

**WP. 34**  
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**  
(Bilbao, Spain, 2-4 December 2009)

Topic (vi): Case studies

**MICRODATA PERTURBATION PRESERVING MULTIVARIATE RELATIONS OF  
TARGET CATEGORICAL VARIABLES IN THE  
HOUSEHOLD ENVIRONMENTAL SURVEY OF EUSTAT**

**Invited Paper**

Prepared by Marta Mas and Anjeles Iztueta-Azkue (Basque Statistics Office), Spain

# Microdata perturbation preserving multivariate relations of target categorical variables in the Household Environmental Survey of Eustat

Marta Mas\* and Anjeles Iztueta-Azkue\*

\*Basque Statistics Office. Vitoria-Gasteiz, Basque Country (SPAIN).  
[marta\\_mas@terra.es](mailto:marta_mas@terra.es); [aiztueta@eustat.es](mailto:aiztueta@eustat.es)

**Abstract:** Advanced and versatile data dissemination formats have been developed over recent years in order to meet user requirements. An increasing number of expert users wish to perform their own data analysis using data banks and original microdata. However, statistics offices should provide safe releases of its dissemination products and, in particular, protect Public Use Files (PUFs) against any indirect identification and record linkages with external files. Microdata protection techniques applied by the Basque Statistics Office (EUSTAT) have traditionally consisted of aggregating categories in the most identifying variables and limiting geographical levels. This paper describes a new approach for applying data perturbation techniques to categorical variables of the Household Environmental Survey conducted by EUSTAT. The goal consists of providing good protection while preserving microdata structure and variable dependencies. Multivariate techniques have first been applied to check variable dependencies and to assist in the choice of variables or combinations to be perturbed. Two perturbation techniques have then been tested: the Invariant PRAM (Post-Randomisation Method) implemented in Mu-Argus software, and a Random Assign function, included in the SAS package, that preserves statistical distributions and allows for treating the structural zeroes of the survey. The paper ends with some conclusions which may prove useful for future microdata perturbation approaches.

## 1 Introduction

Advanced and versatile data dissemination formats have been developed over recent years in order to meet user requirements. An increasing number of experts wish to perform their own data analysis using data banks and original microdata. However, statistics office should provide safe releases of their dissemination products and, in particular, protect Public Use Files (PUFs) against any indirect identification and record linkages with external files.

Microdata protection techniques applied by the Basque Statistics Office (EUSTAT), have traditionally consisted of aggregating categories in the most identifying variables and limiting geographical levels. This paper describes a new approach for applying data perturbation techniques to microfiles for general release. Specifically, we have focused on the categorical variables of the Household Environmental Survey conducted by EUSTAT. The goal is to provide good protection while preserving microdata structure, variable dependencies and statistical properties.

## **2 Microdata description**

The Household Environmental Survey was first conducted in 2008 and collects information on the practices and uses of Basque families in relation to protecting the environment. Aspects, such as household equipment, recycling practices, environmental protection attitudes or mobility, are stored in more than 300 variables. The sample consists of 5,324 households and the microfile includes information on families and individuals (one person per household).

The considerable size of this survey led us to make an initial selection of variables to be included in the study. We decided to focus on the variables referring to individuals. Even though it is not the widest part of the survey, it does provide valuable information about personal attitudes, environmental awareness and mobility. The main thematic blocks addressed by the individual part of the survey are the following:

- Water saving uses
- Mobility and transportation
- Environmental awareness
- Environmental protection activities
- Environmental protection measures
- Socio-demographic characteristics

A total of 36 target variables and 10 socio-demographic characteristics will be included in the microfile of individuals to be released.

Standard aggregations of identifying variables, mainly socio-demographic and geographical indicators, will be applied in order to avoid re-identifications. However, additional protection by means of perturbation of several target variables will be also studied on this occasion.

## **3 Analysis of variable dependencies**

Prior to perturbation, a decision should be taken regarding the variables or combinations of variables to be perturbed. Variable dependencies throughout the file could be seriously affected by perturbation, by generating inconsistencies and breaking relation structure in the file. Therefore, multivariate analysis will be performed to check the strongest relations among target variables and, thus, allow them to be taken into account when applying perturbation techniques.

### 3.1 Bi-variant analysis: independence tests

Socio-demographic characteristics and also variables with structural zeroes<sup>1</sup> were excluded from this first test in order to reduce the dimensions of the problem. However, these variables will be considered in further steps in the dependence analysis and also in the perturbation step. Therefore, a total of 19 target variables, all of which were categorical and almost all dichotomous (yes-no type), were selected for the analysis.

