Potential of Disclosure Limitation Methods for Census Microdata in Japan

Shinsuke Ito[1] Naomi Hoshino[2] Fumika Akutsu[3]

[1] Faculty of Economics, Chuo University, 742-1 Higashinakano, Hachioji-shi, Tokyo, 192-0393 Japan,
[2] National Statistics Center, 19-1 Wakamatsu-cho, Shinjuku-ku, Tokyo, 162-8668 Japan,
[3] Statistics Bureau of Japan, 19-1 Wakamatsu-cho, Shinjuku-ku, Tokyo, 162-8668 Japan,

ssitoh@tamacc.chuo-u.ac.jp, nsaitou2@nstac.go.jp, fakutsu@soumu.go.jp

Shinsuke Ito is a research fellow at the National Statistics Center and conducts research on disclosure limitation methods for microdata in co-ordination with officials at the National Statistics Center and the Statistics Bureau of Japan.

Abstract. Anonymized microdata for seven types of official statistics – including Anonymized microdata from the 2000 and 2005 Population Census conducted by the Statistics Bureau of Japan – are currently made available in Japan under the Statistics Act. For almost all official statistics, only one type of Anonymized microdata is released.

This paper uses geographical thresholds to generate anonymized microdata for smaller geographic areas, and quantitatively assesses data confidentiality and data utility for this data. This research aims to develop approaches for the creation of more detailed Anonymized microdata, which would allow researchers from a variety of fields including economics, sociology, demography, geography and others to conduct more detailed statistical analysis based on Japanese official statistics.

Keywords: Census Microdata, Recoding, Top-Coding, Population Uniques, Distance-Based Information Loss and Geographical Threshold

1 Introduction: Anonymized Census Microdata in Japan

Japan's Statistics Act was revised in April 2007 – the first major revision in sixty years – with the objective of promoting the development and use of

official statistics, and thereby contributing to the development of the national economy and enhancement of the living standards of Japan's citizens. The 'Master Plan Concerning the Development of Official Statistics' was established based on the Statistics Act, and contains a "secondary usage" system that includes the provision of tailor-made tabulations and Anonymized microdata[1]. This was the starting point for the creation and release of Anonymized census microdata in Japan.

The Statistics Bureau has been releasing Anonymized census microdata since 2013. The data is made available five years after each census, so data from the 2000 and 2005 census is currently available. Various disclosure limitation methods such as sampling (at a sampling rate of 1%), recoding, top (bottom) coding, and data deletion are applied to the data before it is released. In order to conduct recoding and top (bottom) coding for risky records, the '0.5% standard' is applied i.e. categories that make up less than 0.5% of the population are recoded and top (bottom) coded, and records which have categories that make up less than 0.5% of the population are suppressed.

In order to promote a broader use of Anonymized official microdata, several empirical studies on the effectiveness of disclosure limitation methods such as microaggregation, additive noise, and data swapping for official microdata have been conducted by the National Statistics Center (Ito and Murata (2011), Ito and Hoshino (2012, 2013, 2014)). The Statistics Bureau of Japan and the National Statistics Center are currently conducting empirical research to prepare for the release of Anonymized microdata from the 2010 Population Census (Ito et al. (2015)).

Small area microdata is an important type of microdata, as the more detailed information contained in them allows researchers from a variety of fields including economics, sociology, demography and geography to use microdata for detailed statistical analysis. Currently, geographical classification contained in the Anonymized census microdata for 2000 and 2005 is limited to prefecture level or municipality level (500,000 persons and more), and for smaller areas only restricted microdata is available. For data from the 2010 census, the Statistics Bureau is researching ways to provide access to anonymized microdata that includes small area results while maintaining data confidentiality.

This paper aims to suggest an approach for the creation of anonymized small area microdata in Japan. Towards this objective, a quantitative assessment of data confidentiality was conducted using the concept of 'geographical threshold to ensure data confidentiality'. Information loss was calculated for several combinations of recoding and top coding using distance-based measures.

---

[1] 'Anonymized microdata' with a capital "A" are defined as individual data 'that is processed so that no particular individuals or juridical persons, or other organizations shall be identified' (Article 36 of the Japanese Statistical law).

