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

Shinsuke Ito[*], Naomi Hoshino[**], Fumika Akutsu[***], Ryo Kikuchi[****]

[*] Faculty of Economics, Chuo University, 742-1 Higashinakano, Hachioji, Tokyo, 192-0393 Japan,
E-mail: ssitoh@tamacc.chuo-u.ac.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 Statistics Bureau of Japan, and National Statistics Center.
[**] National Statistics Center, 19-1 Wakamatsu-cho, Shinjuku-ku, Tokyo, 162-8668 Japan,
E-mail: nsaitou2@nstac.go.jp
[***] Statistics Bureau of Japan, 19-1 Wakamatsu-cho, Shinjuku-ku, Tokyo, 162-8668 Japan,
E-mail: fakutsu@soumu.go.jp
[****] NTT Secure Platform Laboratories, 3-9-11 Midorimachi, Musashino, Tokyo, 180-8585 Japan,
E-mail: kikuchi.ryo@lab.ntt.co.jp
Ryo Kikuchi is a research fellow at the National Statistics Center.

Abstract: Anonymized microdata from the 2000 and 2005 'Population Census' are currently made available under the Statistics Act in Japan. However, only one type of Anonymized microdata is released, and this type does not contain detailed geographical information.

Several empirical studies on the effectiveness of disclosure limitation methods for official microdata such as microaggregation, additive noise, and data swapping have been conducted by the National Statistics Center in cooperation with the Statistics Bureau of Japan, including empirical research to create anonymized microdata that contain more detailed geographical information.

This research investigates new methods for creating anonymized microdata from individual data from the Population Census data. Perturbative methods including data swapping and PRAM (Post Randomization Method) were applied as disclosure limitation methods in order to create anonymized microdata for smaller geographic areas. To evaluate disclosure risk, a quantitative assessment of data confidentiality and data utility was conducted for the anonymized microdata created as part of this research.

1 Introduction

Japan's Statistics Act was revised in April 2007 for the first time in more than sixty years. Among the objectives of this revision was to promote the development and use of official statistics in order to contribute to the development of the national economy and the enhancement of citizen's living standards. Based on the Statistics Act, the 'Master Plan Concerning the Development of Official Statistics' was established. It contains a "secondary usage" system that includes the provision of disclosive data (individual data) and Anonymized microdata[1]. These developments

[1] 'Anonymized microdata' (with a capital "A") are defined as individual data that are 'processed so that no particular individuals or juridical persons, or other organizations shall be identified' (Article 36) and currently

form the starting point for the creation and release of Anonymized census microdata in Japan.

In recent years, 'Statistical Reform' has attracted attention in Japan. Statistical reform focuses on improving the usability of official statistical data as well as big data and administrative data to foster evidence-based policy making by the Japanese national and regional governments. Methods and approaches for creating and releasing microdata with the objective of promoting secondary use of Japanese official statistics are currently discussed.

At present, the Statistics Bureau releases 6 types of Anonymized microdata from Japanese official statistics including the Population Census. Anonymized Census microdata 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. Data swapping is applied as an additional perturbative method to create Anonymized Census microdata.

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)). 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, 2016)).

At present, the geographical classification contained in 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. However, smaller area microdata fulfil an important role as they allow researchers from a variety of fields including economics, sociology, demography and geography to use microdata for detailed statistical analysis. 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 the potential of perturbative methods such as data swapping for this purpose.

Specifically, 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.

released. All other individual data to which disclosure limitation methods have been applied are referred to as 'anonymized official microdata' (with a small "a").

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)).

As part of this research, an empirical study of data swapping that involved record swapping between different areas was conducted using data from the 2010 Population Census. Four sets of data from the 2010 Population Census – each containing a different number of records – were used as test data. The first set was created based on more than 500,000 records of individual data from a certain geographic area within a specific Japanese prefecture. This data set is referred to as "Data A". The second set of data was created based on more than 500,000 records of individual data from another geographic area within the same prefecture. This data set is referred to as "Data B". The third set of data was created based on more than 200,000 records of individual data from a third geographic area within the same prefecture. This data set is referred to as "Data C". The fourth set of data was created based on more than 200,000 records of individual data from a fourth geographic area within the same prefecture. This data set is referred to as "Data D". The detailed number of records contained in each data set are listed in Table 1.