In a first approach, the Chi-square independence tests were performed on all possible 2-dimensional contingency tables of selected variables. Some association measures were also calculated to check the degree of dependence between them.

Table	Chi-square observed - values	Phi-Pearson $\sqrt{\chi^2 / n}$	Q-Yules (-1 to 1)	Odds-Ratio (0 - $\infty$ )
Activity1 x Activity2	296,549.01	0.39	0.94	29.97
Activity3 x Activity4	378,111.71	0.44	0.93	27.36
Activity3 x Activity5	115,974.51	0.25	0.84	11.56
Activity1 x Activity3	114,237.63	0.24	0.78	8.20
Measure3 x Measure5	171,388.46	0.30	0.68	5.18
Measure4 x Measure5	225,813.06	0.34	0.66	4.93
Measure2 x Measure3	221,437.98	0.34	0.65	4.68
Measure6 x Measure8	128,848.23	0.26	0.65	4.68
Measure1 x Measure2	177,200.22	0.30	0.63	4.35
Measure3 x Measure4	202,363.55	0.32	0.62	4.28
Measure2 x Measure4	156,726.66	0.29	0.60	3.98
Measure1 x Measure3	153,246.07	0.28	0.55	3.45
Measure5 x Measure7	119,515.97	0.25	0.54	3.36

**Table 3.1** Highest Chi-square estimates and associate measures for 2-dim contingency tables.

Although almost all the independent tests for 2-dim contingency tables refused the null hypothesis of independence, the highest values in a “chi-square sense” show a stronger dependence between variables pertaining to the same thematic block. Looking at the association measures, the Q-Yules coefficient, which measures the ratio of concordant pairs of observations in 2x2 tables, is significantly high in tables that combine variables related to participation (or not) in environmental protection activities (activity1, activity2, etc.).

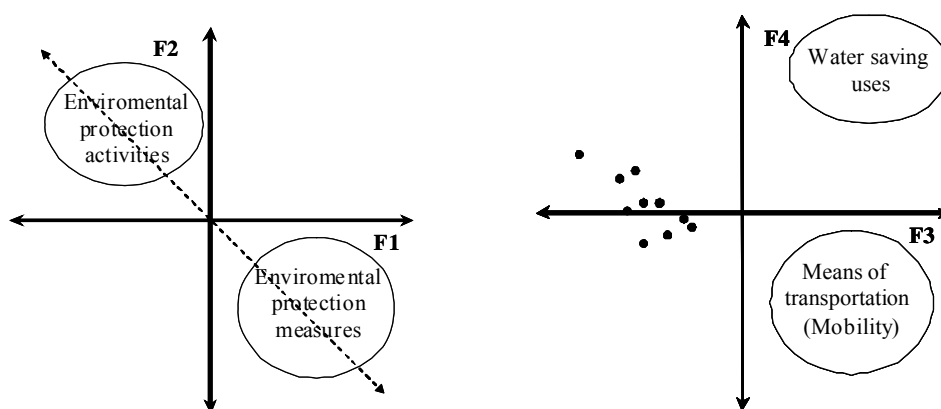
---

<sup>1</sup> Structural zeroes represent null values generated by the very structure of the questionnaire, i.e. it only makes sense to put some questions to a group of individuals meeting certain conditions.

## 3.2 Multi-variant analysis

### 3.2.1 Correspondence Analysis of target variables

The results of a multiple correspondence analysis performed on all variables considered in section 3.1 would confirm the tight relations between variables that already been perceived in the previous analysis. Summary diagrams of the first four factors obtained and the description of axes are given below:



