general research community have shown such an interest in Community Innovation Survey microdata that the problem of the release of a microdata for research file has arisen.

The data with which we work come from a representative sample of Italian manufacturing and services sector enterprises with twenty or more employees. The variables of the Community Innovation Survey can be divided into two sets. The first contains all the general information about the enterprise such as its main economic activity (four digit NACE rev. 1 Classification), geographical area, number of employees (integer ≥ 20), turnover, exports, total expenditure for research and innovation, and group membership. The first three of these variables are public. The variables turnover, exports and total expenditure for research and innovation are for 1996 and are measured in millions of Italian lire. We omit enterprises with zero turnovers or exports because the release of data about these requires special consideration. Also, when expenditures for research and innovation are taken into account, we further omit those enterprises performing innovation at no cost, for the same reasons as above.

We apply disclosure limitation separately to subsets of enterprises that are engaged in the same economic activity. As an example we will begin by considering enterprises performing the following two digit NACE rev. 1 main economic activities: 15 (food and beverage), 18 (clothing manufacture), 28 (metal products) and 36 (other products of manufacturing industries including furniture). The same framework has also been applied to all the possible economic activities (provided the sample size is large enough), and results are reported in Section 6.

The variable geographical area has eight categories: (1) North West, (2) Lombardy, (3) North East, (4) Emilia Romagna, (5) Centre, (6) Lazio, (7) Abruzzo and Molise, and (8) Campania, South, Sicily and Sardinia. These categories are based on the NUTS1 classification, with the three areas Campania, South, and Sicily and Sardinia being combined into one category since relatively few enterprises are situated in these areas.

The second set of variables contains confidential information on a range of issues connected with innovation. For simplicity we shall work with a single innovation variable that indicates whether or not an enterprise is involved in the innovation of products or processes or both.

4 The Protection Model

In order to make identification a difficult task, we intend to release less precise information for all variables that in one way or another may lead to identification. These usually include the publicly available variables geographical area and number of employees. As far as other variables are concerned, we note that, in general, continuous variables carry more risk for disclosure than categorical variables. Knowledge of the value of a continuous variable, even though it is not publicly available, can lead to the identification of an enterprise.

The leading concept of the method is that the more the enterprise is outlying, the higher the risk of it being identified. The quantitative variables mentioned above can reveal the size of

an enterprise. The information on the size, when combined with the others, can be extremely dangerous as it provides clues about the largest and therefore most identifiable enterprises. For example, information about turnover or exports, or the amount spent for research and innovation can lead to the identification of a very large and well-known enterprise among all those involved in a particular main economic activity. Hence these variables will be perturbed.

On the other hand, knowledge of a categorical variable indicating, for example, whether or not the enterprise is a member of a group, or whether or not the enterprise is involved in innovation, does not allow for such identification. This is because both small and large enterprises can belong to a group or carry out innovation.

Protection of the Community Innovation Survey data may therefore be achieved by releasing less precise information about the public variable number of employees, and the continuous variables turnover, exports and total expenditure for research and innovation. Defining broader categories for the public variable geographical area may also help to protect the data set. Information about whether or not the enterprise is a member of a group, or whether or not the enterprise is involved in innovation will be released unchanged. Releasing the innovation information unaltered is particularly appropriate for the Community Innovation Survey. In order to reduce information loss, we will not categorise the quantitative variables.

In Section 4.1 we present a new model based method for disclosure limitation that releases less precise information about the continuous variables number of employees, turnover and exports. Our treatment will follow Franconi and Stander (2002), and hence the variable total expenditure for research and innovation is not introduced into the protection model here, although it will be used in Section 6. However this is not a limitation for the model, and later the variable is explicitly introduced in the same methodological framework. The method builds regression models for the continuous variables to be protected. Some of the fitted values from these models are then shrunk before being released. In Section 4.2 we briefly describe a protection method based on principal components analysis for defining broader categories for the variable geographical area. In order to provide a comparison with existing methods, in Section 5.3 we shall describe the application of microaggregation to the variables number of employees, turnover and exports.

Assessing the performance of a disclosure limitation method is a difficult task. We shall consider this in Sections 5 and 6, with slightly different perspectives. In general terms, there is a balance to be struck between protection offered and error induced. Hence we shall discuss how to quantify the amount of protection offered and the error induced by a disclosure limitation method. We finish Section 5 by presenting results for the four NACE main economic activities under consideration. These lead us to conclude that the variable geographical area should always be protected. Moreover, the new method generally offers better protection than microaggregation, whilst inducing less error.

4.1 Formal description of the protection model

The basic idea is as follows: given a set of variables X_l and $X_{K\setminus l}$, $l \in K' \subseteq K$, $K \subseteq \{1, 2 \ldots, p\}$, X_l being the logarithms of the quantitative variables of Section 3 to be perturbed, Franconi and Stander propose to regress X_l on $X_{K\setminus l}$. Hereafter we always include in the design matrix $X_{K\setminus l}$ the unit vector $X_0 = \mathbf{1}$, in order to allow for an intercept in the regression.

The values to be released are then

$$\tilde{X}_l = \hat{X}_l + a \ , \tag{1}$$

where $\hat{X}_l$ is the fitted value and a is an adjustment factor.

For each branch of economic activity the authors consider $card(K) = 6$ variables, specifically, X_1 = log number of employees, X_2 = log turnover, X_3 = log exports, X_4 = innovation, X_5 = group membership (both dichotomous variables), X_7 = geographical area. A log-transformation for the numeric variables involved in the regressions is adopted, because of the skewed nature of the original data.

The model consists of three separate regressions, one for each of the continuous variables number of employees, turnover and exports requiring protection:

$$X_{1,ij} = \beta_0^{(1)} + \beta_1^{(1)} X_{2,ij} + \beta_2^{(1)} X_{3,ij} + \beta_3^{(1)} X_{4,ij} + \beta_4^{(1)} X_{5,ij} + \alpha_i^{(1)} + \epsilon_{ij}^{(1)} \quad (2)$$

$$X_{2,ij} = \beta_0^{(2)} + \beta_1^{(2)} X_{1,ij} + \beta_2^{(2)} X_{3,ij} + \beta_3^{(2)} X_{4,ij} + \beta_4^{(2)} X_{5,ij} + \alpha_i^{(2)} + \epsilon_{ij}^{(2)} \quad (3)$$

$$X_{3,ij} = \beta_0^{(3)} + \beta_1^{(3)} X_{1,ij} + \beta_2^{(3)} X_{2,ij} + \beta_3^{(3)} X_{4,ij} + \beta_4^{(3)} X_{5,ij} + \alpha_i^{(3)} + \epsilon_{ij}^{(3)} \quad (4)$$

The model is specified for the $j-$th enterprise in the $i-$th area, $j = 1, \ldots, n_i$, $i = 1, \ldots, N = 8$. In order to allow for spatial dependence, all of the regressions contain fixed area effects $\alpha_i^{(k)}$; the latter are constrained to sum to zero: $\sum_{i=1}^{N} \alpha_i^{(k)} = 0$.