2 Quantitative Assessment of Data Confidentiality Based on the Geographical Threshold

In the U.K., Samples of Anonymised Records (SARs) are compiled and released from the 1991 Population Census onwards. For this data, the level of detail for geographical classification, categories of household and individual attributes as well as the sampling rate are determined based on the 'thresholding rule' (Dale (1995), Marsh et al. (1994)). For the creation of Small Area Microdata (SAM) from the 2001 Population Census, Tranmer et al. (2005) conducted an empirical analysis on the disclosure risk for microdata for smaller geographical areas, and compared it to the disclosure risk for 1991 Individual SAR.

In the US, Hawala (2001) identified the relationship between population unique ratio and geographical area size to conduct an empirical assessment of data confidentiality for a geographical threshold of 100,000 persons. This threshold was adopted as one of the standards for creating Public Use Microdata Samples from the US Population Census.

By setting the thresholds for data confidentiality, possible combinations of geographical classification, categories of household, individual attributes and sampling rate can be determined. One way to set the thresholds is to compare data confidentiality for anonymized data[2] and 'Anonymized microdata'. Ito et al. (2015) assessed data confidentiality based on the two thresholds of 'allowable population unique ratio' and 'allowable UUSU rate[3]' using anonymized official microdata with more detailed geographical information that was created from Japanese Population Census data.

In Japan, the '0.5% standard' is used as a threshold for recoding or top (bottom) coding in the creation of Anonymized microdata. In contrast to the 0.5% standard, setting the 'geographical threshold to ensure data confidentiality' allows to appropriately recode categories of household and individual attributes.

Several sets of data from the 2010 Population Census – each containing a different number of records – were used as test data to conduct an empirical assessment of data utility and data confidentiality for different geographical thresholds. In this research, the following geographical area sizes were determined as the geographical thresholds: (1) geographical areas with more than 200,000 persons (including areas which refer to a prefectural capital), (2) geographical areas with more than 100,000 persons, (3) geographical areas with

---

[2] The 'anonymized data' with a lower-case "a" are defined as microdata to which disclosure limitation methods have been applied as part of this research.

[3] UUSU rates are percentages which are defined as the number of records which are both population uniques and sample uniques divided by the number of records which are sample uniques.

more than 50,000 persons, (4) geographical areas with more than 30,000 persons, (5) geographical areas with more than 20,000 persons, (6) geographical areas with more than 10,000 persons, (7) geographical areas with more than 5,000 persons, and (8) geographical areas with more than 1,000 persons. 20 areas which correspond to one of the geographical thresholds and are located within one specific Japanese prefecture were selected for this research.

The ratio of population uniques for anonymized data from all areas was calculated using the following 10 variables:

Gender (2 categories)
Marital Status (5 categories)
Nationality (2 categories)
Type of (Work) Activity (6 categories)
Employment Status (6 categories)
Type and Tenure of Dwelling (6 categories)
Type of Building and Total Number of Floors (4 categories)
Age
Industry
Occupation

For the variables of gender, marital status, nationality, type of (work) activity, employment status, type and tenure of dwelling, type of building and total number of floors recoding was applied the same way as when creating Anonymized microdata. Age, industry and occupation were recoded and/or top coded based on the following patterns:

Age (9 patterns)
(1) One-year age brackets
(2) One-year age brackets and top coding for 85 years and above
(3) One-year age brackets and top coding for 90 years and above
(4) One-year age brackets and top coding for 95 years and above
(5) Five-year age brackets and top coding for 85 years and above (the same categories as for Anonymized microdata)
(6) Five-year age brackets and top coding for 90 years and above
(7) Five-year age brackets and top coding for 95 years and above
(8) Ten-year age brackets and top coding for 90 years and above
(9) Ten-year age brackets and top coding for 100 years and above

Industry (3 patterns)
(1) 21 categories (categories from original data)
(2) 17 categories (categories recoded using ‘0.5% standard’)
(3) 14 categories (almost same categories as for Anonymized microdata)

Occupation (3 patterns)
(1) 12 categories (categories from original data)
(2) 10 categories (categories recoded using '0.5% standard')
(3) 8 categories (almost same categories as for Anonymized microdata)

Population uniques were calculated for all 81 possible combinations of the patterns for age, industry and occupation, and a quantitative assessment of data confidentiality was conducted.