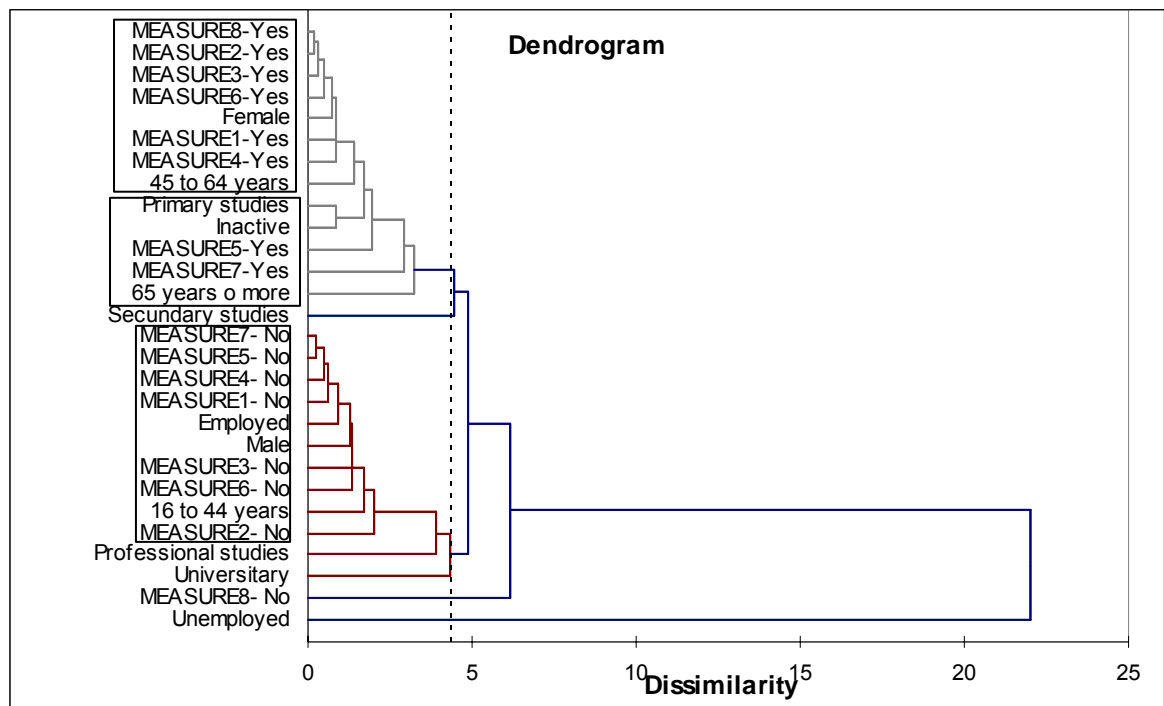
**Fig 3.2** Factorial diagrams F1-F2, F3-F4 for the multiple correspondence analysis and descriptions of axes.

Looking at contributions to the axes and at the indicators of the quality of representation, we can conclude that each factorial axis describes one thematic block of variables. These clear dependences within each topic of the survey will enable us to study relations in each thematic block in greater detail. In addition, socio-demographic characteristics will be added to such a study in order to look for specific population groups (if any) linked to certain environmental behaviours.

### 3.2.2 Correspondence Analysis and Hierarchical clustering by thematic blocks

Multiple Correspondence Analysis will be repeated, but separately for each block of variables. On this occasion, socio-demographic characteristics (age, sex, level of education and relation to activity) are actively involved in the analysis. Finally, a clustering method for each topic is applied using the information provided for all the factorial factors, in order to detect groups of individuals with similar environmental opinions or attitudes.

At the end of this stage, only the block of variables related to different environmental protection measures seems to define significant socio-demographic groups. The dendrogram produced by the hierarchical clustering method used is shown below:



**Fig 3.3** Dendrogram of hierarchical classification – Groups of variables: Measures to protect the environment.

Two male-female clusters discriminate the population with respect to certain environmental protection measures. Middle-age females seem to agree with measures related to recycling and renewable energies (measure1 and measure6) while young males do not agree with measures that increase taxes on fuel or restrict the use of private transport (measure3 and measure4). On further steps in the hierarchical structure, the 65 or over age group get closer to measures like paying more for renewable energies (measure5) or for an environmental tax to be imposed on tourism (measure7).

Variables sex and age pay an important role when defining groups with different opinions about environmental protection. Therefore, these clear relations should be taken into account when perturbing any of the variables of this block.

#### **4 Choice of variables, methods and parameters.**

As a result of the dependence analysis, some decisions on the variables to be perturbed in each thematic block should be taken. In addition, other aspects such as sensitivity of the variables or the structural zeroes will also influence the final decision.

#### 4.1 Sensitive variables

As we have shown in previous sections, variables included in the thematic block of environmental protection activities are strongly related. In both dependence studies, jointly with all the target variables and by thematic blocks, these variables define a specific group of individuals that actively takes part in environmental protection activities, belongs to environmental organizations or participates in demonstrations for environmental reasons. These activities might frequently be linked to political trends, which is considered to be a sensitive topic in our context. Accordingly, this group of variables will be discarded and not released.

#### 4.2 Compounded variables and Invariant PRAM

The area of environmental protection measures is another large block of closely-related variables. Furthermore, as we have seen in section 3.3, these variables, along with socio-demographic characteristics, define homogeneous population groups, which are for or against applying certain measures to protect the environment. These relations should be taken into account during the perturbation process in order to avoid over-distorting the dependence structure in the file.

