Experiences on model based disclosure limitation[1],[2]

Luisa Franconi, Alessandra Capobianchi, Silvia Polettini and Giovanni Seri
Istat, Metodologia di base per la produzione statistica, Via C. Balbo 16, 00184 Roma, Italy
E-mail: franconi@istat.it; capobian@istat.it; polettin@istat.it; seri@istat.it

Abstract. National statistical institutes routinely apply imputation methods based on statistical models to survey nonresponses. This area of research is very important because it is at the basis of the production of economic data which are as accurate as possible. The idea is to take stock of the experiences gathered in the field of imputation methodology and to try to bridge the gap between this area of research and statistical disclosure limitation. In this paper we review our experiences on model based disclosure limitation techniques. In general, these techniques substitute the observed value of a certain variable with the estimated value via a statistical model. In particular, we discuss the problems encountered and the possible solutions found with two different models: a regression tree model [2] for a categorical variable [17] and a hierarchical model for a continuous variable [9].

Keywords: Business microdata, confidentiality, hierarchical models, regression trees

1. Introduction

Currently in Italy the only possibility for researchers to analyse business microdata from official statistical sources is through the on site facilities at Istat. Certainly this is not a satisfactory situation; alternative solutions involving both limiting access and implementing new disclosure limitation techniques should be pursued. In this paper we consider only the latter situation.

By business microdata we mean both data from large and small size enterprises. Due to the different structure of the enterprises, the origins of the data sets are twofold: for small businesses data sets are based on a sample from the population whereas for large size enterprises they are collected through a census. For this reason and due to the high recognisability of large size enterprises most of the disclosure limitation problems in releasing business microdata are encountered for large size enterprises (see [4]).

Several perturbation methods have been proposed to avoid disclosure of confidential data. In the broad family of matrix masking methods [3] we can mention the addition of independent noise [20], data swapping [5], and microaggregation [6,7]. However, disclosure limitation of business microdata is a difficult task.

For the noise addition method Winkler [22] reports failure to produce safe and useful data, whereas the use of data swapping may severely distort business microdata.

[1] This work was partially supported by the European Union project IST-2000-25069 CASC on Computational Aspects of Statistical Confidentiality.

[2] The views expressed are those of the authors and do not necessarily reflect the policies of the Istituto Nazionale di Statistica.

In the recent past, experiments at Istat with single axis microaggregation have allowed the creation of microaggregated data sets from the system of enterprise accounts. However, from the point of view of the final user, this family of techniques is not completely satisfactory. This is mainly due to the fact that it may cause some units to change their economic nature so much that they become unrepresentative of the original enterprise. For this reason the release of microaggregated data is considered mainly to be a starting experiment. Further studies at Istat in collaboration with the University of Plymouth have explored the possibility of developing disclosure limitation techniques for business microdata based on *ad hoc* statistical models. By model based disclosure limitation we mean the process of substituting the true value for a certain variable with an estimated value calculated using a statistical model.

National statistical institutes routinely apply imputation methods to incomplete questionnaires and to questionnaires for which no data is available at all. This area of research is very important because it is at the basis of the production of economic data which are as accurate as possible. In our view most, if not all, imputation methods for nonresponses are based on statistical models [14]. The idea is to take stock of the experiences gained in the field of imputation methodology and to try to bridge the gap between the two areas of research. Obviously this step implies the solution of several computational and methodological problems. Lately NSIs have been faced with the challenge of multiple imputations [18] and ways to include such methodology in the production process of official statistics. Kennickel [15] has reported experiences on the application of multiple imputations for disclosure limitation. Although the results are not completely satisfactory, the development of these ideas seems a promising area of research [11].

In this paper we briefly report on the experiences gathered at Istat in the field of model based disclosure limitation. In particular, we review the work by Romano and Seri [17] which proposes a regression tree model [2] for disclosure limitation of Community Innovation Survey data. We also review the work of Franconi and Stander [9] who suggested a hierarchical model in a Bayesian framework with random area effects. A simplified approach which considers simple regressions for the variables to be protected from disclosure is presented in Franconi and Stander [10].

The common factor underlying all these models is the simplicity of the approach. This should allow to investigate the possibilities that such methods can offer in the field of disclosure limitation, and it should also allow for an easy use and a straightforward implementation in the software μ-Argus [21] as part of the European funded project CASC (Computational Aspects of Statistical Confidentiality).

In Section 2 we present the different approaches to model based disclosure limitation which arise from the type of business variables involved in the survey. In Section 3 we discuss the regression tree model and in Section 4 we present the hierarchical model. Section 5 contains the conclusions and suggestions for further work.

