Data Swapping as a More Efficient Tool to Create Anonymized Census Microdata in Japan

Shinsuke Ito[1,2] Naomi Hoshino[1]

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

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

Abstract. Multiple disclosure limitation methods including non-perturbative methods such as deletion of unique records are currently used to create Anonymized microdata from the 'Population Census' in Japan. Data swapping is a possible alternative to non-perturbative methods, as it can significantly reduce the need for deletion of records. This paper explores the potential of data swapping as a more efficient tool for the creation of anonymized Census microdata. Using test data created based on individual data from the 'Population Census', several types of data swapping are conducted at different swapping rates, and disclosure risk and data utility are calculated and compared for the swapped data.

Keywords: Data Swapping, Census Microdata, Unique Records, Matching, R-U Confidentiality Map

1 Introduction

In Japan, following the revision of the Statistics Act, Anonymized microdata from official statistics have been released since April 2009[1]. Currently, seven types of

[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 released. All other individual data to which disclosure

adfa, p. 1, 2014.

Anonymized microdata from official statistics such as the ‘Survey on Time Use and Leisure Activities’ and the ‘Employment Status Survey’ are available, including Anonymized microdata from the 2000 and 2005 ‘Population Census’ conducted by the Statistical Bureau of Japan.

The Anonymized microdata from the ‘Population Census’ that are currently made available contain more detailed geographical information than other Anonymized microdata released in Japan. However, this information is available only for prefectures and municipalities with a population of 500,000 or more. Also, data is available only at the level of household unit, and based on a sampling rate of 1%[2].

To create Anonymized microdata from the ‘Population Census’, the Statistical Bureau of Japan uses non-perturbative methods such as sampling, recoding, top-coding and bottom-coding as well as deletion of direct identifiers such as individual names or addresses. For example, individuals’ age is recoded and therefore available only in five-year brackets. Unique records, records which correspond to unique cells within tables and records which correspond to cells with a frequency count of 2 within tables are deleted. In addition, a specific method of data swapping[3] is applied.

However, the deletion of records requires significant time and effort to identify the cells in question. Several types of data swapping not currently used by the Statistical Bureau can reduce the need for deletion of records, and therefore offer a potential alternative to non-perturbative methods. An additional benefit of these types of data swapping is the ability to create Anonymized microdata with more detailed categories of attributes or more detailed geographical information, which would allow e.g. the provision of Anonymized microdata for small areas.

This paper explores the potential of data swapping as a more efficient tool for the creation of Anonymized official microdata from the ‘Population Census’. Towards this objective, several types of data swapping were performed at different swapping rates, and disclosure risk and data utility were calculated and compared for the swapped data.

2 Methodology

Two sets of data from the 2005 Population Census which differ in the number of records were used to examine the influence of record numbers on disclosure risk and data utility. The first set consists of approximately 50,000 records of individual data from a certain geographic area within a specific Japanese prefecture. This area is referred to as “Area A”. The second set of data consists of approximately 10,000 rec-

---

limitation methods have been applied are referred to as ‘anonymized official microdata’ (with a small “a”).

2 This procedure is in accordance with the “Guidelines for Creation and Release of Anonymized Microdata”.

3 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), 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 in Japan.

ords of individual data from another geographic area within the same prefecture. This area is referred to as "Area B". The test data is based on household units, but contains only records for the head of household.

'Age' was recoded, resulting in two types of test data: Five-year age brackets and one-year age brackets. Anonymized microdata currently released contains only five-year age brackets.

Sampling was conducted at the rate of 10%, and several types of data swapping were performed.

The detailed process for data swapping was as follows: First, the number of sample uniques was calculated based on the following key variables:

Gender (2 categories)
Age (25 categories for five-year age brackets and 122 categories for one-year age brackets)
Marital Status (5 categories)
Nationality (13 categories)
Type of (Work) Activity (9 categories)
Employment Status (8 categories)
Industry (19 categories)
Occupation (10 categories)
Type and Tenure of Dwelling (9 categories)
Type of Building and Total Number of Floors (5 categories) (30 sub-categories in the case of 'apartment house or flat')
Floor Number of Household (30 categories in the case of 'apartment house or flat')

Second, records that correspond to unique cells for the various combinations of the 11 key variables were selected as target records. In order to determine the degree of priority for data swapping, cross-tabulation was conducted for all combinations of key variables. The number of times a specific record corresponds to a unique cell for every combination of cross-tabulations was calculated, and this score was added to every record in the test data. Records for which the score was high were regarded as 'risky' records with a higher priority for data swapping (Elliot *et al*. (2002))[4].