Data swapping was conducted as follows: First, population uniques for every combination of patterns for the following key variables were calculated, and 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:

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)

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

Table 1 The Characteristics of Data A, B, C and D.

Occupation (7 categories)

Second, 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)).

Third, 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[2]. Donor file records were selected from different areas in order of descending area size. Each donor file record was used only once.

Fourth, 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.

Record linkage between target records and donor file records was conducted based on the calculated distance. Details of the record linkage technique (Domingo-Ferrer and Torra, 2001) used in this research are as follows:

i (= 1, …, m) and j (= 1, …, n) respectively were set as the index of the record to be swapped and the index of the donor file, where m and n are the number of swapped and donor file records, respectively, and k (= 1, …, 10) was set as the index of the key variable. The classification category of the key variable k for the ith record was designated as Cs_{ki}, and the classification category of the key variable k for the jth donor file record as Cd_{kj}. The distance Sd_{kij} between categorical variables for the key variable k between i and j was defined as Eq. (1) (Domingo-Ferrer and Torra, 2001).

[2] Among records with identical scores, records that refer to special uniques based on age, industry and/or occupation were given priority to be selected as target records.

$$Sd_{kij} = \left| Cs_{ki} - Cd_{kj} \right| \tag{1}$$

In this research, if $\left| Cs_{ki} - Cd_{kj} \right| > 0$, then $Sd_{kij} = 1$.

To convert the distance for categorical variables to a score, Sd_{kij} was divided by the classification category C_k of the *k*th key variable, and Eq. (2) was used to determine $Score_{kij}$ for the *k*th key variable.

$$Score_{kij} = \frac{1}{C_k} \cdot Sd_{kij} \tag{2}$$

By calculating the aggregate score for each key variable, a general index D_{ij} was derived for the distance between the *i*th and *j*th records for all key variables by Eq. (3).

$$D_{ij} = \sum_k Score_{kij} \tag{3}$$

The general index D_{ij} was calculated for the respective record distances between swapped records and donor files, and the swapped record replaced with the donor file for which D_{ij} is smallest. (Domingo-Ferrer and Torra, 2001; Takemura, 1999)[3]

Table 2 contains the results of data swapping adopted for Data A, B, C, and D. The results show that in the case of Data A, the rate of targeted records to swapped records using donor files from Data B is 70.2%. This result illustrates 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.

In this research, data utility was defined as the average absolute distance per tabulation cell, and therefore an indicator of distance that measures distortion to the distribution based on Shlomo et al. (2010). Data utility (DU) is given as:

[3] For multiple donor files with the same minimum distance, one of the files was selected randomly.

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

Table 2 The Result of data swapping adopted for Data A and B.

$$DU = \frac{\sum_c \left| T^P(c) - T^O(c) \right|}{n_T} \tag{4}$$

$T^O(c)$ is the cell frequency in the tabulation using original data and $T^P(c)$ is the cell frequency in the tabulation using swapped data, where n_T is the number of cells in the tabulation.

According to Shlomo et al. (2010), the measure of disclosure risk (DR) is given as:

$$DR = \frac{\sum_c I\left(T^O(c) = 1, T^P(c) = 1\right)}{\sum_c I\left(T^O(c) = 1\right)} \tag{5}$$

$\sum_c I\left(T^O(c) = 1\right)$ is the number of unique cells contained in the tabulation using original data. Also, $\sum_c I\left(T^O(c) = 1, T^P(c) = 1\right)$ is calculated as the number of perturbed unique cells in the tabulation.

Table 3 and 4 contain the information loss for swapped data. The results show that DU for five-year age groups and top coding for 90 years and older, 3 categories of industry and 2 categories of occupation is smaller than for broader categories of industry and occupation. This result suggests that if swapping involving donor file records that are clustered based on ten-year age groups, categories of industry and 2 categories of occupation is conducted, information loss within these clusters is likely to be small.