Appendix Table 1 contains the population unique ratios for the one geographical area with more than 200,000 persons ("Area A"). Appendix Table 1 shows that the population unique ratio calculated based on the original categories of key variables for Area A is 18.43%. When only age is recoded, the population unique ratio calculated based on the recoded categories such as five-year age brackets and top coding for 85 years and above (the same categories as for Anonymized microdata) for Area A is 8.83%. When only industry is recoded, the population unique ratio calculated based on the recoded categories (14 categories and almost the same categories as for Anonymized microdata) for Area A is 17.47%. When only occupation is recoded, population unique ratio calculated based on the recoded categories (10 categories recoded using the '0.5% standard') for Area A is 18.33%. These results suggest that the population unique ratio is smaller for recoding of age than for recoding of industry or occupation. Therefore, more detailed categories for industry and occupation can be used for the creation of Anonymized microdata without impacting population unique ratio and data confidentiality.

Figure 1 illustrates the relationship between geographical area size and average population unique ratios for each area. Figure 2 illustrates the relationship between geographical area size and population unique ratios for five-year age brackets and top coding for ages 85 years and above, and for 17 categories of industry and 10 categories of occupation. Figure 3 illustrates the relationship between geographical area size and population unique ratios for ten-year age brackets and top coding for ages 90 years and above, 14 categories of industry and 8 categories of occupation. These figures show that for smaller geographical areas, population unique ratios are larger. These results suggest that there is a relation between geographical area size and population unique ratio, i.e. as the size of a geographical area increases, the population unique ratio tends to decrease. This result suggests that the optimal geographical thresholds can be determined based on the population unique ratio for a specific pattern of categories such as age, industry and occupation.

In order to establish the influence of recoding and top coding for age, industry, occupation and geographical area on data confidentiality, a multiple regression model was created. Model 1 is a model on the population unique ratio.

Fig. 1. Relationship between area size and population uniques: average population unique ratio

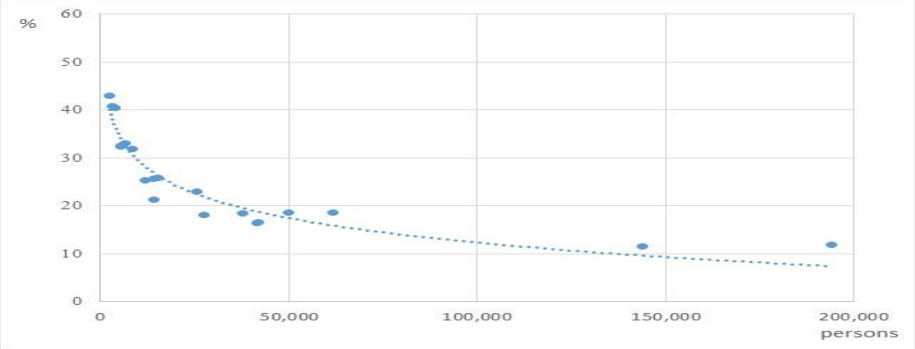


Fig. 2. Relationship between area size and population uniques: population unique ratio for five-year age brackets and top coding for 85 years and above, 17 categories of industry and 10 categories of occupation