Third, data swapping was performed for records with a score of 1 or higher. Data swapping was performed as (1) targeted data swapping, (2) random data swapping and (3) a combination of targeted and random data swapping at a swapping rate p.

[4] The averages of the scores for 10% sampled test data were calculated and descriptive statistics for the averages of the scores calculated for every combination of cross-tabulations. In the case of five-year age brackets, for "Area A" the average score was 556, the median score was 478, the maximum score was 1,772 and the minimum score was 21, whereas for "Area B", the average score was 749, the median score was 707, the maximum score was 1,828 and the minimum score was 58. In the case of one-year age brackets, for "Area A" the average score was 655, the median score was 602, the maximum score was 1,818 and the minimum score was 42, and for "Area B" the average score was 859, the median score was 830, the maximum score was 1,858 and the minimum score was 114.

Targeted data swapping was performed for records that correspond to the top p% (p=1, 5, 10) of the group, and in order of descending score. Random data swapping was performed for target records randomly selected from records with a score of 1 or higher. For the combination of targeted and random data swapping, first targeted data swapping was performed for records that correspond to the top 1/2 p% of the group, and then random data swapping was performed based on a swapping rate 1/2 p% for target records randomly selected from records except those for which targeted swapping had already been performed.

Donor file records were selected from an area referred to as "Area C" that is different to "Area A" and "Area B", but still located within the same prefecture. Approximately 5,000 records were selected from this file.

In order to perform targeted data swapping and random data swapping, the distance between each target record and all donor file records was calculated, and the nearest record in the donor file was swapped. In case of multiple records with identical distances, the partner record was randomly selected from among these records. Based on the distance calculated, record linkage (Domingo-Ferrer and Torra (2001)) between target records and donor file records was performed. Details of the procedure for calculating the distance are as follows:

First, the degree of matching between target records and donor file records determined. For nominal variables except age, the score was calculated so that the score is 1 if the values of the key variables in the target records match the values in the donor file records, otherwise it is 0. This score was then divided by the number of categories for the key variables.

For ordinal variables, the score was calculated by subtracting the values of the target records from the values of the donor file records. The absolute values of these results were then divided by the number of categories.

Second, this score was calculated for each of the above 11 key variables, and the results were added to determine the distance between target records and donor file records[5].

Table 1 contains the descriptive statistics for the distances between target records and donor file records. For random data swapping, the minimum distance is close to 0. This result shows that target records are swapped with donor file records with largely identical values for key variables. For targeted data swapping, the minimum distance is larger than for random data swapping. This result shows that for targeted data swapping, target records are swapped with donor file records with more distant values of key variables than for random data swapping. This suggests that in order to ensure optimum effectiveness, the type of data swapping should be selected based on the distance between target records and donor file records.

---

5 This process is basically identical to Ito and Hoshino (2012) and Ito and Hoshino(2013).

Table 1. Descriptive Statistics about the Distances between Target Records and Donor File Records

Five-Year Age Brackets

	Random Data Swapping											
Swapping Rate	1%				5%				10%			
Descriptive Statistics	Average	Median	Maximum	Minimum	Average	Median	Maximum	Minimum	Average	Median	Maximum	Minimum
5000 Records	0.136	0.109	0.527	0.000	0.136	0.105	0.733	0.000	0.141	0.113	0.789	0.000
1000 Records	0.081	0.060	0.240	0.000	0.074	0.050	0.315	0.000	0.077	0.051	0.398	0.000

	Targeted Data Swapping											
Swapping Rate	1%				5%				10%			
Descriptive Statistics	Average	Median	Maximum	Minimum	Average	Median	Maximum	Minimum	Average	Median	Maximum	Minimum
5000 Records	0.433	0.421	0.882	0.122	0.337	0.315	0.896	0.039	0.274	0.250	0.896	0.003
1000 Records	0.335	0.328	0.579	0.155	0.235	0.213	0.619	0.033	0.195	0.177	0.619	0.010