2. Different approaches for different surveys

The application of any disclosure limitation method has to be carefully tailored to the type of variables present in the business survey. We first discuss different types of variables. Subsequent limitation strategies heavily depend on the variables concerned. Next, we use the comparison between the Community Innovation Survey (CIS) and the System of Enterprises Accounts Annual Survey to clarify the differences between other possible approaches.

First of all, there are variables that seem impossible to perturb because this would completely change the structure of the phenomenon under study. These are the NACE classification and the geographical area of the enterprise. Given their importance, the only way to limit disclosure on these variables is to reduce the amount of their information content by applying global recoding. So, for example, instead of releasing the complete five digit NACE classification, only the two digit level could be released. This, of course, depends on the number of enterprises belonging to this level and therefore on the structure of the economy. As for geographical area, users usually would ask for the most detailed regional information but, again, the level of detail possible for release depends on the number of enterprises present at the desired level of aggregation. There is an evident trade-off between NACE classification and geographical area as far as disclosure limitation is concerned.

In general, the NACE classification and the geographical area are seen mainly as stratification variables for the application of the various disclosure limitation methods. However, whereas most disclosure limiting methods would implement an *a priori* aggregation pattern irrespective of the various structural differences among economic fields, in the case of geographical area, model based methods can easily suggest possible aggregations via fixed area effects [10] or more complex random area effects [9].

Broadly speaking protecting business data via imputation like methods can be pursued in various ways. It is possible to keep the structural variables pertaining to the enterprises untouched, i.e. the publicly available variables such as the number of employees, and simulate all the other variables that are present in the survey. Such an approach is for example suggested by Rubin [19] by means of multiple imputation. The idea is to maintain the real framework, i.e. the true structural information on the enterprises, but to release only simulated data for all the confidential information. In this way, although the identification of the enterprises could be a straightforward task by means of matching techniques, the result of such matching would be harmless for the respondents. This type of approach would be most suitable when the survey encompasses variables that are mainly quantitative. This is because, intuitively, simulation (as well as perturbation) has less an impact on the information content of quantitative variables. Experiments are being set up to implement such an approach to the System of Enterprises Accounts Annual Survey, which collects data referring to yearly balance sheets; the variables involved in the survey are mainly of quantitative nature.

On the other hand, if most of the variables present in a business survey are categorical, a more parsimonious approach can be implemented. In fact, categorical variables carry less risk for disclosure than quantitative variables. As an alternative then, a model based perturbation process can be applied to all the variables that can lead directly or indirectly to the identification of an enterprise. Such variables are all the publicly available variables and the sensitive quantitative variables that can give clues on the size of the enterprise. In fact, knowledge of variables such as turnover, exports and costs together with other public variables, can lead to disclosure.

As discussed in the next two sections, such an approach has been tested on data from the Community Innovation Survey, a survey on technological innovation in European manufacturing and services sector enterprises; CIS is a typical example of surveys encompassing this kind of variables. In particular, for each enterprise in the sample, questions are posed on the most important economic variables and on a range of issues pertaining to innovation. Many of the questions on innovation that enterprises are asked allow an answer that takes the form of a personal view on a subject rather than a precise numerical value. For example, for the question about the objectives of innovation, possible replies are 0 for not relevant, and 1, 2 and 3 according to the degree of importance of particular objectives in a given list. As a consequence, most of the answers of interest can hardly lead to any identification. The value added of this type of approach is that the categorical variables pertaining to innovation, i.e., the variables of interest of the user, are left unchanged.

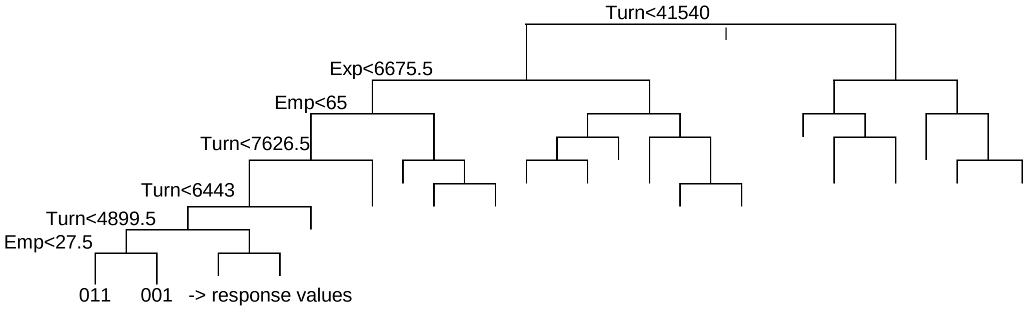

Fig. 1. Scheme of a classification tree: pruned tree with 21 nodes.