As the formulas in (2)–(4) make evident, the protection procedure consists of performing a regression model for each X_l to be protected. Not all variables in the regressions are necessarily protected, and this was the reason for using the notation $l \in K' \subseteq K$.

The protection procedure is motivated by the form of prediction intervals. Each of the models in (2)–(4) has its own predicted values, $\widehat{X}_{l,ij} = \widehat{\mu}_{ij}^{(k)}$, $l, k = 1, \ldots, 3$. A $100(1-\xi)\%$ prediction interval for the logarithm of the response variables X_l of model k for the $j-$th individual in region i takes the form

$$(\widehat{\mu}_{ij}^{(k)} - s^{(k)} \cdot t_{\xi/2,n-12},\, \widehat{\mu}_{ij}^{(k)} + s^{(k)} \cdot t_{\xi/2,n-12})$$

where n is the number of enterprises in the NACE under consideration, $\widehat{\mu}_{ij}^{(k)}$ is the fitted value from the $k-$th regression model for the $j-$th individual in region i , $t_{\xi/2,n-12}$ is such that $\mathrm{P}(T \leq t_{\xi/2,n-12}) = 1 - \xi/2$ in which T follows a t-distribution with $n-12$ degrees of freedom, and $s^{(k)}$ is the predictive standard error from the $k-$th regression that depends upon the covariates measured on individual j. Following the form of the released values given by (1) , to protect the response variable X_l in the $k-$th model, we release $\widehat{X}_{l,ij} + s^{(k)} \cdot F_{ij}$ instead of $X_{l,ij}$, where F depends on the rank r_{ij} of $X_{l,ij}$. For a fixed value $q \in (0, 0.5)$, F_{ij} is taken to decrease linearly from a given value $F_{\max}$ to 0 as the rank of $X_{l,ij}$ increases from 1 to $[qn]$, to be 0 for values of the rank between $[qn]+1$ and $n - [qn]$, and to decrease linearly from 0 to $-F_{\max}$ as the rank increases from $n - [qn] + 1$ to n, where $[qn]$ signifies the nearest integer less than qn. Throughout q is set to 0.25 and $F_{\max}$ to 2, although results similar to the ones presented here were obtained using $F_{\max} = 3$ and 4. With this choice of q the released values of the first (last) quartile are inflated (deflated) with respect to the corresponding fitted values, with the more extreme values receiving the more extreme inflation or deflation. Later this *tail shrinkage* is not applied to values that would already be shrunk if the fitted values were to be released; such a *restricted* tail shrinkage is used in order to reduce the error induced.

The method is therefore designed to modify the marginal distributions, redistributing the tail units in the central body of the distribution. This mechanism clearly changes the marginals, but its additional aim is to allow the users to build regression models and to draw almost the same inferential conclusions as they would from the real data. Information about the original marginal distributions and tables can be recovered from the tables published by the NSIs in aggregate form, or can otherwise be supplied to the user.

4.2 Protecting the variable geographical area

In Franconi and Stander (2002) it is proposed to release the values of the variable geographical area using two broader categories. We proceed by performing a principal components analysis on standardized versions of the variables number of employees, turnover and exports. We then calculate the value of the first principal component for each enterprise. We take the average of these values over enterprises in each of the $N = 8$ geographical areas to obtain an overall effect A_i for each area $i = 1, \ldots, N$. Let $\bar{A} = \frac{1}{N} \sum_{i=1}^{N} A_i$ and $\widetilde{A}_i = A_i - \bar{A}$. Our two broader categories are defined using the $\widetilde{A}_i$ values, with areas with positive $\widetilde{A}_i$ being placed in one category, and the remaining areas being in another.

The extension of this approach to more than three variables and to more than two broader categories is straightforward. Moreover, it would also be possible to multiply the standardized variables

before performing the principal components analysis by an appropriate weight if all variables were not considered to be equally important. The results that we shall present in Section 5 changed very little indeed when we used the log-transformed variables instead of the original ones.

5 Assessing the performance of a disclosure limitation method

There are two aspects to assessing the performance of a disclosure limitation method. The first involves quantifying the level of protection achieved for the file by the method and will be discussed in Section 5.1, while the second concerns estimating the error induced and will be discussed in Section 5.2. The results obtained when both the new method and microaggregation are applied to microdata from the Community Innovation Survey are presented in Section 5.4.

As already noted, any disclosure limitation method involves a balance between protection offered and error induced. An explicit use of this balance in evaluating methods is discussed in Duncan and Mukherjee (2000) in the area of statistical disclosure limitation of databases. The framework for assessing the amount of protection offered is usually formalised by means of linkage techniques; see, for example, Duncan and Lambert (1989). These techniques are especially appropriate when the variables being protected are continuous, as in our case. In such a framework the level of protection is then measured by the number of linked enterprises, that is by the number of enterprises that are recognisable.

Several methods have being proposed to measure the effect of perturbation method on the quality of released data; see Duncan and Mukherjee (2000) and references therein. The measure that we propose is based on the reliability of inferences obtained from the released data. We believe that our measure is able to give an indication of the utility of the released data for further studies.

5.1 Quantifying the amount of protection offered

Our approach to quantifying the amount of protection offered by a disclosure limitation method is to check whether it would be possible to recognise a unit in the released data if we were to have all the information available from the original data. In this sense our measure of protection is somewhat conservative in that the intruder is assumed to have all available information for compromising confidentiality. To calculate our measure we begin by stratifying the whole data set by the variables X_4(innovation) and X_5 (group membership), these variables being released unchanged, and by the released geographical area. We next define the distance d between enterprise $\widetilde{j}$ in the released data and enterprise j in the original data as follows:

$$d(\widetilde{j}, j) = \delta(\text{employees}\,\widetilde{j}, \text{employees}\,j) + \delta(\text{turnover}\,\widetilde{j}, \text{turnover}\,j) + \delta(\text{exports}\,\widetilde{j}, \text{exports}\,j),$$

where, for example,

$$\delta(\text{employees}\,\widetilde{j}, \text{employees}\,j) = |\,\text{rank}(\text{employees}\,\widetilde{j}) - \text{rank}(\text{employees}\,j)|$$

in which rank(employees $\widetilde{j}$) (rank(employees j)) is the rank of the value of the number of employees for enterprise $\widetilde{j}$ (enterprise j) in the released (original) data among all values of number of employees in the stratum.

