Methods of Confidentiality Protection

John M. Abowd
April 2005
General Principles of Disclosure
Review and Control

• Protection of the confidentiality of the underlying micro data
  – Avoiding identity disclosure
  – Avoiding attribute disclosure

• Traditional methods
  – Suppression
  – Coarsening
  – Adding noise (explicit or via swapping)

• New methods
  – Partially-synthetic data
  – Fully-synthetic data
Guidance on Traditional Methods

• Use the Federal Committee on Statistical Methodology working paper series
  • http://www.fcsm.gov/working-papers/wp22.html
  • http://www.fcsm.gov/working-papers/spwp22.html
Tabular Methods

• Tables summarize counts, magnitudes, ratios, differences

• Micro data are aggregated to cells
Coarsening

• Modifying the cell boundaries so that they are broader

• Example: basic Census geography
  – Block
  – Block Group
  – Tract
  – Political divisions
  – Public-use microsample areas
Primary Suppression

• Suppression rule: do not publish a data item if it is based on too few entities or if it is dominated by too few entities

• Example: $n,k$ rules
  – Suppress the data item if it is base on fewer than $n$ entities or if fewer than $n$ entities account for $k$ percent of the total
Complementary Suppression

• Suppress additional data items so that the primary suppressions cannot be recovered from other published items

• Example: if a small county is suppressed then either the margin across all counties or an additional county must be suppressed to prevent recovering the suppressed value from the published values
Noise Addition

• Adding noise to the published item or to the underlying micro data to disguise the true value
• Example: confidentiality edit of Decennial Census
• Example: QWIs
The Quarterly Workforce Indicator System

- Multiplicative noise infusion system
- Establishment level micro data are distorted according to a permanent distortion factor
- Distortion factor always moves the fuzzed item away from the actual item by a minimum and maximum percentage
- All release data are computed from the fuzzed items
Fuzz Factor Distribution

\[ p(\delta_j) = \begin{cases} 
\frac{(b-\delta)}{(b-a)^2}, & \delta \in [a,b] \\
\frac{(b+\delta-2)}{(b-a)^2}, & \delta \in [2-b,2-a] 
\end{cases} \]

\[ F(\delta_j) = \begin{cases} 
0.5 + \frac{[b-a]^2 - (b-\delta)^2}{2(b-a)^2}, & \delta \in [a,b] \\
\frac{[\delta+b-2]^2}{2(b-a)^2}, & \delta \in [2-b,2-a] 
\end{cases} \]
Types of Data Distortion

• Counts: $B, E, M, F, A, S, H, R, FA, FS, W_1, W_2, W_3, NA, NH, NR,$ and $NS$
• Ratios: $ZW_2, ZW_3, ZWA, ZWS, ZNA, ZNH, ZNR,$ and $ZNS$
• Differences: $JF, JC, JD, FJF, FJC, FJD, DWA, DWFA, DWS, DWFS$
Multiplicative Noise Addition

- $B_{jt}^* = \delta_j \times B_{jt}$

- $ZW_{2jt}^* = \frac{W_{2jt}^*}{E_{jt}} = \frac{\delta_j \times W_{2jt}}{E_{jt}}$

- $JF_{kt}^* = G_{kt} \times \bar{E}_{kt}^* = JF_{kt} \times \frac{\bar{E}_{kt}^*}{E_{kt}}$

- $Z\Delta WA_{kt}^* = \frac{\Delta WA_{kt}}{A_{kt}} \times \frac{A_{kt}^*}{A_{kt}}$

- B is beginning of quarter employment; E is end of period; E-bar is the average.
- ZW2 is end of quarter employee earnings, W2 is total payroll for end of quarter employees.
- JF is net job flows.
- Asterisk indicates distorted values.
Weighting

- Each fuzzed micro data item is weighted by the QWI final weight before aggregation.
- This means that all input data are real numbers (not integers).
- Final disclosure control formulas must reflect rounding of the counts.
Interpreting the Algorithm

• Based on the public use version of the spreadsheet
## Analytic Validity of Time Series Properties

<table>
<thead>
<tr>
<th>Quantile</th>
<th>Beginning of Quarter Employment</th>
<th>Accessions</th>
<th>Separations</th>
<th>Full Quarter Employment</th>
<th>Net Job Flows</th>
</tr>
</thead>
<tbody>
<tr>
<td>99%</td>
<td>0.07894</td>
<td>0.07153</td>
<td>0.06711</td>
<td>0.06644</td>
<td>0.01104</td>
</tr>
<tr>
<td>95%</td>
<td>0.04338</td>
<td>0.04253</td>
<td>0.04070</td>
<td>0.03465</td>
<td>0.00503</td>
</tr>
<tr>
<td>90%</td>
<td>0.02610</td>
<td>0.03043</td>
<td>0.02826</td>
<td>0.01972</td>
<td>0.00314</td>
</tr>
<tr>
<td>75%</td>
<td>0.00946</td>
<td>0.01387</td>
<td>0.01326</td>
<td>0.00718</td>
<td>0.00124</td>
</tr>
<tr>
<td>50%</td>
<td>-0.00043</td>
<td>0.00103</td>
<td>0.00004</td>
<td>-0.00003</td>
<td>0.00000</td>
</tr>
<tr>
<td>25%</td>
<td>-0.01026</td>
<td>-0.01271</td>
<td>-0.01179</td>
<td>-0.00641</td>
<td>-0.00096</td>
</tr>
<tr>
<td>10%</td>
<td>-0.02520</td>
<td>-0.03012</td>
<td>-0.02592</td>
<td>-0.01720</td>
<td>-0.00281</td>
</tr>
<tr>
<td>5%</td>
<td>-0.03695</td>
<td>-0.04100</td>
<td>-0.03569</td>
<td>-0.02806</td>
<td>-0.00471</td>
</tr>
<tr>
<td>1%</td>
<td>-0.06984</td>
<td>-0.06863</td>
<td>-0.06645</td>
<td>-0.06185</td>
<td>-0.01038</td>
</tr>
</tbody>
</table>

### Distribution of the Error in the First Order Serial Correlation Coefficient Due to Multiplicative Noise Distortion ($r^* - r$)
Aggregating Formulas

• Counts and magnitudes:
  – Add

• Ratios and differences
  – Multiply by released base
  – Aggregate numerator and denominator separately
  – Add

• Job creations and destructions
  – Handle like counts but understand that there is an inherent loss of information
Handling the Suppressions

• Type 2 suppressions
  – Randomly insert 1 or 2

• Type 3 suppressions
  – Must be treated as missing data and estimated

• Type 4 suppressions
  – Use the aggregating formulas
LEHD Research Papers

- Review the LEHD Disclosure Review memo
- A non-confidential version was included with today’s materials
Business and Establishment-Based Models

• Review the RDC disclosure guide and associated programs

• A non-confidential version was included with today’s materials
Special Tabulations

• **Review the DRB memo on special tabulations**

• A non-confidential version was included with today’s materials