WP.9
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Manchester, United Kingdom, 17-19 December 2007)

Topic (i): Microdata

THE USE OF PROTECTED MICRODATA IN TABULATION: A CASE OF SDC-METHODS, MICROAGGREGATION AND PRAM

Supporting Paper

Prepared by Janika Konnu (Statistics Finland)

The use of protected micro data in tabulation: A case of SDC-methods, microaggregation and PRAM

Janika Konnu[*]

[*] Statistics Finland, P.O. Box 5V, FI-00022 Statistics Finland, Finland, janika.konnu@stat.fi

Abstract: Development on Statistical Disclosure Control (SDC) methods has been fast and focused on micro data. However, the main problem in statistical agencies is, how to produce safe tabular data. It is commonly known that the most suitable and efficient SDC method can be used only if the attributes of the micro data and its variables are properly taken into account. We chose to take deeper look in two of the methods, microaggregation and the Post RAndomization Method (PRAM). These methods were used to protect a personal data. In our study the main objective was to analyse micro data protection from the data user's point of view. We compiled several tables using both original and protected data in the process. One of the interests of the study was to see whether there are significant differences in some basic tables of frequencies - and when there is, which are the parameter values that lead us to acceptable differences.

1 Introduction

Statistical agencies have to take good care of the data they have collected. In case of register based data, it is easy to have access to all kind of data agencies need; but when it comes to surveys, it depends on respondents willingness, whether they give information needed or not. If respondents can trust in statistical agencies that their data will be used properly and there will be no risk of disclosure they are more willing to provide accurate information. In the case where respondents suspect their information is in risk of disclosure, it is only natural to refuse to answer or provide inaccurate information.

In section 2 we introduce the SDC methods and the data we were using in the study. The results are described in section 3 and the conclusion is in section 4.

2 Methods and data

We used two of the methods for statistical disclosure control. First one is microaggregation, which is normally applied to continuous variables, and the second one is the Post RAndomization Method. In the study we applied these SDC methods using software µ-Argus. Statistics Netherlands originally developed the software but upgrading the software was one of the main tasks in two European Union projects:

Computational Aspects of Statistical Confidentiality (CASC) and a CENtre of EXcellence for Statistical Disclosure Control (CENEX-SDC). As result the software has now more methods included and the software can be downloaded as a freeware from the projects’ web pages.

When data is protected using µ-Argus, it is important that metadata is specified very carefully. Software is using these specifications to estimate the disclosure risk. The actual protection is applied based on these estimates. Software has property of generating safe data when at the end of protection final suppressions are allowed. In our research we wanted analyse methods and no suppressions were allowed.

2.1 Microaggregation

SDC method called microaggregation is based on counting averages and releasing those instead of original values of a record. This method has been proposed over a decade ago and it is in use in many European countries and in Eurostat, but still its usability has been under discussion. Microaggregation is a method originally developed for continuos data, but as we will show, it is possible to modify it to be used in case of categorical data.

Microaggregation is one of the SDC methods available in software µ-Argus. In the software fixed size groups are formed using MDAV algorithm (Maximum Distance to Average Vector). This means that the average values for all variables in data are counted, and records are grouped using the difference from these averages. When MDAV algorithm is used, all the similar records in data form groups. This way it is possible to try minimising the information loss that releasing averages instead of actual values can entail.

There have been attempts to modify the method for categorical data. Thought trial and error, we noticed that it was possible already with the software available, if we change the codes of the categories so that they seem like continuos values.

2.2 The Post RAndomization Method

The Post RAndomization Method (PRAM) is an SDC method that is based on misclassification and it can be applied to categorical data only (de Wolf & van Gelder, 2004). In PRAM the values of the variables are changed based on a chosen probability distribution. It has been told that data protected by PRAM has to be analysed taking into account the PRAM matrix used in protection. If the matrix is forgotten, the results one gets might be far from ones that user would get using the original data.

We were interested in analysing the change that actually occurs when PRAM is applied. We wanted to see, whether it is possible to use PRAM in such a case, that researcher gets strongly protected data to plan the analysis. Then the final results are derived from original data by the staff of NSI. In Statistics Finland there is at least one example of this kind of procedure. Finnish Longitudinal Employer-Employee Data has so sensitive with information on both companies and their employees that researchers can have access to strongly protected parts of it.

