Development of Micro-Aggregation System for Confidential Data

On the behalf of EUROSTAT, under SUP-COM programme

L. Béchet, T. Le Tri, Dr P. Loosveldt

Duration: 1 year
Micro-aggregation system for confidential data

I. PRESENTATION OF EUROSTAT

II. PRESENTATION OF ARIANE II

III. PRELIMINARY STUDY

IV. SOFTWARE

V. FURTHER DEVELOPMENTS
Micro-aggregation system for confidential data

I. PRESENTATION OF EUROSTAT
The Statistical Office of the European Union

Theme 1: General statistics
Theme 2: Economy and finance
Theme 3: Population and social conditions
Theme 4: Energy and industry
Theme 5: Agriculture, forestry and fisheries
Theme 6: External trade
Theme 7: Transport
Theme 8: Environment
Theme 9: Research and development
II. PRESENTATION OF ARIANE II

- European approach
- Skills centers
- Our market
- SAS applications
PRESENTATION OF ARIANE II

A European approach

Employees: >550
Turnover in 1998: 1.2 Billion Bef
Turnover in 1999: 1.8 Billion Bef
PRESENTATION OF ARIANE II

Skills centers

Information System Analysis and development
Client/Server applications (Oracle, Powerbuilder, visual basic, etc.)
Internet/Intranet (Silverstream, WebPB, Java, etc.)
3 LG (Cobol, Natural, C, etc.)

System and network Administration and find-user support

Methodology and computer indeed software engineering

Datawarehouse and Statistical analysis (Oracle Express, Business Object, SAS)
PRESENTATION OF ARIANE II

Our Market...

- INDUSTRY
- BANKING AND INSURANCE COMPANIES
- COLLECTIVE ADMINISTRATIONS
- INTERMEDIATE PARTNERS
- SERVICES
- DISTRIBUTION
- TECHNICAL PARTNERS
PRESENTATION OF ARIANE II

SAS Applications...

- STATISTICAL MODELS AND CLINICAL TRIALS
- SALES AND MARKETING
- DATA WAREHOUSING AND INFOCENTER
- DATAMINING AND EXECUTIVE INFORMATION SYSTEM
- MANAGEMENT APPLICATION
- QUALITY CONTROL
- ORGANIZATION OF TRAINING
III. PRELIMINARY STUDY

- Objectives of micro-aggregation
- What is micro-aggregation?
- Example
- Micro-aggregation methods
- Evaluation criteria
Objectives of the project/working program

- Review experience in the use of different approaches to micro-aggregation
- Development of algorithms to implement micro-aggregation techniques
- Develop a computer program to apply micro-aggregation methods

- Inventory of micro-aggregation methods
- Definition of evaluation criteria
- Studies, results and existing software
- Simulations on data
- Software development
What is micro-aggregation?

- Protecting microdata sets from disclosure by intruders
- By means of statistical methods

\[ \Rightarrow \text{Perturbation methods} \]

2 objectives:
- Maximisation of confidentiality
- Maintenance of data structure
What is micro-aggregation?

- Replacing the values of the observations by values of small groups of units
- Obtaining the most homogeneous units
  - ranking methods
  - aggregation methods
Example: single variable method

1) Choice of a variable

2) Ranking of observations according to this variable

3) Forming small groups of “k” elements

4) Calculating mean of these groups

5) Replacing data by sub-groups mean
Micro-aggregation methods

- **Numeric variables**
  - single axis
    - Single variable
    - First principal component
    - Sum of scores
  - individual methods
    - Individual ranking
    - Weighted moving average
  - partitioning techniques
    - Ward’s method
    - Hanani

- **Nominal and ordinal variables**
  - heuristics
  - entropy
  - simple sorts
Evaluation criteria

1. Maintenance of confidentiality
2. Respect of data structure
3. Performance of software

Corresponding to objectives
Confidentiality criteria

1. Value of threshold
2. Concentration rule
3. Uniqueness
4. Data perturbation
Criteria of data quality

Information loss

Comparison of statistics

Further processing ability

Numeric variables (dist)

Nominal variables (entropy)

Position

Dispersion

Relation between variables

for econometric models

for factorial data analyses
Conclusions

- Theoretic issues
- Available results and studies
- Simulations of methods applied to actual data

Recommendations about features of the software
Micro-aggregation system for confidential data

IV. SOFTWARE

- Characteristics of the software
- Software approach
- Process flow
- Screens
- Choice of SAS system
Characteristics of the software