	Targeted and Random Data Swapping											
Swapping Rate	1%				5%				10%			
Descriptive Statistics	Average	Median	Maximum	Minimum	Average	Median	Maximum	Minimum	Average	Median	Maximum	Minimum
5000 Records	0.300	0.260	0.869	0.004	0.271	0.227	0.970	0.000	0.248	0.210	0.979	0.000
1000 Records	0.231	0.200	0.559	0.004	0.180	0.159	0.607	0.000	0.153	0.139	0.626	0.000

One-Year Age Brackets

	Random Data Swapping											
Swapping Rate	1%				5%				10%			
Descriptive Statistics	Average	Median	Maximum	Minimum	Average	Median	Maximum	Minimum	Average	Median	Maximum	Minimum
5000 Records	0.103	0.066	0.475	0.000	0.109	0.076	0.609	0.000	0.113	0.076	0.700	0.000
1000 Records	0.062	0.044	0.209	0.002	0.067	0.039	0.338	0.000	0.069	0.040	0.370	0.000

	Targeted Data Swapping											
Swapping Rate	1%				5%				10%			
Descriptive Statistics	Average	Median	Maximum	Minimum	Average	Median	Maximum	Minimum	Average	Median	Maximum	Minimum
5000 Records	0.451	0.436	0.930	0.113	0.373	0.342	0.964	0.045	0.325	0.292	0.964	0.023
1000 Records	0.320	0.309	0.561	0.136	0.239	0.218	0.621	0.021	0.198	0.181	0.621	0.008

	Targeted and Random Data Swapping											
Swapping Rate	1%				5%				10%			
Descriptive Statistics	Average	Median	Maximum	Minimum	Average	Median	Maximum	Minimum	Average	Median	Maximum	Minimum
5000 Records	0.287	0.254	0.857	0.001	0.256	0.213	0.953	0.000	0.238	0.200	0.964	0.000
1000 Records	0.209	0.164	0.552	0.002	0.167	0.148	0.604	0.000	0.148	0.135	0.621	0.000

3 Matching Swapped data with Records Which Corresponds to Unique Cells from the Population Census

A important concept when it comes to data confidentiality for official microdata is the disclosure risk of personal information contained in the microdata.

Identification can occur if an identification file and a microdata file are used to conduct one-to-one matching based on the key variables included in both files, which can lead to a matched record being identified as referring to a specific individual ((Bethlehem *et al*. (1990), Müller *et al*. (1995)). Empirical studies to assess the risk of identification are frequently conducted in countries where anonymized official microdata are released.

In the case of data from the 'Population Census', there is a potential risk of personal information being identified when special unique cells are contained in the publicly released data tables, as the tables can be used as external information and matched with anonymized official microdata created from the same data source. This section focuses on determining the disclosure risk for anonymized official microdata created from the Population Census by matching them with publicly released tables from the same data source. Specifically, swapped data was matched with records which correspond to unique cells contained in the tables from the 'Population Census' (referred as 'unique cell records').

Identifying unique cell records for all publicly released tables from the 'Population Census' requires significant time and effort. Therefore, this research was limited to the following six attributes. Unique cell records were determined based on tables created using all combinations of these attributes[6].

Gender
Type and Tenure of Dwelling
Type of Building
Total Number of Floors
Floor Number of Household
Area of floor space of dwelling (14 categories)

In order to determine whether unique cell records decrease after data swapping, matching was conducted between swapped data and unique cell records, and unique records were counted. 776 unique records were found among the 50,000 records for "Area A" and 260 unique records were found among the 10,000 records for "Area B".

6 Five of these attributes (excluding "area of floor space of dwelling") were used as key variables for data swapping, and "area of floor space of dwelling" was used as a non-key variable.

Table 2 and Table 3 contain the results of matching swapped data for five-year age brackets and one-year age brackets with unique cell records. The results show that for higher swapping rates, the percentage of unmatched records is also higher. The results also show that the percentage of unmatched records for targeted data swapping is higher than for random data swapping, whereas the percentage of one-to-one matched records which are not swapped with unique cell records is lower for targeted data swapping than for random data swapping.

