STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

COMMISSION OF THE EUROPEAN COMMUNITIES

CONFERENCE OF EUROPEAN STATISTICIANS

EUROSTAT

Joint ECE/Eurostat Work Session on Statistical Data Confidentiality
(Skopje, The former Yugoslav Republic of Macedonia, 14-16 March 2001)

Working Paper No. 15
English only

Topic II: Impact of new technological developments in software, communications and computing on SDC

EXPERIENCES ON MODEL-BASED DISCLOSURE LIMITATION

Contributed paper

Submitted by Istat, Italy [1]

Abstract: National statistical institutes routinely apply imputation methods based on statistical models to survey non-responses. Such an area of research is very strong because is at the basis of the production of as accurate as possible economic data. The idea is to borrow strength from the experiences carried out 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 estimated value via a statistical model for the observed value of a certain variable. In particular, we discuss the problems encountered and the opportunities found with two different models: a regression tree model (Breiman *et al.*, 1984) for a categorical variable (Romano and Seri, 2000) and a hierarchical model for a continuous variable (Franconi and Stander, 2000).

Key words: Business microdata, Confidentiality, Hierarchical models, Regression Trees.

I. INTRODUCTION

1. Currently in Italy the only possibility for researchers to analyse business microdata from official statistics 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 are pursued. In this paper we consider only the latter situation.

2. For business microdata we mean data from both large and small size enterprises. Due to the different structure of the enterprises, the nature of the data set is twofold: for small businesses is a sample from the population whereas for large size enterprises it is 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 Cox (1995).

3. Several perturbation methods have been proposed to avoid disclosure of confidential data. In the broad family of matrix masking methods (Cox, 1994) we can mention the addition of independent noise, Tendick (1991), data swapping, Dalenius and Reiss (1982), microaggregation (Defays and Nanopoulos, 1992; Domingo Ferrer and Mateo Sanz, 1998). However, disclosure limitation of business microdata is a difficult task. For the noise addition method, Winkler (1999) reports failure to produce safe and useful data, whereas the use of data swapping may severely distort business microdata.

[1] Prepared by Luisa Franconi, Alessandra Capobianchi, Silvia Polettini and Giovanni Seri.

4. Recently, experiments at Istat with single axis microaggregation have allowed the creation of microaggregated data sets from the system of enterprises account. 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 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. For model-based disclosure limitation, we mean the substitution of the true value for a certain variable by that obtained from the process of estimation of a statistical model.

5. National statistical institutes routinely apply imputation methods to uncompleted questionnaires and to completely missing enterprises. Such an area of research is very strong because it is at the basis of the production of as accurate as possible economic data. In our view most, if not all, imputation methods for non-responses are based on statistical models (Kalton and Kasprzyk, 1986). The idea is to borrow strength from the experiences matured 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 imputation (Rubin, 1987) and ways to include such methodology in the production process of official statistics. Kennickel (1999) has reported experiences on the application of multiple imputation for disclosure limitation. Although the results are not completely satisfactory the development of these ideas seems a promising area of research, Fienberg *et al*. (1998).

6. In this paper we briefly report on the experiences carried out at Istat in the field of model-based disclosure limitation. In particular, we review the work by Romano and Seri (2000) that proposes a regression tree model (Breiman *et al.*, 1984) for disclosure limitation of Community Innovation Survey data. We also review the work of Franconi and Stander (2000) who suggested a hierarchical model in a Bayesian framework with random area effects. A simplified approach that considers simple regressions for the variables to be protected is presented in another paper in this workshop: Franconi and Stander (2001).

7. The common factor underlying all these models is simplicity of the approach. The reason is twofold: first to investigate the opportunities that such methods can offer in the field of disclosure limitation and second to allow for an easy use and a straightforward implementation in the software µ–Argus (Willenborg and Hundepool, 1999) as part of the European founded project CASC (Computational Aspects of Statistical Confidentiality).

8. In Section II we present the different approaches to model-based disclosure limitation which arise on the type of business variables involved in the survey. In Section III we discuss the regression tree model and in Section IV we present the hierarchical model. Section V contains the conclusions and suggestions for further work.

II. DIFFERENT PERSPECTIVES FOR DIFFERENT SURVEYS

9. The application of any disclosure limitation methods has to be carefully tailored to the type of variables present in the business survey. We first distinguish between different type of variables, for disclosure issues and subsequent limitation strategies heavily depend on the variables under study. Next, the comparison between the Community Innovation Survey (CIS) and the System of Enterprises Accounts Annual Survey will clarify the differences between possible approaches.

