A Quantitative Assessment of Data Confidentiality and Data Utility to Create Anonymized Census Microdata in Japan

Shinsuke Ito, Chuo University, Japan

Naomi Hoshino, National Statistics Center, Japan

Fumika Akutsu, Statistics Bureau of Japan

1. Introduction: Anonymized Census Microdata in Japan
2. Quantitative Assessment of Data Confidentiality Based on the "Allowable Threshold"
3. Quantitative Assessment of Data Utility Based on Entropy
4. The Relationship between Sampling and Recoding from a Perspective of Data Confidentiality and Data Utility
5. Conclusion and Outlook

1. Introduction: Anonymized Census Microdata in Japan

Japan’s Statistics Act was revised in April 2007, and Anonymized microdata from official statistics have been released in Japan since April 2009.

The Statistics Bureau has been providing Anonymized census microdata since 2013.

Current Situation for Population Census Data:

- 2000 and 2005 census are currently available.
- Limited geographical information (prefecture level and geographical areas which have 500,000 people and above)

Current Anonymization Methods for Population Census Data:

- Sampling based on household units (at a sampling rate of 1%)
- Non-perturbative methods incl. deletion of direct identifiers, recoding, top and bottom coding
- Deletion of unique records
- Data swapping

1. Introduction: Anonymized Census Microdata in Japan

Small area results are a very important type of microdata and used in research across a variety of fields.

Currently, geographic information in the Anonymized census microdata is limited to prefecture level and geographical areas with 500,000 people and above.

This paper suggests an approach for the creation of anonymized smaller area microdata in Japan.

1) Three sets of data were created from the 2000 Population Census.
2) Data confidentiality was quantitatively assessed based on the concept of ‘threshold of data confidentiality’.
3) Data utility was assessed based on entropy-based measures.
4) Data utility and data confidentiality were compared for different sampling rates.

2. Quantitative Assessment of Data Confidentiality Based on the “Allowable Threshold”

(1) By setting the allowable threshold for data confidentiality in this research, possible combinations of geographical classification, categories of household, individual attributes and sampling rate can be determined[1].

(2) One way to determine the threshold is to compare data confidentiality of anonymized microdata[2] to that of Anonymized microdata[3].

[1] Based on Dale (1995), Marsh *et al*. (1994), Tranmer *et al*. (2005)

[2] ‘anonymized microdata’ with a lower-case “a” is defined as microdata to which disclosure limitation methods have been applied as part of this research.

[3] ‘Anonymized microdata’ with a capital “A” are defined as official microdata ‘that is processed so that no particular individuals or juridical persons, or other organizations shall be identified’.

2. Quantitative Assessment of Data Confidentiality Based on the “Allowable Threshold”

Three sets of data from the 2000 Population Census were created and used as test data.

Set 1: Based on more than 500,000 records of individual data from a certain geographic area within a specific Japanese prefecture (“Area A”).

Set 2: Based on more than 100,000 records of individual data from another geographic area within the same prefecture (“Area B”).

Set 3: Based on more than 50,000 records of individual data from a third geographic area within the same prefecture (“Area C”).

2. Quantitative Assessment of Data Confidentiality Based on the “Allowable Threshold”

Key Variables:

- Gender
- Marital Status
- Nationality
- Type of (Work) Activity
- Occupation
- Type and Tenure of Dwelling
- Type of Building and Total Number of Floors

Recoding was applied identically to Anonymized microdata

- Age
- Employment Status
- Industry

Recoded and/or top coded

➔ Age (9 patterns)

➔ Employment status (3 patterns)

➔ Industry (3 patterns)

Population uniques were calculated for all 81 possible combinations of the patterns for age, employment status and industry.

2. Quantitative Assessment of Data Confidentiality Based on the “Allowable Threshold”

Population uniques were calculated and a quantitative assessment of data confidentiality was conducted based on the following steps:

Step 1: The ‘decrease rate of population uniques’ for Area A was calculated for the above key variables.

Step 2: The ‘allowable population unique ratio’ was calculated. This ratio was used as the ‘threshold of allowable data confidentiality’.

Step 3: Recoding and top coding were applied to age, employment status and industry for both household data and individual data from Areas B and C.

Step 4: The population unique ratios for anonymized individual data and anonymized household data for Areas B and C were calculated and compared to the ‘allowable population unique ratio’.