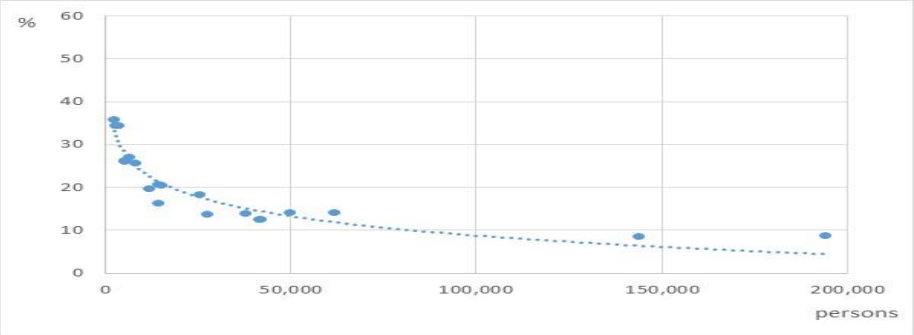
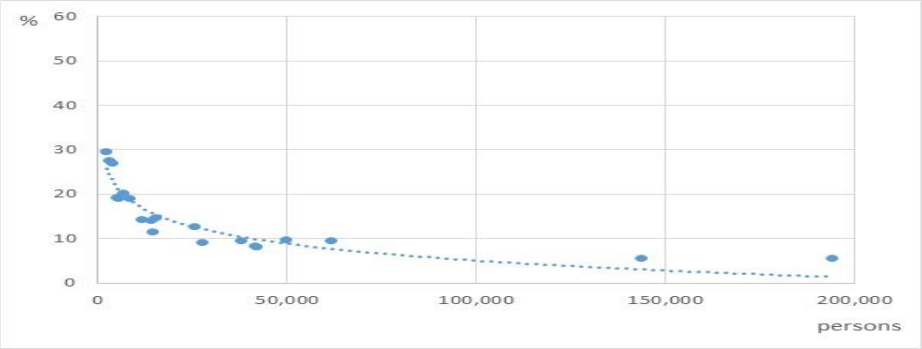


Fig. 3. Relationship between area size and population uniques: population unique ratio for ten-year age brackets and top coding for 90 years and above, 14 categories of industry and 8 categories of occupation



The variables used as dependent variables are the population unique ratios, while the variables used as independent variables are age (9 patterns as above), industry (3 patterns as above), occupation (3 patterns as above), and the logarithm of area size. The original categories for age, industry and occupation were used as a reference group.

Table 1 contains the list of definitions of the variables used in Model 1. Table 2 contains the results for Model 1. The results show that the coefficients for age, industry, occupation and area size are significantly negative. For age, as more categories are recoded and top coded, the absolute values of the coefficients tend to be significantly higher even when controlled for industry, occupation and geographical area size. The absolute value of the standardized partial regression coefficient (see the column 'beta' in Table 3) for the logarithm of area size is largest. This result confirms that area size has a more negative effect on population unique ratio than age, industry, or occupation.

Table 1. Definition of the independent variables used in Model1

Variables	Explanation of variables
Age Category 1	If the age pattern coresponds to one-year age brackets, the variable is 1. Otherwise it is 0.
Age Category 2	If the age pattern coresponds to one-year age brackets and top coding for 85 years and above , the variable is 1. Otherwise it is 0.
Age Category 3	If the age pattern coresponds to one-year age brackets and top coding for 90 years and above , the variable is 1. Otherwise it is 0.
Age Category 4	If the age pattern coresponds to one-year age brackets and top coding for 95 years and above , the variable is 1. Otherwise it is 0.
Age Category 5	If the age pattern coresponds to five-year age brackets and top coding for 85 years and above , the variable is 1. Otherwise it is 0.
Age Category 6	If the age pattern coresponds corespont to five-year age brackets and top coding for 90 years and above , the variable is 1. Otherwise it is 0.
Age Category 7	If the age pattern coresponds to five-year age brackets and top coding for 95 years and above , the variable is 1. Otherwise it is 0.
Age Category 8	If the age pattern coresponds to ten-year age brackets and top coding for 90 years and above , the variable is 1. Otherwise it is 0.
Age Category 9	If the age pattern coresponds to ten-year age brackets and top coding for 100 years and above , the variable is 1. Otherwise it is 0.
Industry Category 1	If the industry pattern coresponds to 21 categories, the variable is 1. Otherwise it is 0.
Industry Category 2	If the industry pattern coresponds to 17 categories, the variable is 1. Otherwise it is 0.
Indsutry Category 3	If the industry pattern coresponds to 14 categories, the variable is 1. Otherwise it is 0.
Occupation Category 1	If the industry pattern coresponds to 12 categories, the variable is 1. Otherwise it is 0.
Occupation Category 2	If the industry pattern coresponds to 10 categories, the variable is 1. Otherwise it is 0.
Occupation Category 3	If the industry pattern coresponds to 8 categories, the variable is 1. Otherwise it is 0.
Logarithm of Area Size	Logarithm of area size for 20 areas

