Measures for Information Loss in Protected Data

Mrs. Mira Todorova

Joint UN/ECE EUROSTAT Work Session on Statistical Data Confidentiality
Skopje, 20-22 September 2017
Masking Methods for Microdata Protection

- Perturbative
- Non-perturbative
  OR
- Continuous
- Categorical
Perturbative Methods

\[ V' = AVB + C \]

Where:

- \( V \) - original microdata set,
- \( V' \) - masked microdata set,
- \( A \) - record-transforming mask,
- \( B \) - variable-transforming mask and
- \( C \) - displacing (noise) mask.
Non-perturbative Methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Continuous data</th>
<th>Categorical data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Global re-coding</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Top and bottom coding</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Local suppression</td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>
Information Loss Measures for Continuous data

- Covariance matrices: COVo on O and COVp on P;
- Variance matrices: VARo on O and VARp on P;
- Correlation matrices: Ro on O and Rp on P.
Information Loss Measures for Continuous data

- Correlation matrices $R_{W_o}$ (respectively $R_{W_p}$) between the $W$ variables and the $W$ factors obtained through principal component analysis;
- Factor score coefficient matrices: $F_{o}$ on $O$ and $F_{p}$ on $P$;
- Commonalities $C_{W_o}$ (respectively $C_{W_p}$) between each $W$ variable and the first principal component.
Information Loss Measures for Categorical Data

- Direct comparison of categorical values;
- Comparison of contingency tables and
- Entropy-based measures.
DSC Practice in the Macedonian statistical system

MICRODATA ACCESS

- 5 of 7 institutions provide access to microdata
- “Safe Room” for access to microdata
- Specific surveys have access to microdata.
DSC Practice in the Macedonian statistical system

SDC MEASURES FOR INFORMATION LOSS

No procedures have been developed for calculation of disclosure risk on produced statistics, nor for information loss.
Conclusions for the Macedonian statistical system

State Statistical Office should put efforts for developing Statistical Disclosure Methods for the whole statistical system, since

ONLY USED DATA ARE USEFUL DATA.