Investigating New Methods for Creating Anonymized Microdata Based on Japanese Census Data

Shinsuke Ito, Chuo University, Japan

Naomi Hoshino, National Statistics Center, Japan

Fumika Akutsu, Statistics Bureau of Japan

Ryo Kikuchi, NTT Secure Platform Laboratories

1. Introduction: Anonymized Census Microdata in Japan
2. The Methodology of Data Swapping
3. Data Swapping for Japanese Census Microdata
4. Applicability of PRAM to Create Anonymised Census Microdata
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 6 types of Anonymized microdata from Japanese official statistics including the Population Census.

Current Situation for Population Census Data:

- 2000 and 2005 census are currently available.
- Limited geographical information (prefecture level and geographical areas with 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

・In order to promote a broader use of Anonymized official microdata, several empirical studies on the effectiveness of disclosure limitation methods for official microdata have been conducted by the National Statistics Center (Ito and Murata (2011), Ito and Hoshino (2012, 2013, 2014)) and by the Statistics Bureau of Japan and the National Statistics Center (Ito et al. (2015, 2016)).

・For data from the 2010 Census, the Statistics Bureau is aiming to provide access to anonymized microdata that include smaller area data while maintaining data confidentiality.

・This research investigates new methods for creating anonymized microdata from Japanese census data, and focuses on perturbative methods including data swapping and PRAM (Post RAndomization Methods). The effectiveness of each method was determined by calculating data confidentiality and information loss.

2. The Methodology of Data Swapping

・Studies on the potential of data swapping as a disclosure limitation method for microdata include Dalenius and Reiss (1978), Moore (1996), Gomatam and Karr (2003), Nin et al. (2008) and Shlomo et al. (2010).

・Takemura (2002), Ito and Hoshino (2012) and Ito and Hoshino (2013) have conducted empirical research on the effectiveness of data swapping specifically for Japanese microdata.

・The U.S. Census Bureau has developed the methodology of n-cycle swapping and conducted empirical research about this methodology (Depersio et al. (2012)).

2. The Methodology of Data Swapping

Four sets of data from the 2010 Population Census were created and used as test data for this research.

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

Set 2: Based on more than 500,000 records of individual data from another geographic area within the same prefecture (“Data B”).

Set 3: Based on more than 200,000 records of individual data from a third geographic area within the same prefecture (“Data C”).

Set 4: Based on more than 200,000 records of individual data from a fourth geographic area within the same prefecture (“Data D”).

Table 1: The Characteristics of Data A, B, C and D from the 2010 Population Census

	Number of Records
Data A	277,665
Data B	257,451
Data C	85,640
Data D	76,442

2. The Methodology of Data Swapping

Data swapping was conducted based on the following steps:

Step 1: Population uniques for every combination of patterns for the following key variables were calculated.

Key Variables*:

- Type and Tenure of Dwelling (5 categories)
- Type of Building and Total Number of Floors (3 categories)
- Sex (2 categories)
- Marital Status (4 categories)
- Nationality (2 categories)
- Type of (Work) Activity (7 categories)
- Employment Status (5 categories)
- Age (19 categories)
- Industry (16 categories)
- Occupation (7 categories)

*The categories of the 10 key variables for which the percentage of population uniques was identical or almost identical to those for the categories of the 10 key variables included in the Anonymized microdata from the 2000 and 2005 Population Census were selected.

2. The Methodology of Data Swapping

Step 2: Records that correspond to unique cells for the various combinations of the 10 key variables were selected as target records for data swapping.

In order to determine the degree of priority for data swapping, cross-tabulation was conducted for all combinations of the 3 key variables*.

*The number of times a specific record corresponded to a unique cell was counted for each combination of cross-tabulations, and this score was added to each record in the test data. Records for which the score was high were classified as 'risky' records with a higher priority for data swapping (Elliot *et al*. (2002)).

2. The Methodology of Data Swapping

Step 3: targeted data swapping was performed for records with a score of 1 or higher. Targeted data swapping was performed for records that corresponded to the top p% (p=0.1 in this study) of the group, and was performed in order of descending score.

*__Donor file records were selected from different areas in order of descending area size__.

**__Each donor file record was used only once__.

The distances between each target record and all donor file records were calculated within groups clustered based on 10-year age groups, 3 categories of industry and 2 categories of occupation, and the nearest donor file record was swapped within these groups. In case of multiple records with identical distances, the donor file record was selected randomly from among these records.

Calculate the Distance between a Swapped Record and its Nearest Donor File Records

Record Linkage Technique

(1) Determine degree to which target records for data swapping and donor file records match.

(A) The score is 1 if the value of the key variable in the target record match the value in the donor file record, otherwise it is 0.

(B) The score is divided by the number of categories for the key variable.

(2) Scores for each of the above 10 key variables were calculated, and results were added to calculate the distance between target records and donor file records.

Table 2: The results of data swapping for Data A, B, C, and D.

Data A	Number of Records	Number of swapped Records (Total)	Number of Swapped Records using Donor Files from Data B	Number of Swapped Records using Donor Files from Data C	Number of Swapped Records using Donor Files from Data D
	277,665	278	195	40	43

Data B	Number of Records	Number of swapped Records (Total)	Number of Swapped Records using Donor Files from Data C	Number of Swapped Records using Donor Files from Data D	Number of Records used as Donor Files for Data Swapping in Data A
	257,451	257	129	96	32

For Data A, the rate of targeted records to swapped records using donor files from Data B is 70.2%. This shows that swapping took place not only for Data B but also Data C and D.