	Data A	Data B
Distance	248	400
DU	0.12	0.19

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	10	22
DU	0.08	0.17

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

Appendix Table 1 contains the DR for every combination of 3 variables selected among the 10 key variables. The 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 Anonymised 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. This procedure is referred to as perturbation. 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) independently developed privacy preserving OLAP (Online Analytical Processing) via retention-replacement perturbation, which is an instantiation of PRAM.

In addition to perturbation, there is a method to reconstruct the distribution of original data from perturbed data. This procedure is referred to as reconstruction. Van der Hout (1999) and Agrawal and Srikant (2000) have established a method for reconstructing cross tabulation.

In the past, it was difficult to compare PRAM with other methods with respect to privacy. However, PRAM has been proven to satisfy some security notions, such as conditional differential entropy (Agrawal and Aggarwal (2001)) and (s, ρ_1, ρ_2)-privacy breach. Among such notions, Lin et al. (2012) and Ikarashi et al. (2015) showed PRAM can satisfy ϵ-differential privacy (DP) (Dwork (2006)), and Ikarashi et al. (2015) also showed PRAM can satisfy Pk-anonimity (Ikarashi et al (2015)), which is the

probabilistic version of k-anonymity (Sweeny (2002)). This has made it possible to compare PRAM with other anonymization methods through ϵ-DP and k-anonymity.

Since PRAM is a probabilistic method, the utility of released data is evaluated through empirical studies. There have been several empirical studies to evaluate the utility of released data by using distance, such as Euclidian distance. Examples include de Wolf et al. (1998) and de Wolf and van Gelder (2004). Recently, a theoretical analysis of accuracy in a variant of PRAM was shown in Hasegawa et al. (2016) in which the confidence interval of accuracy of cross tabulation was provided.

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

4.1 PRAM

For simplicity, it was assumed that the original data is microdata and each record contains attributes corresponding to one individual. $\mathbb{V}$ is the set of possible values of records. PRAM changes data according to a transition probability matrix A. The transition probability matrix consists of probabilities in which each value in a private table will be changed into other specific (or the same) values. $A_{u,v}$ denotes the probability $u \in \mathbb{V}$ is changed into $v \in \mathbb{V}$. For example, $A_{\text{male,female}}$ means the probability of "male -> female".

PRAM is a fairly general method. Invariant PRAM (Kooiman et al, (1998)), retention-replacement perturbation (Agrawal et al. (2005)) are known as instantiations of it.

4.2 Retention-replacement Perturbation

In retention-replacement perturbation, individuals' data is probabilistically replaced with random data using given retention probability ρ. This perturbation is also known as "the fully filled matrices with equal off-diagonal element" (de Wolf and van Gelder (2004)). First, data are retained with ρ, and if the data are not retained, they will be replaced with a uniformly random value chosen from the attribute domain. Even if data are not retained, there still is a possibility that the data will not be changed, because the data value is included in the attribute domain as well as other values. 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 1/2×1/2=1/4 probability. Eventually, the probability that "male" changes into "female" is 1/4, and the probability that it does not change is 3/4. The lower the retention probability, the higher the level of privacy that is preserved. On the contrary, the lower the retention probability, the lower is data utility. These probabilities yield the following transition probability matrix.

$$\begin{pmatrix} 0.75 & 0.25 \\ 0.25 & 0.75 \end{pmatrix}$$

Generally, the transition probability matrix **A** is written as

$$A_{v,v'} = \begin{cases} \rho + \frac{1-\rho}{|\mathbb{V}|} & \text{if } v = v' \\ \frac{1-\rho}{|\mathbb{V}|} & \text{if } v \neq v' \end{cases},$$

where for any $v, v' \in \mathbb{V}$ and ρ is the retention probability.

4.3 Reconstruction

From a perturbed table and retention probability matrix A, the statistics of the original microdata can be estimated. Kooiman et al. (1998) and Agrawal (2000) proposed reconstruction to estimate the cross tabulation of the original microdata from the one of a released microdata. The retention-replacement perturbation changes the cross tabulation close to uniform distribution, so the original cross tabulation was estimated by increasing gaps among the values of cross tabulation. The algorithm of reconstruction is shown in the Appendix.