Table 1: Population unique ratio and threshold of data confidentiality for Areas A, B, and C.

	Population Unique ratio		Decrease Rate of Population Uniques
	Original Categories	Recoded Categories	
Area A	13.46%	4.20%	31.20%

	Unit	Population Unique Ratio for Original Categories	Allowable Population Unique Ratio
Area B	Individual	16.97%	5.30%
	Household	26.72%	8.35%
Area C	Individual	18.47%	5.77%
	Household	31.31%	9.78%

Table 2: Anonymized Individual Data B: Patterns for recoding and top coding for which the population unique ratio is lower than the allowable population unique ratio (Excerpt)

						Employment Status			Industry			
Five-year age brackets	Five-year age brackets and top coding for 85 years and above	Five-year age brackets and top coding for 75 years and above	Ten-year age brackets	Ten-year age brackets and top coding for 85 years and above	Ten-year age brackets and top coding for 75 years and above	8 Categories	6 Categories	4 Categories	14 Categories	10 Categories	4 Categories	Population unique ratio
					*			*			*	1.79%
			*					*			*	1.79%
				*				*			*	1.82%
			*				*				*	2.44%
					*		*				*	2.45%
				*			*				*	2.46%
			*			*					*	2.61%
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
	*							*		*		4.49%
		*						*	*			4.52%
*								*		*		4.53%
	*							*	*			4.65%
*								*	*			4.69%

<u>Result</u>: The number of patterns of recoding and top coding for which the population unique ratio is smaller than the allowable population unique ratio is 42 for anonymized individual data B.

Table 3: Number of patterns of recoding and top coding for which the population unique ratio is smaller than the allowable population unique ratio

Type of Anonymized Data	The Number of Patterns
Anonymized Individual Data B	42
Anonymized Household Data B	36
Anonymized Individual Data C	42
Anonymized Household Data C	24

Result: For anonymized data consisting of individual records, area size does not impact the allowable combinations of recoding.

3. Quantitative Assessment of Data Utility Based on Entropy

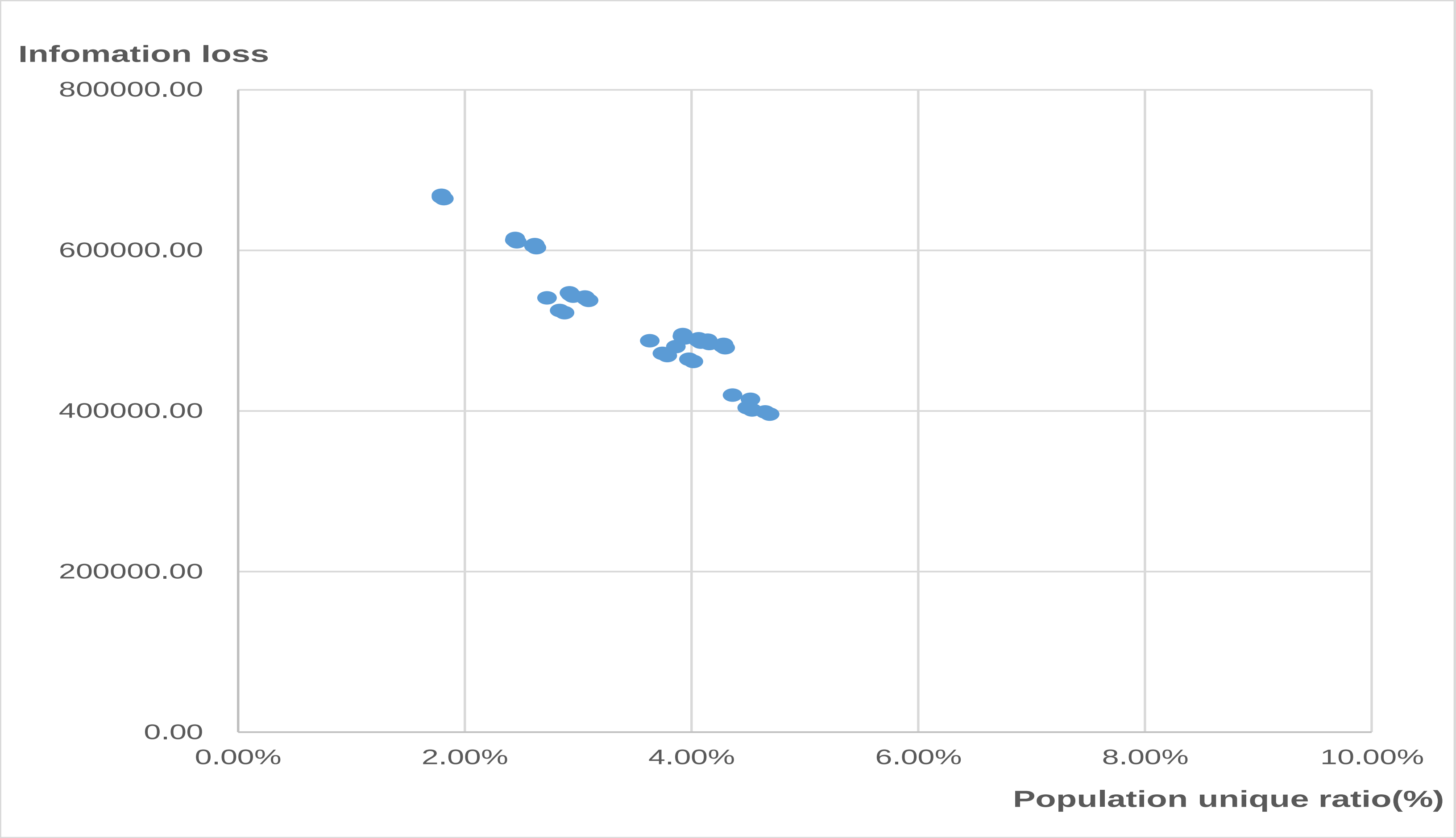
In order to assess data utility of quantitative attributes, information loss was calculated using entropy-based measures*.