3. Model based protection: Regression trees

Classification trees are a technique especially designed to analyse and treat qualitative responses. They belong to the broader class of Classification and Regression Trees (CART) methods proposed by Breiman et al. [2]. CART consist of a set of non-parametric techniques useful to investigate the structure of the data and to solve classification problems when dealing both with categorical and numerical variables.

Such a technique can be exploited in two ways: 1) as a grouping procedure, thus allowing for model based micro-aggregation, and 2) as a classification method, thus resulting in a model based imputation method for categorical data.

The appeal of CART methods over other techniques is that it allows greater flexibility in the way the categorical data are grouped and synthesized.

For the sake of clarity, we briefly introduce CART methods.

Given a population of N individuals, on which M explanatory variables plus 1 response have been observed, a CART is aimed at generating a partition of the population into classes which are homogeneous with respect to the response variable. The classes are built so that some measure of dispersion is minimised within groups and maximised between groups. A partition is produced hierarchically by successive binary splits, each based on a critical value for one explanatory variable at a time. At the first step, all possible binary splits are scanned in order to choose the one which maximises homogeneity of the response in the two generated subgroups. The same procedure is followed iteratively for each subgroup. The algorithm proceeds with splitting until the group size is one, or a stopping criterion is verified.

The splitting algorithm described above can be graphically represented by a binary tree as shown in Fig. 1. In the terminology of trees, the terminal nodes of the tree represent the subgroups into which the sample has been partitioned. Each terminal node is attached to a representative response value: mode, median or mean, depending on the nature of the response variable. Of course, in classification problems, the latter will be a category.

A misclassification error occurs whenever the category observed on an individual differs from the one assigned to the group it belongs to. A terminal node is said to be pure if it contains no misclassifications.

Less fine partitions can be produced by pruning the classification tree. A pruned tree is obtained by deleting a node (the root node excepted) and all of its descendants.

We tested the effectiveness of these ideas by application to national data from the Community Innovation Survey (CIS).

The main aim of the survey was to assess the enterprises' innovation attitudes, roughly categorised by three binary questions concerning introduction of new products, new production processes and improvements in existing products and/or processes, respectively. At least one positive answer identifies an enterprise as innovative. In the CART model that we fitted to CIS data, the response is a Boolean combination of the three innovation variables mentioned above. Non-innovative enterprises (response 000) were excluded from the analysis. As explanatory variables we selected the most important numerical variables of the survey: turnover (Turn), number of employees (Emp) and export amount (Exp).

The aim of the model was twofold.

First, if a microaggregation is to be produced, CART may be used as a clustering procedure. For each of the variables involved in the model and for each of the groups obtained, a synthesis was computed: the mean for explanatory variables and the predicted value for the response. This approach differs from standard microaggregation techniques in the grouping procedure. Indeed CART methods are specifically designed to take into account a categorical response variable.

Secondly, CART techniques may serve as an imputation technique. As before, for the response variable one may impute the predicted value, while for explanatory variables the critical values generating the splits can be exploited to release intervals.

In both approaches, the major drawback is misclassification errors. Moreover, in the first case, aggregated values might be not safe enough to be released; in the second, the intervals suggested by the model may not be directly applicable. Further studies on the comparison between the quality of the released data via regression tree and microaggregated data is reported in Romano and Seri [17].

Application of this method to CIS data resulted in too large an amount of misclassifications (over 25%); this induced us not to pursue the use of CART as a microaggregation procedure; yet, model-based protection is an issue which retains its validity and deserves further investigation. The next example illustrates another experience pursuing the same idea.

4. Model based protection: Hierarchical models

The initial idea in Franconi and Stander [9] was to improve the use of intervals as suggested by the regression tree model approach (the use of intervals as disclosure limitation procedure is not new, see for example Gopal et al. [13]). The new feature of the proposed method is the possibility of releasing an interval based on the predictive distribution associated with the statistical model; for an example of a Bayesian setting with the use of predictive distribution see Duncan and Lambert [8]. We propose a model tailored to the CIS microdata sample; the model is autoregressive normal with response variable log(turnover) and covariates log(exports), log(employees), whether or not the enterprise is involved in product or process innovation, whether or not the enterprise belongs to a group and the associated level of the NACE classification. It also uses the geographical area to which each enterprise belongs. This geographical variable is introduced in the model through both structured and unstructured random effects, the idea being that neighbouring areas should take similar values. This is achieved by adopting a conditional autoregressive scheme as discussed in Besag et al. [1] and Molli é [16], for example. In addition, this method provides further information on how the spatial model reflects the geographical structure underlying the data. This additional knowledge into the area effect suggests a broader categorisation to use when releasing the qualitative public variable geographical area that goes a long way to minimising information loss.