- Modularity
- Interactivity
- Adaptation to non-specialists
- Development in SAS system
- Various micro-aggregation methods
- Decision-making system for data protection
- Mass production / Batch treatments
Software approach

Dataset

Dataset definition → Extrema definition → Statistics

Dataset selection

Dataset generation → Model application

Micro-aggregation: methods and parameters → Indicators

Dataset management

Micro-aggregation

Decision making
“Micro-aggregation session”

1. Create a working data set
2. Declare the working data set
3. Extrema definition
4. Apply micro-aggregation method
5. Study the results-Synthesis
6. Generate final data set
7. Application of a model

Process flow
Main frame

Dataset Definition

Dataset Selection

Test Library
Depart02

μA

Dataset Generation

Model Application

Exit

μAggregation
Dataset definition
Dataset definition

- Assign type of variables
- Characterise variables (identifier)
- Select/drop variables
- Table view
- Graph view
- Statistics for
  - numeric variables
  - categorical variables
Extreme values treatment

- "n" times $\sigma$ criterion

T: total dataset, E: extreme values, N: "normal" values
Extrema definition

For a numeric variable
Definition of extreme elements can combine criteria about several variables (numeric/categorical).

- for numeric variables: selection of observations according to variation to standard deviation
- for categorical variables: display of frequency of categories
Micro-aggregation methods

Ex: First principal component
Micro-aggregation methods

- Implementation of various micro-aggregation methods
  - For numeric/categorical variables
- For each method, the software proposes several parameters.
- After computation of the method, statistics and indicators are calculated for comparison with initial data.
### Decision-making: indicators

<table>
<thead>
<tr>
<th>STAT</th>
<th>MEAN</th>
<th>MEMO</th>
<th>TEST</th>
<th>TESTFCP</th>
<th>BBBBBBBB</th>
<th>TESTFCP1</th>
<th>TESTFCP2</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>2000.00</td>
<td>2000.00</td>
<td>2000.00</td>
<td>2000.00</td>
<td>2000.00</td>
<td>2000.00</td>
<td></td>
</tr>
<tr>
<td>NMSS</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>NUNIQUE</td>
<td>1927.00</td>
<td>664.00</td>
<td>664.00</td>
<td>308.00</td>
<td>308.00</td>
<td>308.00</td>
<td></td>
</tr>
<tr>
<td>MIN</td>
<td>-8631261.00</td>
<td>-523324.87</td>
<td>-523324.87</td>
<td>-523324.87</td>
<td>-523324.87</td>
<td>-523324.87</td>
<td></td>
</tr>
<tr>
<td>MAX</td>
<td>47916710.00</td>
<td>23864764.80</td>
<td>23864764.80</td>
<td>23864764.80</td>
<td>23864764.80</td>
<td>23864764.80</td>
<td></td>
</tr>
<tr>
<td>RANGE</td>
<td>48778971.00</td>
<td>24170779.27</td>
<td>24170779.27</td>
<td>24170779.27</td>
<td>24170779.27</td>
<td>24170779.27</td>
<td></td>
</tr>
<tr>
<td>MEAN</td>
<td>132422.03</td>
<td>132422.03</td>
<td>132422.03</td>
<td>132422.03</td>
<td>132422.03</td>
<td>132422.03</td>
<td></td>
</tr>
<tr>
<td>MEDIAN</td>
<td>8577.50</td>
<td>10752.00</td>
<td>10752.00</td>
<td>11230.80</td>
<td>11230.80</td>
<td>11230.80</td>
<td></td>
</tr>
<tr>
<td>MODE</td>
<td>0.00</td>
<td>4411.00</td>
<td>4411.00</td>
<td>4411.00</td>
<td>4411.00</td>
<td>4411.00</td>
<td></td>
</tr>
<tr>
<td>SUM</td>
<td>26045403.00</td>
<td>26045403.00</td>
<td>26045403.00</td>
<td>26045403.00</td>
<td>26045403.00</td>
<td>26045403.00</td>
<td></td>
</tr>
<tr>
<td>CSS</td>
<td>4511740203209899.00</td>
<td>3318908348678920.00</td>
<td>3318908348678920.00</td>
<td>3318908348678920.00</td>
<td>3318908348678920.00</td>
<td>3318908348678920.00</td>
<td></td>
</tr>
<tr>
<td>STD</td>
<td>128321.03</td>
<td>128321.03</td>
<td>128321.03</td>
<td>128321.03</td>
<td>128321.03</td>
<td>128321.03</td>
<td></td>
</tr>
<tr>
<td>STDEPR</td>
<td>33952.28</td>
<td>74843.33</td>
<td>74843.33</td>
<td>74843.33</td>
<td>74843.33</td>
<td>74843.33</td>
<td></td>
</tr>
<tr>
<td>CV</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>VARP</td>
<td>23068933022.00</td>
<td>150878651789.81</td>
<td>150878651789.81</td>
<td>150878651789.81</td>
<td>150878651789.81</td>
<td>150878651789.81</td>
<td></td>
</tr>
<tr>
<td>KURTOSIS</td>
<td>0.20</td>
<td>337.70</td>
<td>337.70</td>
<td>337.70</td>
<td>337.70</td>
<td>337.70</td>
<td></td>
</tr>
<tr>
<td>SHIIANESS</td>
<td>24.10</td>
<td>17.78</td>
<td>17.78</td>
<td>17.78</td>
<td>17.78</td>
<td>17.78</td>
<td></td>
</tr>
<tr>
<td>NNM%</td>
<td>9.65</td>
<td>9.65</td>
<td>9.65</td>
<td>9.65</td>
<td>9.65</td>
<td>9.65</td>
<td></td>
</tr>
<tr>
<td>NSLP%</td>
<td>9.65</td>
<td>9.65</td>
<td>9.65</td>
<td>9.65</td>
<td>9.65</td>
<td>9.65</td>
<td></td>
</tr>
<tr>
<td>MED</td>
<td>26.52</td>
<td>26.52</td>
<td>26.52</td>
<td>26.52</td>
<td>26.52</td>
<td>26.52</td>
<td></td>
</tr>
<tr>
<td>INFOLoss</td>
<td>0.37</td>
<td>0.37</td>
<td>0.37</td>
<td>0.37</td>
<td>0.37</td>
<td>0.37</td>
<td></td>
</tr>
<tr>
<td>INFOSV%</td>
<td>17.45</td>
<td>17.45</td>
<td>17.45</td>
<td>17.45</td>
<td>17.45</td>
<td>17.45</td>
<td></td>
</tr>
</tbody>
</table>
Decision-making: indicators