Table 2. Results of multiple regression analysis of the population unique ratio

Variables	Coefficient	S.E.	t-value	Beta	Significance
Patterns of Age Categories<Age Category 1>					
Age Category 2	-0.007	0.003	-2.232	-0.018	**
Age Category 3	-0.003	0.003	-1.121	-0.009	
Age Category 4	-0.001	0.003	-0.478	-0.004	
Age Category 5	-0.140	0.003	-46.274	-0.368	***
Age Category 6	-0.139	0.003	-45.956	-0.365	***
Age Category 7	-0.139	0.003	-45.829	-0.364	***
Age Category 8	-0.184	0.003	-60.597	-0.481	***
Age Category 9	-0.184	0.003	-60.533	-0.481	***
Patterns of Industry Categories<Industry Category 1>					
Industry Category 2	-0.003	0.002	-1.645	-0.011	
Indsutry Category 3	-0.008	0.002	-4.368	-0.030	***
Patterns of Occupation Categories<Occupation Category 1>					
Occupation Category 2	-0.001	0.002	-0.564	-0.004	
Occupation Category 3	-0.007	0.002	-3.873	-0.027	***
Logarithm of Area Size	-0.072	0.001	-120.441	-0.718	***
Intercept	1.044	0.006	164.414		***
Adj.R^2	0.943				
F-value	2043.248				
N	1620				

Note 1: *** = 1% significance level, ** = 5% significance level, * = 10% significance level.

Note 2: Reference group in <brackets>.

3 Quantitative Assessment of Distance-Based Information Loss

Calculating information loss using entropy-based measures in order to assess data utility of quantitative attributes was first proposed by Kooiman et al. (1998) and Domingo Ferrer and Torra (2001). De Waal and Willenborg (1999) calculated entropy-based measures of information loss for anoymized data that was created using recoding. Ito et al. (2015) calculated entropy-based measures of information loss in order to determine the optimal combinations of recoding and top coding for which information loss is lowest.

In this research, data utility is defined as the average absolute distance per tabulation cell, and therefore as an indicator of distance that measures distortion

to the distribution based on Shlomo et al. (2010). The information loss (IL) indicator is defined as:

$$IL = \frac{\sum_{c} \left| T^{R}(c) - T^{O}(c) \right|}{n_T} \quad (1)$$

where $T^{O}(c)$ is the cell frequency contained in the tabulation using original data. $T^{R}(c)$ is the cell frequency contained in the tabulation using recoded data, and n_T is the number of cells in the tabulation using original data. For $T^{R}(c)$, the cell frequencies in the table tabulated based on recoded categories are divided by the number of original categories.

Table 3 contains descriptive statistics on information loss for the 81 patterns of recoding and top-coding for each of the 20 areas. The results show that for smaller geographical areas, both average information loss and standard deviation of the information loss are smaller. This is due to the fact that cell frequency tends to be small for smaller geographical areas.

Figure 4 illustrates the relationship between geographical area size and average information loss within each area. Figure 5 presents the relationship between geographical area size and information loss for five-year age brackets and top coding for ages 85 years and above, 17 categories of industry and 10 categories of occupation. Figure 6 presents the relationship between geograph-

Table 3. Descriptive statistics about information loss for 20 areas

Area name	Geographical threshold	Average	S.D.	Min	Max	N
Area A	more than 200,000 persons	1.764	0.721	0	3.042	81
Area B	more than 100,000 persons	1.620	0.636	0	2.704	81
Area C	more than 50,000 persons	0.732	0.287	0	1.198	81
Area D	more than 50,000 persons	0.610	0.238	0	0.989	81
Area E	more than 30,000 persons	0.631	0.239	0	1.015	81
Area F	more than 30,000 persons	0.624	0.231	0	0.979	81
Area G	more than 20,000 persons	0.521	0.198	0	0.836	81
Area H	more than 20,000 persons	0.401	0.160	0	0.654	81
Area I	more than 20,000 persons	0.322	0.134	0	0.527	81
Area J	more than 10,000 persons	0.220	0.089	0	0.355	81
Area K	more than 10,000 persons	0.290	0.104	0	0.444	81
Area L	more than 10,000 persons	0.204	0.088	0	0.339	81
Area M	more than 10,000 persons	0.221	0.078	0	0.340	81
Area N	more than 5,000 persons	0.151	0.060	0	0.236	81
Area O	more than 5,000 persons	0.127	0.050	0	0.199	81
Area P	more than 5,000 persons	0.113	0.045	0	0.174	81
Area Q	more than 5,000 persons	0.099	0.040	0	0.156	81
Area R	more than 1,000 persons	0.076	0.033	0	0.120	81
Area S	more than 1,000 persons	0.058	0.026	0	0.093	81
Area T	more than 1,000 persons	0.046	0.022	0	0.074	81