A quantitative assessment of data utility was conducted for combinations of recoding and top coding for key variables for Areas B and C where the population unique ratio was lower than the allowable population unique ratio.

Result: The pattern of five year age brackets, original categories for industry, and four categories for employment status shows the lowest information loss.

* Based on Kooiman et al. (1998), Domingo Ferrer and Torra (2001), and De Waal and Willenborg (1999).

Figure 1: R-U Confidentiality Map for Anonymized Individual Data B



R-U confidentiality map based on population unique ratios which are lower than the “allowable population unique ratio”, and information loss for individual records and household records for Area B.

<u>Result</u>: Patterns with higher population unique ratios tend to have lower information loss.

4. The Relationship between Sampling and Recoding from a Perspective of Data Confidentiality and Data Utility

- Sampling (or resampling) is used as a disclosure limitation method for the creation of Anonymized microdata.

- As the sampling rate impacts data confidentiality and data utility, a quantitative assessment of data confidentiality and data utility for different sampling rates was conducted as part of this research.

4. The Relationship between Sampling and Recoding from a Perspective of Data Confidentiality and Data Utility

Step 1: UUSU rates* were calculated for 1% sampled anonymized data from Area A using the same key variables that were used for calculating population uniques. The allowable UUSU rate was defined as the ‘threshold of data confidentiality’ for the sampled data.

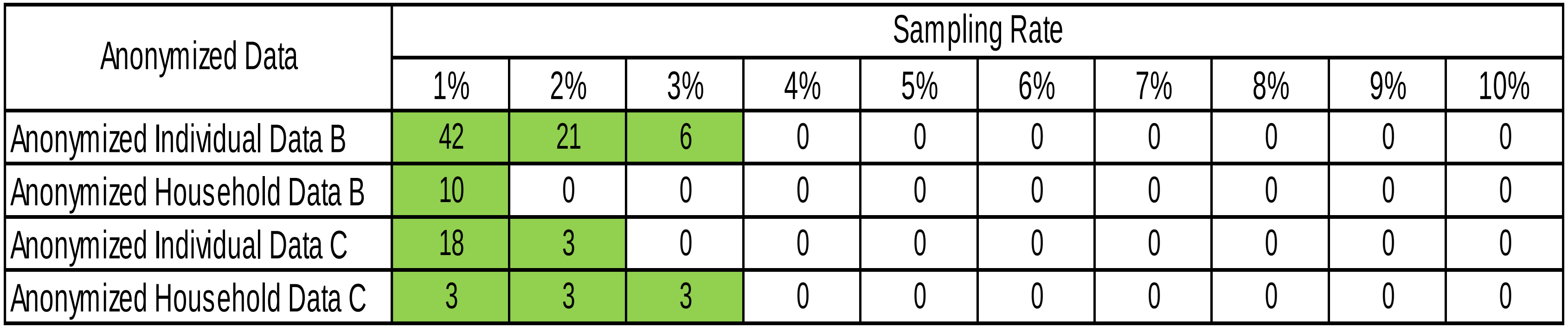
Step 2: UUSU rates were calculated based on a sampling rate p % (p=1, 2 ,3 ,4 ,5 ,6 ,7 ,8 ,9 ,10) for anonymized individual data and anonymized household data for Areas B and C.