We say that enterprise $\widetilde{j}$ in the released data is matched if

$$\widetilde{j} = \underset{j \in \text{stratum}}{\arg\min}\; d(\widetilde{j}, j).$$

The total number of matches then provides us with a measure of protection.

5.2 Estimating the error induced

Estimating the error induced by a disclosure limitation method is a difficult problem, and here we offer only a partial solution. We use the percentage errors in estimating regression parameters as a measure of the quality of the protected data. In particular we consider the results that would be

obtained from a simple regression that users are likely to perform when investigating how turnover depends on the number of employees:

$$X_{2,ij} = \beta_0 + \beta_1 X_{1,ij} + \alpha_i + \epsilon_{ij} \tag{5}$$

where α_i is an area effect as before and $\epsilon_{ij} \sim N(0, \tau^2)$ independently, with τ unknown. We can fit this model using the original data, the data protected by the new method with restricted and unrestricted shrinkage, and the data protected by microaggregation. We define the percentage error involved in estimating β_0, say, when the data are protected using the new method as

$$100 \left(\frac{\widehat{\beta}_0^{\text{new}} - \widehat{\beta}_0^{\text{original}}}{|\widehat{\beta}_0^{\text{original}}|} \right) \%.$$

5.3 Microaggregation

As just mentioned, we will compare the method that we are proposing with the existing microaggregation approach.

The basic idea of microaggregation (Defays and Anwar, 1998) is to combine units into small homogeneous groups. The original data values of each variable being protected are then replaced by the corresponding group mean. For a recent discussion of microaggregation, see Domingo-Ferrer and Mateo-Sanz (2002).

To apply microaggregation to the three continuous variables number of employees, turnover and exports we first stratified the enterprises by the eight geographical areas. Within each stratum, a principal components analysis was performed on standardised versions of these three variables. The enterprises were then ordered according to the values of the first principal component. This is why this form of microaggregation is often referred to as single axis microaggregation. Aggregation takes place by replacing the individual values of these variables by the average over groups of ordered enterprises of size g. If the number of enterprises in a stratum is not a multiple of g, then the last step of this procedure will consider a group of size less than g. This group is combined with the preceding one, the averages of the three variables being taken over this extended group. With $g = 1$ the variables are given no protection. As g increases, the amount of perturbation seems to increase, although the differences are not large. We therefore choose to work with $g = 3$ from now on.

We applied microaggregation separately to enterprises in each of the four NACE main economic activities under consideration. For main economic activity 18 (clothing manufacture) the results of the above procedure for the variable turnover are shown in the right panel of Figure 1. We see that some protected values are lower (higher) than the true values of turnover beyond the left (right) vertical lines. This makes identification of these enterprises easier, which is not the case for the new method. As the value of g is constant, the form of microaggregation that we have considered is sometimes known as fixed group microaggregation. Other forms of microaggregation such as hierarchical clustering microaggregation are discussed in Domingo-Ferrer and Mateo-Sanz (2002), for example. We do not consider these more complicated forms of microaggregation further as they offer less protection than fixed group microaggregation. However, in Section 5 we will briefly mention a very simple type of microaggregation known as individual ranking microaggregation. For this each variable is ranked and aggregated separately. We will see that individual ranking microaggregation performs badly.

5.4 Results

Table 1 presents the number of matches obtained when the data are protected using six different protection methods. These are the new method with restricted shrinkage, the new method with unrestricted shrinkage, and microaggregation, all three being considered without and with the protection of geographical areas.

It is clear that the protection of geographical area leads to considerably fewer matches. Moreover, when geographical area is not protected, the number of matches can be very high. We therefore recommend always protecting the variable geographical area.

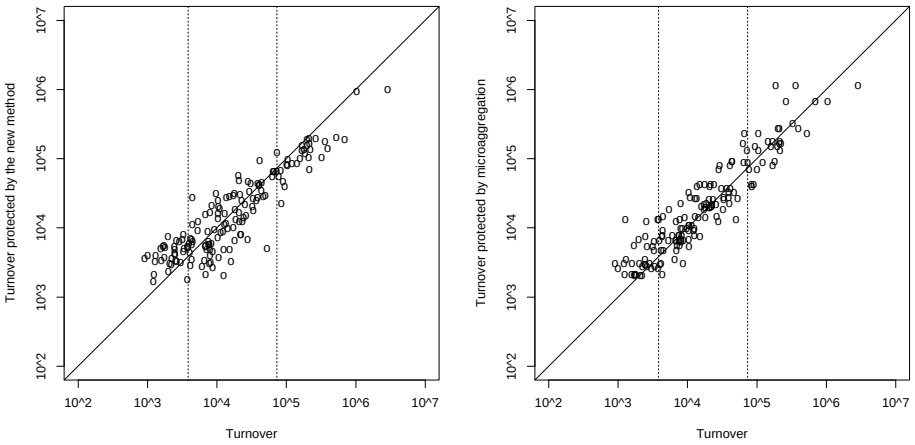

Fig. 1. Original and protected values of the variable turnover based on data from NACE main economic activity 18 (clothing manufacture). The left panel shows the result obtained from the new method. Any value lying on the diagonal line would be released without change. A logarithmic scale is used on all axes. The right panel was obtained by applying microaggregation with groups of size g to the same data.

The new method with unrestricted shrinkage leads to far fewer matches than the new method with restricted shrinkage. However, we will shortly see that the price of this is a considerable increase in induced error.

We also implemented individual ranking microaggregation as mentioned in Section 5.3. We found that the level of perturbation imposed by this method was not sufficient to protect the data even when the variable geographical area was protected. In fact, the percentage of matches achieved were 92%, 74%, 91% and 90% for NACEs 15, 18, 28 and 36 respectively. The group size g had to be increased very considerably to reduce these percentages to any great extent.

Table 2 presents the percentage errors involved in estimating the parameters of Model (5) when the data are protected using the new method with unrestricted shrinkage, the new method with restricted shrinkage and microaggregation. All three methods include the protection of geographical areas. Rather similar results were obtained when geographical area was not protected.

Both the new method with restricted shrinkage and microaggregation give an estimate of τ that is less than that obtained using the original data for all four NACE. This means that these methods have led to a reduced estimate of residual variation. The new method with restricted shrinkage estimates τ better than microaggregation.

