Quantitative Methods to Assess Data Confidentiality and Data Utility for Microdata in Japan

Shinsuke Ito, Meikai University, Japan

Mariko Murata, Statistical Information Institute for Consulting and Analysis, Japan

1. Background: Disclosure Limitation Methods in Japan

2. Microaggregation as a Disclosure limitation Method for Microdata in Japan.

3. Quantitative Assessment of Data Utility and Data Confidentiality for Microdata Created Using Different Disclosure Limitation Methods.

4. Conclusion and Outlook

1. Background: Disclosure Limitation Methods in Japan

Current Disclosure Limitation Methods for Official Microdata:

- Deletion of identifiers (e.g. name, address etc.)
- Top/bottom coding (e.g. annual household income etc.)
- Recoding (e.g. age bracket)
- Resampling (e.g. 80% resampling rate)
- Limitation of geographical information (e.g. “Tokyo/Osaka/Nagoya areas” vs. “others”)

Perturbative methods such as additive noise and data swapping including microaggregation are not currently adopted.

Only few empirical studies on disclosure limitation methods, disclosure risk and information loss have been conducted in Japan.

2. Microaggregation as a Disclosure Limitation Method

Overview of Microaggregation:*

- Records with common values for all types of qualitative attributes based on multi-dimensional tabulation are created.
- Records with common values for qualitative attributes are sorted and divided into groups larger than a specific minimum size.
- The value of each quantitative attribute for records is replaced with a measure of central tendency (ex. average value etc.) within each group.

Usage Potential and Degree of Confidentiality for Micro-aggregated Data:**

- Suitability of microaggregation is examined for several types of micro-aggregated data generated from the Japanese 'National Survey of Family Income and Expenditure'.

* Based on research by Defays and Anwar (1998) , Domingo-Ferrer and Mateo-Sanz (2002), Ito *et al*. (2008), Ito (2009) and Ito and Takano (2011).

** Based on research by Ito *et al*. (2008), Ito (2009) and Ito and Takano (2011).

3. Quantitative Assessment of Data Utility and Data Confidentiality

Step 1: Create masked data using disclosure limitation methods

Step 2: Assess data utility and degree of confidentiality

Step 1: Disclosure Limitation Methods Used:*

(1) Microaggregation:
- Using the individual ranking method
- Using the sum of Z-scores method

(2) Noise Addition:
- Addition of Gaussian noise

(3) Categorization:
- 10-quantile
- 20-quantile

(4) Combined use of disclosure limitation methods :
- Individual ranking method and sum of Z-scores method
- Microaggregation and categorization
- Noise addition and categorization

* Survey Data: Original microdata from the January 2009 'Family Income and Expenditure Survey' (4,220 households)

3. Quantitative Assessment of Data Utility and Data Confidentiality

Step 2: Assessment of Data Utility and Degree of Confidentiality

- Data Utility: Information loss based on mean square error and mean variation.

- Degree of Confidentiality: Distance-based record linkage
  - One-to-one true match
  - False match
  - n:m match

Data utility and degree of confidentiality is assessed for the following six attributes:

- Wages/salaries and consumption expenditure in month of survey
- Yearly household income
- Household savings
- Household liabilities
- Dwelling size

Before applying perturbation to these quantitative attributes, records are clustered within each category by type of tenure of dwelling.

Table 1: Data Utility for Microdata from the 'Family Income and Expenditure Survey'

	Information Loss	
	Mean Square Error	Mean Variation
MicroIR	0.000039	0.020757
MicroZscore	0.025357	0.736120
AddNoise0.01	0.000000	0.000708
AddNoise0.05	0.000002	0.003517
AddNoise0.10	0.000012	0.009386
AddNoise0.16	0.000053	0.020264
AddNoise0.20	0.000110	0.029471
AddNoise0.30	0.000432	0.058772
AddNoise0.50	0.002264	0.131800
CTG10	0.002139	0.107590
CTG20	0.001198	0.079517
MicroIRZscore	0.013403	0.543650
MicroIRCTG10	0.000078	0.039800
MicroZscoreCTG10	0.007535	0.124640
AddNoise0.10CTG10	0.000078	0.034988
AddNoise0.16CTG10	0.000088	0.035931
AddNoise0.30CTG10	0.000186	0.040967
AddNoise0.50CTG10	0.000690	0.066631

Table 2: Degree of Data Confidentiality for Microdata from the 'Family Income and Expenditure Survey'

	One-to-One True Match		False Match	n:m Match
MicroIR	4,203	99.60%	0	17
MicroZscore	860	20.38%	611	2,749
AddNoise0.01	4,218	99.95%	0	2
AddNoise0.05	4,214	99.86%	0	6
AddNoise0.10	4,165	98.70%	1	54
AddNoise0.16	3,980	94.31%	15	225
AddNoise0.20	3,748	88.82%	39	433
AddNoise0.30	3,076	72.89%	199	945
AddNoise0.50	1,838	43.55%	556	1,826
CTG10	3,558	84.31%	6	656
CTG20	3,934	93.22%	1	285
MicroIRZscore	1,633	38.70%	435	2,152
MicroIRCTG10	3,800	90.05%	2	418
MicroZscoreCTG10	2,968	70.33%	68	1,184
AddNoise0.10CTG10	3,780	89.57%	4	436
AddNoise0.16CTG10	3,695	87.56%	6	519
AddNoise0.30CTG10	3,302	78.25%	56	862
AddNoise0.50CTG10	2,657	62.96%	206	1,357

Conclusion and Outlook

・Perturbative methods such as additive noise and microaggregation are not currently adopted for Japanese official microdata, and few empirical studies on disclosure limitation methods have been conducted in Japan.

・This paper proposes methods to assess information loss and degree of true match based on record linkage for masked data created from Japanese official microdata using disclosure limitation methods.

・Based on an empirical study, this paper then suggests methods for the relative measurement of data utility and data confidentiality of masked data in Japan.

・ This research increases the number of available disclosure limitation methods in Japan, and points towards a potential improvement in the usability of anonymized official microdata in case the disclosure limitation methods suggested in this research are adopted in Japan.