2.3 Statistical methods

The purpose of the study was to get an idea how microaggregation or PRAM changes the properties of the data as they are used to protect it. Our interest was on the tables one forms using the protected data. How much the cell values change? If the values are changing, how large is the change for interpretation of the table?

2.4 Data

In our study of Statistical Disclosure Control methods we wanted to test proposed methods on some typical data that researchers want to have for their research. Detailed data on enterprises can only be studied in premises of Statistics Finland. In that case the restrictions and SDC methods have been thought quite carefully. In Statistics Finland most of the problems arise when researcher wants to have very detailed personal data and use it out of our premises. There is need for general guidelines how to protect this kind of data, and when the data is considered so sensitive that researcher can have access to in only in our research laboratory.

Data we use in our research contains information on teachers in Finland. As the main focus the study was on SDC methods, only part of this large data was used. Finally we decided that data containing high school teachers N=7798 fits for our purposes. We chose to protect identifying variables, even though if the data were to be used in actual research, it would have been easier to protect variables containing information on teachers' proficiency.

3 Results

Our study is only a beginning of wider research, where we try to analyse the use of statistical disclosure control methods proposed in literature. Data we used in this part of study is hard to handle in every way, but it gives very good idea how useful these methods are in practice.

Microaggregation was applied to categorical data using changes described in section 2. PRAM was applied in two different ways and thinking that protected data would be used without PRAM matrix. First we tested PRAM without any bandwidth restrictions. This lead to quite rough changes in frequencies even with small changing probabilities, so we tried to overcome this problem by choosing bandwidth of 2.

3.1 Results of microaggregation

In literature it is suggested that microaggregation is not strong enough to protect a data when it is applied one variable at the time, and if it is the only protection data has. Because of this, in our study microaggregation was applied to three variables at once. The variables we protected where the age, position of the teacher and the school level teacher teaches at. Here we see the changes in frequencies of teacher's school level. Later in this chapter you can see how the frequencies of this same variable, school level, changes when it was protected by PRAM.

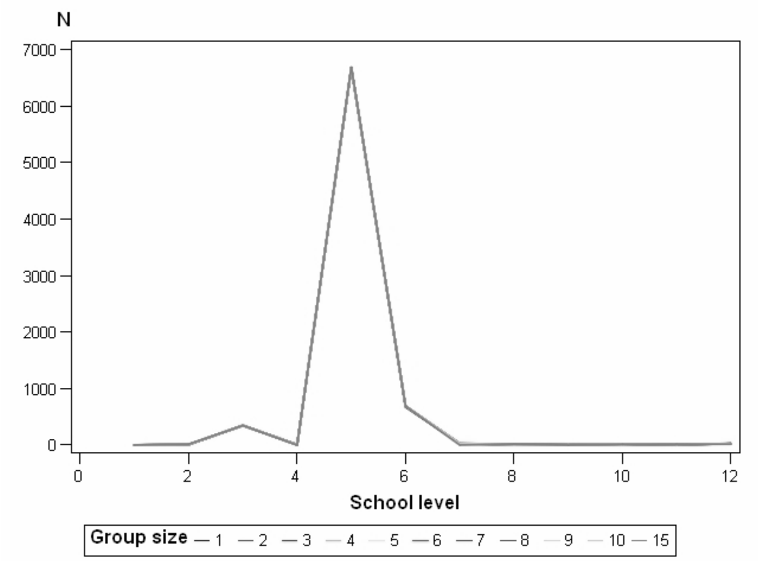

Fig 3.1 Changes in frequencies when data protected with microaggregation.

	1	2	3	4	5	6	7	8	9	10	15
1	1	0	0	0	0	0	0	0	0	0	0
2	10	8	9	8	10	6	7	8	9	10	15
3	343	346	345	344	345	348	350	344	351	340	345
4	7	6	6	4	5	0	0	8	0	10	0
5	6672	6672	6673	6678	6673	6676	6671	6670	6673	6668	6688
6	686	684	684	684	680	684	693	672	693	690	690
7	22	24	24	20	30	30	21	24	18	30	0
8	7	8	9	12	5	6	7	8	9	0	15
9	7	4	6	4	10	6	7	8	0	10	0
10	16	18	15	16	15	12	14	16	18	10	15
11	3	4	3	4	0	6	0	8	9	10	15
12	24	24	24	24	25	24	28	24	18	20	15

Table 3.1 Changes in frequencies when data protected with microaggregation.