When the data are protected using the new method with unrestricted shrinkage the estimate of τ is greater than that obtained using the original data for NACEs 15, 28 and 36; for these NACE the estimate of residual variation has been increased. This suggests that the amount of inflation or deflation produced by the method has lead to a larger dispersion in the protected than in the original data. For NACEs 15, 28 and 36 using the new method with restricted shrinkage induces less error than microaggregation when estimating the parameters β_0 and β_1. For these two parameters the new method with unrestricted shrinkage induces a much higher error than the other two protection techniques. We reached broadly similar conclusions when considering other regressions: that both the new method with restricted shrinkage and microaggregation led to a reduced estimate of residual variation; that very often - but by no means always - the new method with restricted shrinkage induces less error than microaggregation; that the new method without restricted shrinkage induces much more error than the other two protection techniques. The results presented in Table 1 and Table 2 make concrete the trade off between protection offered and error induced by a disclosure limitation method.

6 Further Experiments on the CIS survey

The methodology described in Section 4.1 was applied to the whole Italian sample from the CIS survey. A separate set of regressions plus restricted shrinkage has been fitted for each economic

Table 1. The number of matches obtained using six different protection methods: the new method with restricted shrinkage, the new method with unrestricted shrinkage, and microaggregation, all three being considered without and with the protection of geographical areas. For each NACE main economic activity the percentage of the total number of enterprises is given in brackets.

Method	NACE 15 $n = 236$	NACE 18 $n = 158$	NACE 28 $n = 369$	NACE 36 $n = 294$
New method Restricted shrinkage Areas not protected	84 (36%)	63 (40%)	103 (28%)	104 (35%)
New method Restricted shrinkage Areas protected	46 (19%)	37 (23%)	40 (11%)	49 (17%)
New method Unrestricted shrinkage Areas not protected	41 (17%)	25 (16%)	41 (11%)	35 (12%)
New method Unrestricted shrinkage Areas protected	12 (5%)	12 (8%)	13 (4%)	16 (5%)
Microaggregation Areas not protected	125 (53%)	89 (56%)	156 (42%)	125 (43%)
Microaggregation Areas protected	80 (34%)	49 (31%)	70 (19%)	74 (25%)

Table 2. Percentage errors involved in estimating the model parameters τ, β_0 and β_1 when the data are protected using the new method with restricted shrinkage, the new method with unrestricted shrinkage, and microaggregation. All three methods include the protection of geographical areas.

	New method Restricted shrinkage Areas protected	New method Unrestricted shrinkage Areas protected	Microaggregation Areas protected
τ			
NACE 15	−17.2	13.6	−41.2
NACE 18	−35.8	−2.8	−46.2
NACE 28	−20.5	13.4	−36.1
NACE 36	−25.5	4.1	−40.8
β_0			
NACE 15	−4.0	25.8	−6.6
NACE 18	−22.5	42.7	−8.3
NACE 28	−6.2	18.1	−8.4
NACE 36	−1.5	38.2	−9.2
β_1			
NACE 15	5.7	−35.0	9.1
NACE 18	21.1	−36.5	11.4
NACE 28	5.3	−21.3	8.2
NACE 36	3.2	−35.4	12.3

activity as defined by the two digit NACE classification. In this application the variable total expenditure for research and innovation has been added to each model; in order to protect this variable, an additional regression model has been introduced. Setting $K = 7$, $K' = \{1, 2, 3, 6\}$ and denoting by X_6 the logarithm of total expenditure for research and innovation, the protection model consists of the following four normal regression models:

$$X_{1,ij} = \beta_0^{(1)} + \beta_1^{(1)} X_{2,ij} + \beta_2^{(1)} X_{3,ij} + \beta_3^{(1)} X_{4,ij} + \beta_4^{(1)} X_{5,ij} + \beta_5^{(1)} X_{6,ij} + \alpha_i^{(1)} + \epsilon_{ij}^{(1)} \quad (6)$$

$$X_{2,ij} = \beta_0^{(2)} + \beta_1^{(2)} X_{1,ij} + \beta_2^{(2)} X_{3,ij} + \beta_3^{(2)} X_{4,ij} + \beta_4^{(2)} X_{5,ij} + \beta_5^{(2)} X_{6,ij} + \alpha_i^{(2)} + \epsilon_{ij}^{(2)} \quad (7)$$

$$X_{3,ij} = \beta_0^{(3)} + \beta_1^{(3)} X_{1,ij} + \beta_2^{(3)} X_{2,ij} + \beta_3^{(3)} X_{4,ij} + \beta_4^{(3)} X_{5,ij} + \beta_5^{(3)} X_{6,ij} + \alpha_i^{(3)} + \epsilon_{ij}^{(3)} \quad (8)$$

$$X_{6,ij} = \beta_0^{(4)} + \beta_1^{(4)} X_{1,ij} + \beta_2^{(4)} X_{2,ij} + \beta_3^{(4)} X_{3,ij} + \beta_4^{(4)} X_{4,ij} + \beta_5^{(4)} X_{5,ij} + \alpha_i^{(4)} + \epsilon_{ij}^{(4)} \quad (9)$$

Analogously to what was previously stated in Section 4.1, for $l \in K' = \{1, 2, 3, 5\}$ the released values of each variable X_l take the form

$$\widetilde{X}_{l,ij} = \widehat{\mu}_{ij}^{(k)} + s^{(k)} F_{ij}$$

with the same pattern for the F_{ij}s as before. For geographical area, we follow the approach of Section 4.2, though computing the principal components of the four variables to be protected, including the total expenditure for research and innovation.

6.1 Comments on the Protection Method

In this section we discuss the strengths and weaknesses of the protection method discussed in Section 4 in light of the application of this technique to the Italian section of the CIS survey.

In particular, we analyze the following issues: flexibility of the procedure; protection achieved; validity of the data in terms of estimating means, covariances, and regressions.