The percentage of unmatched records is similar for one-year age brackets and five-year age brackets. For example, in the case of 50,000 records and for a swapping rate of 10%, the results were 75.90% (five-year age brackets, targeted data swapping) and 75.64% (one-year age brackets, targeted data swapping) and 18.94% (five-year age brackets, random data swapping) and 13.66% (one-year age brackets, random data swapping). For the combination of targeted and random data swapping, the results were 57.99% (five-year age brackets) and 56.06% (one-year age brackets)[7].

These results show that using data swapping to create anonymized official microdata results in a reduced disclosure risk for higher swapping rates even when unique cell records are contained in the publicly released tables. This suggests that data swapping has potential to replace the deletion of unique cell records. These results also show that data confidentiality for one-year age brackets is almost identical to that for five-year age brackets[8]. This suggests that if anonymized microdata from the 'Population Census' that contain one-year age brackets were to be released, this would likely not result in an increased disclosure risk.

---

[7] As part of this research, data matching for the above attributes excluding "area of floor space of dwelling", which was not used as a key variable for data swapping was also conducted. In the case of 50,000 records and for a swapping rate of 10%, the percentages of unmatched records were 94.55% (five-year age brackets, targeted data swapping) and 92.73% (one-year age brackets, targeted data swapping) and 28.18% (five-year age brackets, random data swapping) and 14.55% (one-year age brackets, random data swapping). These results are similar to those for data matching for all six attributes, which indicates that data swapping has the potential to lower the disclosure risk associated with unique cells in cross-tabulations of key variables and non-key variables.

[8] As part of this research, swapped donor file records for both one-year age brackets and five-year age brackets were identified. The results show that in the case of 50,000 records and for a swapping rate of 10%, the number of identical donor file records was 361 in both age brackets. This result suggests that the effectiveness of data swapping remains unchanged for different age brackets.

Table 2. Results of Matching between Swapped Data and Unique Cell Records from the Population Census (Five-Year Age Brackets)

50,000 records ("Area A")

	Targeted Data Swapping			Random Data Swapping			Targeted and Random Data Swapping		
Swapping Rate	1%	5%	10%	1%	5%	10%	1%	5%	10%
Percentage of One-to-One Matched Records Which Are Swapped to Unique Cell Records	0.13%	0.00%	0.13%	0.00%	0.00%	0.13%	0.00%	0.00%	0.00%
Percentage of One-to-One Matched Records Which Are Not Swapped to Unique Cell Records	81.96%	43.81%	21.39%	96.78%	87.24%	77.96%	88.14%	59.41%	39.18%
Percentage of One-to-Two Matched Records One of Which Is Swapped to Unique Cell Records	0.52%	0.26%	0.13%	0.64%	0.90%	1.03%	0.77%	0.39%	0.26%
Percentage of One-to-Two Matched Records All of Which Are Swapped to Unique Cell Records	0.00%	0.13%	0.00%	0.00%	0.00%	0.13%	0.00%	0.13%	0.00%
Percentage of 1:n Matched Records All of Which Are Swapped to Unique Cell Records (n is 3 or more.)	0.77%	1.93%	2.45%	0.00%	0.90%	1.80%	0.64%	1.93%	2.58%
Percentage of Unmatched Records to Unique Cell Records	16.62%	53.87%	75.90%	2.58%	10.95%	18.94%	10.44%	38.14%	57.99%

10,000 records ("Area B")

	Targeted Data Swapping			Random Data Swapping			Targeted and Random Data Swapping		
Swapping Rate	1%	5%	10%	1%	5%	10%	1%	5%	10%
Percentage of One-to-One Matched Records Which Are Swapped to Unique Cell Records	0.00%	0.77%	1.92%	0.00%	0.77%	1.15%	0.00%	0.38%	0.77%
Percentage of One-to-One Matched Records Which Are Not Swapped to Unique Cell Records	83.46%	43.08%	25.00%	98.46%	91.92%	80.77%	90.00%	59.23%	36.92%
Percentage of One-to-Two Matched Records One of Which Is Swapped to Unique Cell Records	2.31%	1.92%	1.92%	1.15%	2.31%	3.85%	1.54%	2.31%	2.69%
Percentage of One-to-Two Matched Records All of Which Are Swapped to Unique Cell Records	0.00%	1.92%	1.15%	0.00%	0.00%	0.00%	0.00%	0.38%	1.92%
Percentage of 1:n Matched Records All of Which Are Swapped to Unique Cell Records (n is 3 or more.)	0.00%	2.31%	4.23%	0.00%	0.00%	0.77%	0.00%	1.54%	2.31%
Percentage of Unmatched Records to Unique Cell Records	14.23%	50.00%	65.77%	0.38%	5.00%	13.46%	8.46%	36.15%	55.38%