3. Data Swapping for Japanese Census Microdata

Disclosure risk exists not only for matching with external data, but also for records that are special uniques. Data utility and disclosure risk were calculated as part of this research.

- The indicators of data utility (DU) and disclosure risk (DR) were calculated and compared for different data*.

- DU and DR were determined based on tables of all possible three-variable combinations of the 10 key variables.

*Based on Shlomo *et al*. (2010)

Table 3: Information loss for five-year age groups and top coding for 90 years and older, 16 categories of industry and 7 categories of occupation

	Data A	Data B
Distance	248	400
DU	0.12	0.19

Table 4: Information loss for five-year age groups and top coding for 90 years and older, 3 categories of industry and 2 categories of occupation

	Data A	Data B
Distance	10	22
DU	0.08	0.17

Result suggests that if swapping is conducted involving donor file records clustered based on ten-year age groups, categories of industry and 2 categories of occupation, information loss within these clusters is likely to be small.

Table 5: The Result of DR, Data A (Excerpt)

Type of Tenure and Dwelling	Type of Building and Total Number of Floors	Sex	Marital Status	Nationality	Type of (Work) Activity	Employed Status	Age	Industry	Occupation	Number of Population Uniques	Number of Swapped Records	DR
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
*					*		*			67	41	0.612
*					*			*		13	9	0.692
*					*				*	6	4	0.667
*						*	*			30	21	0.700
*						*		*		18	16	0.889
*						*			*	3	3	1.000
*						*		*		18	16	0.889
*						*			*	3	3	1.000
*								*	*	29	20	0.690
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
					*		*	*		68	36	0.529
					*		*		*	19	16	0.842
					*			*	*	31	12	0.387
						*	*	*		110	60	0.545
						*	*		*	34	24	0.706
						*		*	*	36	16	0.444
							*	*	*	198	66	0.333

Results indicate that the rate of swapped unique cells to special unique cells contained in the tabulation is different according to the type of variables selected.

4. Applicability of PRAM to create Anonymized Census Microdata

Post RAndomization Method (PRAM) was first proposed by Kooiman et al (1998). It changes values within a data set on a transition probability matrix.

・In official statistics, PRAM has been used as a method for statistical disclosure control. A detailed description and empirical study of PRAM appears in de Wolf et al. (1998) and de Wolf and van Gelder (2004).

・PRAM-like approaches have been studied in the context of privacy preserving data mining. Agrawal and Srikant (2000) and Agrawal et al. (2005) have independently developed privacy preserving OLAP (Online Analytical Processing) via retention-replacement perturbation, which is an instantiation of PRAM.

・In this paper, PRAM with retention-replacement perturbation was applied to official microdata from the Population Census.

The Characteristics of PRAM in this reserarch

(1) Retention-replacement Perturbation

In retention-replacement perturbation, individuals' data is probabilistically replaced with random data using given retention probability ρ.

For example, for the attribute "sex," when ρ=0.5, "male" is retained with 1/2 probability, and with the remaining 1/2 probability, it is replaced with a uniformly random value, namely, a value "female" and a value "male," which is the same as the original, both with $1/2 \times 1/2 = 1/4$ probability.

(2) Reconstruction

・Reconstruction was conducted to estimate the cross tabulation of the original microdata from the one of a released microdata.

・The original cross tabulation was estimated by increasing gaps among the values of cross tabulation.

Application of PRAM for Census Microdata

In this experiment, PRAM was conducted to create anonymized Census microdata.

・PRAM was applied for each age, industry and occupation in this research.

- ・Attributes of age were clustered into 10-year age groups and the attributes within each age group were perturbed.
- ・Attributes of industry were clustered into 3 categories and the attributes within each industry group were perturbed.
- ・Attributes of occupation were clustered into 2 categories and the attributes within each occupation group were perturbed.

The retention probabilities ρ, 0.95, 0.90, 0.85, and 0.80 were applied.

*__**Reconstruction was not used in this experiment.**__

Table 6 Information loss for five-year age groups and top coding for 90 years and older, 16 categories of industry and 7 categories of occupation, Data A

Retention Probability	0.95	0.90	0.85	0.80
Distance	1808	2985	4019	4868
DU	0.84	1.39	1.87	2.27

Table 7 Information loss for five-year age groups and top coding for 90 years and older, 3 categories of industry and 2 categories of occupation, Data A

Retention Probability	0.95	0.90	0.85	0.80
Distance	205	361	491	564
DU	1.54	2.71	3.69	4.24

5. Conclusion and Outlook

(1) This paper assesses the effectiveness of data swapping and PRAM as perturbative methods for creating anonymized microdata from Japanese Population Census data.

(2) The results demonstrate that conducting data swapping among four areas using the nearest donor file records allows to achieve low disclosure risk and low information loss.

(3) The results also show that information loss for data swapping is lower than information loss for PRAM for smaller areas or more detailed categories for age, occupation and industry. This result demonstrates that data swapping is suitable for creating smaller area microdata within the setting described in this research.

(4) More precise comparison between data swapping and PRAM requires unifying security or utility levels as well as additional experiments with different parameters.

5. Conclusion and Outlook

(5)PRAM also could have potential for decreasing the risk of personal information being identified when special unique cells are contained in the publicly released data tables. Therefore, when creating anonymized official microdata, disclosure limitation methods should be applied according to the type of risky records.

(6) This research provides an approach for determining the most effective perturbative method for a particular data set, and thereby minimize disclosure risk for Japanese Census microdata including smaller area microdata.

It is hoped that the results from this research will contribute to the creation of a broader variety of anonymized Census microdata in Japan.