Using microaggregation in case of categorical variable seems to have no significant effect on frequencies as seen in figure 4.1. When you think about this worrying only about information loss, it is very nice result. However, if data is protected by microaggregation only this leads to problem with disclosure risk. It is obvious that there are only few changes in the values of a record, and so this procedure brings hardly any uncertainty when it comes to identification of a record. Changes in frequencies are so small that they are hard to see from figure 4.1. To get better idea the actual values can be found in table 4.1.

3.2 Results of PRAM

As mentioned before, our data wasn't nice in any way, and it was expected that our SDC methods would fail in some ways. In this case, one must definitely question whether PRAM should be used to this data at all. There are more than 6500 records in one of the categories in our data, and then there are categories that have nearly none observations. This lead to situation where our distribution tends to smooth when protection is applied as demonstrated in figure 4.2. Actual values can be found in table 4.2 in case you are interested to see in detail how those small frequencies tend to increase.

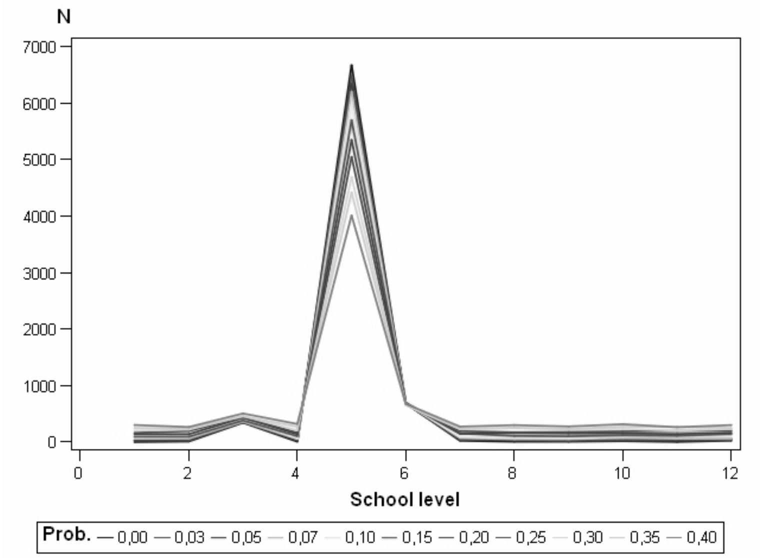

Fig 3.2 Changes in frequencies when data protected with PRAM and no bandwidth was used.

If we look at changing probability of 0,10 or more, it is clear that the results from this data won't coincide with the ones from the original data unless probability matrix is taken into account. But if we are interested using PRAM for some kind of demonstration data, changes that occur with 0,10 probability could be acceptable.

	0,00	0,03	0,05	0,07	0,10	0,15	0,20	0,25	0,30	0,35	0,40
1	1	22	30	59	74	94	142	164	200	246	301
2	10	25	45	61	79	92	139	194	221	241	265
3	343	357	361	359	395	382	430	437	459	455	505
4	7	28	37	57	61	104	140	165	206	272	319
5	6672	6487	6347	6202	5978	5701	5350	5049	4693	4418	4011
6	686	679	678	693	684	697	684	663	648	671	674
7	22	39	61	72	89	150	145	192	228	232	272
8	7	36	33	49	77	108	158	172	250	247	299
9	7	24	44	51	71	101	152	177	229	231	275
10	16	36	64	73	89	119	152	188	233	272	315
11	3	27	43	60	99	110	142	196	201	266	262
12	24	38	55	62	102	140	164	201	230	247	300

Table 3.2 Changes in frequencies when data protected with PRAM and no bandwidth used.

When we got these results, we were a little disappointed with this smoothing effect on distribution. We decided to use bandwidth of 2 and were expecting it to help. Unfortunately protection of our data, were the frequencies differ so much, didn't get any better with this either. Restricting the change has effect when it comes to categories that aren't next to category 5 that most of the cases fall in. Categories next to 5 had the same remarkable increase as in case without any bandwidth as one can see from figure 4.3 and table 4.3.