Table 3. Results of Matching between Swapped Data and Unique Cell Records from the Population Census (One-Year Age Brackets)

50,000 records ("Area A")

	Targeted Data Swapping			Random Data Swapping			Targeted and Random Data Swapping		
Swapping Rate	1%	5%	10%	1%	5%	10%	1%	5%	10%
Percentage of One-to-One Matched Records Which Are Swapped to Unique Cell Records	0.00%	0.00%	0.00%	0.00%	0.00%	0.26%	0.00%	0.26%	0.00%
Percentage of One-to-One Matched Records Which Are Not Swapped to Unique Cell Records	82.86%	44.20%	21.39%	97.68%	90.59%	83.63%	88.66%	61.86%	41.24%
Percentage of One-to-Two Matched Records One of Which Is Swapped to Unique Cell Records	0.39%	0.13%	0.26%	0.77%	0.90%	0.64%	0.52%	0.52%	0.13%
Percentage of One-to-Two Matched Records All of Which Are Swapped to Unique Cell Records	0.13%	0.00%	0.00%	0.00%	0.00%	0.13%	0.00%	0.26%	0.13%
Percentage of 1:n Matched Records All of Which Are Swapped to Unique Cell Records (n is 3 or more.)	0.90%	2.32%	2.71%	0.00%	1.16%	1.68%	0.90%	1.68%	2.45%
Percentage of Unmatched Records to Unique Cell Records	15.72%	53.35%	75.64%	1.55%	7.35%	13.66%	9.92%	35.44%	56.06%

10,000 records ("Area B")

	Targeted Data Swapping			Random Data Swapping			Targeted and Random Data Swapping		
Swapping Rate	1%	5%	10%	1%	5%	10%	1%	5%	10%
Percentage of One-to-One Matched Records Which Are Swapped to Unique Cell Records	0.00%	0.38%	2.31%	0.00%	0.00%	0.00%	0.00%	0.00%	0.77%
Percentage of One-to-One Matched Records Which Are Not Swapped to Unique Cell Records	83.08%	42.31%	25.38%	98.85%	92.31%	88.46%	91.92%	61.92%	39.62%
Percentage of One-to-Two Matched Records One of Which Is Swapped to Unique Cell Records	1.15%	3.08%	2.69%	0.38%	1.54%	2.69%	0.38%	2.31%	3.46%
Percentage of One-to-Two Matched Records All of Which Are Swapped to Unique Cell Records	0.00%	1.15%	1.54%	0.00%	0.00%	0.00%	0.00%	0.77%	1.54%
Percentage of 1:n Matched Records All of Which Are Swapped to Unique Cell Records (n is 3 or more.)	0.77%	2.69%	3.46%	0.00%	0.38%	0.77%	0.38%	1.15%	3.08%
Percentage of Unmatched Records to Unique Cell Records	15.00%	50.38%	64.62%	0.77%	5.77%	8.08%	7.31%	33.85%	51.54%

4 Empirical Assessment of Effectiveness of Data Swapping

In order to assess the effectiveness of data swapping for unique cell records, determining data utility and disclosure risk is required. For this research, the indicators of data utility and disclosure risk were calculated, and a comparison of data utility and disclosure risk using the R-U map was conducted according to Ito and Hoshino (2013). Data utility is 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). The indicator of Data utility (DU) is given as:

$$DU = \frac{\sum_c \left| T^P(c) - T^O(c) \right|}{n_T} \tag{1}$$

$T^O(c)$ 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 indicator of disclosure risk (DR) is given as:

$$DR = \frac{\sum_c I\left(T^O(c)=1, T^P(c)=1\right)}{\sum_c I\left(T^O(c)=1\right)} \tag{2}$$