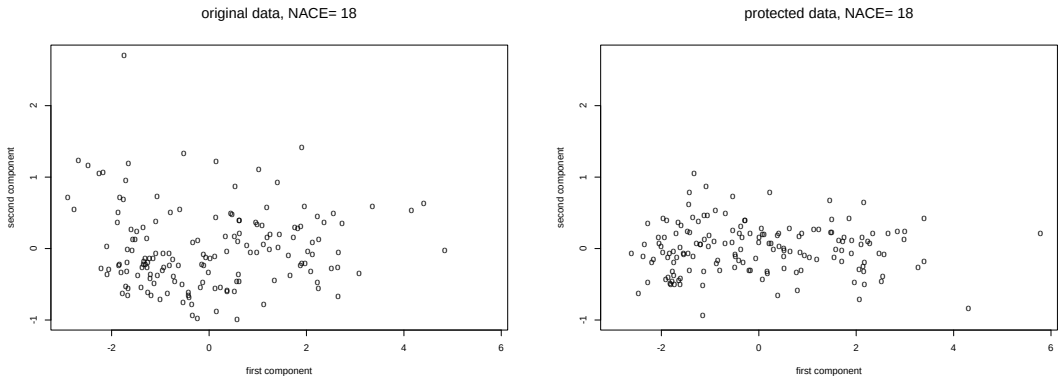

Fig. 2. Effect of the protection method on outlying enterprises. The data have been projected onto their first two principal components. Many outliers have disappeared, although some still remain

– **Flexibility**
  First of all, the main feature of the protection model above is flexibility. For example, if we assume that the risk of disclosure is proportional to the distance to the mean of the data, units may be protected according to their risk. Moreover, in formula (1), which specifies the protection model, adjustment terms a can also be constructed such that particular units receive a higher level of protection.
  Another positive characteristic of the method is its wide applicability, the core of the method being a set of regression models. Further, such a protection is easily implementable in a fully automated fashion.

– **Protection**
By construction, the model is designed to shrink the units towards the mean: in fact, use of the fitted values from the regression models and subsequent modification of the tails both serve to shrink the values towards the mean. This mechanism will in general modify the values of the outlying units.
Figure 2 illustrates the application of principal components analysis to the original and released data. It can be seen that units that are clearly visible in the first graph are effectively moved towards the centre of the data. However, the possibility that some units that fall outside the main body of the released data stay as outliers still remains. In effect, protecting some units may expose others to the risk of disclosure. In cases like this, the best strategy would probably be to re-run the model, using a higher value for q, or otherwise re-run the model on the previously protected data.
– **Data Validity: Means**
Linear regression has the property that means are left unchanged. Moreover, the use of symmetric adjustments a maintains this feature. Note though that when restricted shrinkage is adopted, the symmetry just mentioned may disappear, and the mean values might change. However this effect should in general not be dramatic. The computations carried on the real data show a good agreement between the means of the real data and the protected data. Table 3 shows the results for two selected economic activities, NACE rev.1 categories 18 (clothing manufacture) and 24 (chemical products), and for the whole sample of all economic activities.

Table 3. Effect of protection method on the means - selected economic activities

	NACE 18		NACE 24		all	
variable	original	protected	original	protected	original	protected
turnover	10.78	10.75	11.28	11.26	10.08	10.03
exports	8.06	7.99	9.63	9.59	8.38	8.30
# employees	4.57	4.56	5.22	5.22	4.48	4.47
R & I	3.79	3.85	5.60	5.67	3.87	4.94

– **Data Validity: Variance and Correlation**
In general, one side effect of the regression model plus tail shrinkage protection strategy is a certain amount of reduction in the variability. This effect is clearly visible in Figure 3.
From the point of view of an analyst wanting to use the released data to build and estimate regression models, this effect turns into a reduction of the residual variance and therefore results in an apparently increased precision of the estimates. A similar effect is seen for the correlation matrix. Table 4 show some results for food enterprises (NACE 15) and for the whole sample. In general, the regression acts to strengthen the linear relationships between variables, the only exception being with expenditure for research and innovation (R & I); this is probably due to the presence of structural zeroes for those enterprises which, being not engaged in innovation, do not present any expenditure for R & I.
– **Data Validity: Regression**
As in Franconi and Stander (2002), a simple test model for the variable turnover (here denoted by X_2) has been applied to the protected data, divided according to the NACE classification. The design matrix $X_{K\setminus 2}$ contains as explanatory variables the number of employees and geographical area. For reasons of comparability, the geographical aggregation produced by using principal components analysis has been used in both models. For enterprises performing the same two digit NACE rev.1 main economic activity, the test model takes the following form:

$$X_2 = X_{K\setminus 2}\beta + \epsilon \qquad \epsilon \sim N(0, \sigma_\epsilon^2 I)$$

for the real data, and

$$\tilde{X}_2 = \tilde{X}_{K\setminus 2}\tilde{\beta} + \eta \qquad \eta \sim N(0, \sigma_\eta^2 I)$$

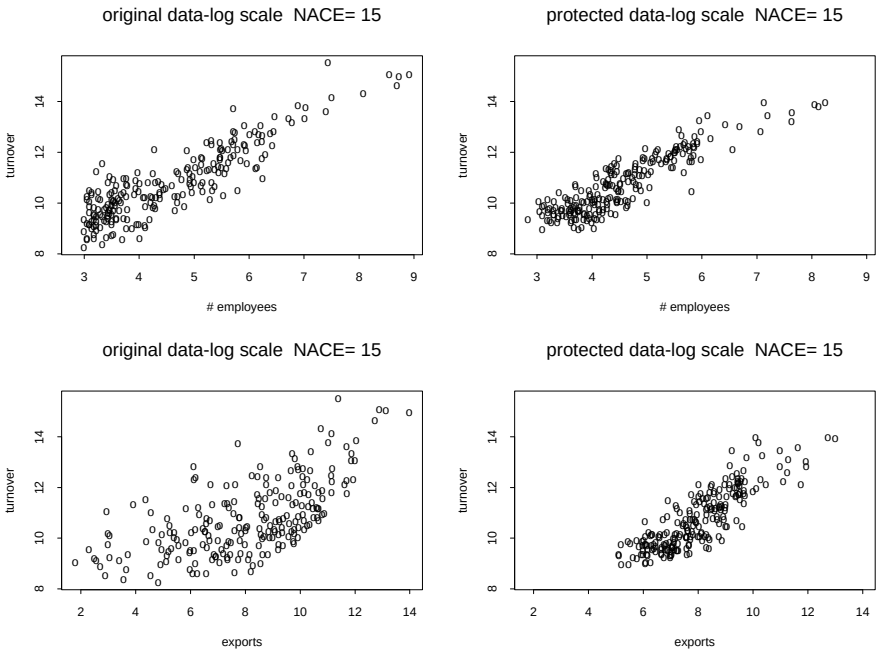

Fig. 3. Effect of the protection method on variability and bivariate relationships

Table 4. Correlation matrices. Correlations for the original data are above the diagonal, correlations for the protected data are below the diagonal.

