

**Statistical Disclosure Control of Microdata at  
Statistics Netherlands**

**Ardo van den Hout**

**Statistic Netherlands ([www.cbs.nl](http://www.cbs.nl))  
Utrecht University**

**Contents:**

- 1. Statistical Disclosure Control**
- 2. Microdata**
- 3. Global Recoding and Local Suppression**
- 4. Post Randomisation Method**
- 5. Conclusion**

# **1. Statistical Disclosure Control (SDC)**

## **Objective:**

**Protecting the privacy of respondents when data are released to external users.**

## **Procedure:**

- 1. Detecting unsafe information in data.**
- 2. Processing the data to make the data safe.**

**Methods for *microdata* differ from methods for *tabular data*.**

## **2. Microdata**

**Series of records, each containing information on an individual unit (≡ data matrix).**

**(Tabular data: aggregated microdata)**

### **Availability of microdata:**

- 1. On site: researchers come to Statistic Netherlands.**
- 2. Remote access: using the internet.**
- 3. Release under contract: data are made safe against spontaneous recognition.**

## Making microdata safe

Identifying variables: variables that can be used to identify a respondent.

Direct identifiers (name, address etc.) are deleted from the microdata.

Indirect identifiers.

### Example:

Variables: Place of Residence,  
Gender,  
Profession.

In the microdata one combination of the scores:

*Volendam × female × notary*

### **3. Global Recoding and Local Suppression**

**Currently used at Statistics Netherlands  
(Software:  $\mu$ -Argus)**

#### **Global recoding**

**Replace *Volendam* by the name of the district:**

*Noord-Holland* × *female* × *notary*

#### **Local Suppression**

**Replace *Volendam* by a missing:**

× *female* × *notary*

**Current practice: first recoding, then suppression**

#### **4. Post Randomisation Method (PRAM)**

**PRAM is applied to microdata**

**PRAM concerns the identifying variables  
in unsafe combinations**

**Description:**

**Conditional on the original scores of the variable, the scores are re-classified using a prescribed probability mechanism that allows of misclassification.**

**The probability mechanism is provided along with the released microdata.**

**References:**

**Kooiman et al. (1997), Warner (1965),  
and Rosenberg (1979)**

## Example of PRAM

**A file with 100 respondents, several variables, and the binary variable  $A$ .**

**Assume, there is exactly one respondent with score  $a = 2$ .**

**Misclassification on purpose:**

$a = 1$	$\rightarrow$	$a^* = 1$ :	<b>90%</b>
$a = 1$	$\rightarrow$	$a^* = 2$ :	<b>10%</b>
$a = 2$	$\rightarrow$	$a^* = 2$ :	<b>80%</b>
$a = 2$	$\rightarrow$	$a^* = 1$ :	<b>20%.</b>

**The PRAM matrix  $P = (p_{kl})$  where  $p_{kl} = P(A^* = l | A = k)$  is given by:**

$$P = \begin{pmatrix} 9/10 & 1/10 \\ 2/10 & 8/10 \end{pmatrix}.$$

**After applying PRAM, the file is safe  
(in a certain sense):**

$$\mathbf{P (A = 2 | A^* = 2) \cong 0.075.}$$

**The identity of the respondent with score 2  
is protected by possible misclassification of  
the score 2 as score 1 *and* by possible  
misclassification of scores 1 as scores 2.**

**In general: protection by “outflow” en “inflow”.**

## **Analysis: estimation of frequencies**

**Let**

$$\mathbf{T} = \begin{pmatrix} \mathbf{T}_1 \\ \mathbf{T}_2 \end{pmatrix} = \begin{pmatrix} \text{\# scores with value 1} \\ \text{\# scores with value 2} \end{pmatrix}$$

**and**

$$\mathbf{T}^* = \begin{pmatrix} \text{\# scores with value 1 after PRAM} \\ \text{\# scores with value 2 after PRAM} \end{pmatrix}.$$

**Then**

$$\mathbf{E}[\mathbf{T}_1^*] = p_{11}\mathbf{T}_1 + p_{21}\mathbf{T}_2$$

$$\mathbf{E}[\mathbf{T}_2^*] = p_{12}\mathbf{T}_1 + p_{22}\mathbf{T}_2.$$

**In matrix notation:**

$$\mathbf{E}[\mathbf{T}^*] = (\mathbf{P})^t \mathbf{T}.$$

**The equality**

$$E [T^*] = (\mathbf{P})^t T,$$

**provides a unbiased *moment estimator*:**

$$\hat{T} = (\mathbf{P}^{-1})^t T^*.$$

**(Kooiman et al. , 1997)**

**Variances can be computed.**

**The estimation of tables is possible with this method**

## **5. Conclusion**

### **Regarding PRAM:**

- Complex data analysis demands complex adjustment.**
- It is statistically sound.**
- It is an alternative to recoding and suppression (not a substitution).**
- More detail, but same safety level.**