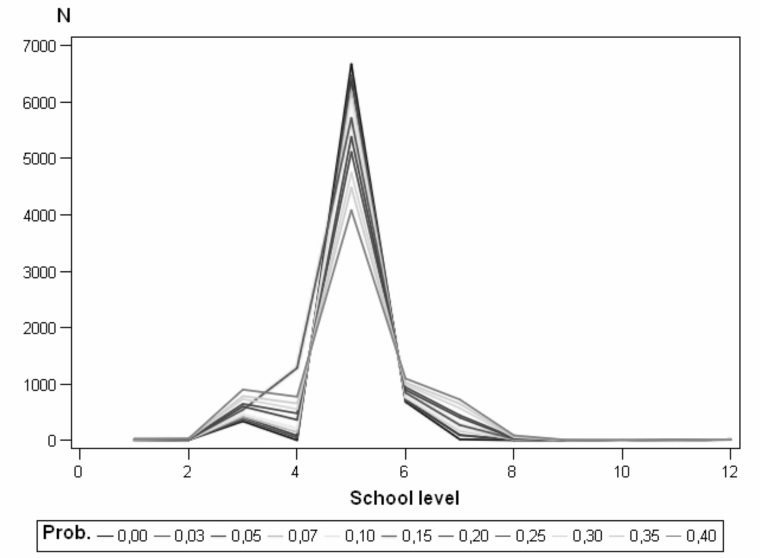

Fig 3.3 Changes in frequencies when data protected with PRAM and bandwidth was 2.

As one can see in figures 4.2 and 4.3, it depends not only on the data, but also on the attributes how well an SDC method performs. If the protector of the data gets carried away and forgets to check the information contend in data, the results from the protected data might differ greatly from the ones from original data.

1	1	2	3	5	9	18	25	26	31	28	29
2	10	15	12	14	20	23	19	26	41	35	36
3	343	376	417	431	469	547	605	651	735	789	904
4	7	50	88	150	218	1287	368	483	559	660	778
5	6672	6475	6377	6182	6047	5717	5382	5114	4748	4485	4083
6	686	711	736	763	767	849	899	946	995	1031	1100
7	22	97	99	175	194	276	408	449	575	654	728
8	7	13	15	27	25	30	40	54	61	68	92
9	7	6	8	8	7	8	8	3	10	10	9
10	16	17	16	17	17	18	16	20	19	17	11
11	3	4	3	4	4	7	4	6	5	4	13
12	24	22	24	22	21	18	24	20	19	17	15

Table 3.3 Changes in frequencies when data protected with PRAM and bandwidth was 2.

We concentrated on usefulness of data, so reader must consider briefly whether the proposed values of the attributes yield to situation when all the individuals are protected against disclosure. It is also good to keep in mind, that since PRAM is based on a probability distribution, the protected data user gets is different in every protecting. This means that our results are an example how this method performs, not as actual truth of its usefulness.

4 Conclusion and future study

This part of research was only the beginning of more extensive research. At this moment it has no actual measures of information loss or identification risks. This is severe weakness but we have made some assumptions for measures of identification risk and information loss. In our case, we were more interested in general usability than some figures.

It is clear that when data is protected using PRAM, researcher have to use PRAM matrix in order to have correct results. However, there aren’t so many researchers that are willing to do some extra work because of protection. And then there are some researchers that aren’t even capable to do this. In our opinion PRAM is quite promising method when we want add some uncertainty for identification. It is possible to see that PRAM can be used it is, if data protector is working with some kind of demonstrative data. This type of data can be given to researcher in case we can’t allow researcher to have access to data even in NSI’s premises. This method can have potential in future when researchers are more familiar with statistics and mathematics.

We used microaggregation for categorical data even if it was supposed to be used to continuous data only. That led us to some problems, but it is the same case in reality when you only have limited options to choose from. In our opinion the usefulness of microaggregation lies on numerical data. Still we can see some possibilities using microaggregation for categorical data too, but in that case data must have some other protection applied.

References

A CENtre of EXellence for Statistical Disclosure Control. http://neon.vb.cbs.nl/cenex/

Computational Aspects of Statistical Confidentiality. http://neon.vb.cbs.nl/casc/

de Wolf, Peter-Paul, van Gelder, Ilan (2004). An empirical evaluation of PRAM. Discussion paper 04012. Statistics Netherlands, Voorburg/Heerlen.

Domingo-Ferrer, J. & Torra, V. (2001). A Quantitative Comparison of Disclosure Control Methods for Microdata. In Doyle, P., Lane, J., Theeuwes, J., Zayatz, L., eds.: *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies.* Elsevier, North-Holland. 111-133.

Gouweleeuw, J., Kooiman, P., Willenborg, L., and de Wolf, P. (1998). Post Randomisation for Statistical Disclosure Control: Theory and Implementation. *Journal of Official Statistics.* Vol. 14, No.4, 463-478.