NACE 15				
turnover	-	0.64	0.88	0.45
# employees	0.85	-	0.57	0.30
exports	0.90	0.88	-	0.48
R & I	0.31	0.23	0.34	-
all NACEs				
turnover	-	0.73	0.90	0.46
# employees	0.81	-	0.67	0.37
exports	0.91	0.84	-	0.46
R & I	0.23	0.19	0.23	-

for the protected data.
The above mentioned drop in variability is again present, as can be seen from Table 5. The table shows estimates of the variance for the following economic activities: 15 (food and beverage), 18 (clothing manufacture), 24 (chemical products) and 28 (metal products).

Table 5. Variance estimates for selected economic activities

NACE	$\widehat{\sigma_\epsilon^2}$	$\widehat{\sigma_\eta^2}$
15	0.67	0.52
18	0.90	0.61
24	0.53	0.42
28	0.49	0.38

An alternative approach that follows the above framework would replace the fitted values by the fitted values plus a random term drawn from the estimated error distribution might provide a solution to the problem of reduced residual variability. Note however that, due to the explained variation, the variability of $\hat{X}_l$ as predicted by a regression model is generally smaller than the variability of X_l. We therefore suggest that a variance inflation factor for the estimators should be released together with the data.
The reduction in variance will of course be present only for the imputed variables. The survey specific variables will be released unchanged.
Having modified the data, one cannot expect the parameter estimates for test regressions based on the released data to be the same as the estimates based on the original data. However, we hope that, on average, the fitted values based on the protected data will be the same as those based on the original data.

7 Concluding Remarks and Further Research

In the previous sections we have discussed a model based framework for disclosure protection methods and presented the results of an extensive study concerning the application of a protection technique based on regression models (Franconi and Stander, 2002). We highlight positive features, limitations, and issues to be further investigated.

Throughout the discussion we saw that the protection method acts in the sense of strengthening the relationships predicted by the regression models used. Consequently, we must be careful to include in the regression the important predictors, and especially any nonlinear relationship, as the imputation strategy tends to throw away any structure not included in the one imposed by the model itself. A good fit is of course a basic requirement that the protection model should meet in order to be sensible and effective; a model exhibiting a poor fit might in fact preserve only a negligible portion of the information carried by the data.

We may use a form of the perturbation function F that differs from the piecewise linear one that we have used. Moreover, in its present form, F depends on the units only through their ranks; using a function of the residuals might link the shrinkage to the (local) fit of the model. This link should be based on optimization algorithms whose formalization is a difficult task.

The normal regression model is sensitive to the presence of outliers. When analyzing data like the business microdata, which are characterized, by their nature, by the presence of outliers, use of robust models such as least absolute deviation regression or trimmed regression would be advisable. Indeed, these methods depend less heavily on outlying points, and more importantly, on influential observations.

Evaluation of a variance inflation factor for the estimators is another point deserving further investigation. The major difficulty is the need to devise a factor taking into account also the effect of the shrinkage component.

Provided that the limitations encountered in the application are overcome, we are confident that the method will provide a valuable solution to the problem of the dissemination of business microdata.

Part III
Information Preserving Statistical Obfuscation

8 Introduction

The model-based work by Franconi and Stander (2002) had the disadvantage that the perturbations suggested for achieving protection result in an unintended corruption of the model-related information contained in the data. While the release of perturbed data inevitably corrupts certain information, Dr Burridge suggested that the information related to the assumed model for the data - such as a simple regression model for instance - could be explicitly preserved by using ideas of sufficiency and conditional sampling from mathematical statistics. The resulting method of perturbing data has been termed Information Preserving Statistical Obfuscation (IPSO). A special case of the method produces new data with the same mean and covariance structure as the original data set. The idea has been applied to the following situations:

- multivariate continuous data
- mixed continuous and discrete data
- discrete categorical data

Progress on these topics is described in the next few sections. The next paragraph describes the basic idea.

Dr Burridge's work considers the situation where a survey consists of information gathered on continuous variables for a set of companies or individuals (the "respondents"). The information is assumed to consist of two parts for each respondent:

- public data $y = (y_1, \ldots, y_p)$
- specific survey data $x = (x_1, \ldots, x_s)$.

It is assumed that the intention is to release, for a subset of respondents, perturbed data (y', x) in place of the true data (y, x). Thus the method assumes that the intention is to release the specific survey data unchanged, but the public data will be changed in some way. The aim of this procedure is to preserve as many features of the data as possible while maintaining the identity of the respondents. For example, it might be decided to disclose only the means across all respondents of the public data y. In practice it is only possible to reach a compromise between the two objectives. The compromise investigated in Dr Burridge's work is based on considering a model for the conditional distribution of $y|x$, for example a multivariate regression model for the y variables with explanatory variables represented by the x variables. The "information" contained in the data y is summarized by a sufficient statistic T. The proposal is to produce a sample value, y' say, from the conditional distribution of $Y|(T, x)$ and to disclose (y', x). This procedure will reproduce the sufficient statistic T if information on all respondents is requested. Hence, information has been preserved while achieving some protection. The level of protection achieved will be data dependent. The particular cases investigated are described in the next few paragraphs.

9 Multivariate Continuous Data

In this section of the work it has been assumed that the, possibly transformed, public data are well approximated by a multivariate normal distribution after conditioning on the survey data. The IPSO method developed by Dr Burridge produces a new set of data with exactly the same sample properties (means, covariances, slopes, etc) as the original data. In its simplest form ("random IPSO") the method will produce a new random set of data with these properties. In this case the method produces perturbed data very easily. It is also possible to search, in a systematic manner,

for a new sample having additional properties such as a maximum level of protection ("purposive IPSO"). Computer code has already been developed within S-PLUS and FORTRAN for doing random IPSO, but still needs to be developed for purposive IPSO. Full details of this work are planned to be reported as a contribution by Dr Burridge to a special issue of the statistics journal *Statistics and Computing*.

10 Mixed Continuous and Discrete Data

This part of the work assumes that the public data consist of discrete counts and that the survey data are either discrete or continuous. The method developed so far assumes that the public data variables are observations from independent Poisson log-linear models. A FORTRAN computer program has been written which produces new samples consistent with the log-linear model estimated from the true data. A potential difficulty with the method is that the production of new samples can be computationally intensive. During the development of this program it was realized that the procedure of producing a new sample was formally similar to certain procedures developed for Monte Carlo "exact" tests of models for contingency tables, a survey of which is given by Agresti (1992). The connection between sampling of contingency tables and the needs of statistical disclosure has also been discussed by Fienberg *et al.* (1998). Dr Burridge's work on this is described briefly next.