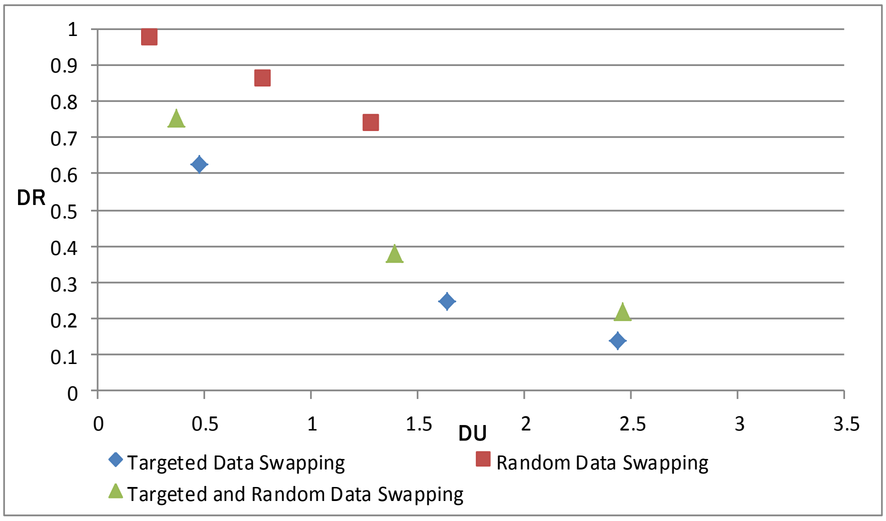
$\sum_c I\left(T^O(c)=1\right)$ is the number of unique cells contained in the tabulation using original data. $\sum_c I\left(T^O(c)=1, T^P(c)=1\right)$ is calculated as the number of unperturbed unique cells in the tabulation.

In this research, DU and DR were calculated based on all possible two-variable combinations of the 6 key variables.

Figure 1 shows the R-U confidentiality map created based on the average values of DU and DR for targeted data swapping, random data swapping and the combination of targeted data swapping and random data swapping in the case of “Area A”. The results show that DU tends to increase as the swapping rate increases, and tends to be higher for targeted data swapping than for random data swapping. This indicates that data utility for targeted data swapping is lower than for random data swapping. DR tends to decrease as the swapping rate increases, and tends to be lower for targeted data swapping than random data swapping. This indicates that disclosure risk for targeted da-

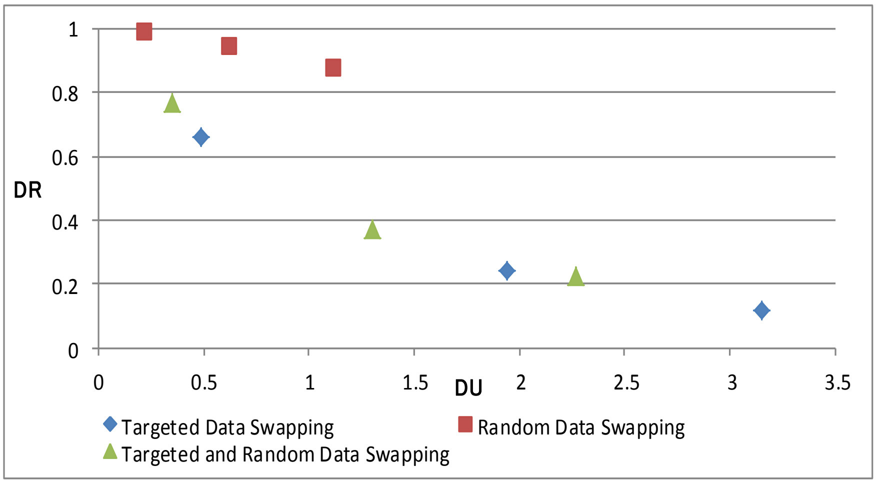
Fig. 1. R-U Confidentiality Map with Data Utility (DU) and Disclosure Risk (DR), The Case of "Area A"

Five-year age brackets



Note: Results are the averages of the values calculated for each of the 10 files of sampled data.

