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. 22
English only

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

A MODEL-BASED DISCLOSURE LIMITATION METHOD FOR BUSINESS MICRODATA

Contributed paper

Submitted by University of Plymouth, U.K. and Istituto Nazionale di Statistica, Italy[1]

Summary. We outline a new method for disclosure limitation, illustrating our approach on microdata from the Community Innovation Survey. We describe microaggregation and explain how it does not offer protection to the variable geographical area, whereas the new method suggests how to define broader categories for releasing this variable. We discuss how to assess both the amount of protection offered and the error induced by a disclosure limitation method. We find that for the NACE main economic activity considered the new method offers more protection than microaggregation, and can sometimes lead to a smaller error.

Keywords: Area effects; Community Innovation Survey; Performance assessment for disclosure limitation methods; Prediction intervals

I. Introduction

I.1 Background

1. The disclosure of information is often associated with the concept of identification. Identification occurs when it is possible to recognise the name of a unit and leads to the disclosure of all confidential information pertaining to that unit. Although a released microdata file would not contain direct identifiers such as name or fiscal code, identification may still be made by using both *a priori* knowledge about the inclusion of a unit in the sample and other data generally available from public registers.

2. This paper is motivated by the disclosure limitation of microdata from the Community Innovation Survey of manufacturing and services sector enterprises. 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, 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 is available. In this case identification and hence disclosure is accomplished without too much difficulty.

[1] Prepared by Julian Stander, Department of Mathematics and Statistics, University of Plymouth and Luisa Franconi, Servizio della Metodologia di Base per la Produzione Statistica, ISTAT.

3. For the reasons just mentioned disclosure limitation of business microdata requires the use of methods that perturb the original data. In this paper we develop a new disclosure limitation methodology motivated by prediction intervals and apply it to the Community Innovation Survey. We also illustrate how single axis microaggregation (Defays and Nanopoulos, 1992) can be applied to the same data.

I.2 Data available from the Community Innovation Survey

4. 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. 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, geographical area, number of employees, turnover, export, and group membership. As already mentioned, the first three of these variables are public. The variables turnover and exports 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.

5. We propose applying disclosure limitation separately to subsets of enterprises that are engaged in the same economic activity. In this paper, we will consider enterprises involved in clothing manufacture, that is with NACE main economic activity code 18.

6. The variable geographical area has eight categories: North West (NW), Lombardy (LOM), North East (NE), Emilia Romagna (ER), Centre (CEN), Lazio (LAZ), Abruzzo and Molise (ABMO), and Campania, South, Sicily and Sardinia (SOU+ISL). 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.

7. The second set of variables contains confidential information on a range of issues connected with innovation. We shall work with a single innovation variable that indicates whether or not an enterprise is involved in product or process innovation.

I.3 Disclosure limitation approaches for the Community Innovation Survey and outline of the paper

8. In order to make identification a difficult task, we intend to release less precise information for all variables that may lead to identification. These certainly include the publicly available variables geographical area and number of employees. In addition turnover and exports need to be perturbed because information about these variables can lead to the identification of a very large and well-known enterprise. Since 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, these variables will not be perturbed. Releasing the innovation information unaltered is particularly appropriate for the Community Innovation Survey.

9. In Section II we apply microaggregation to the variables number of employees, turnover and exports. As the microaggregation procedure does not suggest how to release the variable geographical area, background knowledge about the general economic conditions in Italy is used to define broader categories for this variable. In Section III we present a new method for disclosure limitation that not only releases less precise information about the variables number of employees, turnover and exports, but also suggests how to define broader categories for the variable geographical area. It is planned to include this new method in μ-Argus, a software package for producing safe microdata files discussed by Willenborg and Hundepool (1999).

10. We consider the difficult task of assessing the performance of a disclosure limitation method in Section IV. We discuss how to quantify the amount of protection offered and the error induced, and

illustrate our measures by presenting results for the Community Innovation Survey data. In Section V we outline our conclusions.

II. MICROAGGREGATION

11. To apply microaggregation to the variables number of employees, turnover and exports we first stratified the enterprises by the eight geographical areas. Within each stratum, a principal component analysis was performed based on the correlation matrix of these three variables. The enterprises were then ordered according to the first principal component. The individual values of these variables were replaced by the average over groups of ordered enterprises of size g. As g increases, the amount of perturbation seems to increase, although the differences are not large. We therefore chose to work with $g = 3$. The results of applying microaggregation to the variable turnover are shown in the left panel of Figure 1.