Step 3: The combinations of sampling rate and recoding for which the UUSU rate is lower than the allowable UUSU rate were determined and the highest possible sampling rate that meets the threshold for data confidentiality was selected.

Result: The UUSU rate for anonymized data A was 12.32 %. This rate was set as the “allowable UUSU rate”.

* The UUSU rate is defined as the number of records which are both population uniques and sample uniques divided by the number of records which are sample uniques.

Table 4: Number of patterns for recoding and top coding at different sampling rates for which the UUSU rate is lower than the allowable UUSU rate



Anonymized Data	Sampling Rate									
	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%
Anonymized Individual Data B	42	21	6	0	0	0	0	0	0	0
Anonymized Household Data B	10	0	0	0	0	0	0	0	0	0
Anonymized Individual Data C	18	3	0	0	0	0	0	0	0	0
Anonymized Household Data C	3	3	3	0	0	0	0	0	0	0

Result: When using the ‘allowable UUSU rate’ as an additional threshold for data confidentiality, and for 3% sampled data for individual records from Area B, there were six patterns of recoding for which the UUSU rate was lower than the allowable UUSU rate.

4. The Relationship between Sampling and Recoding from a Perspective of Data Confidentiality and Data Utility

In cases where the UUSU rate is lower than the allowable UUSU rate, adding perturbative methods such as data swapping can maintain data confidentiality and improve data utility.

In this research, entropy-based measures[*] for 3% sampled data and 5% sampled and swapped data were calculated and compared for the six patterns of recoding for which the UUSU rate was lower than the allowable UUSU rate.

[*] Entropy-based measures were calculated based on conditional entropy.

Table 5: Comparison of information loss between 3% sampled data and 5% sampled and swapped data

Ten-year age brackets	Ten-year age brackets and top coding for 85 years and above	Ten-year age brackets and top coding for 75 years and above	Employment Status	Industry			Sampling Rate of 3%		Sampling Rate of 5%		Sampling Rate of 5% +Swapping		
			4 Categories	14 Categories	10 Categories	4 Categories	UUSU Rate	Entropy-Based Measure	UUSU Rate	Entropy-Based Measure	The Number of Population Unique and Sample Unique	The Number of Swapped Records of Population Unique and Sample Unique	Entropy-Based Measure
		*	*			*	11.65%	4.1886	14.09%	4.2694	123	15	4.2719
*			*			*	11.30%	4.1739	13.83%	4.2573	122	13	4.2586
	*		*			*	11.41%	4.1563	13.73%	4.2424	121	12	4.2430
*			*		*		12.28%	3.3923	15.92%	3.4670	199	45	3.4777
*			*	*			12.24%	3.3661	16.38%	3.4332	208	52	3.4411
	*		*	*			12.31%	3.3489	16.25%	3.4186	206	50	3.4300

Result: There is little difference in information loss between 5% sampled data and 5% sampled and swapped data. This means that the latter presents an additional option for the creation of anonymized microdata.

5. Conclusion and Outlook

(1) This paper uses anonymized official microdata created from Japanese Population Census data to assess data confidentiality and data utility based on the two thresholds of ‘allowable population unique rate’ and ‘allowable UUSU rate’.

(2) This paper shows that for individual data it is possible to create anonymized microdata with more detailed geographical information. It also identifies the combination of recoding for categories of individual and household attributes to maintain data confidentiality at the same level as Anonymized census microdata that is currently available.

(3) This research contributes to the creation and release of different types of Anonymized microdata that will allow researchers from a variety of fields including economics, sociology, demography, geography etc. to conduct more detailed statistical analyses based on official statistics in Japan.