10. First of all there are variables that seem impossible to perturb because they will completely change the structure of the phenomenon under study. These are the NACE classification and the geographical area of the enterprise. Given the importance of such variables the only way to limit disclosure through them is by reducing the amount of their information content 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 far as geographical area is concerned, usually users would

ask for the most detailed regional information but, again, the level of detail that is possible to 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.

11. Therefore, 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, as far as the geographical area is concerned, whereas most disclosure limiting methods would implement an *a priori* aggregation pattern irrespective of the different structural differences among economic fields, model-based methods can easily suggest possible aggregations via fixed area effects (Franconi and Stander, 2001) or more complex random area effects (Franconi and Stander, 2000).

12. Broadly speaking, protecting business data via imputation-like methods can be pursued in various ways. It is possible to keep untouched the structural variables pertaining to the enterprises, i.e. the publicly available variables such as the number of employees, and simulate all the other variables that are present of the survey. Such an approach is suggested, for example, by Rubin (1993) 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.

13. On the other hand, if most of the variables present in a business survey are categorical, a more parsimonious approach could 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 the enterprise. Such variables are all the publicly available variables and the quantitative sensitive variables that can give clues on the size of the enterprise. In fact, knowledge of variables such as turnover, exports and costs together with the public variables, can lead to disclosure. The Community Innovation Survey, a survey on technological innovation in European manufacturing and services sector enterprises, is a typical example of this second class of survey. 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 the 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.

III. MODEL-BASED PROTECTION: REGRESSION TREES

14. The use of a procedure specially designed for the treatment and analysis of qualitative responses that would allow for flexibility in the way the categorical data are grouped and synthesised are the remarks that motivate the introduction of classification trees techniques (Breiman *et al.* 1984):

i) as a grouping procedure;
ii) as a classification method, viewed as a tool for categorical data imputation.

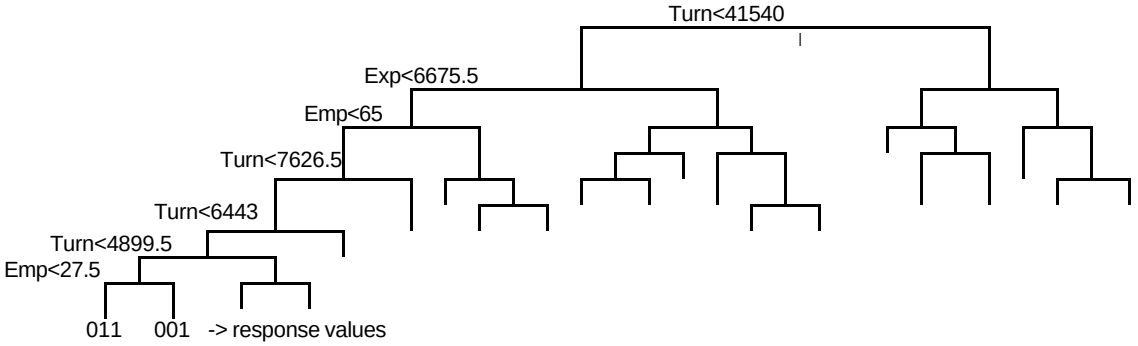
We tested the effectiveness of these ideas by application to national data from the Community Innovation Survey (CIS).

15. For the sake of clarity, we briefly introduce Classification and Regression Trees (CART) methods, a set of non-parametric techniques useful to investigate the structure of the data and to solve classification problems when dealing with both categorical and numerical variables.

16. 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 on each subgroup. The algorithm proceeds with splitting until the group size is one, or a stopping criterion is verified.

17. The splitting algorithm described above can be represented by a binary tree as shown in Figure 1. The terminal nodes of the tree represent the classes of the target partition. Each terminal node is attached a representative response value (e.g. group mode, median or mean, depending on the nature of the response variable), a category in classification problems. 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.

Figure 1: *Scheme of a classification tree: pruned tree with 21 nodes*



18. The main objective of the survey was to asses the enterprise's 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 we fitted to CIS data (non-innovative enterprises being excluded from the analysis) the response is a Boolean combination of the three innovation variables above. As explanatory variables we selected the main numerical variables: turnover (Turn), number of employees (Emp) and export amount (Exp).