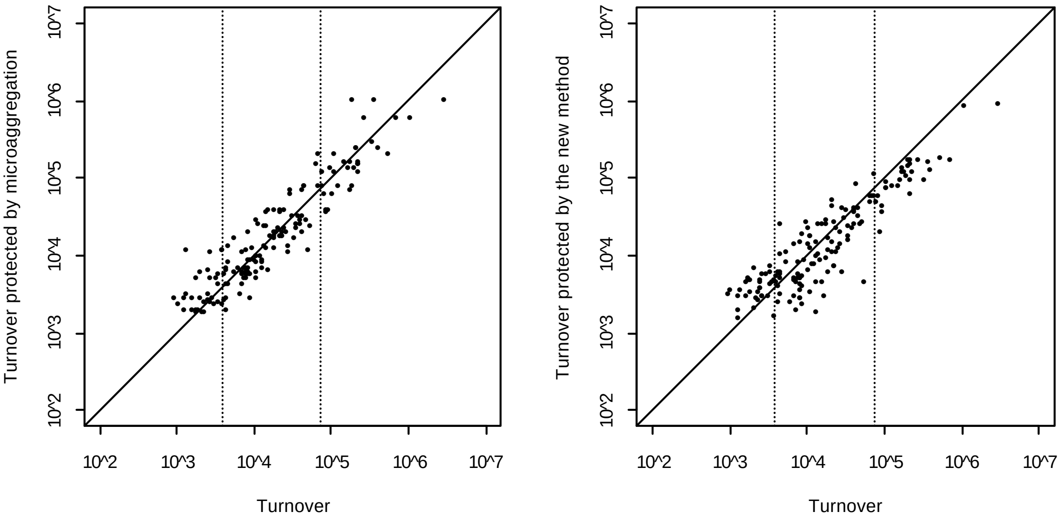

Fig. 1. Original and protected values of the variable turnover. The left panel was obtained using microaggregation. The right panel was obtained using the new protection method. Values lying on the diagonal line are released without change. The vertical lines are referred to in Section 3. A logarithmic scale is used on all axes.

12. As this microaggregation procedure does not suggest how to release the variable geographical area, we use background knowledge about the general economic conditions in Italy to define broader categories for this variable. In particular, the four areas North West, Lombardy, North East and Emilia Romagna are aggregated into one large area, while the remaining four areas Centre, Lazio, Abruzzo and Molise, and Campania, South, Sicily and Sardinia are aggregated into another.

III. A NEW METHOD FOR DISCLOSURE LIMITATION

13. Our new method for disclosure limitation is based on three regressions, one for each of the variables number of employees, turnover and exports requiring protection:

$$x_{1,ij} = n^{(1)} + a^{(1)} x_{2,ij} + b^{(1)} x_{3,ij} + g^{(1)} x_{4,ij} + d^{(1)} x_{5,ij} + a_i^{(1)} + e_{ij}^{(1)}$$
$$x_{2,ij} = n^{(2)} + a^{(2)} x_{1,ij} + b^{(2)} x_{3,ij} + g^{(2)} x_{4,ij} + d^{(2)} x_{5,ij} + a_i^{(2)} + e_{ij}^{(2)}$$
$$x_{3,ij} = n^{(3)} + a^{(3)} x_{1,ij} + b^{(3)} x_{2,ij} + g^{(3)} x_{4,ij} + d^{(3)} x_{5,ij} + a_i^{(3)} + e_{ij}^{(3)}$$

in which $x_{1,ij} = \log(\text{ number of employees }_{ij})$ is the logarithm of the number of employees for the j^{th} enterprise in the i^{th} area, $j = 1,\ldots,n_i$, $i = 1,\ldots,N = 8$, $x_{2,ij} = \log(\text{ turnover }_{ij})$, $x_{3,ij} = \log(\text{ exports }_{ij})$, $x_{4,ij} = 1$ if the enterprise is involved in innovation and 0 otherwise, $x_{5,ij} = 1$ if the enterprise belongs to a group and 0 otherwise, $a_i^{(k)}$ is a fixed factor for the i^{th} area in the k^{th} regression, $k = 1,2,3$. The $a_i^{(k)}$'s are constrained to sum to zero: $\sum_{i=1}^{N} a_i^{(k)} = 0$. We decided to employ the logarithmic transformations by considering a standard residual analysis.