Perturbing a compounded variable whose categories are all the possible combinations of the categories of the related variables is a way to keep these relations invariant. During the perturbation process, one “super-category” might be changed by another, but without affecting the internal dependences.

Post Randomisation Method (PRAM), see Gouweleeuw et al. (1998a and 1998b) is a method that is able to deal with this type of variables. This perturbation technique for categorical variables allows the data protector to assign probabilities of change to each variable category in order to misclassify them deliberately. However, the transition probabilities should be released along with the microfile in order to get unbiased estimates from the perturbed microfile. This is not a desirable option, neither for us nor for our users, at present.

However, a variant of the method, the Invariant PRAM, allows unbiased estimation of original distributions without the need of a probability matrix. Therefore, the probability of changing a category  $k$  into another  $l$  ( $p_{kl}$ ) should be computed as follows:

$$p_{kl} = \begin{cases} 1 - (\theta T_x(K)/T_x(k)) & \text{if } l=k \text{ (probability of score } k \text{ to be left unchanged)} \\ (\theta T_x(K)/(K-1)T_x(k)) & \text{if } l \neq k \text{ (probability for score } k \text{ to be changed into score } l) \end{cases} \quad (4.1)$$

Where  $0 < \theta < 1$  and

$T_x(k)$  = Total number of records for which variable  $X = k$

$K$  = Number of categories of variable  $X$

$T_x(K)$  = Total number of records for which variable X = K assuming that  $T_x(k) \geq T_x(K) \geq 0$  and  $k=1, \dots, K$

Having decided on the method and the parameters, we apply Invariant PRAM for the compounded variable: SEX x AGE (groups) x MEASURE2 x MEASURE3. In order to check the “best”  $\theta$ , we will check several probabilities patterns computed as in (4.1). Nevertheless, knowing the variable distribution and fixing a value for  $\theta$ , we can compute the expected number of changes of scores for variable X (see Gouweleeuw et al., 1998a and 1998b):

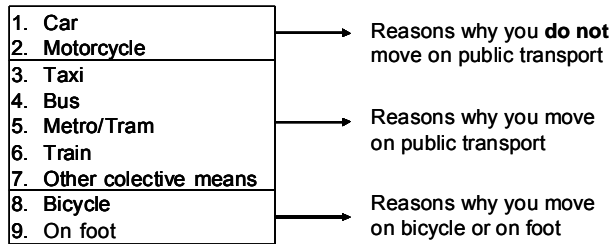
<b>Invariant PRAM Estimates</b> <i>SEX x AGE (groups) x MEASURE2 x MEASURE3</i>		
	Expected number of changes	Percentage of change
( $\theta=0.1$ )	122.4	2.3%
( $\theta=0.2$ )	244.8	4.6%
( $\theta=0.3$ )	367.2	6.9%
( $\theta=0.4$ )	489.6	9.2%
( $\theta=0.5$ )	612.0	11.5%
( $\theta=0.6$ )	734.4	13.8%
( $\theta=0.7$ )	856.8	16.1%
( $\theta=0.8$ )	979.2	18.4%
( $\theta=0.9$ )	1.101.6	20.7%

Given these estimates, a first decision on  $\theta$  's value could be taken before perturbation. In our case, we will make a conservative choice, as it is the first perturbation approach at EUSTAT and we really do not want to overly distort data. Several checks for  $\theta=0.1$  and  $\theta=0.2$  have been performed. Definitive results and some comments on the method are given in section 5.

### 4.3 Structural zeroes treatment

How to deal with the structural zeroes of the survey is an important aspect to consider before perturbation. If we change a score of a variable which generates a lot of structural zeroes, it is likely to get a considerable amount of inconsistencies in the perturbed file, as a result. This is the case of the survey variable TRANS (mean of transportation frequently used) which generates a variety of structural zeroes depending on its values:

Means of transportation – Questionnaire structure



**Fig 4.2** Questionnaire structure for variable TRANS

Perturbing within groups of categories which derive the same zeroes structure, would be a possible solution to disturb the variable TRANS, preserving the structure of the questionnaire. That means, in this case, that category “1.Car” could only be replaced by “2.motorcycle”, or category “8.bicycle” only by “9.on foot”, and so on.