Fig. 4. Relationship between area size and information loss: average information loss

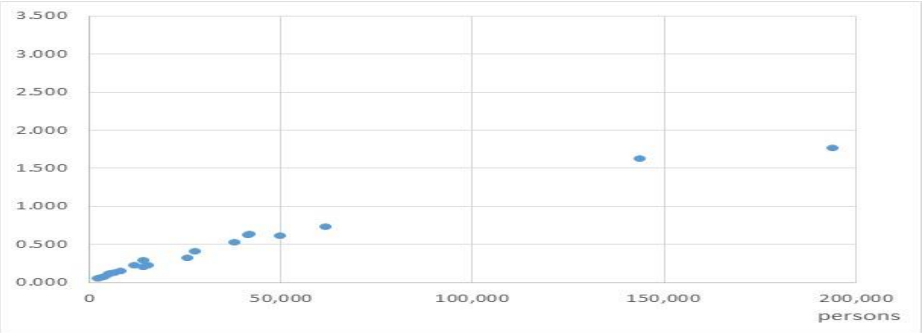


Fig. 5. Relationship between area size and information loss: information loss for five-year age brackets and top coding for 85 years and above, 17 categories of industry and 10 categories of occupation

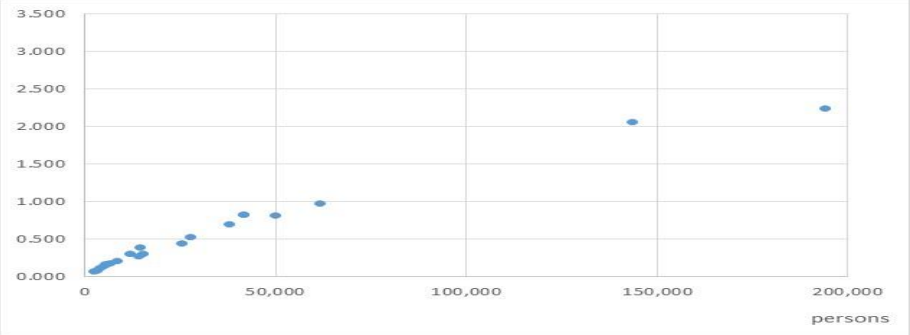
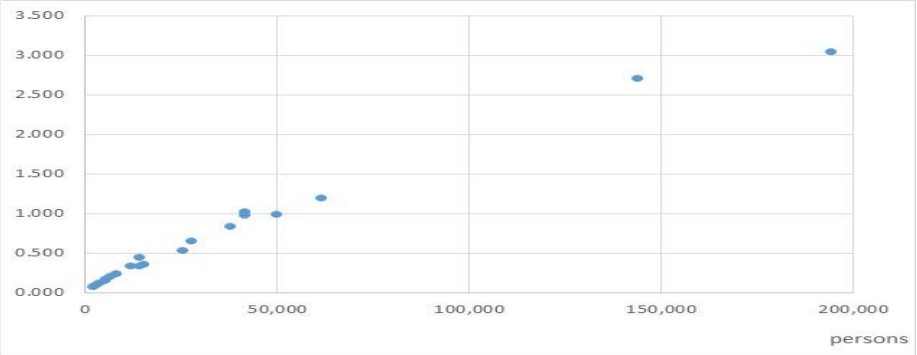


Fig. 6. Relationship between area size and information loss: information loss for ten-year age brackets and top coding for 90 years and above, 14 categories of industry and 8 categories of occupation



ical area size and information loss for ten-year age brackets and top coding for ages 90 years and above, 14 categories of industry and 8 categories of occupation. These results show that information loss tends to be higher for larger geographical areas. However, a comparison of the results in Figure 5 and Figure 6 for smaller geographical areas shows that the difference in information loss is comparatively narrow. This result suggests that when releasing data for smaller geographical areas, more broadly recoded categories for age can be adopted without impacting information loss and therefore usability of the anonymized microdata[4].