19. 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 a synthesis (the mean for explanatory variables and the predicted value for the response) can be computed in each of the groups obtained. This approach differs from standard microaggregation procedures for the grouping procedure, designed for taking specifically into account a categorical response variable. Secondly, CART techniques may serve as an imputation technique. As before, the response may be assigned the predicted value, while for explanatory variables the critical values generating the splits can be exploited to release intervals.

20. 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, intervals as suggested by the model may be not 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 (2000). Application to CIS data resulted in too large an amount of misclassifications (over 25%); this induced us not to pursue use of CART as a microaggregation procedure; though, model-based protection is an issue which retains its validity and deserves further investigation. The next example illustrates another experience pursuing the same idea.

IV. MODEL-BASED PROTECTION: HIERARCHICAL MODELS

21. The initial idea in Franconi and Stander (2000) 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, Goes and Garfinkel, 1999). 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 by Duncan and Lambert (1986). The proposed 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.* (1991) and Mollié (1996), for example. A by-product of the method is the insight that the spatial model gives into the geographical structure underlying the data. This insight 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.

22. To make inferences from the model Franconi and Stander (2000) 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.* (1996), for example. The main reason for this choice is simplicity of implementation. We obtained through the Gibbs sampler a sequence of $G = 1000$ vectors $\boldsymbol{q}^{(i)}$ of parameters for our model. We throw away the first $B = 500$ to remove the effect, on the process, due to the starting value. Inference is then based upon this sequence. The values that will be released are based on the predictive density $p(y^{\text{new}} \mid \text{data})$, 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 $p(y^{\text{new}} \mid \boldsymbol{q}^{(t)})$ for each $t = B+1, \ldots, G$. In this way for each of the original observations we obtain a vector $(y_{ij}^{(B+1)}, \ldots, y_{ij}^{(G)})$ of realisations from the corresponding predictive density.
A $(1-g)\%$ predictive interval can be obtained from this vector by sorting it and taking the $\text{floor}\left\{\frac{g}{2}(G-B)\right\}^{\text{th}}$ and the $\text{ceiling}\left\{\left(1-\frac{g}{2}\right)(G-B)\right\}^{\text{th}}$ elements, where $\text{floor}\,(x)$ ($\text{ceiling}\,(x)$) returns the nearest integer below (above) x.

23. 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.

24. 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 the method have shown better results than those obtained by single axis microaggregation when only the variable turnover was considered. However a matching experiment would involve also the publicly available variables 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 were 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 (2001) suggests.

V. CONCLUSIONS AND FURTHER RESEARCH

25. In this paper we discuss the use of model-based disclosure limitation and argue on different protection strategies. In general, outlying values of quantitative variables 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 only in part this type of problem. 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.

26. 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 imputation. 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.

Acknowledgements

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

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

References

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

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

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 *Business Survey Methods* (eds. B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge and P. S. Kott), pp. 443–473. New-York: Wiley.

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

Defays, D. and Nanopoulos, P. (1992) Panels of enterprises and confidentiality: the small aggregates method. *Proceedings of Statistics Canada Symposium 92, Design and Analysis of Longitudinal Surveys*, 195–204.

Domingo Ferrer, J. and Mateo Sanz, J.M. (1998). Practical data-oriented microaggregation for statistical disclosure control. *Report de Recerca*, DEI-RR-98-005, Departament d'Enginyeria Informàtica, Universitat "Rovira i Virgili", Spain.

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

Franconi, L. and Stander, J. (2000) Model based disclosure limitation for business microdata. *Proceedings of the International Conference on Establishment Surveys-II*, June 17–21, 2000, Buffalo, New York. In Press.

Franconi, L. and Stander, J. (2001) Microaggregation and model based disclosure limitation and their application to business microdata. Submitted for publication.

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

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

Gopal, R., Goes, P. and Garfinkel, R. (1999) Confidentiality via camouflage: the CVC approach to database query management. *Proceedings of the Conference on Statistical Data Protection*, March, 25–27, 1998, Lisbon, pp. 19–28.

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

Kennickell, A. B. (1999) Multiple imputation and disclosure protection, *Proceedings of the Conference on Statistical Data Protection*, March, 25–27, 1998, Lisbon, pp. 381–400.

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

Romano, D. and Seri, G. (2000). 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. In Press.

Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

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

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

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

Winkler, W. E. (1999) Re-identification methods for evaluating the confidentiality of analitically valid microdata. *Proceedings of the Conference on Statistical Data Protection*, March, 25–27, 1998, Lisbon, pp. 319–335.