4.4 Difficulty of Managing Reconstructed Cross Tabulation

Although reconstruction can raise the precision of cross tabulation, it can also cause problems. In this research, classification, perturbation, and reconstruction were performed iteratively: records were classified by 10-year age into several groups, each group was perturbed, the cross tabulations were reconstructed from perturbed ones, the cross tabulations were converted into a tuple of microdata, they were concatenated into unified microdata, and then classified by job category and so on. In the above steps, cross tabulation was converted into tables. This is easy when the cross tabulation contains integers only. However, after reconstruction, the cross tabulation can contain decimals and therefore cannot be converted into microdata.

A naïve solution to manage the decimal is rounding, but this causes a change in the number of records. Another solution is computing the sum of decimals by $N - \sum_p \lfloor x_p \rfloor$ and then adding 1 to a value of the cross tabulation $N - \sum_p \lfloor x_p \rfloor$times, in order from the value whose decimal is larger. However, this solution causes the change of non-perturbed attributes. Therefore, reconstruction was not applied as part of this research.

4.5 The Result of PRAM for Census Microdata

In this research, PRAM was conducted to create anonymized Census microdata. PRAM was applied for each age, industry and occupation. 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. For retention probability ρ, 0.95, 0.90, 0.85, and 0.80 were applied.

Appendix Tables 2 to Appendix Table 7 contain the information loss for perturbed data. The results show that DU for five-year age groups and top coding for 90 years and older, 16 categories of industry and 7 categories of occupation is smaller than when broader categories for age, industry or occupation are used.

The information loss of perturbed data is larger than that of swapped data. In the experiments of swapping and PRAM, the number of swapped records were 0.1% of the total records, while the number of perturbed records were approximately 15% even for $\rho = 0.95$. Therefore, the difference in information loss may be caused by the difference in the guaranteed security level. In future research we will aim to unify the security level between these two methods.

5 Conclusion

This paper assessed the effectiveness of data swapping and PRAM as perturbative methods for creating anonymized microdata from Population Census data.

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. 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.

However, the security level of swapped/perturbed data was not unified, and reconstruction was not used in this research. Therefore, there exists a possibility that PRAM can create microdata with less information loss than data swapping. 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.

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.

References

Agrawal, R., Srikant, R. Privacy-Preserving Data Mining, SIGMOD 2000.

Agrawal, D., and Aggarwal, C. C. On the design and quantification of privacy preserving data mining algorithms. PODS. ACM, 2001.

Agrawal, R., Srikant, R., and Thomas, D. (2005) Privacy preserving OLAP. SIGMOD, ACM.

Dalenius, T and Reiss, S. P. (1978) Data-swapping: a technique for disclosure control (extended abstract). In Proc. Section on Survey Research Methods, pp. 191-194. Washington, D.C.: American Statistical Association.

DePersio, M, Lemons, M., Ramanayake, K. A., Tsay, J., Zayatz, L.(2012) n-Cycle Swapping for the American Community Survey, In *Privacy in Statistical Databases UNESCO Chair in Data Privacy International Conference, PSD 2012 Palermo, Italy, September, 2012 Proceedings*(eds Domingo-Ferrer, J. and Tinnirello, I.), Springer, pp.143-164.

de Wolf, P.P., Gouweleeuw, J.M., Kooiman, P., and Willenborg, L.C.R.J.(1998) Reflections on PRAM, Statistical Data Protection.

de Wolf, P.P. and van Gelder, I.(2004) An empirical evaluation of PRAM, Discussion paper 04012, Statistics Netherlands.

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* (eds Doyle P. et al.), pp. 91-110. Amsterdam: Elsevier Science.

Dwork, C. Differential privacy. ICALP, 2006.

Elliot, M., Manning, A. M., and Ford, R. W. (2002) A computational algorithm for handling the special uniques problem. *Int. J. Uncertainty Fuzziness Knowledge Based Syst*., 10(5), 493-509.

Gomatam, S. and Karr, A. F. (2003) Distortion measures for categorical data swapping. Technical Report, No. 131, National Institute of Statistical Sciences.

Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J., and de Wolf, P.P.(1997) Post Randomisation for Statistical Disclosure Control: Theory and Implementation, Research paper no. 9731, Statistics Netherlands.