A random assign function, included in the SAS statistical package, will be used to perturb TRANS variable. This function generates random values for a discrete variable given a probability distribution. The syntax for this function is as follows:

$$\text{RAND}(\text{'table'}, p_1, p_2, \dots, p_n) \text{ where 'table' is a fix parameter and } \sum_{i=1}^n p_i \leq 1$$

The function is applied to each group of categories and probabilities are estimated by the frequency distribution in the group. At the end of the process, some scores have been perturbed, but preserving the distribution of the categories in the group.

It is also possible to apply PRAM in this case but not as it is implemented in Mu-Argus 4.2 (only groups of categories of the same size are allowed). Applying PRAM to a compounded variable which includes all the zeroes structure, would be an alternative. This might be feasible if there were not too many variables with structural zeroes, otherwise such a variable could be intractable.

## 5 Results.

Some of the results derived from the application of the perturbation methods explained in preceding sections are then shown. The observed results of applying PRAM to a compounded variable for several probability patterns are given in the table below. Perturbation impact in each variable separately is also included:

	$(\theta=0.1)$		$(\theta=0.2)$	
	Observed number of changes	Percentage of change	Observed number of changes	Percentage of change
SEX	68	1.28%	114	2.14%
AGE	80	1.50%	161	3.02%
MEASURE2	57	1.07%	123	2.31%
MEASURE3	56	1.05%	118	2.22%
COMPOUNDED	115	2.16%	225	4.23%

**Table 5.1** Invariant PRAM results for compounded variable *sex x age x measure2 x measure3*

Perturbation results for TRANS variable using random assign function of SAS, provide much larger perturbation (around 18 %). However, changes occur only within groups of categories which are similar in two senses: they generate the same structural zeroes and represent homogeneous concepts (private transport, public transport and others). Therefore, these two aspects will not be affected by perturbation. Changes observed in each group of categories are given below:

Groups of categories	Number of changes	Percentage of change
Private transport (1.car, 2. motorcycle)	99	4.67%
Public transport (3.taxi, 4.bus, 5.tram/metro, 6. train, 7.other colective)	750	51.80%
Bicycle/On foot (8. Bicycle, 9. On foot)	140	7.97%
Total variable TRANS	989	18.58%

**Table 5.2** Random assign function results for variable TRANS

Comparing both methods and the obtained results, some conclusions are derived. Invariant PRAM not only enables us to get unbiased estimates of a perturbed variable but it also provides a parameter ( $\theta$  's values) to control the degree of perturbation. Nevertheless, it is difficult to deal with a large number of variables when they should be perturbed simultaneously.

On the contrary, SAS programming facilitates control of ranges of categories in which the random assign function is applied, and provides a solution to treat complex zero-structures. However, the percentage of perturbation cannot be controlled by the data protector as it has been applied in this study, and depends directly on variable distribution.

## 6 Conclusions and future work.

Prior to perturbation, several dependence analysis have been performed on categorical variables of the Household Environmental Survey. Different perturbing solutions have been proposed in order to preserve variable relations detected. This work represents a first approach for data perturbation at EUSTAT. Several microfiles will be ready for public use next year and the integration of perturbing techniques is an option to be considered, along with the traditional aggregation of categories.

However, some relevant aspects should be taken into account for future approaches. Questions such as, what we consider a good level of perturbation in terms of quality and safety, or how perturbation affects to weighted frequencies, should be answered in order to properly apply perturbation techniques.

## References

- Agresti A. (1990). *Categorical data analysis*. John Wiley & Sons, New York.
- Gentle, J.E. (1998) *Random Number generation and Monte Carlo Methods*. Springer-Verlag: New York
- Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J. & de Wolf, P.-P., (1998). *Post Randomisation for Statistical Disclosure Control*. Journal of Official Statistics, Vol.14, No. 4, pp. 463-478.
- Herman J. (1986). *Analyse de données qualitatives. 1. Traitement d'enquêtes, échantillons, répartition, associations*. Masson, Paris.
- Hundepool, A. et al.(2008) *Mu-Argus 4.2 User's Manual*. Statistics Netherlands.
- Lebart, L., Morineau, A, & Piron, M., (2000) *Statistique exploratoire multidimensionnelle*. Dunod. (3ème ed.) Paris.
- Lebart, L., Morineau, A. & Warwick, K.M. (1984) *Multivariate Descriptive Statistical Analysis. Correspondence Analysis and Related Techniques for Large Matrices*. John Wiley & Sons.
- de Wolf, P.-P., Gouweleeuw, J., Kooiman, P., and Willenborg, L. (1998). *Reflections on PRAM*. In *Statistical Data Protection, Luxembourg*, pp. 337–349. Office for Official Publications of the European Communities.