One-year age brackets



Note: Results are the averages of the values calculated for each of the 10 files of sampled data.

ta swapping is lower than for random data swapping.

DU for the combination of targeted and random data swapping tends to be lower than for targeted data swapping and higher than for random data swapping, whereas DR for the combination of targeted and random data swapping tends to be higher than for targeted data swapping and lower than for random data swapping. This indicates that a combination of targeted and random data swapping has the potential to provide a sufficient balance between data utility and disclosure risk.

To further explore this issue, additional research about the impact of different combinations of targeted and random data swapping on disclosure risk and data utility is needed.

5 Conclusion

This research assesses the potential of data swapping as a more efficient tool to create Anonymized census microdata in Japan. The analysis is conducted by matching swapped data created from Japanese Population Census microdata with records which correspond to unique cells from the Population Census tables. The results show that data swapping lead to a reduced disclosure risk at higher swapping rates, even when unique cell records are contained in the publicly released tables.

Data confidentiality for one-year age brackets is almost identical to that for five-year age brackets, which suggests that if anonymized microdata from the 'Population Census' that contain one-year age brackets were to be released, this would likely not result in an increased disclosure risk.

This research calculates data utility and disclosure risk for data swapping based on the R-U map. Results suggest that the combination of targeted and random data swapping has can offer a sufficient balance between data utility and disclosure risk. Further research is required to explore this issue more fully.

Data swapping is a methodology for creating anonymized microdata that can avoid the use of data deletion, and the findings of this research show that it can minimize disclosure risk and increase data utility if an appropriate swapping rate and swapping strategy are chosen. This makes data swapping an efficient tool to create Anonymized Census microdata in Japan, and it is hoped that the results of this research will contribute to the creation of a wider variety of Anonymized microdata in Japan in the future.

Note

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

References

1. Bethlehem, J. M., Keller, W. J., Pannekoek, J.(1990) “Disclosure Control of Microdata”, *Journal of American Statistical Association,* Vol. 85, pp.38-45.
2. Dale, A. and Elliot, M. (2001) “Proposal for 2001 Samples of Anonymized Records: An Assessment of Disclosure Risk”, *Journal of the Royal Statistical Society,* Series A, Vol.164, No.3, pp.427-447.
3. Dalenius, T and Reiss, S. P. (1978) “Data-Swapping: A Technique for Disclosure Control (Extended Abstract)”, in Proceedings of the Section on Survey Research Methods, American Statistical Association, Washington, D.C., pp.191-194.
4. Domingo-Ferrer, J. and Torra, V. (2001) “Disclosure Control Methods and Information Loss for Microdata”, Doyle *et al*. (eds.) *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies,* Elsevier Science, Amsterdam, pp. 91-110.
5. Elliot, M., Manning, A. M., Ford, R. W. (2002) “A Computational Algorithm for Handling The Special Uniques Problem”, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,* Vol. 10, No.5, pp.493-509.
6. Gomatam, S. and Karr, A. F. (2003) “Distortion Measures for Categorical Data Swapping”, *Technical Report,* No.131, National Institute of Statistical Sciences.
7. 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.
8. 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.
9. 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.
10. 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..

11. Müller, W., Blien, U., Wirth, H.(1995) "Identification Risks of Micro Data: Evidence from Experimental Studies", *Sociological Methods and Research*, Vol.24, No.2, pp.131-157.
12. Nin, J., Herranz, J., Torra, V. (2008) "Rethinking Rank Swapping to Decrease Disclosure Risk", *Data and Knowledge Engineering*, Vol. 64, No. 1, pp. 346-364.
13. Shlomo, N., Tudor, C., Groom, P. (2010) "Data Swapping for Protecting Census Tables", Domingo-Ferrer, J. and Magkos, E.(eds) *Privacy in Statistical Databases UNESCO Chair in Data Privacy International Conference, PSD 2010 Corfu, Greece, September, 2010 Proceedings,* Springer, pp.41-51.
14. Takemura, A. (2002) "Local Recoding and Record Swapping by Maximum Weight Matching for Disclosure Control of Microdata Sets", *Journal of Official Statistics*, Vol.18, No.2, pp.275-289.
15. Willenborg, L. and De Waal, T. (2001) *Elements of Statistical Disclosure Control*, Springer, New York.