4 Conclusion

This paper uses geographical thresholds to create anonymized official microdata from Japanese Population Census data, and assesses population unique ratio and information loss for this data. The results empirically show that there is a trade-off between geographical area size and number of population uniques. This research also confirms that if area size is larger, information loss tends to be higher, while the influence of recoding and top coding on information loss for anonymized microdata differs based on the characteristics of the variables for which recoding is performed.

It is hoped that this research will contribute to the provision of different types of Anonymized microdata e.g. with more smaller area information, that will allow researchers from a variety of fields including economics, sociology, demography, geography etc. to conduct more detailed statistical analysis based on official statistics in Japan.

Note

The opinions expressed in this paper do not necessarily reflect those of organizations to which the authors belong or those of the Statistics Bureau of Japan or the National Statistics Center.

References

1. Dale, A. (1995) “Samples of Anonymised Records from the 1991 Census for Great Britain” *IASSIST Quarterly*, pp.5-12.

[4] Entropy-based information loss was calculated based on Ito et al. (2015), and the relationship between geographical area size and entropy-based information loss was evaluated in the same way as Figures 4,5, and 6. The results showed almost no difference to the results obtained based on the average absolute distance per tabulation cell used in this research.

2. De Waal, T. and Willenborg, L. (1999) “Information Loss Through Global Recoding and Local Suppression”, *Netherlands Official Statistics (special issue on SDC)*, Vol.14, pp.17-20.
3. Domingo-Ferrer, J. and Torra, V. (2001) “Disclosure Control Methods and Information Loss for Microdata”, In *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies* (P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz, eds.), Elsevier Science, Amsterdam, pp. 91-110.
4. Hawala, S.(2001) “Enhancing the "100,000 rule" On the Variation of the Per Cent of Uniques in A Microdata Sample and the Geographic Area Size Identified on the File”, Proceedings of the Annual Meeting of the American Statistical Association.
5. Ito, S. and Murata, M. (2011) “Quantitative Methods to Assess Data Confidentiality and Data Utility for Microdata in Japan”, Paper presented at Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Tarragona, Spain, pp.1-10.
6. Ito, S. and Hoshino, N.(2012) “The Potential of Data Swapping as a Disclosure Limitation Method for Official Microdata in Japan: An Empirical Study to Assess Data Utility and Disclosure Risk for Census Microdata” Paper presented at Privacy in Statistical Databases 2012, Palermo, Sicily, Italy, pp.1-13.
7. Ito, S. and Hoshino, N.(2013) “Assessing the Effectiveness of Disclosure Limitation Methods for Census Microdata in Japan” Paper presented at Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Ottawa, Canada, pp.1-10.
8. Ito, S. and Hoshino, N.(2014) “Data Swapping as a More Efficient Tool to Create Anonymized Census Microdata in Japan”, Paper presented at Privacy in Statistical Databases 2014, Ibiza, Spain, pp.1-14.
9. Ito, S., Hoshino, N., Akutsu, F. (2015) “A Quantitative Assessment of Data Confidentiality and Data Utility to Create Anonymized Census Microdata in Japan”, Paper presented at Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Helsinki, Finland, pp. 1-14.
10. Kooiman, P., L. Willenborg and J. Gouweleeuw (1998) “PRAM: A Method for Disclosure Limitation of Microdata”, Research Paper, No. 9705, Statistics Netherlands, Voorburg.
11. Marsh, C., Dale, A., Skinner, C.(1994) “Safe Data versus Safe Settings: Access to Microdata from the British Census”, *International Statistical Review*, Vol.62, No.1, pp.35-53.
12. Shlomo, N., Tudor, C., and Groom, P. (2010) Data swapping for protecting census tables. In *Privacy in Statistical Databases UNESCO Chair in Data Privacy International Conference, PSD 2010 Corfu, Greece, September, 2010 Proceedings* (J. Domingo-Ferrer and E. Magkos, eds), New York: Springer, pp. 41-51.
13. Tranmer, M., Pickles, A., Fieldhouse, E., Elliot, M., Dale, A., Brown, M.(2005) “The Case for Small Area Microdata”, *Journal of Royal Statistical Society A*, Vol.168, pp.29-49.