14. As we apply the same protection procedure to all three variables, we will illustrate it by considering turnover. We release perturbed values of the variables of interest based on the form of prediction intervals. A $100(1-x)\%$ prediction interval for $x_{2,ij} = \log(\text{ turnover }_{ij})$ at $(x_{1,ij}, x_{3,ij}, x_{4,ij}, x_{5,ij})$ based on the second regression takes the form
$$(\hat{m}_{ij} - t_{x/2,n-12}\, s, \quad \hat{m}_{ij} + t_{x/2,n-12}\, s),$$

where $n = 158$ is the number of enterprises, $\hat{m}_{ij}$ is the fitted value from the second regression at $(x_{1,ij}, x_{3,ij}, x_{4,ij}, x_{5,ij})$, $t_{x/2,n-12}$ is such that $P(T \le t_{x/2,n-12}) = 1 - x/2$ in which T follows a t-distribution with $n-12$ degrees of freedom, and s is the predictive standard error from the second regression that depends upon $(x_{1,ij}, x_{3,ij}, x_{4,ij}, x_{5,ij})$. To protect $x_{2,ij}$, we release $\hat{m}_{ij} + K_{ij}s$ instead of $x_{2,ij}$, where K_{ij} depends on r_{ij} the rank of $x_{2,ij}$. For a fixed value of $p \in (0, 0.5)$, we take K_{ij} to decrease linearly from $K_{\max}$ to 0 as r_{ij} increases from 1 to $[pn]$, to be zero for values of r_{ij} between $[pn]+1$ and $n-[pn]$, and to decrease linearly from 0 to $-K_{\max}$, as r_{ij} increases from $n-[pn]+1$ to n, where $[pn]$ signifies the nearest integer less that pn. In this way the released values of the smallest (largest) $100p\%$ of the $x_{2,ij}$'s are inflated (deflated) with respect to the corresponding fitted values $\hat{m}_{ij}$, with the more extreme values receiving the more extreme inflation or deflation. We do not however apply this inflation (deflation) to values that would already be inflated (deflated) if the fitted value $\hat{m}_{ij}$ itself were to be released. Throughout we set $p = 0.25$, so offering further protection to the first and last quartiles of the variables being protected. We take $K_{\max} = 2$, although similar results to those that we present here were obtained using $K_{\max} = 3$ and 4.

15. In the right panel of Figure 1 we plot the values of turnover released by the new methods against the true values. We see that the released values are lower (higher) than the true values for high (low) values of turnover, that is for values of turnover beyond the right (left) vertical line. In this way, the value of turnover is reduced (increased) for enterprises with very large (small) values of turnover, so rendering identification more difficult for these. This effect is not seen for microaggregation.