To make inferences from the model, Franconi and Stander [9] make use of the Gibbs sampler. The Gibbs sampler is an example of a Markov chain Monte Carlo algorithm; for further details see Gilks et al. [12]. The main reason for this choice is simplicity of implementation. We obtained through the Gibbs sampler a sequence of $G = 1000$ vectors $\theta^{(i)}$ of parameters for our model. We delete the first $B = 500$

to remove the effect on the process due to the starting value. We then made an inference based upon this sequence. The values that will be released are based on the predictive density $p(y^{\text{new}}|\text{data})$, where y^{new} is a predicted value of the vector of log(turnover). Realisations from this predictive density can easily be obtained by simulating a vector from $p(y^{\text{new}}|\theta^{(t)})$ for each $t = B + 1, \ldots, G$. In this way for each of the original observations we obtain a vector $\left(y_{ij}^{(B+1)}, \ldots, y_{ij}^{(G)}\right)$ of realisations from the corresponding predictive density.

A $(1 - \gamma)\%$ predictive interval can be obtained from this vector by sorting it and taking the floor$\left\{\frac{\gamma}{2}(G - B)\right\}^{\text{th}}$ and the ceiling$\left\{\left(1 - \frac{\gamma}{2}\right)(G - B)\right\}^{\text{th}}$ elements, where floor(x) (ceiling(x)) returns the nearest integer below (above) x.

In order to protect the true value of turnover and hence to reduce the possibility that an enterprise is identified, we propose releasing these intervals instead. Of course, given a predictive interval in which the turnover may lie, one could estimate the true value by the midpoint for example. It may be felt more appropriate to release a point summary of the predictive density instead of an interval. Examples of such point summaries would be the predictive mean, and the predictive median.

The model has been applied to the Italian sample of CIS microdata corresponding to two different NACE sectors: sector 18 (clothing manufacture) and sector 28 (metal product manufacture). Studies on the protection offered by this method have shown better results than those obtained by single axis microaggregation when only considering the variable turnover. However a matching experiment would have to involve also the publicly available variable number of employees. A possibility for releasing such variable in this framework would be to use again an interval instead of the true value. The results so far are encouraging but not completely satisfactory. Results improve with the use of one model for each of the variable to be protected as the work by Franconi and Stander [10] suggests.

5. Conclusions and further research

In this paper we discussed the use of model based disclosure limitation and argued on different protection strategies. In general, outlying values of quantitative variables do create severe problems for disclosure limitation. Both extremely large and very small values are easily recognisable by experts of the field. The use of model based disclosure limitation addresses this type of problem only in part. A practical approach would suggest applying model based disclosure limitation and then, on the few enterprises for which the level of safety is not completely satisfactory, applying further disclosure limitation techniques. However, it would be advisable to create a general framework that is able to treat automatically all the problems that are present in business microdata. Related issues that are of vital interest to disclosure limitation are the assessment of the level of safety of the released file and the quantification of the possible distortion and information loss in the protected data. Both issues are going to be pursued further as part of the CASC project. The first one involves the study of more sophisticated record linkage techniques and the second the study of a framework for the evaluation of different perturbation techniques.

The studies carried out at Istat have shown possibilities, issues and limits of model based protection. They have also suggested different and more radical ways of protecting business microdata files. In fact, the release of a single protected file via a model based disclosure limitation methods or any other perturbative technique will always produce underestimates of the original variability in the data set. However, to be able to recover such information, national statistical institutes have to be ready to implement complex simulation methods and the users have to be ready to accept the release of several simulated data sets from the same survey. Further experiments are being set up to explore the creation of pseudo micro-data files via multiple imputations. This is to verify how much it is possible to gain from a simulation approach and how much is the burden for the final user.

References

[1] J. Besag, J. York and A. Mollié, Bayesian image restoration, with two applications in spatial statistics (with discussion), *Annals of the Institute of Statistical Mathematics* **43** (1991), 1–59.

[2] L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone, *Classification and regression trees,* Wadsworth International Group, 1984.

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

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

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

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

[7] J. Domingo Ferrer and J.M. Mateo Sanz, Practical data-oriented microaggregation for statistical disclosure control, *IEEE transaction on Knowledge and Data Engineering* (2001), in press.