Appendix Table 1. Population unique ratios for different patterns for recoding and top coding: Area A

One-year age brackets	One-year age brackets and top coding for 85 years and above	One-year age brackets and top coding for 90 years and above	One-year age brackets and top coding for 95 years and above	Five-year age brackets and top coding for 85 years and above	Five-year age brackets and top coding for 90 years and above	Five-year age brackets and top coding for 95 years and above	Ten-year age brackets and top coding for 90 years and above	Ten-year age brackets and top coding for 100 years and above	Industry			Occupation			Population unique ratio
									21 Catego-ries	17 Catego-ries	14 Catego-ries	12 Catego-ries	10 Catego-ries	8 Catego-ries	
*									*			*			18.43%
*									*				*		18.33%
*									*					*	17.64%
*										*		*			18.21%
*										*			*		18.11%
*										*				*	17.41%
*											*	*			17.47%
*											*		*		17.35%
*											*			*	16.63%
	*								*			*			18.24%
	*								*				*		18.14%
	*								*					*	17.45%
	*									*		*			18.02%
	*									*			*		17.92%
	*									*				*	17.22%
	*										*	*			17.28%
	*										*		*		17.16%
	*										*			*	16.43%
		*							*			*			18.37%
		*							*				*		18.26%
		*							*					*	17.58%
		*								*		*			18.15%
		*								*			*		18.04%
		*								*				*	17.34%
		*									*	*			17.40%
		*									*		*		17.28%
		*									*			*	16.56%
			*						*			*			18.41%
			*						*				*		18.31%
			*						*					*	17.62%
			*							*		*			18.19%
			*							*			*		18.09%
			*							*				*	17.39%
			*								*	*			17.45%
			*								*		*		17.33%
			*								*			*	16.60%
				*					*			*			8.83%
				*					*				*		8.75%
				*					*					*	8.27%
				*						*		*			8.69%
				*						*			*		8.61%
				*						*				*	8.13%
				*							*	*			8.13%
				*							*		*		8.04%
				*							*			*	7.56%

Note '*' denotes the combination of recoding and top coding selected in this research.

Appendix Table 1. Population unique ratios for different patterns for recoding and top coding: Area A (Continued)

One-year age brackets	One-year age brackets and top coding for 85 years and above	One-year age brackets and top coding for 90 years and above	One-year age brackets and top coding for 95 years and above	Five-year age brackets and top coding for 85 years and above	Five-year age brackets and top coding for 90 years and above	Five-year age brackets and top coding for 95 years and above	Ten-year age brackets and top coding for 90 years and above	Ten-year age brackets and top coding for 100 years and above	Industry			Occupation			Population unique ratio
									21 Catego-ries	17 Catego-ries	14 Catego-ries	12 Catego-ries	10 Catego-ries	8 Catego-ries	
					*				*			*			8.86%
					*				*				*		8.78%
					*				*					*	8.30%
					*					*		*			8.72%
					*					*			*		8.64%
					*					*				*	8.16%
					*						*	*			8.16%
					*						*		*		8.07%
					*						*			*	7.59%
						*			*			*			8.87%
						*			*				*		8.79%
						*			*					*	8.31%
						*				*		*			8.73%
						*				*			*		8.64%
						*				*				*	8.17%
						*					*	*			8.17%
						*					*		*		8.08%
						*					*			*	7.60%
							*		*			*			6.45%
							*		*				*		6.38%
							*		*					*	6.01%
							*			*		*			6.34%
							*			*			*		6.27%
							*			*				*	5.91%
							*				*	*			5.88%
							*				*		*		5.80%
							*				*			*	5.45%
								*	*			*			6.45%
								*	*				*		6.38%
								*	*					*	6.02%
								*		*		*			6.34%
								*		*			*		6.27%
								*		*				*	5.91%
								*			*	*			5.88%
								*			*		*		5.81%
								*			*			*	5.45%

Note '*' denotes the combination of recoding and top coding selected in this research.