16. In Figure 3 the estimated area effects $\hat{a}_i^{(k)}$, $i = 1,\ldots,8$, $k = 1,2,3$, are shown.

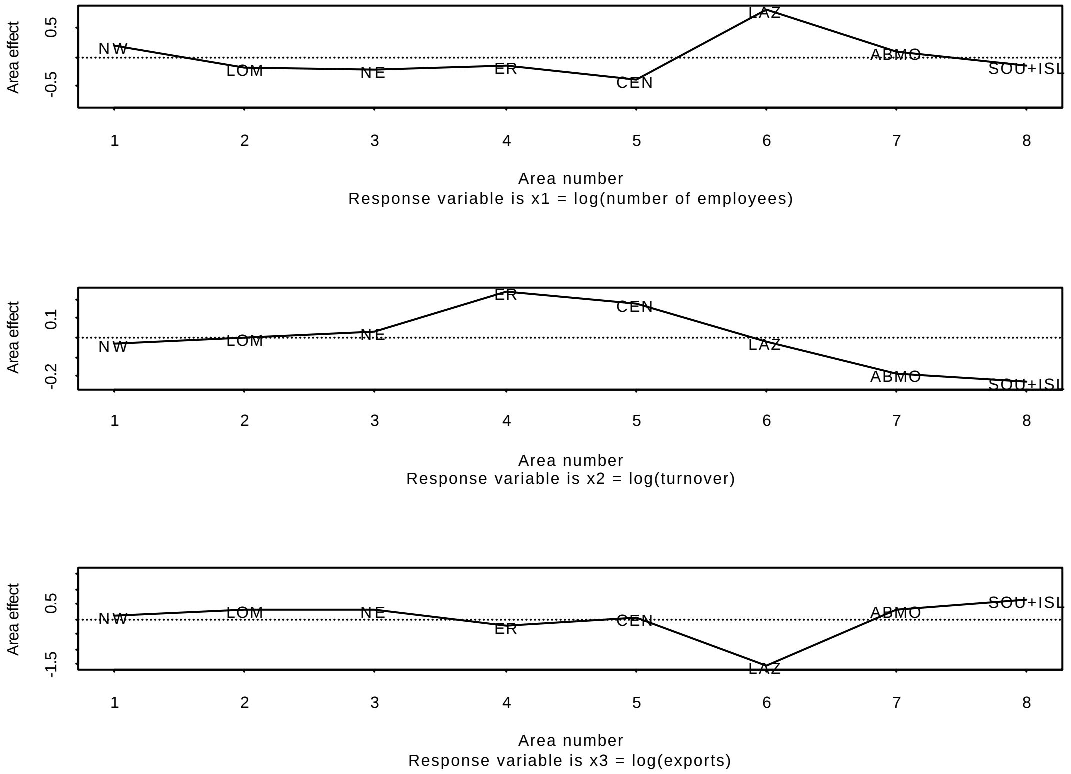

Fig. 3. The estimated area effects $\hat{a}_i^{(k)}$, $i = 1,\ldots,8$, $k = 1,2,3$.

16. We propose releasing the values of the variable geographical area using two broader categories based on these area effects. A simple approach could be based on the area effects from just one of the regressions, say the first. For example, the two broader categories could correspond to whether $\hat{a}_i^{(1)}$ were positive or negative, yielding the North West, Lazio and Abruzzo and Molise in one category, and Lombardy, North East, Emilia Romagna, Centre, and South, Sicily and Sardinia in the other. We could write this broader categorisation as $A^{(1)} = (1,0,0,0,0,1,1,0)$. Note that the opposite assignment $\overline{A}^{(1)} = (0,1,1,1,1,0,0,1)$ provides an equivalent categorisation. We could produce categorisations $A^{(2)}$ and $A^{(3)}$ for the second and third regressions in a similar way, with $\overline{A}^{(2)}$ and $\overline{A}^{(3)}$ representing the opposite assignments.

17. A more sophisticated approach would combine the estimated area effects from all three regression. To achieve this we define a measure of the overall dissimilarity between any three categorisations $B^{(1)}$, $B^{(2)}$ and $B^{(3)}$ to be

$$D(B^{(1)}, B^{(2)}, B^{(3)}) = d(B^{(1)}, B^{(2)}) + d(B^{(1)}, B^{(3)}) + d(B^{(2)}, B^{(3)}),$$

in which $d(B^{(k)}, B^{(l)}) = \sum_{i=1}^{N} \left| B_i^{(k)} - B_i^{(l)} \right|$, where $B_i^{(k)}$ is the i^{th} element of $B^{(k)}$. We then minimise $D(B^{(1)}, B^{(2)}, B^{(3)})$ over the four categorisation triples $(A^{(1)}, A^{(2)}, A^{(3)})$, $(A^{(1)}, \overline{A}^{(2)}, A^{(3)})$, $(A^{(1)}, A^{(2)}, \overline{A}^{(3)})$ and $(A^{(1)}, \overline{A}^{(2)}, \overline{A}^{(3)})$, calling the minimising categorisation triple $(B^{(1)*}, B^{(2)*}, B^{(3)*})$. We do not need to consider the other four categorisation triples because of symmetry. Finally we summarise this categorisation triple by averaging to obtain $\left(B^{(1)*} + B^{(2)*} + B^{(3)*}\right)/3$ and rounding $1/3$

down to zero and $2/3$ up to one. This yields a broader categorisation based on all three regressions. From now on we use this second approach.

IV. ASSESSING THE PERFORMANCE OF MICROAGGREGATION AND OF THE NEW METHOD FOR DISCLOSURE LIMITATION

18. There are two aspects to assessing the performance of a method for disclosure limitation. The first involves quantifying the amount of protection offered by the method and will be discussed in Section IV.1, while the second concerns estimating the error induced by the method and will be discussed in Section IV.2. The results obtained for both microaggregation and the new method are presented in Section IV.3.