Hasegawa, S., Masaki, S., Hamada, K., and Kikuchi, R. (2016) A theoretical analysis on accuracy of reconstruction in probabilistic k-anonymization, Symposium on Cryptography and Information Security.

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.

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.

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.

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.

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.

Ito, S., Hoshino, N., Akutsu, F. (2016) Potential of Disclosure Limitation Methods for Census Microdata in Japan, Paper presented at Privacy in Statistical Databases 2016, Dubrovnik, Croatia, pp.1-14.

Ikarashi, D., Kikuchi, R., Chida, K., and Takahashi, K. k-anonymous Microdata Release via Post Randomisation Method, IWSEC 2014.

Kooiman, P., L. Willenborg and J. Gouweleeuw (1998) PRAM: A Method for Disclosure Limitation of Microdata, Research Paper, No. 9705, Statistics Netherlands, Voorburg.

Lin, B.R., Wang, Y., and Rane, S. A framework for privacy preserving statistical analysis on distributed databases. WIFS, 2012.

Moore, R. A. (1996) Controlled data-swapping techniques for masked public use microdata sets. Statistical Research Division Report Series, RR 96-04, U.S. Bureau of the Census, Statistical Research Division, Washington, D. C.

Müller, W., Blien, U., and Wirth, H. (1995) Identification risks of micro data: evidence from experimental studies. *Socio. Meth. Res*., 24(2), 131-157.

Nin, J., Herranz, J., and Torra, V. (2008) Rethinking rank swapping to decrease disclosure risk. *Data Knowl. Engineering*, 64(1), 346-364.

Shlomo, N.(2007) "Statistical Disclosure Control Methods for Census Frequency Tables", S3RI Methodology Working Papers M07/04, pp.1-40.

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* (eds J. Domingo-Ferrer and E. Magkos), pp. 41-51. New York: Springer.

Sweeney, L.(2002) k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5):557–570.

Takemura, A. (2002) Local recoding and record swapping by maximum weight matching for disclosure control of microdata set. *J. Off. Stat*., 18(2), 275-289.

Zayatz, L. (2007) "Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update", *Journal of Official Statistics*, Vol.23, No.2, pp.253-265.

Willenborg, L. and De Waal, T. (2001) *Elements of Statistical Disclosure Control*, New York: Springer.

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
*	*	*								0	0	-
*	*		*							0	0	-
*	*			*						1	0	0.000
*	*				*					2	0	0.000
*	*					*				1	0	0.000
*	*						*			10	4	0.400
*	*							*		5	1	0.200
*	*								*	2	0	0.000
*		*	*							0	0	-
*		*		*						0	0	-
*		*			*					0	0	-
*		*				*				1	1	1.000
*		*					*			5	4	0.800
*		*						*		3	2	0.667
*		*							*	1	1	1.000
*			*	*						0	0	-
*			*		*					4	3	0.750
*			*			*				3	2	0.667
*			*				*			11	6	0.545
*			*					*		9	7	0.778
*			*						*	3	3	1.000
*				*	*					2	1	0.500
*				*		*				4	2	0.500
*				*			*			10	6	0.600
*				*				*		5	3	0.600
*				*					*	3	1	0.333
*					*	*				2	1	0.500
*					*		*			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
	*	*	*							0	0	-
	*	*		*						0	0	-
	*	*			*					0	0	-
	*	*				*				0	0	-
	*	*					*			6	5	0.833
	*	*						*		3	2	0.667
	*	*							*	1	1	1.000
	*		*	*						0	0	-
	*		*		*					5	4	0.800
	*		*			*				4	2	0.500
	*		*				*			20	8	0.400
	*		*					*		14	10	0.714
	*		*						*	3	3	1.000
	*			*	*					2	1	0.500
	*			*		*				3	1	0.333
	*			*			*			13	7	0.538
	*			*				*		5	4	0.800
	*			*					*	3	2	0.667
	*				*	*				3	2	0.667
	*				*		*			74	43	0.581
	*				*			*		17	11	0.647
	*				*				*	6	5	0.833
	*					*	*			41	25	0.610
	*					*		*		25	17	0.680

Note '*' denotes the combination of variables selected in this research.
Appendix Table 1 The Result of DR, Data A