11 Discrete Categorical Data

The most recent part of Dr Burridge's work has focussed on applying the IPSO method to categorical data which are commonly presented in the form of contingency tables. Such data are usually analyzed using Poisson or multinomial log-linear models involving factors. The difficulty of producing new random samples from certain log-linear models, as required by the IPSO method, has long been recognized and recent work (e.g. Forster, McDonald and Smith, 1996; Diaconis and Sturmfels, 1998) has concentrated on the development of Markov chain Monte Carlo methods for performing the sampling. However, when the model has the special structure of a decomposable graphical model (Lauritzen, 1996; Whittaker, 1990) more straightforward algorithms are possible (Kreiner, 1987). Dr Burridge, jointly with Dr Colin Christopher of the Department of Mathematics and Statistics, has developed a new algorithm for using a decomposable graphical model to produce data with given marginal totals from multi-way contingency tables. The algorithm is very fast and has been implemented in a FORTRAN program. The algorithm is presently being extended to make the user interface more transparent. Comparisons with the algorithms developed by Svend Kreiner have yet to be performed. A written report is currently being prepared.

Part IV
Microdata protection via simulation

12 Use of simulation techniques for disclosure protection

Traditionally, every protection technique tries to balance between protection achieved and data quality. Most protection strategies are based on perturbation, and by construction perturbation enhances protection while lowering data quality; the trade-off between disclosure protection and information loss in this case is evident. These procedures generally perturb the data to an extent that preserves confidentiality of the respondents, whereas the information loss should be addressed after the data has been transformed. Under this respect, we believe that simulation can provide an attractive alternative for data protection. Indeed, whereas artificial units are not protected by law, so that confidentiality is guaranteed, proper choice of the model to simulate from allows us to preserve data quality. Therefore in this case the protection model can be primarily tailored to data quality. The proximity of artificial units to the original data can be given additional checks, although although the concept of establishing a link between simulated and original data is not fully meaningful, as the simulated sample size may not coincide with the observed sample size. This issue will not be discussed here. Some approaches to establishing a link between real and simulated data are discussed in Dandekar, Domingo-Ferrer and Sebé (2002a).

As far as disclosure protection is concerned, our view is that in general the model which is to be used for simulation will not allow identification of individual traits. This may not be true for categorical data; however in this paper we deal with business microdata, which consist mainly of continuous variables. Concerning information loss, the point is that from a statistician's perspective, what really is of interest are *aggregate* parameters, not individual characteristics, so that release of a model that preserves these parameters will suffice for the analyst needs.

We stressed before that simulation allows to work primarily on data quality, as protection is in some sense guaranteed; this justifies our approach to disclosure limitation. As already stated, further checks for protection can be conducted before releasing the data.

13 The Maximum Entropy Approach

As highlighted in the literature (Winkler, 1998), among the desirable properties a protection procedure should exhibit, one is the agreement with some prescribed characteristics of the distribution generating the original data, e.g. the means, the covariance structure, and so on. Such an issue is clearly connected with the analytic validity of the file. Once agreed on use of simulation, restriction of the model to the class of distributions preserving these characteristics seems therefore a sensible approach to disclosure avoidance. The latter aspect is connected with the fit of the model to be used for simulation. In fact, imposing characteristics to the target distribution amounts to imposing a structure to the model in order to get as close an approximation as possible to the true distribution.

If we were to know the true underlying distribution, the ideal procedure would restrict attention to the class of models having a set of characteristics (the first M moments, say) equal to those of the original distribution. As we can only infer those parameters, the best we can do is estimate them and set $\mathcal{G}$ as the class of distributions having the prescribed characteristics equal to their empirical counterparts.

Once the class of distributions with prescribed moments or characteristics has been selected, still the problem of picking one model out of the family has to be solved. In the class $\mathcal{G}$ we propose to choose the *maximum entropy distribution*, that is known from the literature (e.g. Ihara, 1993) to exist and be unique under general conditions.

Besides providing a way to select one model in a class of distributions, the maximum entropy approach is here adopted for its information-theoretic implications. The entropy measures "uniformity" or spreadness of a distribution, which provides a measure of minus information, therefore the maximum entropy distribution in the class $\mathcal{G}$ is the element of $\mathcal{G}$ which exhibits the maximum homogeneity (e.g. uniformity) compatible with the given constraints. Resorting to the maximum entropy distribution hence in a sense tantamounts to adding no extra information other than the characteristics fixed in advance. For this reason the maximum entropy distribution in the class $\mathcal{G}$ can be described as a non-informative distribution.

Here the set of characteristics (not necessarily coinciding with the moments) represents the information to be used for the purpose of *approximating* the original distribution. If we can build a model that is sufficiently close to the original distribution, we can simulate synthetic units from it, something parallel to the parametric bootstrap. In fact, the procedure can be described as a semiparametric bootstrap. Choice of the maximum entropy distribution is further justified by the need to be as uninformative as possible about the distributional properties not accounted for. The number and type of constraints clearly affect the fit of the resulting MaxEnt distribution; the more (independent) constraints are added, the more closed to the original will be the MaxEnt distribution.

Under this approach, constraints are imposed on the expected values of *features*. We denote the features by $\phi^1(\cdot), \ldots, \phi^M(\cdot)$. These are instrumental functions of the random variable that generates the data. The expected values of features have been previously termed *characteristics*. Once the relevant characteristics and hence the corresponding features are chosen, constraints are imposed on the characteristics. In practice, empirical constraints are used, so that the expected values of the features are constrained by their empirical averages.

Under general conditions, for given constraints on such characteristics, the parametric form of the density of the maximum entropy distribution is given by:

$$f(x; \boldsymbol{\lambda}) = \frac{\exp\{-\sum_{j=1}^{M} \lambda_j \phi^j(x)\}}{Z(\boldsymbol{\lambda})} \tag{10}$$

where $Z(\boldsymbol{\lambda})$ is the normalising constant and the parameters $\lambda_1, \ldots, \lambda_M$ have to be determined so that the imposed constraints are satisfied. Since the parametric form of such model is known to belong always to an exponential family with number of parameters depending on the imposed constraints, a synthetic sample can easily be drawn for example using MCMC techniques. In this case, use of the Metropolis-Hastings is suggested to avoid numerical computation of the normalising constant.