IV.1 Quantifying the amount of protection offered by a disclosure limitation method

19. 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 offered may be somewhat pessimistic. 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. Then for each stratum we find the number of enterprises in the released data set that have themselves as 'nearest neighbours' amongst enterprises in the original data set in the same released geographical area. We refer to this as the number of matches, and this will be our measure of protection. In order to define 'nearest neighbour', we prescribe the distance d between enterprise j in the released data and enterprise j^* in the original data as follows:

$$\begin{aligned} d(\text{released enterprise } j, \text{original enterprise } j^*) = {} & d(\text{employees } j, \text{employees } j^*) \\ & + d(\text{turnover } j, \text{turnover } j^*) \\ & + d(\text{exports } j, \text{exports } j^*), \end{aligned}$$

where, for example,

$$d(\text{employees } j, \text{employees } j^*) = \left| \text{rank}(\text{employees } j) - \text{rank}(\text{employees } j^*) \right|,$$

in which $\text{rank}(\text{employees } j)$ ($\text{rank}(\text{employees } j^*)$) is the rank of the value of the number of employees for enterprise j (j^*) in the released (original) data among all values of number of employees in the stratum and released geographical area.

IV.2 Estimating the error induced by a disclosure limitation method

20. In order to estimate the error induced by a disclosure limitation method we considering the results that would be obtained from some simple regressions that users are likely to perform. The first regression that we consider is:

$$x_{2,ij} = n + b\, x_{1,ij} + a_i + e_{ij}, \tag{1}$$

where a_i is an area effect and $e_{ij} \sim N(0, t^2)$ independently, with t unknown. We can fit this model using the original data, the data released by the new method, and the data released by the microaggregation method. We define the percentage error involved in estimating n, say, when the data are protected using the new method as

$$100\left(\frac{\hat{n}^{\text{new}}-\hat{n}^{\text{original}}}{\hat{n}^{\text{original}}}\right)\%,$$

where $\hat{n}^{\text{new}}$ ($\hat{n}^{\text{original}}$) is the estimate of n based on data protected using the new method (the original data); the percentage error when the data are protected using microaggregation is defined similarly.

21. In addition to regression (1) above, we shall consider these two simple regressions:

$$x_{2,ij} = n + g\, x_{3,ij} + a_i + e_{ij}, \qquad (2)$$

$$x_{2,ij} = n + b\, x_{1,ij} + g\, x_{3,ij} + a_i + e_{ij}, \qquad (3)$$

IV.3 Results

22. We found the amount of protection offered by the new method to be greater than that offered by microaggregation, since with the new method 36 (23%) matches were obtained, while with microaggregation this figure rose to 43 (27%).

23. Table 1 presents the percentage errors involved in estimating the model parameters when the data are protected using the new method and microaggregation. Both methods always give an estimate of t that is less than that obtained using the original data. This means that both methods have led to reduced residual variation. For all regressions the new method estimated t better than microaggregation. For the other parameters, there is no clear winner between the two methods, with microaggregation performing better than the new method for four out of the seven parameters.

	Protection method	n	b	g	t
Regression (1)	New Microaggregation	–26.8 **-9.7**	23.3 **11.7**		**-36.6** –45.9
Regression (2)	New Microaggregation	**-28.8** –38.5		**40.0** 47.5	**-39.1** –45.2
Regression (3)	New Microaggregation	–20.3 **-16.9**	7.6 **0.5**	**27.7** 31.3	**-33.2** –49.0

Table 1. Percentage errors for estimating the model parameters from data protected using the new method and microaggregation. **Bold** indicates better performance.

V. CONCLUSIONS

24. In this paper we have proposed a model based disclosure limitation method. We have illustrated our approach on data arising from the Community Innovation Survey of manufacturing and services sector enterprises. For each quantitative variable to be protected we build a simple regression model with fixed area effects. Our protection procedure is motivated by prediction intervals. For the smallest (largest) values the fitted values from the regression are altered by the addition of an inflation (deflation) factor that depends on the predictive standard errors. These values are then released. This inflation (deflation) is not applied to values that would already be inflated (deflated) if the fitted value itself were to be released. For the remaining middle values, the unaltered fitted values are released. In this way protection is increased for the most recognisable enterprises. We have also reviewed an approach to disclosure limitation based on microaggregation. The microaggregation procedure does not itself offer protection to the variable geographical area, whereas the new method suggests how to define broader categories for releasing this variable.

25. We also discuss the difficult task of assessing the performance of a disclosure limitation method. In particular, we consider how to quantify the protection offered and how to estimate the error induced by the method. We find that the new method offers more protection than microaggregation, and sometimes

leads to a smaller error.

Acknowledgements

This work was partially supported by the EU TMR network ERB-FMRX-CT96-0095 on 'Statistical and computational methods for the analysis of spatial data' and by the EU project IST-2000-25069 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 or the University of Plymouth.

References

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.

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.

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.