[8] G.T. Duncan and D. Lambert, Disclosure-limited data dissemination (with discussion), *Journal of the American Statistical Association* **81** (1986), 10–28.

[9] L. Franconi and J. Stander, Model based disclosure limitation for business microdata, *Proceedings of the International Conference on Establishment Surveys-II,* Buffalo, New York, June 17–21, 2000, pp. 887–896.

[10] L. Franconi and J. Stander, A model based method for disclosure limitation of business microdata, Submitted for publication, 2001.

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

[12] W.R. Gilks, S. Richardson and D.J. Spiegelhalter, Introducing Markov chain Monte Carlo, in: *Markov Chain Monte Carlo in Practice,* W.R. Gilks, S. Richardson and D.J. Spiegelhalter, eds, Chapman & Hall, London, 1996, pp. 1–19.

[13] R. Gopal, P. Goes and R. Garfinkel, Confidentiality via camouflage: the CVC approach to database query management, *Proceedings of the Conference on Statistical Data Protection,* Lisbon, March 25–27, 1998, pp. 19–28.

[14] W. Kalton and D. Kasprzyk, The treatment of missing survey data, *Survey Methodology* **12** (1986), 1–16.

[15] A.B. Kennickell, Multiple imputation and disclosure protection, *Proceedings of the Conference on Statistical Data Protection,* Lisbon, March 25–27, 1998, pp. 381–400.

[16] A. Mollié, Bayesian mapping of disease, in: *Markov Chain Monte Carlo in Practice,* W.R. Gilks, S. Richardson and D.J. Spiegelhalter, eds, Chapman & Hall, London, 1996, pp. 359–379.

[17] D. Romano and G. Seri, L'uso delle tecniche di regressione ad albero per la protezione di dati elementari di impresa, *XL Riunione Scientifica della Società Italiana di Statistica,* Firenze, 26–28 Aprile 2000, 175–178.

[18] D.B. Rubin, *Multiple Imputation for Nonresponse in Surveys,* Wiley, New York, 1987.

[19] D.B. Rubin, Discussion of statistical disclosure limitation, *Journal of Official Statistics* **9** (1993), 461–468.

[20] P. Tendick, Optimal noise addition for the preservation of confidentiality in multivariate data, *Journal of Statistical Planning and Inference* **27** (1991), 342–353.

[21] L. Willenborg and A. Hundepool, ARGUS: software from the SDC project, *Statistical Data Confidentiality: Proceedings of the Joint Eurostat/UN-ECE Work Session on Statistical Data Confidentiality,* Thessaloniki, March 8–10, 1999, pp. 87–98.

[22] W.E. Winkler, Re-identification methods for evaluating the confidentiality of analytically valid microdata, *Proceedings of the Conference on Statistical Data Protection,* Lisbon, March 25–27, 1998, pp. 319–335.

Ms. Luisa Franconi has an MA in Mathematics from the University of Perugia (Italy), an MSc in Computational Statistics from the University of Bath (UK) and a PhD in Statistics from the University of Trento (Italy).

From 1992 to 1994 she was a Research Officer at the School of Mathematical Sciences, University of Bath. Since 1994 she has been with the Servizio Studi Metodologici, Istituto Nazionale di Statistica (ISTAT) working on Statistical disclosure control (SDC) methodology. She has been head of the SDC methodology unit at ISTAT since 1998.

Luisa Franconi was a committee member of the Esprit Project no 20462 on SDC and was responsible for testing the ARGUS software and developing methodology for SDC. Luisa is the Istat representative on the Eurostat Task Force on Methodological Aspects of Confidentiality. Her publications include SDC methods in various conference proceedings and journals.

Ms. Alessandra Capobianchi graduated in Statistics at the University of Rome "La Sapienza" (Italy) and she holds a PhD in Methodological Statistics from the University "La Sapienza" of Rome (Italy). Since 1998 she has been with the Servizio della Metodologia di base per la Produzione Statistica (MPS), Istituto Nazionale di Statistica (ISTAT), working on Statistical Disclosure Control (SDC) methodology.

Ms. Silvia Polettini has a MA in Statistics and a PhD in Statistics from the University of Rome (Italy). Since October 2000 she has joined ISTAT as a researcher and has been working on statistical methods for data confidentiality.

Mr. Giovanni Seri obtained his degree in Statistics at the University of Rome "La Sapienza". He is currently researcher at the Italian National Statistical Institute (ISTAT) and he has been working on Statistical disclosure control methodology since 1996.