Part V

14 Feasibility of automated procedures

All the procedures discussed so far have been put into practice by using prototype codes. Most of these methods are still under study and need to be thoroughly tested by application to real data. Such testing will be performed in Workpackage 5 and will be dicussed in Deliverable 5-D5, due at the end of the project. At the moment actual implementation of such protection strategies into μ-Argus can therefore be considered for the regression-based model of Franconi and Stander (2002) only. Generally speaking, we believe that all these models should be considered for use by experts only, mainly because the set up and details of the protection model may change across different data. In particular, as far as the implementation into $\mu-$Argus of the procedure discussed in part II is concerned, we recall that the protection method proposed is the result of data analysis and re-analysis, an involved process that cannot be fully automated. The applications show that use of a model exhibiting a poor fit to generate perturbed or synthetic data can have a dramatic effect on the quality of the released data themselves. Choice of the variables to be included in the model is another point that cannot be automated, it is heavily dependent on the objects of the survey and moreover cannot be determined in advance.

Another issue is the size of the sample in the subdomains where separate models have to be fitted. Sometimes the subdomains are too small to allow meaningful analysis. In that case no solution other than working on subsamples resulting from the aggregation of similar domains is envisaged.

References

Agresti, A. (1992). A survey of exact inference for contingency tables, *Statistical Science* **7**: 131–177. (with discussion).

Brand, R. (2002). Microdata protection through noise addition, *in* J. Domingo-Ferrer (ed.), *Inference Control in Statistical Databases*, Vol. 2316 of *Lecture Notes in Computer Science*, Springer, pp. 97–116.

Cox, L. H. (1994). Matrix masking methods for disclosure limitation in microdata, *Survey Methodology* **20**: 165–169.

Cox, L. H. (1995). Protecting confidentiality in business surveys, *in* B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge and P. S. Kott (eds), *Business Survey Methods*, Wiley, pp. 443–473.

Cox, L. H. (2000). Towards a bayesian perspective on statistical disclosure limitation, *in* E. I. George (ed.), *ISBA 2000 - The Sixth World Meeting of the International Society for Bayesian Analysis*, International Society for Bayesian Analysis, pp. 91–98.

Dalenius, T. and Reiss, S. P. (1982). Data-swapping: a technique for disclosure control, *Journal of Statistical Planning and Inference* **6**: 73–85.

Dandekar, R. A., Domingo-Ferrer, J. and Sebé, F. (2002a). LHS-based hybrid microdata vs rank swapping and microaggregation for numeric microdata protection, *in* J. Domingo-Ferrer (ed.), *Inference Control in Statistical Databases*, Vol. 2316 of *Lecture Notes in Computer Science*, Springer, pp. 153–162.

Dandekar, R., Cohen, M. and Kirkendall, N. (2001). Applicability of Latin Hypercube Sampling to create multi variate synthetic micro data, *ETK-NTTS 2001 Pre-proceedings of the Conference*, Crete, pp. 839–847.

Dandekar, R., Cohen, M. and Kirkendall, N. (2002b). Sensitive micro data protection using Latin Hypercube Sampling technique, *in* J. Domingo-Ferrer (ed.), *Inference Control in Statistical Databases*, Vol. 2316 of *Lecture Notes in Computer Science*, Springer, pp. 117–125.

Defays, D. and Anwar, M. N. (1998). Masking microdata using micro-aggregation, *Journal of Official Statistics* **14**: 449–461.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B* **40**: 1–38.

Diaconis, P. and Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions, *Annals of Statistics* **26**: 363–397.

Domingo-Ferrer, J. and Mateo-Sanz, J. M. (2002). Practical data-oriented microaggregation for statistical disclosure control, *IEEE Transactions on Knowledge and Data Engineering* **14**: 189–201.

Duncan, G. T. and Lambert, D. (1989). The risk of disclosure for microdata, *Journal of Business and Economic Statistics* **7**: 207–217.

Duncan, G. T. and Mukherjee, S. (2000). Optimal disclosure limitation strategy in statistical databases: deterring tracker attacks through additive noise, *Journal of the American Statistical Association* **95**: 720–729.

Duncan, G. T. and Pearson, R. W. (1991). Enhancing access to microdata while protecting confidentiality: prospects for the future, *Statistical Science* **6**: 219–239.

Fienberg, S. E., Makov, U. and Steele, R. J. (1998). Disclosure limitation using perturbation and related methods for categorical data, *Journal of Official Statistics* **14**: 485–502. (with discussion).

Forster, J. J., McDonald, J. W. and Smith, P. W. F. (1996). Monte carlo exact conditional tests for log-linear and logistic models, *Journal of the Royal Statistical Society, Series B* **58**: 445–453.

Franconi, L. and Stander, J. (2000). Model based disclosure limitation for business microdata, *Proceedings of the International Conference on Establishment Surveys-II*, Buffalo, New York, pp. 887–896.

Franconi, L. and Stander, J. (2002). A model based method for disclosure limitation of business microdata, *Journal of the Royal Statistical Society, Series D* **51**: 51–61.

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

Grim, J., Boček, P. and Pudil, P. (2001). Safe dissemination of census results by means of interactive probabilistic models, *ETK-NTTS 2001 Pre-proceedings of the Conference*, Crete, pp. 849–856.

Ihara, S. (1993). *Information Theory for Continuous Systems*, World Scientific Pub.

Kennickell, A. B. (1998). Multiple imputation and disclosure protection, *Proceedings of the Conference on Statistical Data Protection*, Lisbon, pp. 381–400. 1999 edition.

Kim, J. (1986). A method for limiting disclosure of microdata based on random noise and transformation, *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 370–374.

Kreiner, S. (1987). Analysis of multi-dimensional contingency tables by exact conditional tests: techniques and strategies, *Scandinavian Journal of Statistics* **14**: 97–112.

Lauritzen, S. L. (1996). *Graphical Models*, Clarendon Press, Oxford.

Little, R. J. A. (1993). Statistical analysis of masked data, *Journal of Official Statistics* **9**: 407–426.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*, John Wiley & Sons, New York.

Polettini, S. and Franconi, L. (2002). Simulation methods in data protection: an approach based on maximum entropy, *International Conference of the Royal Statistical Society*, Plymouth.

Polettini, S., Franconi, L. and Stander, J. (2002). Model based disclosure protection, *in* J. Domingo-Ferrer (ed.), *Inference Control in Statistical Databases*, Vol. 2316 of *Lecture Notes in Computer Science*, Springer, pp. 83–96.

Rubin, D. B. (1993). Discussion of "Statistical disclosure limitation", *Journal of Official Statistics* **9**: 461–468.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*, John Wiley & Sons, Chichester.

Winkler, W. E. (1998). Re-identification methods for evaluating the confidentiality of analytically valid microdata, *Research in Official Statistics* pp. 87–104.