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
	*					*			*	4	2	0.500
	*						*	*		118	51	0.432
	*						*		*	40	28	0.700
	*							*	*	37	28	0.757
		*	*	*						0	0	-
		*	*		*					1	1	1.000
		*	*			*				0	0	-
		*	*				*			5	3	0.600
		*	*					*		0	0	-
		*	*						*	0	0	-
		*		*	*					0	0	-
		*		*		*				1	1	1.000
		*		*			*			4	3	0.750
		*		*				*		1	1	1.000
		*		*					*	0	0	-
		*			*	*				0	0	-
		*			*		*			19	13	0.684
		*			*			*		1	1	1.000
		*			*				*	0	0	-
		*				*	*			9	7	0.778
		*				*		*		3	3	1.000
		*				*			*	0	0	-
		*					*	*		29	25	0.862
		*					*		*	8	5	0.625
		*						*	*	6	3	0.500
			*	*	*					3	1	0.333
			*	*		*				2	1	0.500
			*	*			*			5	3	0.600
			*	*				*		8	3	0.375
			*	*					*	2	1	0.500
			*		*	*				1	1	1.000
			*		*		*			42	24	0.571
			*		*			*		3	2	0.667
			*		*				*	2	1	0.500
			*			*	*			22	14	0.636
			*			*		*		12	8	0.667
			*			*			*	1	0	0.000
			*				*	*		65	44	0.677
			*				*		*	29	21	0.724
			*					*	*	21	11	0.524
				*	*	*				0	0	-
				*	*		*			17	10	0.588
				*	*			*		8	3	0.375
				*	*				*	2	0	0.000
				*		*	*			16	9	0.563
				*		*		*		6	5	0.833
				*		*			*	3	1	0.333
				*			*	*		32	20	0.625
				*			*		*	17	11	0.647
				*				*	*	13	7	0.538
					*	*	*			13	10	0.769
					*	*		*		8	6	0.750
					*	*			*	2	1	0.500
					*		*	*		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

Note '*' denotes the combination of variables selected in this research.
Appendix Table 1 The Result of DR (Continued), Data A

Retention Probability	0.95	0.9	0.85	0.8
Distance	1808	2985	4019	4868
DU	0.84	1.39	1.87	2.27

Appendix Table 2 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	0.9	0.85	0.8
Distance	1343	2453	3421	4263
DU	1.19	2.17	3.03	3.77

Appendix Table 3 Information loss for ten-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	0.9	0.85	0.8
Distance	205	361	491	564
DU	1.54	2.71	3.69	4.24

Appendix 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

Retention Probability	0.95	0.9	0.85	0.8
Distance	1757	2811	3899	4859
DU	0.82	1.31	1.82	2.26

Appendix Table 5 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 B

Retention Probability	0.95	0.9	0.85	0.8
Distance	1357	2297	3297	4280
DU	1.20	2.03	2.92	3.79

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

Retention Probability	0.95	0.9	0.85	0.8
Distance	216	339	542	544
DU	1.62	2.55	4.08	4.09

Appendix 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 B

Algorithm of reconstruction

The iterative Bayesian technique for the algorithm of reconstruction. The algorithm is in fact the maximum likelihood estimation by using expectation-maximization algorithm. Let y be the cross tabulation of released table, y_q be the q-th element in y, and x^i be the i-th estimated cross tabulation. The cross tabulation of the original table is estimated by

$$x^{i+1} = \sum_{q<M} y_q \frac{A_{p,q} x_p^i}{\sum_{r<M} A_{r,q} x_r^i},$$

where M be the length of the cross tabulation.

Let $\cdot$ denote the inner product, / denote the element-wise division. The algorithm that computes the above iterative update [] is as follows.

$x_0 \coloneqq y$

$i \coloneqq 0$

repeat

$x^{i+1} \coloneqq x^i \cdot (A(y/(x^i A)))$

$i \coloneqq i + 1$

until $\mathrm{cnv}(x^i, x^{i-1}) = \mathrm{true}$

return x^i,

where $\mathrm{cnv}(x^i, x^{i-1})$ is $[\sum_{p<M} |x^i - x^{i-1}| < \epsilon N]$ with preliminary-defined constant ϵ.