For each variable:
Comparison of indicators obtained with initial data and various micro-aggregated data

Indicators: Statistics
Confidentiality criteria
Correlations
Dataset generation
Displays every data set the user worked with:
- initial data set,
- “extreme” part of data set,
- “normal” part of data set,
- all micro-aggregated data set,

For each variable, user selects how they are treated.
Generation of synthetic indicator

μA - Dataset Generation Options

- Information loss (global) 100
- Information loss/Variable 0
- % obs. with a perturbation degree < 0.01 0
- % obs. with a perturbation degree > 10.00 0
- % Variance variation 0
- % Median variation 0
- Concentration coefficient 0

Default  Refresh  Treat variables individually  Exit
Generation of synthetic indicator

User selects:

- Micro-aggregation criteria used to rank the different results calculated.
- Weights assigned to each selected indicator

- This selection generates a synthetic indicator, adapted to the user’s needs
Model application
Each step of micro-aggregation process is listed in “model application” frame

This model can be applied to any data set with identical structure.

Each step can modified, or skipped.
Choice of SAS system

Advantages...
- Portability across platforms
- Modular programming
- Used in (almost) all National Statistical Institutes
- Statistical procedures

But...
- Certain procedures have different results according to versions of SAS system
- Adaptation to Unix: re-customisation of the screens (size, buttons, …)
Micro-aggregation system for confidential data

V. FURTHER DEVELOPMENTS

- Statistical aspects
- Software aspects
Statistical aspects

• More complex methods on:
  – numeric variables (Ward)
  – categorical variables (multiple correspondence…)

• New statistical criteria on confidentiality measure
  – more “operational” indicator
  – assuring more confidence in micro-aggregation methods

• Treatment of hierarchical variables (recoding)

• Treatment of missing data
Software aspects

- Towards a more automated application
- Complete reporting of micro-aggregation session
- On-line help (based on key-words)
- Datasets import and export (other formats)
- Optimization of complex statistical methods
- Recoding hierarchical variables
CONCLUSION

E-mail addresses:

- Dr Pascal Loosveldt
  pl@arianeii.com
- Laurence Béchet
  laurence.bechet@co.arianeii.be
- Thien Le Tri
  thien.letri@co.arianeii.be

Web site:

http://www.arianeii.com